



Actes du colloque CORESA 2026

COmpression et REprésentation des
Signaux Audiovisuels

Nantes, 27-29 mai 2026

Table des matières

Session 1 : Représentation et codage 1	5
Codes économiques réexaminés : généralisations, bornes et algorithmes <i>Steven Pigeon</i>	5
Evaluation du désordre d'un pavage par la transformée Mojette <i>Jeanpierre Guédon</i>	9
Méthode basée histogramme pour extraire une image balise d'un plan de film <i>Benjamin Serva, Frédéric Comby, William Puech, Loig Le Bihan, Olivier Strauss</i>	12
Comparaison de la géométrie 3D des molécules pour la conception de nouvelles méthodes de codage pour le stockage sur ADN <i>Raphaël Lançon-Ville, Marc Antonini, Frédéric Payan</i>	17
Session 2 : Images volumétriques et technologies immersives	20
Protection de rendus d'objets 3D texturés par obscurisation de la carte de texture <i>Axel Dubar, William Puech, Gérard Subsol</i>	20
Quand l'encombrement visuel sert de mécanisme de simplification de maillage en réalité virtuelle <i>Rivo Andriamanalina, Mohamed-Chaker Larabi, Steven Le Moan</i>	25
Une représentation 3DGS de rangs faibles auto-sélectionnés pour les avatars multi-identités <i>Oumaima Badi, Xiaoran Jiang, Luce Morin, Mårten Sjöström</i>	29
Zone-Adaptive Depth Encoding for Volumetric Video Compression : YUV420p Analysis <i>Abdul Rehman, Francois Bouille</i>	33
Session 3 : Représentation et codage 2	37
A Framework for Event-Level Representation of Wearable Healthcare Time Series <i>Wenda Liu, Fabien Picarougne, Guillaume Raschia, Vincent Ricor- del</i>	37
Compression et Diffusion de Nuages de Points Dynamiques Issus d'un Capteur Mobile	40

Mohammad Dika, Jalal Possik, Petra Bilane, Jad Nassar, Roger Achkar, Charles Yaacoub

Débruitage d'images secrètes à partir d'images partagées compressées en JPEG	44
<i>Valentin Noyé, Pauline Puteaux, William Puech</i>	

Quand le mouvement émerge des co-activations : modélisation probabi- liste des dépendances spatio-temporelles pour les flux de caméras événementielles	48
<i>Wessim Omezzine, Lionel Fillatre</i>	

Session 4 : Traitement d'images et deep learning 52

Adaptation multi-échelle de résolution par apprentissage léger pour VVC 52	
<i>Kra Tchimbé Cyrille Landry Koffi, Thomas Amestoy, Elie Mora, Anissa Mokraoui</i>	

Finetuning Minimax de la règle de décision d'un réseau de neurones pour la classification de lésions cutanées avec une fonction de coût spécifique	56
<i>Younes Habbal, Cyprien Gilet, Lionel Fillatre</i>	

Investigation des effets de dégradations appliquées aux entrées des gé- nérateurs image-vers-image	60
<i>Ahmed Ibrahim, Mohamed-Chaker Larabi, Abderrezzaq Sendjasni</i>	

Rééclairage de visage et harmonisation des teintes dans l'espace latent pour un faceswapping réaliste basé sur les modèles de diffusion	64
<i>Thu Hien Le, Christophe Charrier, Emmanuel Giguët, Maxime Bé- rubé</i>	

Session Poster 68

CAFe-GS : Densification orientée compacité par guidage fréquentiel pour le 3D Gaussian Splatting	68
<i>Léo-Paul Huar, Pierre Hellier, Christine Guillemot, Neus Sabater, Gustavo Sandri</i>	

Détection de la double compression des vidéos HEVC par analyse du domaine compressé et attention neuronale	72
<i>Syamantak Sarkar, Emmanuel Giguët, Christophe Charrier, Kiran Raja</i>	

Egocentric Collision Perceptual Envelopes : Comparing Dynamic and Static Conditions in Immersive visualization.	76
--	----

Hussein Ammar, Toïnon Vigier, Pierre LEBRANCHU, Patrick Le Cal-
let

Évaluation de la Qualité d'Expérience et de l'impact environnemental
de systèmes immersifs 80
Emerik Aji, Amar Tious, Toïnon Vigier

From Vision to Perception : A Pipeline for Assessing Walkability in
Urban Environments 84
Abderrahim Menadi, Vincent Tourre, Jean-Philippe Rivière, My-
riam Servières

Fusion de YOLO à multi-échelles pour la détection d'objets à tailles
très variable 88
Alvin Vedel, Jérôme Pasquet, Romain Thévenoux, Jérôme Cortet,
Cécile Villenave

Hand Movement Tracking for Motor Pattern Identification in Learners
with Trisomy 21 : Preliminary Results 92
Caio Franco Porto, Deise Santana Maia, Marian Oliveira

Spatially-Grounded Gaussian-Prior Attention for Handwritten Mathe-
matical Expression Recognition 96
Ibtissem Haj Ali, Harold Mouchère

Sensibilité des représentations latentes aux distorsions : implications
pour l'évaluation de la qualité d'image 99
Abderrezzaq Sendjasni, Mohamed-Chaker Larabi

Synchronicity Between Gestures and Speech in People with Trisomy
21 : Preliminary Results and Perspectives 103
Laura Chaves Miranda, Gia-Thuy Pham, Deise Santana Maia, Ma-
rian Oliveira, Vera Pacheco

Codes économiques réexaminés : généralisations, bornes et algorithmes

Steven Pigeon

Département de mathématiques, informatique et génie
Université du Québec à Rimouski
300 Allée des Ursulines, Rimouski, Québec
steven_pigeon@uqar.ca

Résumé

Présentés par Eastman *et al.*, les codes économiques permettent d'encoder des valeurs $0 \leq v < n$, indépendamment identiquement uniformément distribuées lorsque n n'est pas une puissance de deux. Bien que d'apparence simple, ce cas particulier des codes de Huffman n'a jamais été complètement caractérisé. Dans cette communication, nous généralisons les codes économiques aux bases $b \geq 2$, donnons les longueurs espérées des codes, leur inefficacité maximale ainsi que le nombre espéré de chiffres sauvés. Finalement, nous présenterons les procédures d'encodage et de décodage.

Mots clefs

Codes de Huffman, codes économiques, compression de données, distribution uniforme.

1 Introduction

Il n'est pas rare que nous devions encoder des entiers $0 \leq v < n$, où n n'est pas une puissance de deux et pour lesquels nous n'avons pas de bonnes raisons de croire qu'ils soient tirés d'une distribution nettement différente de la distribution uniforme. La méthode naïve dans ce cas est d'utiliser un nombre fixe de bits, $\lceil \log_2 n \rceil$, pour encoder les valeurs. Cette approche est rapide, mais peu efficace, car lorsque $2^k < n \ll 2^{k+1}$, nous gaspillons presque un bit. Une meilleure solution serait d'utiliser un code de longueur variable, mais aussi simple que possible. Les codes économiques, présentés par Eastman *et al.*, exploitent un cas particulier des codes de Huffman pour résoudre ce problème [1]. Cependant, bien que mentionnés à plusieurs reprises dans la littérature (par ex. [2, 3, 4, 5]), ces codes n'ont jamais été décrits avec précision.

Dans cet article, nous décrivons la généralisation de ces codes pour une base $b \geq 2$. Certaines expressions, déjà connues pour le cas particulier $b = 2$, comme la longueur espérée des codes [3] et l'inefficacité maximale [5] seront étendues à toutes les bases. Nous abordons aussi de nouvelles considérations comme le gain spécifique et le gain espéré, qui n'ont jamais été caractérisés. Finalement, nous discuterons des procédures d'encodage et de décodage.

2 Codes économiques généralisés

Dans ce qui suit, nous dirons « base » pour la taille de l'alphabet de codage et « chiffre » pour un symbole de cet alphabet.

2.1 Structure des codes

Pour coder optimalement $b^k \leq n < b^{k+1}$ valeurs distinctes, il nous faut construire un arbre de codes complet où les feuilles se trouvent, au plus, sur deux profondeurs adjacentes (voir fig. 1). Pour ce faire, nous réutiliserons la procédure de Huffman [6]. Pour la construction l'arbre de codes b -aire complet pour n valeurs, nous utiliserons la décomposition

$$n = b^k + r,$$

qui est unique si $0 \leq r < b^{k+1} - b^k = b^k(b - 1)$.

Nous savons que si $r \neq 0$, l'arbre de code ne pourra pas être parfait, car un arbre parfait aurait exactement b^k feuilles ; or, il nous en faut $b^k + r$. Donc certaines feuilles seront à la profondeur k , recevant des codes courts, tandis que les autres seront à la profondeur $k + 1$, recevant des codes longs.

Pour utiliser la procédure de Huffman, c'est-à-dire itérativement fusionner les b feuilles de poids plus faibles pour finir par n'obtenir qu'un seul nœud actif, la racine, nous devons avoir n tel que $b - 1 \mid n - 1$. Il se peut donc que nous ayons à ajuster n pour accommoder cette condition, en posant $n' = n + (n - 1) - (n \bmod b - 1)$.

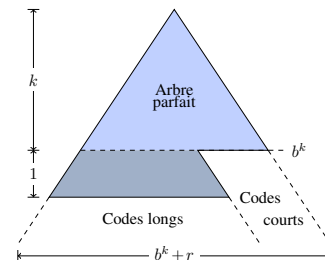


FIGURE 1 – La structure du code.

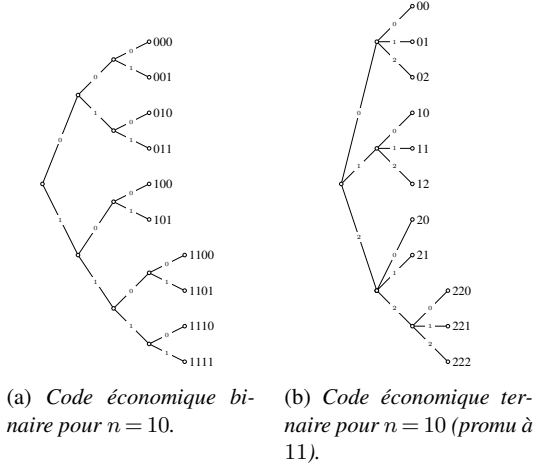


FIGURE 2 – Exemples de codes économiques.

nœuds, tous de poids égal. En utilisant la procédure de Huffman, nous retirons les b nœuds de poids plus faibles, les fusionnons, et ajoutons un nouveau nœud dont le poids est la somme des nœuds fusionnés. À chaque itération, le nombre de nœuds actifs est réduit de $b - 1$. Nous répétons jusqu'à ce que nous ayons b^k nœuds actifs. Le nombre m de fusions nécessaires pour atteindre un arbre parfait est tel que $r = m(b - 1)$, soit encore $m = \frac{r}{b-1}$. Puisque ces m fusions représentent chacune b feuilles, nous aurons $l = bm = \frac{br}{b-1}$ codes longs et $s = n - bm = b^k - \frac{r}{b-1}$ codes courts. Cette valeur s nous servira de seuil pour les procédures rapides d'encodage et de décodage.

La fig. 2 donne deux exemples de codes économiques pour $n = 10$. À la fig. 2a, nous trouvons un code binaire pour $n = 2^3 + 2$, avec un seuil $s = 2^3 - 2 = 6$, tandis qu'à la fig. 2b, nous devons promouvoir n à 11 pour accommoder le code ternaire, avec un seuil de $s = 3^2 - 1 = 8$.

2.2 Longueur espérée

Maintenant que nous connaissons le nombre de codes courts et de codes longs, la longueur espérée s'obtient sans peine :

$$\begin{aligned} \bar{L}_b(n) &= \bar{L}_b(b^k + r) = \frac{sk + l(k+1)}{b^k + r} \\ &= k + \frac{br}{(b-1)(b^k + r)}. \end{aligned} \quad (1)$$

Comme $0 \leq r < b^k(b-1)$, nous pouvons vérifier que la longueur moyenne est $k \leq \bar{L}_B(n) < k+1$.

2.3 Inefficacité maximale

Puisqu'un code économique est un arbre b -aire de codes utilisant un nombre entier de chiffres pour chaque code, il contiendra nécessairement une part d'inefficacité. Un code idéal utiliserait *exactement* $\log_b n$ chiffres pour encoder une valeur, toujours en supposant les valeurs uniformément iid. Avec un code économique, nous avons

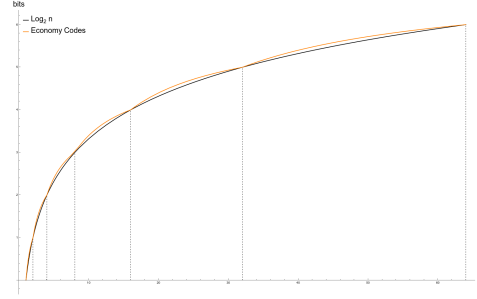


FIGURE 3 – Longueur des codes économiques pour le cas $b = 2$ vs $\log_2 n$.

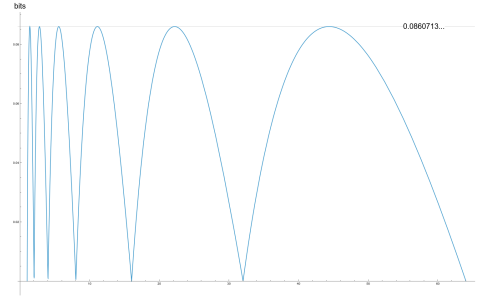


FIGURE 4 – L'inefficacité $\bar{L}_b(n) - \log_b n$ des codes économiques pour le cas $b = 2$.

$\bar{L}_b(n) \geq \log_b n$ avec l'égalité seulement lorsque $r = 0$, et une différence $0 < \bar{L}_b(n) - \log_b n < 1$ autrement. La fig. 3 montre la longueur moyenne d'un code économique pour le cas particulier $b = 2$ vs un code idéal (et hypothétique) utilisant exactement $\log_2 n$ bits, tandis que la fig. 4 montre l'inefficacité $\bar{L}_b(n) - \log_b n$ des codes économiques.

Nous remarquons que $\bar{L}_b(n) - \log_b n$ est convexe en $0 \leq r < b^k(b-1)$, nous pouvons donc trouver le maximum de l'inefficacité en résolvant

$$\frac{d}{dr} (\bar{L}_b(n) - \log_b(n)) = 0 \quad (2)$$

pour r . Nous trouvons que

$$\frac{d}{dr} (\bar{L}_b(n) - \log_b(n)) = \frac{b^{k+1} \ln b - (b-1)(b^k + r)}{(b-1)(b^k + r)^2 \ln b},$$

puis nous résolvons

$$\frac{b^{k+1} \ln b - (b-1)(b^k + r)}{(b-1)(b^k + r)^2 \ln b} = 0 \quad (3)$$

pour r . L'éq. (3) est satisfaite lorsque son numérateur est nul, c'est-à-dire lorsque

$$b^{k+1} \ln b = (b-1)(b^k + r), \quad (4)$$

et nous trouvons que le r qui maximise l'inefficacité est

$$r^* = \frac{b^k(b(\ln b - 1) + 1)}{b-1}. \quad (5)$$

En substituant r^* dans $n = b^k + r$, nous trouvons que le

« pire » $b^k < n < b^{k+1}$ est donné (approximativement) par

$$n^* = b^k + r^* = \frac{b^{k+1} \ln b}{b-1}. \quad (6)$$

Maintenant, en utilisant les éqs. (5) et (6), nous trouvons

$$\begin{aligned} \bar{L}_b(n^*) - \log_b n^* &= \frac{br^*}{(b-1)(b^k + r^*)} - \log_b \left(1 + \frac{r^*}{b^k}\right) \\ &= \frac{1}{b-1} + \frac{\ln(b-1) - \ln \ln b - 1}{\ln b}. \end{aligned} \quad (7)$$

Pour le cas particulier $b=2$, l'éq. (7) nous donne

$$\begin{aligned} \frac{1}{b-1} + \frac{\ln(b-1) - \ln \ln b - 1}{\ln b} \\ &= \frac{1}{2-1} + \frac{\ln(2-1) - \ln \ln 2 - 1}{\ln 2} \\ &= 0.086071332\dots, \end{aligned}$$

un résultat précédemment obtenu [5, p. 102–104], et qui peut être vu comme un cas particulier du résultat de Gallager pour une distribution quelconque [7]. Pour d'autres bases, comme par exemple, $b=3$ et $b=10$, nous trouvons que l'inefficacité maximale est de 0.135084... et 0.268843... chiffres, respectivement. Un examen plus attentif de l'éq. (7) nous révèle que l'inefficacité est strictement croissante en b et bornée supérieurement par 1 lorsque b tends vers l'infini. Cela nous permet de conclure que la base avec la plus petite inefficacité maximale est la base $b=2$.

2.4 Gain spécifique et gain espéré

Pour un code économique en base b pour $n = b^k + r$, avec $0 < r < b^k(b-1)$, nous sauverons, en moyenne

$$\begin{aligned} \lceil \log_b n \rceil - \bar{L}_b(n) &= (k+1) - \left(k + \frac{br}{(b-1)(b^k + r)} \right) \\ &= 1 - \frac{br}{(b-1)(b^k + r)} \end{aligned} \quad (8)$$

chiffres par rapport à un code fixe qui utilise toujours $\lceil \log_b n \rceil = k+1$ chiffres.

Le gain spécifique donné par l'éq. (8) est maximal lorsque $r > 0$ est petit et presque nul lorsque $r = (b^k - 1)(b-1)$, soit $n \approx b^{k+1}$. Le cas $r=0$ correspond à $n = b^k$, auquel cas le code de taille fixe est optimal. Toutefois, puisque nous devons avoir $b-1 \mid n-1$, le plus petit $r \neq 0$ n'est pas forcément 1 : le plus petit $r \neq 0$ est $r = b-1$. En conséquence, le gain maximal est de

$$\begin{aligned} 1 - \frac{br}{(b-1)(b^k + r)} &= 1 - \frac{b(b-1)}{(b-1)(b^k + (b-1))} \\ &= 1 - \frac{b}{b^k + b - 1} \end{aligned}$$

chiffres. Les codes économiques sont les plus intéressants

lorsque $b^k < n \ll b^{k+1}$, car nous sauverons presque un chiffre à chaque code. Par contre, les codes économiques le sont beaucoup moins lorsque n est presque b^{k+1} . Si nous nous questionnons sur l'utilité d'un code économique, nous pouvons établir un seuil d'utilité $0 < \alpha < 1$ tel que si

$$1 - \frac{br}{(b-1)(b^k + r)} \geq \alpha,$$

ou, de façon équivalente, si

$$r \leq b^k \frac{(1-\alpha)(b-1)}{1+\alpha(b-1)},$$

nous jugerons utile l'emploi de codes économiques. Sinon, nous utiliserons simplement un code fixe sur $k+1$ chiffres. Quel est le gain espéré étant donné un $b^k < n < b^{k+1}$ uniformément distribué sur cet intervalle ? Bien qu'Eastman *et al.* nous assurent que nous sauvons en moyenne un demi bit [1, p. 18], voyons ce qu'il en est vraiment. L'espérance de l'éq. (8), en ne considérant que les $b^k < n < b^{k+1}$ valides, c'est-à-dire tels que $b-1 \mid n-1$, est

$$\begin{aligned} \mathbb{E}[\lceil \log_b n \rceil - \bar{L}_b(n)] &= \frac{1}{b^k - 1} \sum_{\substack{r=b-1, 2(b-1), \\ \dots, (b^k-1)(b-1)}} 1 - \frac{br}{(b-1)(b^k + r)} \\ &= 1 - \frac{b}{(b^k - 1)(b-1)} \sum_{j=1}^{b^k-1} \frac{j(b-1)}{b^k + j(b-1)}. \end{aligned} \quad (9)$$

La forme $\sum_{r=1}^u \frac{r}{v+r}$ doit être transformée pour être exploitable. Remarquons que

$$\sum_{r=1}^u \frac{r}{v+r} = u - v \sum_{r=1}^u \frac{1}{v+r},$$

puis que

$$\sum_{j=a}^b \frac{1}{1+js} = \frac{1}{s} \left(H\left(b + \frac{1}{s}\right) - H\left(a - 1 + \frac{1}{s}\right) \right),$$

où $H_m = \psi(m) + \gamma \approx \ln m + \gamma$ est le $m^{\text{ième}}$ nombre harmonique généralisé, $\psi(m)$ la fonction digamma et $\gamma = 0.577215644\dots$ la constante Euler-Mascheroni. Pour suivons l'éq. (9) :

$$\begin{aligned} \mathbb{E}[\lceil \log_b n \rceil - \bar{L}_b(n)] &= 1 - \frac{b}{(b^k - 1)(b-1)} \sum_{j=1}^{b^k-1} \frac{j(b-1)}{b^k + j(b-1)} \\ &= 1 - \frac{b}{(b^k - 1)(b-1)} \times \\ &\quad \left((b^k - 1) - \frac{b^k}{b-1} \left(H\left(\frac{b^{k+1}}{b-1} - 1\right) - H\left(\frac{b^k}{b-1}\right) \right) \right) \\ &= \frac{b^{k+1}}{(b^k - 1)(b-1)^2} \left(H\left(\frac{b^{k+1}}{b-1} - 1\right) - H\left(\frac{b^k}{b-1}\right) \right) - \frac{1}{b-1}. \end{aligned} \quad (10)$$

Si l'éq. (8) nous donne le nombre de chiffres sauvés en

moyenne pour un n donné, l'éq. (10) nous donne le nombre de chiffres que nous pouvons espérer d'un n valide entre deux puissances de b . Sachant

$$\lim_{k \rightarrow \infty} \mathbb{E}[\lceil \log_b n \rceil - \bar{L}_b(n)] = \frac{b \ln b - (b-1)}{(b-1)^2}, \quad (11)$$

nous pouvons trouver le gain espéré en fonction de b . Pour $b=2$, l'éq. (11) nous donne un gain espéré de $2 \ln 2 - 1 = 0.386294\dots$ bits, c'est-à-dire que pour un n quelconque, nous pouvons nous attendre à sauver environ 0.39 bits — pas un demi bit! Pour $b=3$ et $b=10$, nous trouvons des gains espérés de $0.323959\dots$ et $0.173158\dots$ chiffres, respectivement. De plus, nous voyons que l'éq. (11) tend vers zéro alors que b croît, donc que le gain espéré est le plus grand lorsque b est le plus petit, c'est-à-dire lorsque $b=2$.

2.5 Codage et décodage

Les procédures d'encodage et de décodage sont simples et rapides, exploitant la structure des codes pour nous dispenser de la construction explicite de l'arbre de codes. Revenons à la fig. 1. Encore une fois, utilisons la décomposition $n = b^k + r$ avec $0 \leq r < b^k(b-1)$ et n tel que $b-1 \mid n-1$. Les codes courts, au nombre de $s = b^k - \frac{r}{b-1}$, utiliseront k chiffres, tandis que les codes longs en utiliseront $k+1$.

Pour encoder une valeur v , si $v < s$, la valeur correspond à un code court et nous émettons le code « naturel » pour v sur k chiffres. Si $v \geq s$, la valeur correspond à un code long et nous émettrons un code sur $k+1$ chiffres. Les codes pour les valeurs de 0 à $s-1$ étant courts, le premier code long est la valeur s décalée d'un chiffre, c'est-à-dire sb . Comme les valeurs $v \geq s$ sont codées relativement au premier code long, nous émettons $sb + (v - s) = v + s(b-1)$ sur $k+1$ chiffres.

Pour décoder, nous lisons d'abord k chiffres et vérifions si cette valeur est inférieure à s . Si c'est le cas, nous retournons la valeur lue telle quelle. Sinon, il s'agit des k premiers chiffres d'un code long ; il nous faut lire un chiffre supplémentaire qu'on ajoute aux chiffres déjà lus, formant la valeur v . Pour inverser l'encodage, nous retournons $v - s(b-1)$.

Les procédures sont montrées aux algorithmes 2.1 et 2.2. Notons qu'une implémentation optimisée s'abstiendrait de recalculer la décomposition. De plus, pour le cas spécial $b=2$, la méthode se simplifie grandement puisque $b-1$ devient 1.

3 Conclusion

Nous avons présenté une généralisation des codes économiques d'Eastman *et al.* aux bases $b \geq 2$. Nous avons donné les expressions pour la longueur espérée et pour l'inefficacité maximale. Nous avons montré comment généraliser le gain spécifique d'un code économique pour un n particulier au gain que l'on peut espérer d'un $b^k < n < b^{k+1}$ quelconque, mais toujours tel que $b-1 \mid n-1$. Nous avons

```

fonction Encoder( $v, n, b$ )
   $k = \lfloor \log_b n \rfloor$ 
   $r = n - b^k$ 
   $s = b^k - \frac{r}{b-1}$ 
  si  $v < s$  alors
    émettre  $v$  sur  $k$  chiffres
  sinon
    émettre  $v + s(b-1)$  sur  $k+1$  chiffres

```

Algorithme 2.1 – Procédure d'encodage.

```

fonction Décoder( $v, n, b$ )
   $k = \lfloor \log_b n \rfloor$ 
   $r = n - b^k$ 
   $s = b^k - \frac{r}{b-1}$ 
   $v \leftarrow$  lire  $k$  chiffres
  si  $v < s$  alors
    retourner  $v$ 
  sinon
     $v \leftarrow vb + (\text{lire 1 chiffre})$ 
    retourner  $v - s(b-1)$ 

```

Algorithme 2.2 – Procédure de décodage.

montré que l'inefficacité maximale pour le cas particulier $b=2$ est d'au plus $0.086071\dots$ bits, alors que le gain espéré est de $0.396294\dots$ bits, et que le gain espéré est le plus grand lors que $b=2$. Finalement, nous avons montré que les codes économiques sont à la fois très simples et très efficaces, du moins si nous supposons que les valeurs à coder sont tirées d'une distribution uniforme.

Références

- [1] Willard L. Eastman, Abraham Lempel, Jacob Ziv, et Martin Cohn. US Patent 4,464,650 : Apparatus and Method for Compressing Data Signals and Restoring the Compressed Data Signals, Août 1984.
- [2] Willard L. Eastman. US Patent 5,087,913 : Short-Record Data Compression and Decompression System, Février 1992.
- [3] Hidetoshi Yokoo. An Improved Ziv-Lempel Coding Scheme for Universal Source Coding. *Electronics and Communications in Japan, Part I*, 69(12) :12–20, 1986.
- [4] Peter Tischer. A Modified Lempel-Ziv-Welch Data Compression Scheme. *Australian Computer Science Communications*, 9(1) :262–272, 1987.
- [5] Steven Pigeon. *Contributions à la compression de données*. Thèse de doctorat, Département d'informatique et de recherche opérationnelle, Université de Montréal, Décembre 2001.
- [6] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Procs of the IRE*, 49(9) :1098–1101, Septembre 1952.
- [7] Robert G. Gallager. Variations on a Theme by Huffman. *IEEE Trans on Information Theory*, 24(6) :668–674, Novembre 1978.

Evaluation du désordre d'un pavage par la transformée Mojette

Jeanpierre GUEDON
équipe IPI, LS2N
Polytech Nantes, Nantes Université
jeanpierre.guedon@univ-nantes.fr

Résumé

Dans ce papier, nous présentons une variation de la transformée Mojette pour évaluer le désordre d'un pavage 2D. Nous présentons également 2 types de pavages créés par l'auteur, un basé sur 2 tailles de carrés, l'autre à base de triangles de Héron. Les premiers résultats sont encourageants.

Mots clefs

désordre de pavage, pavage 2D irrégulier, transformée Mojette.

1 Introduction

La mesure classique du désordre est l'entropie et sert autant en physique qu'en théorie de l'information (avec une définition différente). Cependant, l'entropie en 2D est généralement définie par produit tensoriel ce qui en limite le pouvoir. Il y a 25 ans, j'ai produit un type de pavage avec 2 carrés de base de côtés entiers a et b , avec $1 < a < b$ et $\text{PGCD}(a,b)=1$. Depuis 5 ans, j'ai également produit des pavages à base de triangles de Héron dont les 3 côtés ainsi que l'aire sont entiers.

Les pavages à partir de différents pavés sont étudiés en mathématiques depuis les célèbres de l'Alhambra à Grenade. La théorie des pavages sépare généralement ceux-ci en pavage régulier, irréguliers et apériodiques. Ces derniers sont ceux qui semblent avoir un désordre à la fois maximal et obligatoire. Dans ce papier, nous sommes seulement intéressés par les pavages irréguliers soit avec un petit nombre de pavés différents (pavage des 2 carrés) soit avec un grand nombre de pavés différents (pavage de triangles de Héron). Il est à noter que peu de travaux s'intéressent à ce dernier cas, sauf la conjecture de Maurice Nivat (1997) [1]. Le but de cette recherche est d'évaluer si le niveau de désordre de certains pavages irréguliers approche ceux des pavages apériodiques comme ceux de Roger Penrose [2] ou de David Smith [3]. Une question complémentaire est de savoir comment le nombre de pavés impacte le désordre.

La transformée Mojette est un outil de projections selon une orientation donnés par 2 points de la grille discrète [4].

Dans la prochaine section, on définit les 2 types de pavages utilisés ainsi que la définition un peu particulière de la transformée Mojette qui va nous servir. La section

suivante présente quelques premiers résultats qui seront discutés dans la dernière section.

2 Définitions

Ici, sont présentés nos 2 types de pavage ainsi que la variation de transformée Mojette utile pour la mesure du désordre.

2.1 Pavage des 2 carrés

Soient a et b les 2 côtés de carrés avec $1 < a < b$ et $\text{PGCD}(a,b)=1$. Pour les figures, on a pris l'exemple original avec $a=3$ et $b=4$. On comprend bien que l'on peut avoir un demi-pavage composé seulement de carré de côté a et l'autre demi-pavage de carrés b , ce qui représente la situation de désordre minimale. Mais en construisant des rectangles de taille $(\text{PPCM}(a,b) \times (a+b))$, on conçoit facilement que le pavage de ces rectangles orientables selon les 2 axes va augmenter le désordre comme le montre la figure 1 :

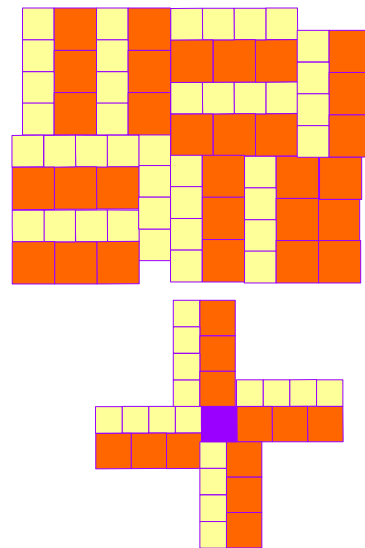


Figure 1 – Pavage de 2 carrés (ici $a=3$ et $b=4$).

Il est simple de paver le plan avec un seul rectangle (dans l'exemple (12×7)) mais on voit sur la figure précédente qu'on peut ajouter (ou pas) un carré a ou un carré b à

Figure 4 – Repérage des transitions sur un pavage à 2 carrés bien mélangé

En reprenant approximativement le même nombre de carrés, mais en augmentant le mélange, on obtient la figure 4. On a calculé les 2 projections avec le nombre de changements de segments. Alors que dans la première situation, nous n'avions qu'un seul changement (de valeur 10), ici on en repère 29 (16+13) dont la plupart possède la valeur 4. Par conséquent, notre outil semble bien affûté pour l'évaluation du désordre de ce type de pavage.

3 Expérimentation sur les pavages de Héron

Même si les triangles de Héron possèdent des cotés entiers, l'inclinaison de ces cotés dans un puzzle n'est pas forcément donnée par une droite discrète sur la grille. Mais on peut restreindre ces pavages de sorte que chacun des points de chaque triangle soit sur la grille comme sur la figure 5 :

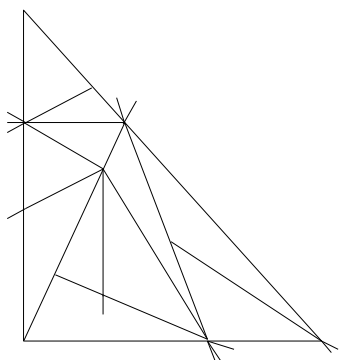


Figure 5 – Repérage des transitions sur un triangle de Héron avec tous les points de la division sur la grille.

Dans ce cas, on obtient au plus qu'un seul point projeté pour chacun des segments décrivant le pavage. Par contre, ils sont projetés quasiment dans des directions toutes différentes et c'est ce qui produit le niveau de désordre visible.

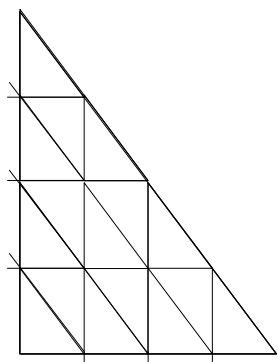


Figure 6 – Repérage des transitions sur un triangle de Héron de type $(3n \ 4n \ 5n)$.

A l'inverse, la figure 6 montre un pavage de Héron périodique bâti à partir des multiples du triangle $(3 \ 4 \ 5)$. Il n'y a plus que 3 projections (à valeurs nulles).

On a donc obtenu des résultats équivalents sur 2 types de pavage irréguliers avec notre mesure de projection. Il nous manque toutefois la comparaison avec l'entropie maximale qui doit être faite sur des supports de différentes tailles afin de pouvoir finir de caractériser le niveau de désordre d'un pavage irrégulier.

4 Conclusion

Dans ce papier, on a d'une part présenté 2 types de pavages irréguliers que l'on a pas retrouvé dans la littérature, et d'autre part utilisé une variante de la transformée Mojette (initialement présentée à CORESA95) pour donner un ordre de grandeur du désordre de ces pavages. Contrairement à l'entropie, les projections donnent une bonne idée du désordre car tout n'est pas condensé dans un seul nombre. On a bien l'équivalent de l'entropie nulle pour un pavage périodique. Toutefois, il reste à comparer ces pavages irréguliers avec les pavages apériodiques grâce à cet outil afin de savoir (selon la dimension du pavage fini donné) comment on s'approche d'une limite, quelle est elle, et comment relier cette limite au maximum d'entropie

Remerciements

Je tiens à remercier particulièrement ici Nicolas Normand qui m'avait assisté dans la première réalisation « physique » du pavage à 2 carrés en carrelant ma salle à manger en 2000. Merci aussi à lui et Imants Svalbe pour leur retours scientifiques sur ce papier.

Références

- [1] Nivat's conjecture and pattern complexity in algebraic subshifts. Jarkko Kari, Etienne Moutot. Theoretical Computer Science, Vol. 777, p. 379-386 . July 2019,
- [2] Roger Penrose, « *The role of aesthetics in pure and applied mathematical research* », *Bulletin of the Institut. for Maths and its Appl.*, 10, , 266-271. 1974.
- [3] David Smith, Joseph Samuel Myers, Craig S. Kaplan et Chaim Goodman-Strauss, « An aperiodic monotile » 2023,
- [4] The Mojette Transform: Theory and Applications, Jeanpierre Guédon Ed., ISTE & WILEY, 2009
- [5] Jp Guédon, D. Barba. Décomposition psychovisuelle et transmission rapide d'images. CORESA 1995, Rennes, Janv 1995.
- [6] R. Guédon, P. Evenou, Jp. Guédon, M. LeTreust, Site PPMOJ.fr, Janv 2026.

Méthode basée histogramme pour extraire une image balise d'un plan de film

Benjamin Serva¹ Frédéric Comby¹ William Puech¹ Loig Le Bihan² Olivier Strauss¹

¹ LIRMM, Univ Montpellier, CNRS, Montpellier France

² RIRRA21, Université de Montpellier Paul-Valéry, Montpellier France

{benjamin.serva, frederic.comby, william.puech, olivier.strauss}@lirmm.fr

{loig.le-bihan}@univ-montp3.fr

Résumé

La croissance des contenus multimédia renforce le besoin de méthodes capables d'extraire automatiquement des images représentatives, appelées ici images balises, à partir de vidéos. Contrairement à la vidéosurveillance, où les cibles sont prédéfinies, le cinéma présente une grande diversité visuelle et narrative, rendant la sélection d'une image balise par plan particulièrement complexe. Nous proposons une méthode fondée sur la détection d'objets et une analyse statistique des histogrammes d'occurrences des classes détectées, visant une sélection sémantiquement cohérente. Cette approche est enrichie par des pondérations spatiales, temporelles et de netteté afin d'optimiser la pertinence de l'image retenue. Une évaluation qualitative menée auprès d'experts en analyse cinématographique montre que notre méthode est préférée aux approches de l'état de l'art et présente une meilleure robustesse sur différents types de plans.

Mots clefs

Image-balises, image clé, analyse de film, statistiques basées sur l'histogramme, détection d'objets.

1 Introduction

L'analyse cinématographique débute généralement par la segmentation d'un film en plans, via la détection automatique ou manuelle des transitions, comme décrit par Halim et Faiza [1]. Une fois segmenté, le film peut être synthétisé en associant à chaque plan une image représentative, que nous appelons une image balise. Ce terme, introduit dans le cadre de nos travaux menés avec des experts en analyse cinématographique, s'est avéré plus adapté que celui d'image clé, généralement associé à un événement ponctuel de la vidéo. L'image balise vise à caractériser un plan dans son ensemble.

Une méthode simple a été proposée par Widiarto *et al.*, consistant à résumer un plan en sélectionnant la première et la dernière image de chaque plan [2]. Bien que cette approche puisse être efficace dans certains cas, elle repose fortement sur le hasard. En effet, la qualité du résultat dépend fortement du contenu des transitions. Dans le contexte du cinéma, où les transitions sont souvent com-

plexes, la première ou la dernière image peut appartenir à la transition elle-même et ne pas représenter fidèlement le plan. Une deuxième approche consiste à privilégier les images contenant des objets en mouvement. Pei propose ainsi d'utiliser le flux optique comme caractéristique discriminante, combiné à des histogrammes locaux (HSV) pour construire des vecteurs de caractéristiques [3]. Ce type de méthode est efficace pour capturer des informations dynamiques, mais peut perdre de sa pertinence lorsque la caractérisation sémantique d'un plan n'est pas transmis par le mouvement.

Le mouvement peut également être utilisé en conjonction avec d'autres caractéristiques visuelles. Par exemple, Zhang *et al.* proposent une méthode de fusion multi-descripteurs qui combine des informations dynamiques et statiques (couleur, structure) à l'aide d'une transformée de Fourier quaternionique [4]. L'objectif est d'obtenir des images saillantes, présentant de fortes variations globales et locales, qui sont susceptibles d'être perçues comme représentatives par l'observateur. D'autres approches intègrent la détection d'objets et l'évaluation de la qualité d'image. Chen *et al.* utilisent un détecteur d'objets pour attribuer une note de qualité à chaque image, combinée à des descripteurs locaux (SIFT/SURF) pour extraire les images clés les plus informatives [5]. Ces approches fonctionnent bien dans des contextes où les cibles sont bien définies, mais soulèvent la question de la généralisation à des contenus cinématographiques plus variés. Cependant, la grande majorité de ces approches a été développée pour des applications spécifiques, en particulier dans le domaine de la vidéosurveillance. Dans ce domaine, la recherche se concentre généralement sur des cibles spécifiques, telles que les personnes ou les véhicules.

Dans cet article, nous présentons une nouvelle méthode pour extraire une image balise unique et représentative à partir d'un plan de film. L'efficacité de cette méthode repose à la fois sur l'analyse statistique des histogrammes, mais aussi sur l'intégration de pondérations temporelles, spatiales et liées à la qualité de l'image afin de mieux caractériser le plan illustré. Nous proposons une comparaison, réalisée par un panel d'experts en analyse cinématographique, de nos images balises à celles générées par des méthodes de l'état de l'art.

2 La méthode d'extraction d'images balise proposée

Dans cette section, nous présentons notre méthode d'extraction d'images balise dédiée aux plans de films cinématographiques. Notre méthode s'inspire des travaux de Chen *et al.* [5]. Pour chaque image du plan courant, nous construisons un histogramme basé sur les classes détectées par un modèle de détection d'objets donné. Il convient de noter que les objets pris en compte sont ceux dont le niveau de confiance de détection dépasse un seuil donné.

Supposons que le modèle de détection considéré ait mis en évidence K classes d'objets. La méthode que nous proposons consiste, dans un premier temps, à associer un histogramme à chaque image de la manière suivante. Pour chaque classe C_k , avec $k \in \{1, \dots, K\}$, nous calculons $S_f(C_k)$, son score associé, comme suit :

$$S_f(C_k) = \sum_{i=1}^{n_k} \rho_i(C_k), \quad (1)$$

où n_k est le nombre d'objets de classe C_k détectés dans la $f^{\text{ème}}$ image et $\rho_i(C_k)$ est le score de confiance associé au $i^{\text{ème}}$ objet c -à- d . la probabilité associée à la prédiction que le $i^{\text{ème}}$ objet appartienne à la classe C_k .

La deuxième étape consiste à calculer l'histogramme moyen parmi toutes les images du plan courant :

$$\bar{S}(C_k) = \frac{1}{N} \sum_{f=1}^N S_f(C_k), \quad (2)$$

où N est le nombre total d'images dans le plan courant. Nous sélectionnons ensuite, comme image balise, l'image dont l'histogramme associé est le plus proche, en distance euclidienne, de l'histogramme moyen, comme expliqué dans la section 2.4.

Cette approche fournit une base initiale pour la sélection des images de référence. Cependant, elle présente certaines limites, notamment en termes de pertinence narrative et de qualité visuelle. Afin de surmonter ces limites et de mieux répondre aux exigences spécifiques de l'analyse cinématographique, nous proposons trois améliorations complémentaires. Ces améliorations ont été déterminées grâce à une étroite collaboration avec des experts en analyse cinématographique, ce qui nous a permis d'identifier diverses règles utilisées pour construire les plans au cinéma. Sur la base de ces règles, nous avons pu améliorer la méthode susmentionnée en ajoutant une pondération spatiale centrée, une pondération temporelle et enfin une pondération de netteté.

2.1 Pondération spatiale centrée

Une pondération spatiale centrée vise à privilégier les objets détectés près du centre de l'image (zone fovéale) par rapport à ceux situés à la périphérie. L'intuition derrière cette pondération repose sur le fait que, dans les films d'auteur, les objets centraux sont généralement plus pertinents

dans la scène. La pondération spatiale est définie par :

$$w_{\text{pos}}(x, y) = \exp \left(-\frac{(x - c_x)^2 + (y - c_y)^2}{2\sigma_s^2} \right), \quad (3)$$

où c_x et c_y sont les coordonnées du centre de l'image, x et y les coordonnées de la boîte englobante de l'objet détecté, et σ_s un paramètre contrôlant la répartition de la pondération. En tenant compte de cette pondération spatiale, le nouveau score individuel $S'_f(C_k)$ est :

$$S'_f(C_k) = \sum_{i=1}^{n_k} \rho_i(C_k) w_{\text{pos}}(x_i, y_i), \quad (4)$$

où (x_i, y_i) est la position du $i^{\text{ème}}$ objet détecté.

2.2 Pondération de netteté

L'objectif de la pondération de la netteté est de privilégier les images nettes par rapport aux images floues. Cela vise à garantir une meilleure qualité visuelle et esthétique pour l'image choisie comme image balise. La netteté de la $f^{\text{ème}}$ image, notée $w_Q(f)$, est basée sur la variance du Laplacien de la $f^{\text{ème}}$ image et normalisée sur l'ensemble de la prise de vue. Dans ce cas, l'équation Eq. (4) devient :

$$S''_f(C_k) = w_Q(f) \sum_{i=1}^{n_k} \rho_i(C_k) w_{\text{pos}}(x_i, y_i). \quad (5)$$

2.3 Pondération temporelle

La pondération temporelle vise à tenir compte de la structure narrative habituelle d'un plan. Dans un plan, le début de la séquence contient rarement des images caractéristiques du récit. Les images pertinentes se trouvent plutôt au milieu ou à la fin de la séquence, là où l'action ou l'intention du scénario atteint généralement son apogée.

La pondération temporelle que nous proposons $w_T(f)$, avec $f \in \{1, \dots, N\}$, est donc définie de manière à donner plus d'importance aux images situées vers le milieu du plan ou légèrement au-delà :

$$w_T(f) = \exp \left(-\frac{(f - \frac{N}{2})^2}{2\sigma_N^2} \right) + \exp \left(-\frac{(f - (N - \epsilon_N))^2}{2\sigma_N^2} \right), \quad (6)$$

où σ_N est un paramètre qui contrôle la largeur de la distribution gaussienne et ϵ_N est un décalage appliqué par rapport à la dernière image afin de positionner la deuxième gaussienne.

Cette pondération temporelle est utilisée pour pondérer la contribution de chaque image à l'histogramme moyen :

$$\bar{S}(C_k) = \frac{1}{\sum_{f=1}^N w_T(f)} \sum_{f=1}^N S''_f(C_k) w_T(f). \quad (7)$$

2.4 Sélection de l'image de balise

Enfin, nous calculons, pour chaque image f , la distance euclidienne entre $\{S''_f(C_k)\}_{k=1\dots K}$ et $\{\bar{S}(C_k)\}_{k=1\dots K}$

comme suit :

$$d(f) = \sqrt{\sum_{k=1}^K \left(S_f''(C_k) - \bar{S}(C_k) \right)^2}. \quad (8)$$

L'indice de l'image balise est $\hat{f} = \operatorname{argmin}_{f \in \{1, \dots, N\}} d(f)$.

3 Résultats expérimentaux

Cette section présente les bases de données utilisées pour l'évaluation des méthodes d'extraction d'images balise ainsi que la comparaison des performances de notre méthode et de celles de l'état de l'art.

3.1 Bases de données et paramètres

La première base de données comprend 105 plans tirés de divers films et annotés par un expert. La deuxième base de données comprend 10 plans tirés de 10 films différents. Pour les deux bases de données, les plans ont été sélectionnés pour la diversité de leur contenu, notamment les différents mouvements de caméra, les entrées et sorties de personnages, ainsi que des extraits de films récents en couleur et de films plus anciens en noir et blanc.

Nous avons comparé notre méthode à quatre méthodes de sélection d'images balise, à savoir Chen *et al.* [5], Pei [3] et Zhang *et al.* [4]. La quatrième méthode consiste à considérer comme image balise l'image située au milieu du plan filmé, que nous appelons la **image centrale**. Les trois premières méthodes, initialement fondées sur un seuil dynamique produisant plusieurs images balise, ont été adaptées pour ne retenir que l'image maximisant le score.

Notre approche repose sur un modèle Faster R-CNN [6] avec un seuil de confiance de 70 %. Ce choix s'appuie sur des analyses basées sur la métrique PDQ [7], indiquant qu'un seuil compris entre 0.7 et 0.8 offre un bon compromis entre qualité de localisation, fiabilité de classification et rappel. L'utilisation de Faster R-CNN sur les plans considérés révèle l'émergence de $K = 91$ classes sur lesquelles nous basons le calcul de nos histogrammes. Pour la pondération spatiale (section 2.1), σ_s est fixé à $\sigma_s = 0.5$, tandis que pour la pondération temporelle (section 2.3), les paramètres ϵ_N et σ_N sont fixés à $\epsilon_N = 0.1N$ et $\sigma_N = 0.15N$, N étant le nombre d'images dans le plan considéré. Les expériences menées sur ces deux bases de données se complètent mutuellement.

3.2 Comparaison qualitative

Dans cette section, nous présentons une comparaison qualitative des images balises obtenues par différentes méthodes sur un plan extrait de la première base de données. Dans ce plan (extrait du film "Man of the West" [8]), le moment clé correspond à la découverte du compagnon blessé. Comme l'illustre la figure 1, seule l'image balise issue de notre méthode capture précisément cet instant décisif, contrairement aux autres approches.

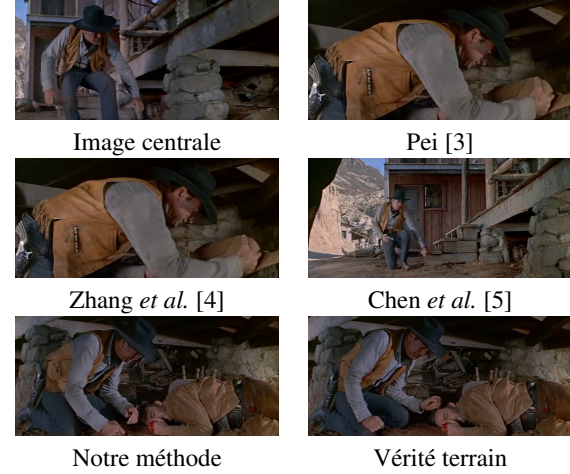


FIGURE 1 – Comparaison des résultats obtenus pour le plan extrait du film "Man of the West" [8]

3.3 Première expérience

Dans cette première expérience sur la première base de données, nous mesurons objectivement la proximité entre l'image sélectionnée par chaque méthode et celle de l'expert. Naïvement, on pourrait penser qu'une méthode pertinente consisterait à évaluer la différence de temps entre l'image fournie par l'expert et celle produite par la méthode évaluée. Cependant, cette différence s'est avérée incompatible avec l'analyse de notre expert. Nous proposons donc de mesurer la proximité entre l'image extraite par une méthode d'estimation d'une image balise et l'image choisie par l'expert, en comparant les histogrammes de ces deux images. Cette méthode a été validée par notre expert. Nous proposons de comparer les histogrammes à l'aide de la distance de Bhattacharyya. Plus la distance de Bhattacharyya est petite, plus le contenu des images est proche, et donc meilleure est la concordance avec la vérité terrain définie par les experts. Pour chacune des cinq méthodes, nous pouvons donc calculer 105 distances de Bhattacharyya et ainsi produire certaines statistiques qui sont présentées dans le tableau 1.

Dans ce tableau, ainsi que dans les suivants, les meilleurs résultats sont en gras, tandis que les deuxièmes meilleurs résultats sont soulignés. Notre méthode obtient la distance moyenne et l'écart type les plus faibles (0.18 et $\sigma=0.1$), ce qui indique une précision et une cohérence supérieures à celles des approches concurrentes. La médiane et l'écart interquartile confirment également la cohérence de la méthode proposée, tandis que la valeur maximale indique sa stabilité. Pour mener à bien cette expérience, nous proposons de calculer un taux d'acceptation pour un seuil d'acceptation donné, *c.-à-d.* le nombre d'images balise sélectionnées dont la distance de Bhattacharyya dépasse ce seuil. Notre méthode obtient l'AUC le plus élevé (0.81), démontrant ainsi des performances supérieures sur une large gamme de seuils. Dans l'ensemble, ces résultats montrent que l'approche proposée surpasse systématiquement

ment les méthodes de pointe pour la sélection d’une image balise qui correspond à la sélection de notre expert.

Méthode	Moy↓	σ ↓	Med↓	IQR↓	Max↓	AUC
Image centrale	0.224	0.137	0.211	0.196	0.701	0.772
Pei [3]	0.321	0.170	0.347	0.214	0.756	0.673
Zhang <i>et al.</i> [4]	0.329	0.184	0.281	0.195	0.774	0.667
Chen <i>et al.</i> [5]	0.228	0.191	0.172	0.239	0.808	0.769
Notre méthode	0.188	0.100	<u>0.176</u>	0.148	0.384	0.810

TABEAU 1 – Statistiques des distances de Bhattacharyya entre les images sélectionnées et les annotations.

3.4 Deuxième expérience

Une faiblesse évidente de notre première expérience est qu’elle ne fait intervenir qu’un seul expert. Nous proposons donc une deuxième expérience impliquant 16 experts différents, spécialistes en littérature et praticiens réguliers de l’analyse cinématographique. La définition de la vérité terrain étant intrinsèquement subjective, car elle dépend de l’interprétation des experts, la constitution d’un ensemble de données de test conduirait donc à des valeurs de vérité multiples et potentiellement contradictoires. Nous proposons donc une seconde expérimentation basée sur les votes des 16 experts à partir de la deuxième base de données. Chaque expert s’est vu présenter les 10 plans de cette base (voir section 3.1), ainsi que les cinq images sélectionnées par les cinq méthodes concurrentes. Nous avons pris soin de ne jamais présenter les images dans le même ordre afin de limiter les biais de sélection. Nous avons demandé à chaque expert de sélectionner toutes les images, parmi les 5 qui lui étaient présentées, qu’il jugeait suffisantes pour identifier ou reconnaître le plan qu’il avait vue. Chaque méthode s’est vu attribuer un score compris entre 0 et 160, correspondant au nombre de sélections jugées pertinentes par les experts. En normalisant ce score par la valeur maximale (160), nous avons obtenu un taux de sélection reflétant l’adéquation de la méthode aux attentes des experts.

Le tableau 2 présente le taux de sélection obtenu pour chaque méthode. Il apparaît clairement que la méthode que nous proposons correspond tout à fait aux attentes des experts, loin devant la deuxième meilleure méthode, qui serait celle de Chen *et al.* [5], suivie de près par la méthode de l’image centrale.

Méthode	Taux de Sélection	Top 1	Top 2	Top 3
Image centrale	0.381	10%	<u>50%</u>	80%
Pei [3]	0.213	10%	40%	50%
Zhang <i>et al.</i> [4]	0.238	0%	30%	50%
Chen <i>et al.</i> [5]	<u>0.388</u>	<u>30%</u>	<u>50%</u>	<u>70%</u>
Notre méthode	0.513	50%	70%	<u>70%</u>

TABEAU 2 – Taux de sélection basé sur l’évaluation de 16 experts et classement pour les 1er, 2e et 3e places.

Pour compléter cette analyse, nous présentons, pour $k = 1, 2, 3$, le pourcentage de sélection dans le Top k . Une méthode est sélectionnée dans le Top k si, pour un extrait donné, elle figure parmi les k images ayant reçu le plus de votes. Une fois encore, notre méthode est la plus efficace, suivie de près par celle proposée par Chen *et al.* [5].

Le choix de l’image centrale, qui est l’une des pratiques les plus courantes, ne semble pas être une si mauvaise solution. En tout état de cause, elle est bien meilleure que les méthodes proposées par Pei ou Zhang *et al.*.

4 Conclusion

Pour faciliter l’analyse cinématographique, nous avons proposé une méthode dédiée à l’extraction d’une image balise pour chaque plan. Elle repose sur une analyse statistique des objets détectés dans les images du plan, enrichie par des pondérations spatiales, temporelles et de netteté issues des échanges avec des experts. Les comparaisons menées avec plusieurs méthodes existantes montrent que notre approche est celle qui fournit le plus souvent des images jugées pertinentes par les spécialistes de l’analyse filmique.

Références

- [1] Benoughidene Abdel Halim et Titouna Faiza. Shot boundary detection : Fundamental concepts and survey. Dans *The 1st International Conference on Innovative Trends in Computer Science, CITSC 2019, Guelma, Algeria, November 20-21, 2019*, volume 2589 de *CEUR Workshop Proceedings*, pages 34–40. CEUR-WS.org, 2019.
- [2] Wisnu Widiarto, Eko Mulyanto Yuniarno, et Mochamad Hariadi. Video summarization using a key frame selection based on shot segmentation. Dans *2015 International Conference on Science in Information Technology (ICSITech)*, pages 207–212, 2015.
- [3] Xiao-Gen Pei. The key frame extraction algorithm based on the indigenous disturbance variation difference video. *Procedia Computer Science*, 183 :533–544, 2021. Proceedings of the 10th International Conference of Information and Communication Technology.
- [4] Yunzuo Zhang, Jiayu Zhang, Ruixue Liu, Pengfei Zhu, et Yameng Liu. Key frame extraction based on quaternion fourier transform with multiple features fusion. *Expert Systems with Applications*, 216 :119467, 2023.
- [5] Mingju Chen, Xiaofeng Han, Hua Zhang, Guojun Lin, et M.M. Kamruzzaman. Quality-guided key frames selection from video stream based on object detection. *Journal of Visual Communication and Image Representation*, 65 :102678, 2019.
- [6] Shaoqing Ren, Kaiming He, Ross Girshick, et Jian Sun. Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(06) :1137–1149, Juin 2017.
- [7] Simon Wenkel, Khaled Alhazmi, Tanel Liiv, Saud Alrshoud, et Martin Simon. Confidence score : The forgotten dimension of object detection performance evaluation. *Sensors*, 21(13), 2021.
- [8] Anthony Mann. Man of the west. Film, 1958.

Remerciements

Merci au projet NUMALYSE. Ce travail a été soutenu par une subvention du gouvernement français gérée par l'Agence nationale de la recherche dans le cadre du programme France 2030, référence ANR-24-CE54-7489.

Nous tenons à remercier **Jean-Philippe Trias** pour son aide dans la création de la base de données et des valeurs de vérité associées, ainsi que pour avoir partagé son expérience en matière d'analyse cinématographique, qui nous a grandement aidés à développer les différentes pondérations.

Comparaison de la géométrie 3D des molécules pour la conception de nouvelles méthodes de codage pour le stockage sur ADN

Raphaël LANÇON-VILLE¹

Marc ANTONINI¹

Frédéric PAYAN¹

¹ Université Côte d'Azur, CNRS, I3S, France

{lanconville, am, fpayan}@i3s.unice.fr

Résumé

Une limite actuelle du stockage de données numériques sur ADN synthétique est l'introduction fréquente d'erreurs au moment du séquençage des brins ADN décrivant les données numériques initiales. Sachant que la forme 3D d'une molécule joue un rôle primordial dans l'interaction avec son environnement, nous cherchons actuellement à répondre à la question suivante : est-ce que la forme géométrique 3D des structures moléculaires constituant les brins d'ADN joue un rôle crucial dans l'apparition d'erreurs lors du séquençage/décodage ? Et si oui, peut-on prendre en compte cela pour améliorer les performances des codeurs dédiés au stockage sur ADN ?

Mots clefs

stockage sur ADN, nuages de points 3D, recalage 3D, similarité moléculaire, codage de données

1 Introduction

En offrant une densité très élevée et un archivage passif potentiellement durable, le stockage sur ADN de synthèse constitue une alternative prometteuse pour stocker les données binaires dites "froides" [1]. Le principe général du stockage sur ADN (voir Figure 1) est de transformer une donnée binaire (décrite par définition à l'aide de 2 symboles 0 et 1) en une séquence de symboles quaternaires (notés A, T, C et G en référence aux nucléotides), pour ensuite la synthétiser en brins d'ADN stockés dans une capsule étanche et résistante aux conditions environnementales.



FIGURE 1 – Principe général du stockage sur ADN.

quente d'erreurs au moment du désarchivage de la donnée numérique initiale, notamment lors du séquençage. La plupart des solutions proposées jusqu'à maintenant pour réduire et/ou corriger les erreurs de stockage sur ADN sont logiquement inspirés des populaires codes correcteurs d'erreurs, le stockage sur ADN pouvant être vu comme une transmission sur un canal bruité.

2 Objectif

Dans le domaine de la chimie ou de la biologie, il est reconnu que la forme géométrique 3D d'une molécule joue un rôle primordial dans l'interaction avec son environnement [2]. L'objectif de notre projet actuel est de répondre à la question suivante : dans le contexte du stockage sur ADN, est-ce que la forme géométrique 3D des structures moléculaires constituant les brins d'ADN joue un rôle crucial dans l'apparition d'erreurs lors du séquençage/décodage ? Plus précisément : i) la forme géométrique des nucléotides influence-t-elle leur probabilité de confusion lors du séquençage ? ii) peut-on exploiter ces caractéristiques géométriques pour améliorer l'encodage ?

Pour bien comprendre la problématique, le premier test que nous avons effectué a été de comparer la géométrie des nucléotides, éléments de base constituant les brins ADN. Pour cela, nous avons utilisé la suite logicielle SENSEAAS [3, 4] qui permet d'aligner des structures moléculaires, pour ensuite évaluer non seulement leurs similarités géométriques mais aussi la localisation de leurs propriétés physico-chimiques.

3 Présentation du logiciel SENSEAAS

Le principe général de SENSEAAS [3], illustré par la Figure 2, est décrit ci-dessous. La première étape consiste à modéliser la surface de van der Waals des deux molécules que l'on souhaite aligner/ comparer à l'aide de nuage de points 3D. Chaque point est décrit par ses coordonnées 3D et une couleur liée à sa nature atomique ou, plus tard dans l'algorithme, à une propriété physico-chimique cartographiée sur la surface. Ces deux nuages sont d'abord superposés en ne considérant que leur géométrie, décrite localement à l'aide de descripteurs FPFH [5]. Cette superposition est ensuite raffinée en tenant compte de la géométrie mais aussi des

Une limite actuelle de cette solution est l'introduction fré-

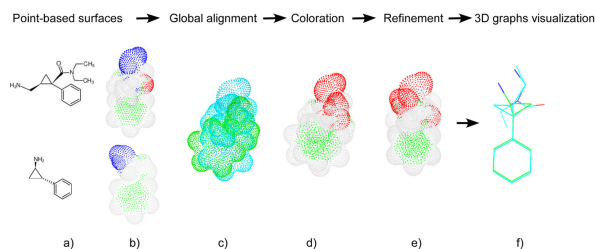


FIGURE 2 – Principe général du logiciel SENSEAAS [3]. a) Molécules d'entrée; b) modélisation de leur surface à l'aide de nuages de points 3D; c) Alignement globale (prise en compte de la géométrie uniquement) d) Coloration des points selon la propriété physico-chimique de l'atome le plus proche; e) Raffinement de l'alignement (prise en compte de la géométrie et des couleurs); f) Alignement final.

propriétés physico-chimiques associées à chaque point [6], qui dépendent de la nature de l'atome le plus proche. La dernière étape consiste à mesurer la similarité des surfaces moléculaires ainsi alignées. Pour cela, SENSEAAS calcule trois scores :

- *gfit* : score de correspondance géométrique, allant de 0 à 1 (meilleure correspondance, *self-matching* géométrique)
- *hfit* : score de correspondance des couleurs/propriétés physico-chimiques, allant de 0 à 1 (meilleure correspondance, *self-matching* colorimétrique)
- *gfit + hfit* : score combinant géométrie et propriétés, allant de 0 à 2 (meilleure correspondance, *self-matching* total).

4 Expérimentations

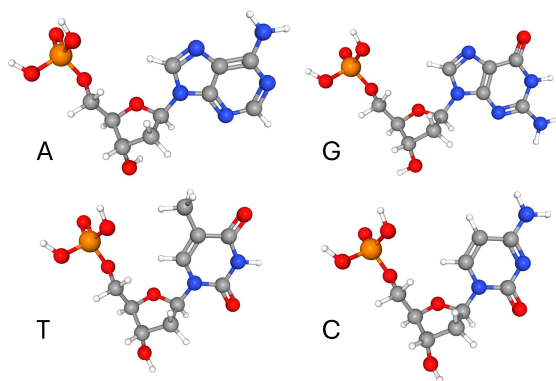


FIGURE 3 – Graphes 3D des nucléotides formant les brins d'ADN synthétique.

La première expérimentation a été de vérifier si l'alignement du graphe 3D de chaque nucléotide sur lui-même donnait le score maximal. La figure 3 montre les graphes

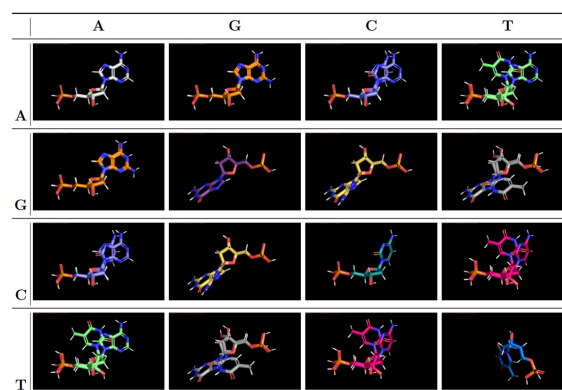


FIGURE 4 – Alignements croisés des quatre nucléotides constituant les brins ADN synthétiques.

3D des quatre nucléotides (A,G,T,C). La Figure 4 montre les alignements obtenus pour chaque nucléotide sur lui-même et sur les trois autres nucléotides. Les 3 scores de similarité "SENSEAAS" obtenus pour tous ces alignements sont indiqués Figure 5.

On peut observer sur la diagonale de la Figure 4 que l'alignement de chaque nucléotide sur lui-même est parfaitement réussi. Ceci est confirmé puisque les scores *gfit+hfit* correspondants sont tous égaux à 2 (diagonale Figure 5). Le self-matching est donc validé pour chaque nucléotide.

Aussi, pour tous les autres alignements obtenus hors self-matching, on constate sur la Figure 4 que la structure moléculaire commune à chaque nucléotide (phosphate et sucre) est systématiquement correctement alignée, et que les nucléotides ne se distinguent donc que par leur base (A,C,G,T). Ceci confirme que SENSEAAS, initialement prévu pour aligner et comparer des molécules médicamenteuses est aussi adapté pour comparer géométriquement des nucléotides et donc a priori des brins d'ADN.

5 Discussion

Ces tests préliminaires montrent que SENSEAAS est capable de mesurer des ressemblances entre les nucléotides. En analysant plus en détails les scores "SENSEAAS", on peut faire quelques observations supplémentaires : la cytosine (C) et la thymine (T) semblent être les nucléotides les plus ressemblants géométriquement (*gfit+hfit* = 1,627). Ceci est cohérent puisque leur base respective est très proche d'un point de vue structural (constituée d'un unique cycle). Aussi, si on calcule la moyenne des scores de chaque nucléotide aligné sur les 3 autres, la thymine obtient le score moyen le plus élevé (*gfit+hfit* = 1,43), alors que les trois autres obtiennent un score *gfit+hfit* légèrement supérieur à 1. On peut donc supposer que la thymine est le nucléotide le plus souvent confondu au moment du décodage - notamment avec le cytosine - de par sa forme et ses propriétés physico-chimiques locales communes qui vont interagir avec la structure moléculaire du nanopore utilisé

Matrice gfit				
	Adenine	Cytosine	Guanine	Thymine
Adenine	1	0,674	0,601	0,57
Cytosine	0,674	1	0,481	0,7
Guanine	0,601	0,481	1	0,692
Thymine	0,57	0,7	0,692	1
Matrice hfit				
	Adenine	Cytosine	Guanine	Thymine
Adenine	1	0,395	0,493	0,544
Cytosine	0,395	1	0,399	0,927
Guanine	0,493	0,399	1	0,867
Thymine	0,544	0,927	0,867	1
Matrice gfit+hfit				
	Adenine	Cytosine	Guanine	Thymine
Adenine	2	1,069	1,094	1,114
Cytosine	1,069	2	0,88	1,627
Guanine	1,094	0,88	2	1,559
Thymine	1,114	1,627	1,559	2

FIGURE 5 – Scores SENSEAAS obtenus.

lors du séquençage.

Ces premiers résultats nous ont mené à nos travaux en cours, qui sont dédiés à la comparaison non pas des nucléotides isolés mais de séquences de nucléotides constituant les mots codes utilisés par certains codeurs spécifiques au stockage sur ADN, tels que celui proposé dans [7]. À terme, nous espérons pouvoir déterminer quels mots codes sont les moins susceptibles d'être confondus lors du séquençage par nanopore, afin d'améliorer la robustesse des (dé)codeurs.

Références

- [1] Melpomeni Dimopoulou et Marc Antonini. Data and image storage on synthetic dna : existing solutions and challenges. *EURASIP Journal on Image and Video Processing*, 2022, 10 2022.
- [2] Sandhya Kortagere, Matthew David Krasowski, et Sean Ekins. The importance of discerning shape in molecular pharmacology. *Trends in pharmacological sciences*, 30 3 :138–47, 2009.
- [3] D. Douguet et F. Payan. Sensesaas : Shape-based alignment by registration of colored point-based surfaces. *Molecular Informatics*, 39 :e2000081, 2020.
- [4] Hamza Biyuzan, Mohamed-Akram Masrour, Lucas Grandmougin, Frédéric Payan, et Dominique Douguet. Sensesaas-flex : a joint optimization approach for aligning 3d shapes and exploring the molecular conformation space. *Bioinformatics*, 40(3) :btac105, 02 2024.
- [5] Radu Bogdan Rusu, Nico Blodow, et Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. Dans *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009.
- [6] Jaesik Park, Qian-Yi Zhou, et Vladlen Koltun. Colored point cloud registration revisited. Dans *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 143–152, 2017.

- [7] Melpomeni Dimopoulou. *Encoding techniques for long-term storage of digital images into synthetic DNA*. Theses, Université Côte d'Azur, Décembre 2020.

Protection de rendus d'objets 3D texturés par obscurisation de la carte de texture

A. Dubar

W. Puech

G. Subsol

Équipe ICAR, LIRMM, Univ Montpellier, CNRS, Montpellier, France

{axel.dubar, william.puech, gerard.subsol}@lirmm.fr

Résumé

La diffusion croissante d'objets 3D texturés dans les environnements virtuels soulève des enjeux de protection visuelle contrôlée. Les approches existantes ciblent principalement la géométrie, alors que les cartes de texture concentrent une part importante de l'information visuelle et sémantique. Nous proposons une méthode de protection des textures compatible avec le rendu, qui altère la carte de texture tout en conservant la géométrie et le paramétrage UV. Un masque binaire dérivé du paramétrage UV restreint les transformations aux régions effectivement projetées sur la surface, garantissant une dégradation visuelle contrôlée. Plusieurs stratégies d'obscurisations sont évaluées sur un ensemble d'objets 3D texturés à l'aide de métriques objectives et d'une analyse sémantique par réseau de neurones convolutionnel. Les résultats montrent que l'approche permet d'atteindre différents niveaux de sécurité visuelle tout en préservant l'intégrité géométrique et la compatibilité avec les moteurs de rendu.

Mots clefs

Sécurité multimédia, Sécurité visuelle, Métaverse, Obscuration d'image, Objet 3D texturé.

1 Introduction et travaux précédents

L'usage croissant des objets 3D texturés dans les plateformes interactives, les environnements virtuels et les applications de type Métaverse soulève de nouvelles problématiques de protection, de confidentialité et de contrôle de diffusion. Au-delà de la simple visualisation, ces objets peuvent être désormais échangés, commercialisés et intégrés dans des environnements à forte valeur économique. Dans ce contexte, garantir la protection des objets 3D devient un enjeu central.

Un objet 3D texturé, illustré en figure 1, est constitué d'un maillage définissant sa géométrie, d'une ou plusieurs cartes de texture décrivant les propriétés visuelles des surfaces lors du rendu, et d'un paramétrage UV assurant la correspondance entre la géométrie et la texture. Cette distinction entre maillage, carte de texture et rendu est essentielle dans le contexte de la protection des objets 3D, les approches existantes s'étant majoritairement concentrées sur la sécurisation du maillage.

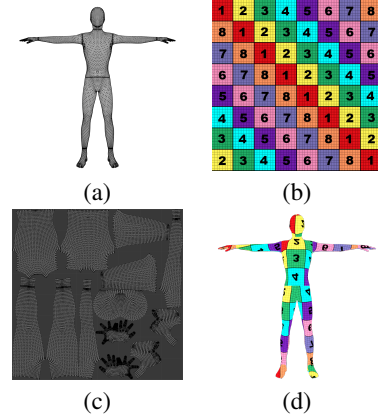


FIGURE 1 – Structure d'un objet 3D texturé : a) Le maillage, avec des sommets, des faces et des arêtes, b) La carte de texture, c) Une illustration du paramétrage UV, d) Un rendu de l'objet 3D texturé.

Ainsi, différentes approches de sécurité ont été proposées pour les objets 3D, notamment le tatouage et le chiffrement appliqués au maillage. Les travaux pionniers sur le tatouage de maillages 3D [1] ainsi que les méthodes de chiffrement sélectif adaptées aux objets 3D [2] ont principalement ciblé les coordonnées des sommets et la structure du maillage. Bien que ces approches permettent d'assurer traçabilité ou confidentialité, elles ignorent la carte de texture, qui concentre pourtant une part significative de l'information visuelle et sémantique.

Certaines approches proposent un chiffrement conjoint du maillage et des cartes de texture [3]. Toutefois, ces méthodes reposent sur un chiffrement intégral, rendant l'objet résultant totalement ininterprétable sans clé et ne permettant pas de contrôle progressif du niveau de dégradation visuelle. Il n'existe alors pas de méthodes permettant une obscurisation avec un contrôle progressif qui tiennent compte de l'information supplémentaire apportée par les cartes de textures. Or, une carte de texture peut être considérée comme une image dont les pixels sont projetés sur le maillage via le paramétrage UV. On peut alors s'intéresser aux techniques d'obscurisation d'images qui ont été largement étudiées [4]. Ces méthodes permettent de moduler le niveau de dégradation visuelle tout en conservant la compatibilité des formats. Cependant, leurs applications directes

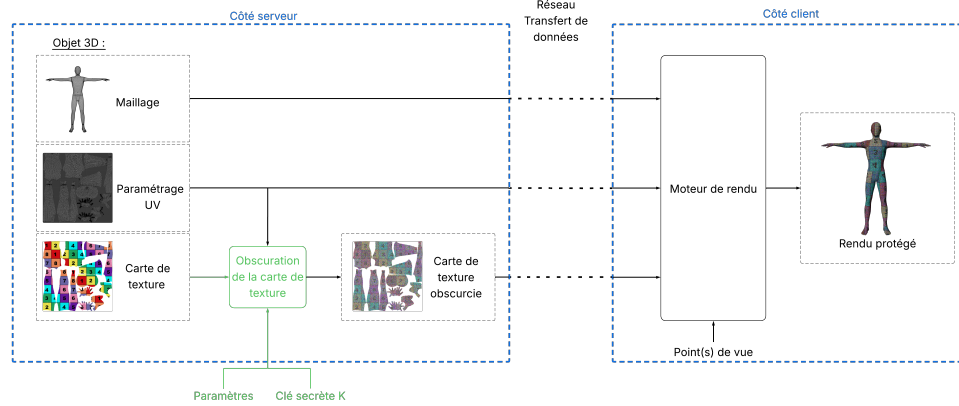


FIGURE 2 – Vue d’ensemble de la méthode proposée. La carte de texture est obscurcie en tenant compte du paramétrage UV avant rendu.

aux cartes de texture pose des difficultés spécifiques liées au paramétrage UV et au processus de rendu.

Afin de caractériser les différents degrés de protection visuelle, nous nous appuyons sur la notion de niveaux de sécurité visuelle (*Visual Security Level VSL*) introduite pour les objets 3D chiffrés [5]. Cette notion permet de distinguer différents niveaux allant de *transparent* à *confidentiel*, en fonction de la capacité d’interprétation visuelle résiduelle.

Dans ce travail, nous proposons une méthode de protection de rendu d’objets 3D texturés basée sur l’obscurcissement de la carte de texture, tenant explicitement compte du paramétrage UV. Cette approche vise à contrôler la dégradation visuelle tout en préservant l’intégrité géométrique et la compatibilité avec les moteurs de rendu standards.

2 Méthode proposée

2.1 Vue d’ensemble

La figure 2 illustre le principe général de la méthode. La protection agit exclusivement sur la carte de texture, tandis que le maillage et le paramétrage UV restent inchangés afin de préserver l’intégrité géométrique et la compatibilité avec le moteur de rendu.

Dans un premier temps, un masque est extrait à partir des coordonnées UV du maillage. Ce masque identifie les régions de la texture effectivement projetées sur la surface de l’objet. Dans un second temps, une stratégie d’obscurcissement est appliquée uniquement sur ces régions. La texture modifiée est ensuite utilisée dans le pipeline de rendu standard pour générer une visualisation protégée.

Cette approche garantit que la dégradation visuelle observée provient uniquement des zones contribuant réellement à l’apparence rendue.

2.2 Extraction du masque UV

La carte de texture d’un objet 3D contient généralement des régions non utilisées correspondant aux zones non couvertes par le paramétrage UV.

Afin de restreindre la transformation aux régions effectivement projetées sur la surface, un masque est construit à partir des coordonnées UV associées aux sommets du maillage. Pour construire le masque, chaque face du maillage est projetée dans l’espace de la texture, et les pixels appartenant aux faces comme valides. Les autres pixels sont exclus du processus d’obscurcissement.

Ce masque garantit que la dégradation visuelle observée après rendu provient exclusivement des régions contribuant réellement à l’apparence de l’objet.

2.3 Stratégies d’obscurcissement

Une fois le masque extrait, différentes stratégies peuvent être appliquées à la carte de texture.

Obscuration globale. Une transformation est appliquée pixel par pixel sur les régions valides. Selon les paramètres choisis, il peut s’agir d’un chiffrement partiel ou d’un filtrage dégradant tel qu’un flou gaussien [6]. L’intensité de la dégradation visuelle dépend directement des paramètres choisis.

Obscuration par bloc. La carte de texture est divisée en blocs carrés de taille fixe. Deux variantes sont alors considérées : les obscurcissements *intra-bloc*, où l’obscurcissement intervient sur les pixels au sein même du bloc, indépendamment des autres blocs ; l’obscurcissement *inter-blocs*, où l’obscurcissement intervient sur la relation que différents blocs ont entre eux, en permutant leurs positions par exemple.

Les opérations sont restreintes aux blocs entièrement contenus dans les régions valides du masque, afin d’éviter toute interaction avec des zones non valides. La taille des blocs et la fonction d’obscurcissement permettent d’ajuster le niveau de dégradation visuelle.

2.4 Reconstruction

Lorsque l’obscurcissement repose sur une transformation réversible paramétrée par une clé, la texture originale peut être reconstruite en appliquant la transformation inverse. Le maillage et le paramétrage UV restant inchangés, le rendu restauré est identique à l’original.

Dans le cas de transformations irréversibles, la texture modifiée constitue une version définitivement altérée destinée à la visualisation contrôlée.

3 Résultats expérimentaux

3.1 Protocole expérimental

Les expérimentations sont menées sur 568 objets texturés issus de la base de données RWTT [7]. Pour chaque objet, 60 points de vue sont générés sur une sphère suivant une distribution de Fibonacci, le rayon de la sphère étant choisi de manière à ce que chaque point vue puisse observer l'entière de la boîte englobante de l'objet. Les cinq vues présentant la plus forte occupation dans l'image sont retenues pour l'évaluation.

L'analyse est effectuée à deux niveaux : au niveau de la texture, afin de mesurer l'altération directe de la carte de texture, et au niveau du rendu, afin d'évaluer l'impact visuel final. Les métriques utilisées incluent le PSNR et le SSIM [8] pour mesurer la perte en qualité des cartes de texture et des rendus. Nous utilisons également des mesures de sécurité telles que l'Edge Similarity Score (ESS) et la Luminance Similarity Score (LSS) [9], qui évaluent respectivement la persistance des contours et la similarité globale de luminance entre une image originale et une image protégée. L'évaluation est complétée par une évaluation sémantique à l'aide d'un classifieur utilisant réseau convolutionnel pré-entraîné sur les classes d'ImageNet [10].

Les évaluations sont effectuées et présentées avec trois types de fonction d'obscurcissement, un flou gaussien qui est une fonction globale non réversible, un chiffrement sélectif fondé sur AES en mode CTR qui est une fonction globale réversible, et une combinaison d'opérations de blocs décrit par Aprilpyone et Kiya [11] qui contient des opérations intra-bloc et inter-blocs.

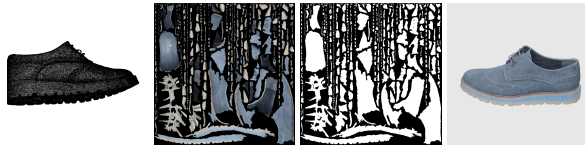


FIGURE 3 – *Objet 3D Chaussure : a) Maillage, b) Carte de texture, c) Masque UV (en blanc, les zones valides), d) Rendu en clair de l'objet*

3.2 Analyse visuelle

La figure 3 présente l'objet Chaussure issu de la base de données RWTT avec son maillage, sa carte de texture et le masque extrait associé, et un rendu en clair. Enfin la figure 4 illustre l'impact des différentes fonction d'obscurcissement que nous évaluons, chacune sur une valeur de paramètre. Le flou gaussien préserve la structure globale et la silhouette de l'objet tout en atténuant les détails fins. Le chiffrement complet supprime la majorité des indices texturaux, rendant l'interprétation visuelle difficile tout en

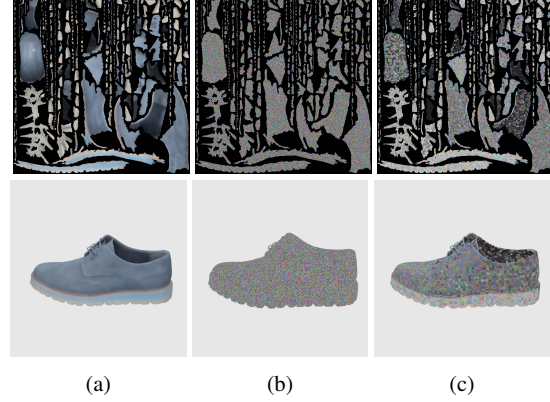


FIGURE 4 – *Obscuration de l'objet 3D texturé Chaussure avec : a) Un flou gaussien, b) Un chiffrement total avec un AES-CTR, c) Une combinaison d'opérations de blocs.*

étant réversible, alors que les opérations de blocs introduisent une rupture spatiale plus marquée, perturbant la cohérence locale tout en conservant la géométrie.

Nous observons que, malgré une altération importante de la texture, la forme géométrique reste perceptible, ce qui confirme que la protection agit principalement sur l'apparence.

3.3 Analyse quantitative

Le tableau 1 présente les résultats obtenus pour les différentes stratégies d'obscurcissement sur l'ensemble des objets de la base RWTT.

Fonction	Param.	Carte de texture				Rendu			
		PSNR	SSIM	ESS	LSS	PSNR	SSIM	ESS	LSS
Flou gaussien	9 px	26.51 ± 3.74	0.74 ± 0.14	0.70 ± 0.09	0.97 ± 0.06	15.26 ± 3.73	0.32 ± 0.14	0.51 ± 0.13	0.56 ± 0.16
	17 px	24.93 ± 3.44	0.70 ± 0.15	0.59 ± 0.11	0.93 ± 0.06	15.12 ± 3.6	0.53 ± 0.14	0.47 ± 0.13	0.56 ± 0.16
	33 px	24.52 ± 3.35	0.69 ± 0.15	0.56 ± 0.12	0.91 ± 0.06	15.09 ± 3.54	0.53 ± 0.14	0.45 ± 0.13	0.56 ± 0.16
AES-CTR	0-6 LSB	13.98 ± 1.03	0.11 ± 0.05	0.61 ± 0.07	0.50 ± 0.13	12.88 ± 2.42	0.26 ± 0.11	0.49 ± 0.11	0.37 ± 0.12
	MSB	9.00 ± 0.00	0.03 ± 0.04	0.58 ± 0.06	0.11 ± 0.08	10.85 ± 2.15	0.17 ± 0.11	0.48 ± 0.11	0.16 ± 0.12
	FULL	8.83 ± 1.24	0.02 ± 0.04	0.53 ± 0.05	0.27 ± 0.14	11.03 ± 2.44	0.16 ± 0.11	0.48 ± 0.10	0.22 ± 0.14
Opérations bloc	8 px	16.81 ± 3.41	0.27 ± 0.11	0.60 ± 0.07	0.60 ± 0.14	13.18 ± 2.93	0.35 ± 0.14	0.51 ± 0.11	0.47 ± 0.14
	16 px	18.94 ± 4.82	0.43 ± 0.13	0.60 ± 0.09	0.71 ± 0.16	13.74 ± 3.2	0.42 ± 0.15	0.53 ± 0.15	0.50 ± 0.14
	32 px	22.16 ± 6.26	0.58 ± 0.14	0.70 ± 0.10	0.82 ± 0.15	14.55 ± 3.86	0.47 ± 0.15	0.54 ± 0.13	0.53 ± 0.15

TABLEAU 1 – *Évaluation de la qualité et de la sécurité au niveau des textures et des rendus des objets obscurcis issus de la base RWTT.*

Concernant le flou gaussien, nous observons qu'il conserve des valeurs plus élevées de PSNR (≈ 25 dB) et de SSIM (≈ 0.7), traduisant la préservation de la structure globale malgré la suppression des détails fins. L'ESS diminue avec la taille du noyau, indiquant une atténuation progressive des contours, tandis que le LSS reste élevé, montrant que la distribution globale de luminance est peu affectée. Le flou correspond ainsi à une dégradation modérée, principalement localisée sur les hautes fréquences.

Le chiffrement AES-CTR des six premiers bits de poids faibles provoque une forte baisse du PSNR (≈ 14 dB) et du SSIM (≈ 0.1) tout en conservant un ESS (> 0.6) relativement élevé, suggérant la persistance de certains contours. Le chiffrement AES-CTR du bit de poids fort (MSB) et le chiffrement complet conduisent à des valeurs plus faibles de PSNR (≈ 9 dB) et SSIM (≈ 0.03), traduisant une altération profonde de la texture. La chute du LSS, particuliè-

rement marquée pour les MSB, confirme une perturbation significative des contrastes. Le chiffrement AES-CTR induit ainsi une obscurisation plus forte au niveau de la carte de texture.

Les opérations de blocs génèrent des résultats se situant entre le chiffrement sélectif et le flou gaussien, avec un PSNR ≈ 14 dB et un SSIM ≈ 0.27 . Il est à noter avec cette fonction une variation importante des résultats en fonction de la taille des blocs. L'augmentation du PSNR et du SSIM avec la taille des blocs ne traduit pas une obscurisation plus faible due à la méthode, mais résultat de la contrainte du masque UV : des blocs plus grands sont plus difficiles à générer sur les zones du masque de petite taille, réduisant la proportion de texture effectivement obscurcie. Des petites tailles de blocs (8 px) produisent donc une altération plus homogène et un niveau d'obscurisation plus marqué.

Les résultats montrent que les méthodes de floutage préservent une partie importante de la structure globale, tandis que les approches basées sur le chiffrement ou la permutation de blocs induisent une dégradation plus marquée.

Les variations des métriques suivent les mêmes tendances entre les évaluations des textures et celles des rendus, mais de manière atténuée. Cela illustre que l'obscurisation de la texture influence directement le rendu final, tout en étant partiellement compensé par les informations apportées par le maillage et par le processus de rendu (éclairage, projection et ombrage) qui contribuent à réduire l'impact visuel de la dégradation apporté à la carte de texture.

3.4 Analyse sémantique

Afin de mesurer l'impact sur l'interprétation sémantique, nous avons sélectionné 10 objets 3D dont les classes sémantiques appartiennent à ImageNet. Les rendus de ces objets sont évalués par un classifieur InceptionV4 [12] pré-entraîné sur ImageNet qui retourne des taux de reconnaissance en Top-1 et Top-5. Le tableau 2 résume les performances moyennes obtenues sur les objets.

Fonction	Param.	Top-1 (%)	Top-5 (%)	VSL
Flou gaussien	33px	54.27	73.75	Suffisant
	17px	51.43	74.76	Suffisant
	9px	57.38	78.33	Suffisant
AES	0-6	42.04	65.37	Transparent
	MSB	20	42.59	Suffisant
	FULL	10.74	28.33	Confidentiel
Opérations bloc	32 px	17.22	36.11	Suffisant/Confidentiel
	16 px	14.81	28.15	Confidentiel
	8 px	16.3	28.52	Confidentiel

TABEAU 2 – Evaluation sémantique par classification automatique et attribution d'un niveau de sécurité visuelle.

Les résultats montrent une diminution progressive des performances de classification lorsque l'intensité de l'obscurisation augmente en suivant les tendances observées lors de l'analyse quantitative en section 3.3

Les taux de reconnaissance obtenus permettent de relier les configurations expérimentales aux différents VSL. Les configurations de flou ou de faible chiffrement sélectif cor-

respondent à un niveau de sécurité visuelle dit *transparent* ou *suffisant*, où l'objet reste identifiable. En revanche, le chiffrement complet ou les permutations de blocs de petite taille entraînent une chute significative du taux de reconnaissance, correspondant à un niveau de sécurité *confidentiel*.

3.5 Rendu en scène complexe

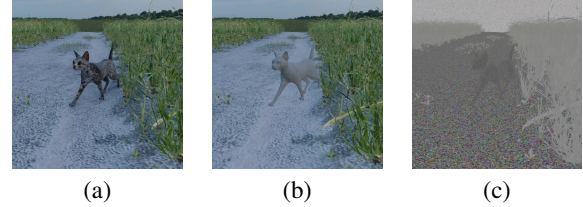


FIGURE 5 – Exemple du rendu d'un objet 3D texturé dans une scène complexe : a) L'objet et la scène ne sont pas protégés, b) L'objet est protégé mais pas la scène, c) L'objet et la scène sont protégés.

Les expérimentations précédentes ont été réalisées dans un cadre contrôlé où les objets étaient rendus sur un fond uniforme afin d'isoler l'impact de l'obscurisation de texture. Dans le contexte des mondes virtuels et Métaverse, les objets 3D texturés sont intégrés dans des scènes complexes comprenant un éclairage dynamique ainsi qu'un arrière-plan structuré, tout en étant en interaction avec d'autres objets.

Dans ce contexte, la perception de la dégradation visuelle peut être influencée par des facteurs supplémentaires tels que les ombres, les réflexions ou le contraste local. Toutefois, la méthode proposée restant indépendante du moteur de rendu et de la configuration de la scène, l'obscurisation appliquée directement à la carte de texture peut s'intégrer à ces scènes. Mais le résultat obtenu reste à déterminer, ce qui pourrait faire place à des expériences futures.

4 Conclusion

Dans cet article, nous avons proposé une méthode de protection de rendu d'objets 3D texturés fondée sur l'obscurisation contrôlée de la carte de texture. En introduisant un masque dérivé du paramétrage UV, la méthode garantit que les transformations affectent uniquement les régions contribuant réellement au rendu, assurant une dégradation visuelle maîtrisée. Les résultats expérimentaux montrent que différentes stratégies d'obscurisations permettent d'atteindre des niveaux de sécurité visuelle variés, tout en préservant l'intégrité géométrique et la compatibilité avec les moteurs de rendu standards. L'analyse sémantique confirme que les configurations les plus agressives réduisent significativement la fuite d'information.

Ces travaux mettent en évidence l'intérêt d'une protection centrée sur la texture pour des scénarios de visualisation contrôlée et de diffusion sécurisée d'objets 3D.

Remerciement

Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet ANR-24-CE39-0987-01 NEXT-LIFE.

Références

- [1] Francois Cayre, Benoit Macq, et Mauro Barni. A survey on watermarking techniques for 3d objects. *IEEE Transactions on Multimedia*, 7(4) :541–555, 2005.
- [2] Sebastien Beugnon, William Puech, et Jean-Pierre Pedebay. From visual confidentiality to transparent format-compliant selective encryption of 3d objects. Dans *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2018.
- [3] Xin Jin, Shuyun Zhu, Le Wu, Geng Zhao, Xiaodong Li, Quan Zhou, et Huimin Lu. Multi-level chaotic maps for 3d textured model encryption. Dans *2nd EAI International Conference on Robotic Sensor Networks : ROSENET 2018*, pages 107–117. Springer, 2020.
- [4] Ayoub Massoudi, Frédéric Lefebvre, Christophe De Vleeschouwer, Benoit Macq, et J-J Quisquater. Overview on selective encryption of image and video : challenges and perspectives. *Eurasip Journal on information security*, 2008(1) :179290, 2008.
- [5] Sébastien Beugnon, Bianca Jansen Van Rensburg, Naima Amalou, William Puech, et Jean-Pierre Pedebay. A 3D Visual Security (3DVS) score to measure the visual security level of selectively encrypted 3D objects. *Signal Processing : Image Communication*, 108 :116832, 2022.
- [6] Giacomo Boracchi et Alessandro Foi. Modeling the performance of image restoration from motion blur. *IEEE Transactions on Image Processing*, 21(8) :3502–3517, 2012.
- [7] Andrea Maggiordomo, Federico Ponchio, Paolo Cignoni, et Marco Tarini. Real-world textured things : A repository of textured models generated with modern photo-reconstruction tools. *Computer Aided Geometric Design*, 83 :101943, 2020.
- [8] Alain Hore et Djemel Ziou. Image quality metrics : PSNR vs. SSIM. Dans *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.
- [9] Yinian Mao et Min Wu. Security evaluation for communication-friendly encryption of multimedia. Dans *2004 International Conference on Image Processing, 2004. ICIP'04.*, volume 1, pages 569–572. IEEE, 2004.
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, et Li Fei-Fei. Imagenet : A large-scale hierarchical image database. Dans *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [11] MaungMaung AprilPyone et Hitoshi Kiya. Block-wise image transformation with secret key for adversarially robust defense. *IEEE Transactions on Information Forensics and Security*, 16 :2709–2723, 2021.
- [12] Christian Szegedy, Sergey Ioffe, et Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

Quand l'encombrement visuel sert de mécanisme de simplification de maillage en réalité virtuelle

Rivo Andriamanalina^{1,2}

Mohamed-Chaker Larabi¹

Steven Le Moan²

¹ Université de Poitiers, XLIM, CNRS UMR 7252, France

² Norwegian University of Science and Technology (NTNU), Colourlab, Norway

Résumé

En réalité virtuelle (RV), la simplification géométrique via niveaux de détail (LoD) est essentielle pour garantir les performances, mais elle introduit des distorsions visibles susceptibles de dégrader la qualité perçue. En vision périphérique, la perception est fortement contrainte par l'encombrement visuel ("visual crowding"), un phénomène psychophysique où la présence d'objets voisins rend une cible plus difficile à analyser. Nous testons l'hypothèse selon laquelle cet effet peut masquer des artefacts de décimation de maillages et ainsi augmenter le niveau de distorsion toléré en périphérie. Nous présentons une série d'expériences psychophysiques contrôlées en RV (HTC Vive Pro Eye) où des observateurs détectent des distorsions induites par une décimation QEM sur une cible présentée à 15° d'excentricité, avec fixation contrôlée par eye-tracking. Nous manipulons (i) l'espacement latéral cible-distracteurs et (ii) la séparation en profondeur via la disparité binoculaire. Les résultats montrent un effet d'encombrement robuste : des distracteurs proches élèvent le seuil de détection (distorsions plus fortes tolérées), et une grande séparation en profondeur atténue mais n'élimine pas l'effet, même au-delà de 200 arcsec de disparité. L'encombrement visuel induit également un coût temporel (temps de réaction plus long). Ces observations ouvrent la voie à des stratégies de LoD "crowding-aware" pour équilibrer performance et qualité perçue en RV.

Mots clefs

Réalité Virtuelle, Vision Périphérique, Visual Crowding, Niveaux de détail, Décimation QEM, Qualité perçue, Psychophysique.

1 Introduction

En réalité virtuelle (RV), les casques imposent des contraintes fortes de calcul et de latence, car le rendu doit être stéréoscopique et couvrir un large champ de vision. Dans les applications interactives (jeu, simulation), les scènes sont majoritairement composées d'objets maillés, et le coût de rendu dépend directement de la complexité géométrique et du niveau de détail (LoD). La réduction du LoD constitue une stratégie essentielle pour préserver

les performances, cependant elle peut introduire des distorsions visibles susceptibles de dégrader la qualité d'expérience [1, 2, 3, 4].

Une approche prometteuse consiste à exploiter les limites du système visuel humain. Au-delà d'une simple baisse d'acuité, la perception périphérique est fortement contrainte par l'encombrement visuel ("visual crowding"), qui rend plus difficile l'identification d'une cible lorsqu'elle est entourée d'éléments voisins [5, 6, 7]. Ce mécanisme est souvent interprété comme une perte de résolution effective ou un masquage des détails, plutôt qu'une invisibilité complète. Ce phénomène pourrait ainsi réduire la visibilité de modifications géométriques locales, tant que l'apparence globale de l'objet reste préservée.

Malgré une littérature abondante en 2D, l'encombrement visuel reste moins étudié en 3D immersive, alors même que la profondeur (disparité binoculaire et profondeur de fixation) est intrinsèque à la RV. Quelques travaux suggèrent que la séparation en profondeur peut affaiblir l'effet d'encombrement jusqu'à l'annuler [8, 9]. Du côté de l'imagerie, lorsque des versions multi-résolutions ne sont pas fournies, les LoD sont souvent générés automatiquement par décimation, notamment par *Quadric Error Metrics* (QEM), une méthode efficace mais susceptible d'altérer des détails et des silhouettes [10, 11, 12, 13, 14].

Dans ce contexte, nous étudions si l'encombrement visuel en RV peut masquer des artefacts de simplification QEM sur des objets 3D en vision périphérique. Nous mesurons l'effet via des seuils de détectabilité de distorsion et analysons son évolution lorsque l'on manipule (i) l'espacement latéral cible-voisins et (ii) la séparation en profondeur (disparité et profondeur de fixation). Nos contributions sont :

- Démonstration que l'encombrement visuel augmente la tolérance aux distorsions QEM en périphérie, par rapport à une condition non encombrée.
- Caractérisation de l'effet en Réalité Virtuelle via des manipulations de l'espacement latéral et de la disparité binoculaire.
- Mise en évidence des conséquences du "crowding" sur la performance des tâches interactives, notamment sur le temps de décision des participants.

Le reste de l'article est organisé comme suit : la section 2 décrit les expériences psychophysiques menées ; les résul-

tats principaux sont discutés en section 3. Enfin, la section 4 conclut et ouvre des perspectives.

2 Méthode

2.1 Matériel et participants aux expériences

Les expériences ont été conduites sur le casque HTC Vive Pro Eye intégrant un dispositif de suivi oculaire. Les scènes ont été générées avec Unity3D 2022.3.62f2 avec le pipeline de rendu par défaut ("Built-in") de Unity. Les paramètres de rendu ont été simplifiés conformément aux recommandations de *Diaz et al.* [15]. 22 volontaires ont participé à l'étude (17 hommes; 21–30 ans, médiane 25). La participation était volontaire et un consentement éclairé a été recueilli avant l'expérience. L'acuité visuelle était normale ou corrigée (20/15 à 20/20); les participants porteurs de correction ont utilisé leurs lunettes dans le casque. Le test d'Ishihara a confirmé qu'ils avaient tous une vision des couleurs normale. Une courte phase d'entraînement précédait l'acquisition afin de familiariser les observateurs avec la tâche et la réponse.

2.2 Stimuli et conditions

Chaque essai présente une configuration de stimuli composée d'un objet **cible** (celui qui doit être jugé) et, selon la condition, de **voisins** ("distracteurs" ou "flankers") destinés à induire l'encombrement visuel. Tous les objets sont des icosphères non texturées afin de contrôler la forme et d'éviter des indices sémantiques, et sont rendus à taille rétinienne constante. Le point de fixation est placé à **15° d'excentricité** par rapport à la cible. En condition non encombrée ("uncrowded"), seule la cible est affichée. En condition "crowded", la cible est entourée de **6 distracteurs** disposés autour d'elle.

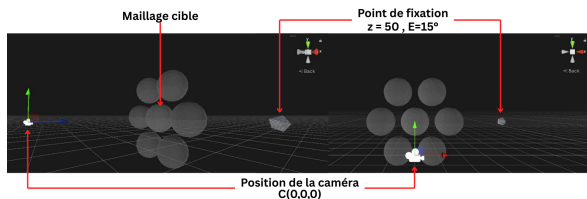


FIGURE 1 – Illustration de la scène de test dans la condition où on a de l'encombrement visuel

La distorsion géométrique est manipulée uniquement sur la cible via une séquence de niveaux de détail (LoD) générés par décimation **QEM**. Le niveau de détail des distracteurs reste fixe afin d'isoler l'effet de masquage dû à l'encombrement. Les LoD de la cible sont indexés par n et suivent une décroissance progressive du nombre de sommets $V(n) = V(0) \cdot 0.95^n$, avec $n = 0$ correspondant au maillage original.

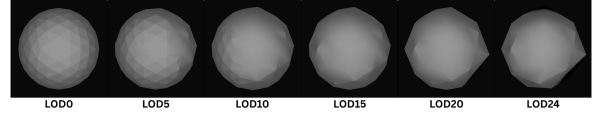


FIGURE 2 – Niveaux de détails du maillage cible

2.3 Variables manipulées

Nous avons étudié deux dimensions spatiales principales :

- **Séparation latérale (plan frontal)**. L'espacement cible–distracteurs est contrôlé via un facteur de Bouma b correspondant au rapport entre l'espacement des stimuli et de l'excentricité de la cible (plus b est faible, plus les distracteurs sont proches). Les valeurs testées sont $b \in \{0.35, 0.40, 0.45, 0.50, 0.55\}$.
- **Séparation en profondeur**. La relation de profondeur entre cible et distracteurs est manipulée via une disparité binoculaire (distracteurs devant/derrière la cible) dans une plage typique de $[-300, +300]$ arcsec, et/ou via la profondeur de fixation selon les conditions expérimentales.

Une condition "uncrowded" (cible seule) sert de référence pour estimer la sensibilité hors encombrement.

2.4 Tâche et procédure

La tâche expérimentale consiste en une détection binaire : à chaque essai, l'observateur répond si une dégradation géométrique est perceptible sur la cible. Le niveau de décimation de la cible est ajusté par une staircase 1-down 1-up avec un pas $\Delta LoD = 1$. La première présentation est le maillage original ($n = 0$), puis la staircase démarre au second essai avec un niveau initial choisi pour être proche de la zone de transition (d'après des essais pilotes). Chaque staircase s'arrête après un nombre fixe de 10 inversions. Le seuil est défini comme la moyenne géométrique des 6 derniers niveaux de renversement.

Afin de garantir l'observation en périphérie, l'observateur doit maintenir sa fixation sur un point qui change de position aléatoirement à chaque essai mais toujours à 15° d'excentricité. Le suivi du regard permet de masquer le stimulus en cas de déviation, assurant que les mesures reflètent bien la vision périphérique.

2.5 Mesures et analyse

Pour chaque condition, nous rapportons :

- un **seuil de détection** de distorsion : le LoD (et le nombre de sommets/faces associé) au seuil, ainsi que la distorsion 2D visible correspondante mesurée sur le viewport (proxies RMSE/SSIM). En effet, les comptes de sommets/faces restent des indicateurs grossiers et ne reflètent pas nécessairement ce qui est visible dans le viewport, notamment en présence d'occlusions. Nous capturons donc des screenshots du viewport et calculons, entre la projection du maillage original et celle du LoD correspondant, deux métriques d'écart d'apparence

simples et standard : RMSE (erreur pixel-à-pixel) et SSIM (similarité structurale). Les métriques sont calculées dans une fenêtre de 6° d'angle visuel centrée sur la cible, afin de capturer les déformations de silhouette observées aux plus forts niveaux de dégradation. La distorsion visible mesurée depuis le viewport varie de façon quasi linéaire avec le niveau de LoD : la RMSE augmente tandis que le SSIM diminue à mesure que la décimation s'intensifie.

- le **temps de réaction moyen** correspondant à la durée nécessaire aux observateurs pour prendre une décision et répondre si oui ou non ils perçoivent une distorsion.

Les comparaisons entre conditions *crowded* et *uncrowded* permettent d'estimer l'impact de l'encombrement visuel. Les effets de la séparation latérale et de la séparation en profondeur sont analysés en fonction des paramètres manipulés (facteur de Bouma, disparité binoculaire, profondeur de fixation) à l'aide de statistiques classiques (tests t et corrélations).

3 Résultats

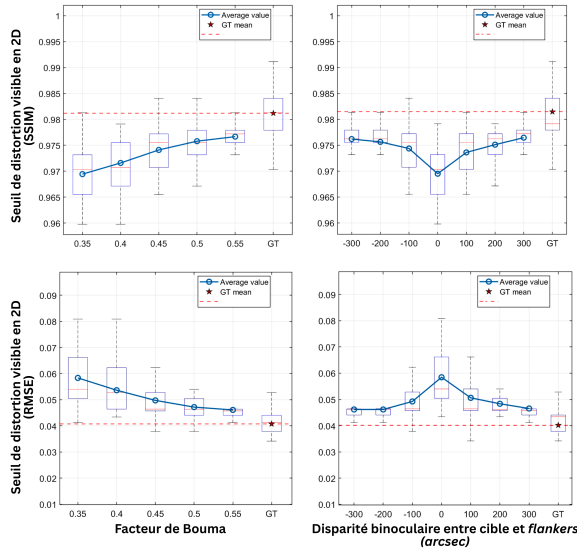


FIGURE 3 – Seuils de perception des distorsions exprimés à l'aide de métriques 2D calculées entre la projection 2D du maillage originale et celle du maillage au moment de la détection. Ligne du haut : seuils exprimés en SSIM. Ligne du bas : seuils exprimés en RMSE. Colonne de gauche : encombrement latéral en fonction du facteur de Bouma. Colonne de droite : encombrement axial en fonction de la disparité binoculaire entre la cible et les voisins (distracteurs). Les boîtes à moustaches montrent la distribution des seuils entre observateurs, tandis que la courbe bleue indique le seuil moyen pour chaque condition. La colonne GT (étoile rouge) et la ligne rouge pointillée représentent le seuil moyen de référence en condition non encombrée (*uncrowded*).

Les résultats des expériences sont présentés dans la figure 3. En exprimant les seuils en termes de distorsion visible dans le viewport (RMSE et SSIM), nous observons le même comportement que lorsque les niveaux sont indexés directement par le LoD, en raison de la relation quasi linéaire entre ces paramètres.

Pour l'**encombrement latéral**, les voisins proches (facteur de Bouma faible) déplacent les seuils vers des distorsions plus fortes : au moment où les participants déclarent percevoir la dégradation, la RMSE est plus élevée et le SSIM plus faible que dans la condition non encombrée. Autrement dit, la présence de distracteurs latéraux limite la perception des détails et retarde la détection des distorsions : il faut une altération plus marquée de l'apparence pour dépasser le seuil. À mesure que le facteur de Bouma augmente (voisins plus éloignés), les seuils se rapprochent progressivement de la référence non encombrée, ce qui traduit une libération graduelle de l'encombrement.

Pour l'**encombrement axial**, l'effet est maximal lorsque la cible et les distracteurs sont au même plan en profondeur (disparité nulle) : les seuils correspondent alors aux distorsions les plus fortes (RMSE maximale, SSIM minimal). Lorsque la séparation en profondeur augmente (disparité positive ou négative), les seuils diminuent et se rapprochent de la condition non encombrée, indiquant que la profondeur atténue l'encombrement. Néanmoins, cette atténuation n'équivaut pas à une suppression complète : même à forte disparité, les seuils restent généralement décalés par rapport à la référence, suggérant un encombrement résiduel.

Enfin, RMSE et SSIM conduisent aux mêmes conclusions, ce qui renforce la robustesse de l'analyse. Le SSIM présente une relation légèrement plus linéaire avec la sévérité de décimation, mais sa dynamique est limitée par une saturation près de 1 ; à l'inverse, la RMSE offre une variation plus étendue. Leur concordance indique que l'effet d'encombrement observé est stable et ne dépend pas du choix du proxy 2D.

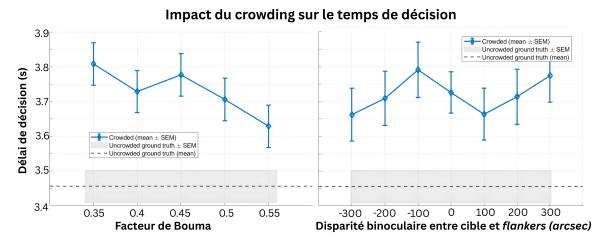


FIGURE 4 – Impact de l'encombrement visuel sur le temps de réaction moyen des observateurs

La figure 4 nous montre que dans toutes les conditions encombrées, le temps de réaction (RT) est systématiquement plus long que dans la condition non encombrée. Cela suggère que l'encombrement visuel n'affecte pas uniquement la visibilité des distorsions, mais augmente aussi l'effort de décision (hésitation/temps de traitement). Pour l'encombrement latéral, on observe une tendance faible mais co-

hérente : lorsque le facteur de Bouma augmente (voisins plus éloignés), le RT moyen diminue légèrement, en accord avec une réduction progressive de l'encombrement. À l'inverse, pour l'encombrement axial, les RT restent élevés pour toutes les disparités sans profil systématique, ce qui indique que la séparation en profondeur ne module pas clairement le coût temporel dans nos données, contrairement à l'espacement latéral. En pratique, cela met en évidence un coût temporel de l'encombrement susceptible d'impacter la réactivité perçue et l'efficacité d'interaction en RV.

4 Conclusion et perspectives

Nous avons étudié comment l'encombrement visuel influence la perception des distorsions géométriques introduites par une décimation de maillages basée sur QEM en RV immersive. Des expériences psychophysiques contrôlées (fixation imposée, taille rétinienne constante) montrent que l'encombrement élève significativement les seuils de détection des distorsions en vision périphérique, et que cet effet se retrouve de manière cohérente pour des configurations latérales et axiales.

Au-delà d'une description en termes de LoD ou de compte de sommets, nous avons reformulé les seuils à partir du point de vue de l'observateur, via des mesures 2D calculées sur le viewport (RMSE et SSIM). Les tendances d'encombrement restent inchangées avec ces proxies, ce qui renforce la robustesse des conclusions et suggère qu'une caractérisation centrée sur l'apparence visible peut faciliter la comparaison entre stratégies de décimation.

Cette étude se limite volontairement à une géométrie simple et contrôlée ; l'extension à des maillages plus complexes, à l'influence de la sémantique et des textures, ainsi qu'à des conditions d'observation plus proches d'usages RV naturels, constitue une suite directe. Globalement, ces résultats positionnent l'encombrement visuel comme un mécanisme perceptif clé pour comprendre l'expérience de la simplification géométrique en RV et pour orienter des des stratégies de rendu conciliant performance et qualité perçue.

Références

- [1] Chaminda TER Hewage et Maria G Martini. Quality of experience for 3d video streaming. *IEEE Communications Magazine*, 51(5) :101–107, 2013.
- [2] Patrick Seeling. Visual user experience difference : Image compression impacts on the quality of experience in augmented binocular vision. Dans *2016 13th IEEE Annual Consumer Communications & Networking Conference*, pages 924–929. IEEE, 2016.
- [3] Jindong Liang, Ping Xie, Xin Yang, et Yanping Wu. Performance analysis of improvemental lod technology under vr headsets. Dans *Third international conference on artificial intelligence, virtual reality, and visualization*, volume 12923, pages 402–409. SPIE, 2023.
- [4] Francisco Lopez, Ramon Molla, et Veronica Sundstedt. Exploring peripheral lod change detections during interactive gaming tasks. Dans *7th Symposium on Applied Perception in Graphics and Visualization*, pages 73–80, 2010.
- [5] David Whitney et Dennis M Levi. Visual crowding : A fundamental limit on conscious perception and object recognition. *Trends in cognitive sciences*, 15(4) :160–168, 2011.
- [6] Dennis M Levi et Thom Carney. Crowding in peripheral vision : Why bigger is better. *Current biology*, 19(23) :1988–1993, 2009.
- [7] Denis G Pelli, Melanie Palomares, et Najib J Majaj. Crowding is unlike ordinary masking : Distinguishing feature integration from detection. *Journal of vision*, 4(12) :12–12, 2004.
- [8] Andrew T. Astle, David P. McGovern, et Paul V. McGraw. Characterizing the role of disparity information in alleviating visual crowding. *Journal of Vision*, 14(6) :1–14, 2014.
- [9] Lisa Valentina Eberhardt et Anke Huckauf. Crowding in depth for binocular and monocular observation. *Attention, Perception, & Psychophysics*, 81 :1951–1961, 2019.
- [10] David Luebke, Martin Reddy, Jonathan D Cohen, Amitabh Varshney, Benjamin Watson, et Robert Huebner. *Level of detail for 3D graphics*. Elsevier, 2002.
- [11] William J Schroeder, Jonathan A Zarge, et William E Lorensen. Decimation of triangle meshes. Dans *19th annual conference on Computer graphics and interactive techniques*, pages 65–70, 1992.
- [12] Michael Garland et Paul S Heckbert. Surface simplification using quadric error metrics. Dans *24th annual conference on Computer graphics and interactive techniques*, pages 209–216, 1997.
- [13] Yan Tang et Qingchen Zhang. Edge-collapse mesh simplification method based on gauss curvature. Dans *2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing*, pages 660–662. IEEE, 2011.
- [14] Kanchan Bahirat, Chengyuan Lai, Ryan P Mcmahon, et Balakrishnan Prabhakaran. Designing and evaluating a mesh simplification algorithm for virtual reality. *ACM transactions on multimedia computing, communications, and applications*, 14(3s) :1–26, 2018.
- [15] Francisco Díaz-Barrancas, Raquel Gil Rodríguez, Florian S Bayer, Avi Aizenman, et Karl R Gegenfurtner. High-fidelity color characterization in virtual reality across head mounted displays, game engines, and materials. *Optics Express*, 32(13) :22388–22409, 2024.

Une représentation 3DGS de rangs faibles auto-sélectionnés pour les avatars multi-identités

Oumaima BADI¹, Xiaoran JIANG¹, Luce MORIN¹, Mårten SJÖSTRÖM²

¹Univ. Rennes, INSA Rennes, CNRS, IETR, UMR 6164, Rennes, France

²Mid Sweden University, Sundsvall, Sweden

¹prenom.nom@insa-rennes.fr ²prenom.nom@miun.se

Résumé

Nous abordons l'apprentissage, à partir de vidéos monoculaires, d'une représentation partagée et compacte d'avatars humains multi-identités. Les approches monoculaires actuelles optimisent en général un modèle par sujet, ce qui entraîne des coûts d'entraînement et de mémoire qui croissent linéairement avec le nombre d'identités. Nous proposons une décomposition Tensor Train (TT) du tenseur global des paramètres des Gaussiennes 3D décrivant les avatars en pose canonique. Afin d'exploiter la géométrie, nous réordonnons les gaussiennes selon une courbe de Hilbert 3D, puis nous réindexons la dimension correspondante sous forme d'un arrangement multi-dimensionnel préservant la proximité spatiale. Enfin, nous introduisons une sélection automatique des rangs TT intégrée à l'optimisation, supprimant le réglage manuel et contrôlant le compromis qualité-complexité. Notre modèle atteint une qualité de reconstruction et d'animation comparable à la méthode de l'état de l'art MIGS, tout en n'utilisant qu'environ 12.3% de ses paramètres, ce qui le rend adapté aux contraintes de déploiement.

Mots clefs

Avatars humains, 3D Gaussian Splatting, Tensor Train, sélection automatique des rangs, multi-identités.

1 Introduction

Apprendre des avatars humains photoréalistes et animables à partir de vidéos monoculaires est essentiel pour de nombreuses applications (AR/VR, cinéma, sport), tout en évitant des dispositifs multi-vues coûteux. La tâche reste toutefois ambiguë (profondeur, occultations) ; en pratique, des a priori cinématiques tels que SMPL [1] ramènent les sujets dans une pose canonique et stabilisent l'apprentissage.

La plupart des approches monoculaires apprennent un modèle par identité, qu'il s'agisse de représentations implicites (NeRF) [2, 3] ou explicites (3DGS) [4, 5]. Si la qualité de rendu est élevée, l'entraînement et la mémoire croissent linéairement avec le nombre d'identités, ce qui limite la mise à l'échelle. Pour améliorer la généralisation inter-identités, deux familles de solutions se dégagent : (i) des priors universels adaptés à chaque sujet [6], et (ii) des représentations partagées de faible rang, exploitant les cor-

rélations entre avatars [7]. MIGS [7] compresse les paramètres 3DGS via une décomposition CP [8], mais impose un rang global unique et n'exploite pas la structure spatiale. Dans cette optique, nous avons proposé MISTA [9], une représentation multi-identités fondée sur une décomposition Tensor Train (TT) [10] avec structuration spatiale du mode gaussien ; toutefois, les rangs sont choisis empiriquement et ne proviennent pas d'une optimisation explicite du coût de rendu. Dans ce travail, nous présentons MISTA-AR, une extension de MISTA qui intègre une sélection automatique des rangs TT durant l'optimisation. Il existe des approches de masquage de rangs, telles que MARS [11], fondées sur une formulation probabiliste et une optimisation stochastique ; ce cadre n'est cependant pas adapté à notre objectif, car nous visons une sélection intégrée et déterministe au sein d'un pipeline 3DGS multi-identités, où la stabilité de l'optimisation et le contrôle direct du compromis qualité-complexité sont critiques. Pour répondre à ces contraintes, nous proposons ainsi une stratégie de sélection automatique des rangs qui supprime le réglage manuel de MISTA. Nos résultats montrent que MISTA-AR maintient une qualité compétitive pour un modèle multi-identités compact.

2 Méthode proposée

2.1 MISTA : tensorisation TT à structure spatiale

Principe. Les paramètres canoniques multi-identités sont organisés en un tenseur $\mathbf{X} \in I \times G \times M$, où I est le nombre d'identités, G est le nombre de gaussiennes pour un avatar, et M représente les paramètres par gaussienne. La manipulation de $\mathbf{X}$ devient prohibitive lorsque I croît. Sous une hypothèse de faible rang, MIGS [7] compresse $\mathbf{X}$ via une décomposition CP à rang global fixe ; nous utilisons une décomposition Tensor Train (TT) à rangs multiples, couplée à une structuration spatiale du mode gaussien.

Initialisation et ré-ordonnancement (Hilbert 3D). L'entraînement débute avec une seule identité : les positions des Gaussiennes sont initialisées depuis un maillage SMPL canonique, définissant un tenseur de dimension $1 \times G \times M$. Nous ré-ordonnons ensuite les Gaussiennes par un parcours de Hilbert 3D [12] afin d'aligner l'indexation avec la proximité spatiale.

Réorganisation du mode gaussien. Comme la TT est plus efficace pour des modes équilibrés [10], nous factorisons G en L sous-modes $G = G_1 \cdots G_L$ et obtenons $\hat{\mathbf{X}} \in I \times G_1 \times \cdots \times G_L \times M$. Nous choisissons la factorisation (et sa permutation) qui préserve au mieux la localité en minimisant

$$A(G_1, \dots, G_L) = \sum_{i=1}^G \sum_{j \in \mathcal{N}(i)} \|p_i - p_j\|_2^2, \quad (1)$$

où $p_i \in R^3$ et $\mathcal{N}(i)$ désigne les voisins induits.

Décomposition Tensor-Train. Nous approximations $\hat{\mathbf{X}}$ par une décomposition TT en cœurs $\mathbf{C}_k \in r_{k-1} \times n_k \times r_k$ ($r_0 = r_d = 1, d = L + 2$) :

$$\hat{\mathbf{X}}(i, g_1, \dots, g_L, m) \approx \mathbf{C}_1(i) \mathbf{C}_2(g_1) \cdots \mathbf{C}_{d-1}(g_L) \mathbf{C}_d(m). \quad (2)$$

La décomposition TT est effectuée une seule fois au début de l'entraînement ; l'optimisation porte ensuite sur les cœurs TT et les poids des MLP.

Extension multi-identité. Le premier cœur $\mathbf{C}_1 \in 1 \times 1 \times r_1$ est associé à l'axe des identités. Pour représenter I personnes, nous étendons ce cœur à $\mathbf{C}_1 \in 1 \times I \times r_1$: chaque identité i sélectionne sa propre tranche $\mathbf{C}_1(i, :, :)$, tandis que les cœurs restants $\mathbf{C}_2, \dots, \mathbf{C}_d$ sont partagés entre toutes les identités. Les avatars canoniques reconstruits sont animés via le pipeline 3DGS-Avatar [5] (déformation rigide, non-rigide, apparence) puis rendus par alpha blending. Ce système complet est détaillé dans [9].

2.2 Sélection automatique des rangs TT

La décomposition TT permet des rangs multiples $\{r_k\}$, mais leur choix contrôle directement le compromis qualité-compacité. Plutôt que de fixer manuellement ces rangs et de les choisir de manière empirique en testant plusieurs configurations et en comparant la taille du modèle à la performance, nous proposons MISTA-AR (AutoRank) une extension de MISTA [9] qui permet d'apprendre des rangs effectifs en introduisant, pour chaque rang interne r_k ($k = 1, \dots, d - 1$), un vecteur de logits apprenables $\vec{\phi}^{(k)} \in R^{r_k}$ qui détermine les dimensions à conserver ou à supprimer.

Masque continu. On définit d'abord un masque continu $\vec{s}^{(k)} \in (0, 1)^{r_k}$, qui sera ensuite binarisé :

$$\vec{s}^{(k)}(t) = \sigma \left(\frac{\vec{\phi}^{(k)}}{T(t)} \right), \quad (3)$$

où t désigne l'itération d'entraînement, $\sigma(\cdot)$ la fonction sigmoïde et $T(t) > 0$ une température décroissante. Lorsque $T(t)$ est grande (début d'entraînement), la transition de $\sigma(\cdot)$ est douce et laisse passer des gradients stables. Quand $T(t) \rightarrow T_{\min}$ (avec $T_{\min} > 0$), $\vec{s}^{(k)}(t)$ se rapproche d'un masque binaire, ce qui rend la sélection de rang plus franche (transition progressive vers une décision discrète). Toutefois, appliquer directement $\vec{s}^{(k)}$ lors de la contraction

TT multiplie chaque composante de rang par une valeur comprise entre 0 et 1, sans jamais atteindre exactement l'une ou l'autre de ces bornes. Toutes les composantes, y compris celles nécessaires à une bonne reconstruction, sont donc partiellement atténuées, ce qui dégrade la qualité du rendu. Pour éviter cela, nous binarisons le masque pendant la passe avant.

Masque binaire avec straight-through estimator. Nous binarisons le masque continu $\vec{s}^{(k)}$ par seuillage :

$$\vec{m}^{(k)}(t) = \mathbf{1} \left[\vec{s}^{(k)}(t) > \tau \right] \in \{0, 1\}^{r_k}, \quad (4)$$

où $\mathbf{1}[\cdot]$ désigne la fonction indicatrice (valant 1 si la condition est vérifiée, 0 sinon) et $\tau = 0,5$ est le seuil de binarisation. Pour conserver la possibilité d'optimiser $\vec{\phi}^{(k)}$ par descente de gradient, nous utilisons l'estimateur *straight-through* (STE) : la passe avant utilise $\vec{m}^{(k)}(t)$, tandis que la passe arrière propage le gradient à travers $\vec{s}^{(k)}(t)$,

$$\frac{\partial \mathcal{L}}{\partial \vec{\phi}^{(k)}} \approx \frac{\partial \mathcal{L}}{\partial \vec{m}^{(k)}} \frac{\partial \vec{s}^{(k)}}{\partial \vec{\phi}^{(k)}}. \quad (5)$$

L'utilisation du masque binaire lors de la passe avant garantit que chaque composante de rang, indexée par $j \in \{1, \dots, r_k\}$, est soit pleinement conservée ($m_j^{(k)}(t) = 1$), soit totalement désactivée ($m_j^{(k)}(t) = 0$) évitant ainsi toute atténuation partielle lors de la contraction TT. Le rang effectif $r_k^* = \sum_j m_j^{(k)}$ est alors déterminé automatiquement par l'optimisation.

Masquage cohérent des cœurs adjacents. Le masque $\vec{m}^{(k)}$ agit sur l'axe de rang partagé entre $\mathbf{C}_k$ et $\mathbf{C}_{k+1}$:

$$\begin{aligned} \tilde{\mathbf{C}}_k(:, :, j) &= \mathbf{C}_k(:, :, j) m_j^{(k)}, \\ \tilde{\mathbf{C}}_{k+1}(j, :, :) &= m_j^{(k)} \mathbf{C}_{k+1}(j, :, :). \end{aligned} \quad (6)$$

Cette opération annule exactement les composantes désactivées sans modifier la contraction TT.

Pénalité de parcimonie et rangs finaux. Pour encourager une solution compacte, nous ajoutons une pénalité ℓ_1 sur les masques continus $\vec{s}^{(k)}$ (plutôt que sur le masque binaire $\vec{m}^{(k)}$) :

$$\mathcal{R}(t) = \lambda(t) \sum_{k=1}^{d-1} \|\vec{s}^{(k)}(t)\|_1, \quad (7)$$

avec un coefficient $\lambda(t)$ activé après une phase d'échauffement, afin de ne pas contraindre la représentation avant stabilisation du rendu. La fonction objectif devient alors

$$\mathcal{L}(t) = \mathcal{L}_{\text{rend}}(t) + \mathcal{R}(t), \quad (8)$$

où $\mathcal{L}_{\text{rend}}$ désigne la fonction de coût de rendu (photométrique et régularisations [5]). En fin d'entraînement, les rangs effectifs sont obtenus par seuillage :

$$r_k^* = \sum_{j=1}^{r_k} \mathbf{1} \left[\sigma(\phi_j^{(k)}) > \tau \right], \quad (9)$$

Méthode	SSIM↑	PSNR↑	LPIPS* ↓	#Param↓
MIGS ($r=100$) [7]	0.9692	30.06	35.34	5.01M
MIGS ($r=10$) [7]	0.9667	29.59	45.24	501k
MISTA (fixe) [9]	0.9691	29.98	36.50	326k
MISTA-AR	0.9681	29.82	38.30	616k

TABEAU 1 – Comparaison sur ZJU-MoCap [13]. $LPIPS^* = LPIPS \times 10^3$. Les MLP étant identiques, #Param compte uniquement les paramètres gaussiens/TT.

ce qui définit le masque binaire $\vec{m}^{(k)}$. Comme illustré en Figure 1, ces masques sont appliqués sur les dimensions de rang adjacentes des cœurs TT (via un produit extérieur, diffusé sur les n_k tranches). Nous supprimons définitivement les dimensions j telles que $m_j^{(k)} = 0$ dans les cœurs TT.

3 Résultats expérimentaux

3.1 Protocole expérimental

Nous évaluons sur le jeu de données ZJU-MoCap [13] (9 sujets, 23 vues). Nous entraînons sur 8 identités et effectuons un *fine-tuning* sur le dernier sujet pour évaluer la généralisation. Comme dans MIGS [7], l’adaptation à une nouvelle identité ajoute une tranche sur l’axe identité du premier cœur TT et n’optimise que cette tranche avec le MLP couleur, le reste étant gelé. Nous entraînons sur la vue 1 et testons sur les 22 autres, en utilisant les 555 premières frames pour tous les sujets. Sur une RTX 2080 Ti, l’entraînement prend environ 4 h 20 min par modèle. Tous les modèles sont entraînés pendant 50×10^3 itérations avec Adam [14] en suivant [5, 7]. Le MLP rigide est actif dès le début ; les modules sont dégelés à 10^3 itérations, sauf le non-rigide activé à 3×10^3 . Nous utilisons 50×10^3 Gaussiennes sans densification, et implémentons la décomposition avec TensorLy [15]. Nous reportons les métriques PSNR, SSIM [16] et LPIPS [17] (avec $LPIPS^* = LPIPS \times 10^3$). La compacité est mesurée par le nombre de paramètres des Gaussiennes (MLP identiques), et nous complétons ces métriques par une évaluation qualitative, car elles ne suffisent pas à capturer certains artefacts visuels.

3.2 Résultats et analyse

Compromis qualité-compacité. Le Tableau 1 met en évidence un équilibre clair entre fidélité visuelle et compacité des représentations : à budget paramétrique comparable, MISTA (fixe) atteint des scores SSIM/PSNR très proches de MIGS ($r=100$) tout en réduisant drastiquement le nombre de paramètres gaussiens/TT, et surpasse nettement la version fortement comprimée de MIGS ($r=10$) en perceptuel (LPIPS). Dans le même temps, MISTA-AR effectue la compression en réalisant un choix automatique et non plus manuel des rangs (et donc adaptable aux données), au prix d’une légère dégradation mesurable sur les métriques, sans rupture brutale de qualité : les valeurs restent proches de MISTA (fixe) et de MIGS ($r=100$), ce qui suggère que l’adaptation du rang capture l’essentiel là où il compte. Cette lecture quantitative se confirme qualita-

tivement dans la Fig. 3 : par rapport à MIGS ($r=100$), MISTA (rangs fixes (1, 51, 73, 37, 32, 1)) et MISTA-AR (rangs effectifs (1, 80, 77, 79, 32, 1)) préservent globalement les contours et l’apparence globale, tandis que l’effet de la compression adaptative se manifeste surtout par de subtiles pertes de micro-détails (textures fines, zones à fort contraste) plutôt que par des artefacts structuraux marqués.

Rangs effectifs sélectionnés. La Fig. 2 montre l’évolution des rangs effectifs r_k^* au cours de l’entraînement à partir d’un rang uniforme $R = 80$. Durant l’échauffement (itérations 0–15 000), les logits $\vec{\phi}$ sont gelés et les rangs restent inchangés. Après activation, seul r_4 décroît nettement et se stabilise à $r_4^* = 32$ (soit -60%), tandis que r_1 , r_2 et r_3 sont largement préservés ($r_1^* = 80$, $r_2^* = 77$, $r_3^* = 79$).

Cette sélection non uniforme reflète le rôle des rangs TT : r_4 ne couple que le bloc des modes amonts avec le dernier mode, dont la structure est plus simple et donc plus compressible. À l’inverse, les rangs internes (r_2 , r_3) portent les corrélations croisées entre plusieurs modes et nécessitent davantage de capacité ; r_1 reste élevé car il capture l’essentiel de la variabilité inter-identités dans le premier cœur partagé. Enfin, le gel des logits après 40 000 itérations stabilise la structure de rang et laisse les cœurs TT s’ajuster en fin d’entraînement.

4 Conclusion

Nous avons présenté une représentation compacte d’avatars gaussiens multi-identités, fondée sur une décomposition Tensor Train, structurée spatialement par un réordonnement de type Hilbert et une réorganisation du mode des Gaussiennes. Nous avons également introduit une sélection automatique des rangs TT intégrée à l’optimisation, permettant d’ajuster la capacité du modèle sans réglage manuel et de mieux contrôler le compromis qualité-compacité. Malgré cette compacité, la décomposition TT peut lisser certains détails fins, notamment les textures vestimentaires et les traits du visage. À terme, nous visons un entraînement sur davantage d’identités afin de renforcer la généralisation du modèle partagé.

5 Remerciement

Cette étude est soutenue par l’ANR dans le cadre du projet PIA EUR DIGISPORT (ANR-18-EURE-0022).

Références

- [1] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, et M. J. Black. Smpl : A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6) :248 :1–248 :16, 2015.
- [2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, et R. Ng. Nerf : Representing scenes as neural radiance fields for view synthesis. Dans *ECCV*, 2020.
- [3] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, et I. Kemelmacher-Shlizerman. Humannerf : Free-

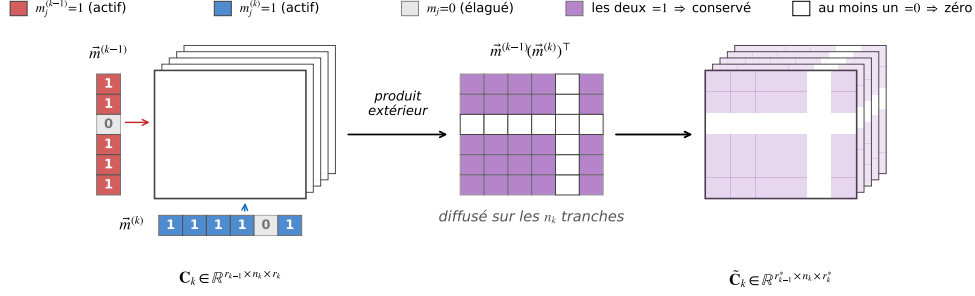


FIGURE 1 – Masquage d’un cœur TT. Les deux masques de rang adjacents sont combinés par produit extérieur, puis appliqués à chaque tranche du cœur. Seules les composantes où les deux masques sont actifs (violet) sont conservées.

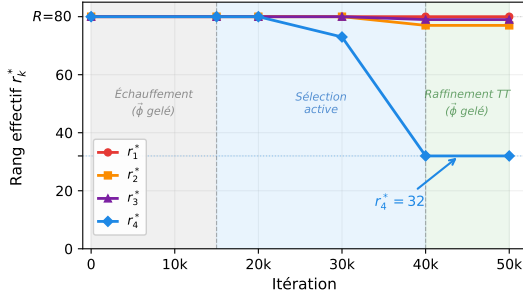


FIGURE 2 – Évolution des rangs effectifs au cours de l’entraînement.

viewpoint rendering of moving people from monocular video. Dans *CVPR*, pages 16210–16220, 2022.

- [4] B. Kerbl, G. Kopanas, T. Leimkuehler, et G. Dretakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4) :139 :1–139 :14, 2023.
- [5] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, et S. Tang. 3dgs-avatar : Animatable avatars via deformable 3d gaussian splatting. Dans *CVPR*, 2024.
- [6] C. Guo, J. Li, Y. Kant, Y. Sheikh, S. Saito, et C. Cao. Vid2avatar-pro : Authentic avatar from videos in the wild via universal prior. Dans *CVPR*, 2025.
- [7] A. Chatziagapi, G. G. Chrysos, et D. Samaras. Migs : Multi-identity gaussian splatting via tensor decomposition. Dans *ECCV*, 2024.
- [8] T. G. Kolda et B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3) :455–500, 2009.
- [9] O. Badi, X. Jiang, L. Morin, et M. Sjostrom. Mista : Compact multi-identity structure-aware tensorized avatars. Dans *ICASSP*, 2026.
- [10] I. V. Oseledets. Tensor-train decomposition. *SIAM J. Sci. Comput.*, 33(5) :2295–2317, 2011.
- [11] M. Kodryan, D. Kropotov, et D. Vetrov. Mars : Masked automatic ranks selection in tensor decompositions. Dans *AISTATS*, pages 3718–3732, 2023.
- [12] D. Hilbert. Über die stetige abbildung einer linie auf ein flächenstück. *Math. Ann.*, 38(3) :459–460, 1891.

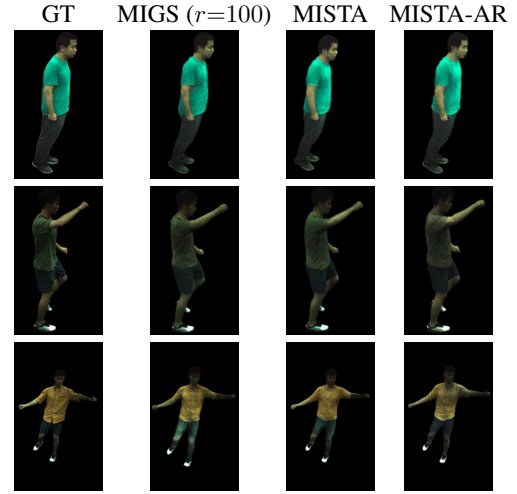


FIGURE 3 – Comparaison qualitative sur trois sujets vus à l’entraînement : synthèse de vue nouvelles

- [13] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, et X. Zhou. Neural body : Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. Dans *CVPR*, pages 9054–9063, 2021.
- [14] D. P. Kingma et J. Ba. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014.
- [15] J. Kossaifi, Y. Panagakis, A. Anandkumar, et M. Pantic. Tensorly : Tensor learning in python. *J. Mach. Learn. Res.*, 20(26) :1–6, 2019.
- [16] Z. Wang, A. C. Bovik, H. R. Sheikh, et E. P. Simoncelli. Image quality assessment : From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4) :600–612, 2004.
- [17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, et O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. Dans *CVPR*, pages 586–595, 2018.

Zone-Adaptive Depth Encoding for Volumetric Video Compression : YUV420p Analysis

A. Rehman¹

F. Bouille²

^{1,2} French Touch Factory, France

{abdul, f.bouille}@frenchtouchfactory.com

(¹ Postdoctoral Researcher, ² Research Director)

Abstract

Volumetric video capture using RGB-D sensors needs an efficient mapping of depth data to color encoding prior to compression, to achieve the parallel, side-by-side export of color and depth data streams. Although standard codecs like H.264 achieve better compression ratios, they introduce reconstruction errors because of YUV420p chroma sub-sampling on color encoded depth data. We propose histogram-based zone-adaptive encoding of depth data into the TURBO colormap to minimize the impact of YUV420p chroma sub-sampling. Cross-dataset validation (TUM RGBD fr1/desk, fr1/room, and Kinect Azure) demonstrates an average MAE of 15.3 depth units (averaged across all datasets), with per-dataset MAE values of 7.8 (desk), 11.1 (room), and 27.0 (Kinect). The proposed TURBO encoding achieves 45% MAE reduction compared to HSV (averaged 27.7 vs 15.3) and 61% compared to LAB (39.5 vs 15.3).

Keywords

Volumetric video, Depth compression, Depth encoding, YUV420p, Video codecs

1 Introduction

RGB-D sensors (Kinect Azure, Orbbec) generate 3.6 Gps of uncompressed data at a resolution of 640×576 with a 9-camera setup. After fusing the depth stream, the data is exported in side-by-side format (color and depth) into an MP4 file. For visualization and downstream 3D reconstruction, the depth data must be encoded into a color representation compatible with standard video codecs. However, during compression, significant errors are introduced due to YUV420p chroma sub-sampling, which discards 50% of chrominance information [1]. In contrast to traditional video compression applications, volumetric video requires precise depth preservation to achieve accurate 3D reconstruction.

We propose a histogram-based zone boundary method that removes the need for manual threshold selection. We also provide a cross-colorspace comparison showing the robustness of TURBO encoding for accurate depth recovery after

H.264/YUV420p compression.

2 Related Work

In the context of volumetric video capture, depth compression has been studied in point cloud and video-based frameworks. MPEG V-PCC [2] projects 3D point clouds onto 2D patches for standard video compression but requires separate custom pipelines. Deep learning-based compression methods [3] give high rate-distortion performance but depend on specialized decoders, constraining hardware deployment.

Depth-aware encoding methods [4] and texture-based approaches [5] have been proposed, but these ignore chroma sub-sampling, introducing reconstruction errors. Multi-view systems using standard codecs [6] further restrict widespread deployment.

YUV420 chroma sub-sampling is the primary source of reconstruction error and at the same time a key factor in compression efficiency. YUV444 preserves chrominance and reduces errors [7], but is rarely supported by consumer standards. YUV420 remains the dominant format for H.264/H.265 due to bandwidth and decoder restrictions [8].

In contrast, we work within standard H.264 and YUV420 constraints, proposing a depth-to-color encoding that minimizes reconstruction errors while maintaining full compatibility with existing hardware.

3 Methodology

3.1 Problem Formulation

We consider a depth sequence $\{D_t\}_{t=1}^T$ where each frame D_t contains depth values in the range $[0, 65535]$ with size $H \times W$. The goal is to minimize the recovery error between the original depth frames D_t and the decoded frames $\hat{D}_t$. The objective function is :

$$\min_{\theta} \mathbb{E}_t \left[\|D_t - \hat{D}_t\|_1 \right] + \lambda \mathbb{E}_t \left[\|\Delta \hat{D}_t - \Delta D_t\|_1 \right], \quad (1)$$

where $\|D_t - \hat{D}_t\|_1$ is the L1 reconstruction error, $\|\Delta \hat{D}_t - \Delta D_t\|_1$ is the temporal consistency constraint, and $\lambda = 0.1$ balances accuracy and smoothness.

3.2 Encoding Pipeline

The encoding of depth data into color-coded video consists of the following steps, as shown in Figure 1.

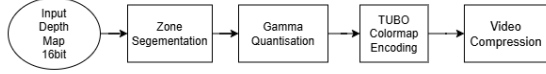


FIGURE 1 – *Depth map encoding pipeline : zone-adaptive processing, TURBO colormap encoding, and H.264 compression.*

Histogram-Driven Zone Segmentation. Manual depth thresholds often fail to capture the actual distribution of a scene. To address this, we compute zone boundaries from the depth histogram at target cumulative proportions, defining near-, mid-, and far-field regions to contain approximately 40%, 35%, and 25% of valid pixels, respectively. For the experimental dataset, the optimized boundaries $b_1 = 1200$ and $b_2 = 2584$ (with implicit $b_0 = 0$ and $b_3 = \max(D)$) yield pixel distributions of 44.4%, 37.8%, and 17.8% for near-, mid-, and far-field zones. Compared to fixed thresholds of (2000, 5000), this histogram-based approach reduces MAE by 19.4%, from 48.0 to 38.7, demonstrating the efficiency of adaptive zone selection. This ablation is measured prior to TURBO colormap application and k-NN decoding.

Zone-Adaptive Gamma Quantization. To minimize quantization error within each zone, we apply a power-law transformation before mapping to the $[0, 255]$ range. The quantization operation is :

$$E(\gamma) = \mathbb{E}_{d \in D_z} \left[(d - \text{Dequant}(\lfloor \text{Norm}(d)^\gamma \cdot 255 \rfloor, \gamma))^2 \right], \quad (2)$$

where D_z denotes depth values within zone z , $\text{Norm}(d) = (d - b_{z-1}) / (b_z - b_{z-1})$ normalizes depth d to $[0, 1]$, and $\text{Dequant}(q, \gamma) = b_{z-1} + (q/255)^{1/\gamma} \cdot (b_z - b_{z-1})$ maps quantized values back to the original depth range. A uniform gamma of $\gamma = 0.45$ is applied across all zones, selected empirically to minimize quantization error across the full depth range.

TURBO Colormap Encoding. We use the TURBO colormap [9] to encode depth into color because of its monotonically increasing luminance and perceptually uniform transitions [10], avoiding the hue-wrapping artifacts inherent to HSV colormaps. After zone-adaptive gamma quantitation, each uint8 depth value is mapped directly to the corresponding TURBO RGB triple via a 256-entry lookup table.

Video Compression. The color-encoded depth frames are compressed using H.264 [11] with CRF 18 and YUV420p chroma sub-sampling. The low CRF value prioritizes reconstruction quality over bitrate [12, 13], minimizing compression-induced depth errors that are critical for 3D reconstruction.

3.3 Decoding Pipeline

The decoding process reconstructs depth values as shown in Figure 2. The YUV420p stream is first decompressed to RGB. Then a tree search for the nearest neighbor k-D tree [14] is performed against the color map lookup table (LUT) with $k = 5$ neighbors. Depth values are estimated via inverse-distance weighting ($w_i = 1/d_i^{1.5}$), yielding a per-pixel quantized index which is finally converted back to metric depth via zone-adaptive de-quantitation.

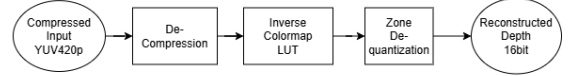


FIGURE 2 – *Depth reconstruction pipeline : YUV decompression, inverse colormap lookup via k-NN ($k = 5$) with IDW weighting, and zone-adaptive quantitation.*

3.4 Temporal Consistency Metrics

The mean temporal jitter $\bar{J}$ and frame-wise MAE variance σ_{MAE}^2 together assess temporal stability across frames :

$$J_t = \left| (\hat{D}_t - \hat{D}_{t-1}) - (D_t - D_{t-1}) \right|, \quad (3)$$

$$\bar{J} = \frac{1}{T-1} \sum_{t=2}^T \mathbb{E}[J_t]. \quad (4)$$

4 Experiments

4.1 Hardware Setup

Experiments were conducted using Kinect Azure sensors at a resolution of 640×576 (16-bit, 30 fps). The side-by-side video containing RGB and color-encoded depth was compressed using H.264 with YUV420p chroma sub-sampling. Evaluated baselines include HSV, a true CIE Lab-space colormap (LAB), and YUV444p (no chroma sub-sampling). Performance is evaluated using MAE, RMSE, SSIM, and the percentage of good pixels (< 30 depth units for cross-dataset comparison; < 5 depth units for zone-level analysis).

4.2 Results

The cross-dataset evaluation (Table 1) shows that TURBO achieves an average MAE of 15.3 across three datasets : 7.8 on TUM desk, 11.1 on TUM room, and 27.0 on Kinect. TURBO outperforms HSV by 41–47% and LAB by 51–67% across all datasets. High SSIM values (0.968–0.981) and good-pixel recovery rates of 90.4–98.6% further confirm strong structural preservation. LAB achieves competitive MAE on TUM datasets (15.9 for desk, 21.3 for room) but its performance degrades significantly on the Kinect data (MAE 81.4, Good 72.9%), due to gamut clipping in the CIE Lab-to-RGB conversion that causes ambiguous color assignments for mid-range depths.

TURBO provides more consistent performance across all datasets and sensor types.

Although LAB achieves lower RMSE than TURBO on TUM desk (64.4 vs 51.1), it degrades severely on Kinect (308.9 vs TURBO 170.0). This, combined with the lowest good-pixel recovery on Kinect (72.9% vs TURBO 90.4%), highlights that RMSE alone is inadequate to evaluate depth reconstruction quality. Overall, TURBO provides stable and consistent performance across datasets.

TABLE 1 – *Cross-Dataset Colorspace Comparison (CRF=18, Good% < 30)*

Dataset	Method	MAE	RMSE	SSIM	Good%
TUM desk	HSV	13.1	67.1	0.966	96.9%
	LAB	15.9	64.4	0.965	94.6%
	TURBO	7.8	51.1	0.975	98.6%
TUM room	HSV	19.1	100.3	0.958	95.7%
	LAB	21.3	89.5	0.957	92.6%
	TURBO	11.1	76.8	0.968	97.9%
Kinect	HSV	50.9	252.3	0.976	84.8%
	LAB	81.4	308.9	0.967	72.9%
	TURBO	27.0	170.0	0.981	90.4%
Average	HSV	27.7	139.9	0.967	92.5%
	LAB	39.5	154.2	0.963	86.7%
	TURBO	15.3	99.3	0.974	95.6%

Bold = best MAE and Good% per dataset.

The zone-specific analysis (Table 2), conducted on a single Kinect frame, shows that Zone 1 (0–1200, 44.4% coverage) achieves 68.7% good pixels and 28.4% perfect reconstruction. Zone 2 (1200–2584, 37.8% coverage) achieves 54.2% good pixels. Zone 3 (>2584, 17.8% coverage) records 41.3% good pixels. These results confirm that histogram-driven zone allocation, which assigns more quantization levels to the dominant near-field region, directly improves reconstruction accuracy.

TABLE 2 – *Zone-Specific Performance (Kinect, single-frame)*

Zone	Cover	Good [†]	Perfect [†]
1 (0–1200)	44.4%	68.7%	28.4%
2 (1200–2584)	37.8%	54.2%	15.7%
3 (>2584)	17.8%	41.3%	8.6%

[†]Threshold : <5 depth units.

The mean temporal jitter is $\bar{J} = 2.6 \pm 1.1$ units (0.04% of range), with frame-wise MAE variance $\sigma_{\text{MAE}}^2 = 3.2$. Zone 1 shows lower jitter (1.7 ± 0.6) compared to Zone 3 (3.9 ± 1.6). Temporal errors primarily occur at sudden depth transitions where YUV420p creates spatial ambiguity. For reference, using YUV444p (no chroma sub-sampling) on a simple baseline encoder reduces MAE to 17.3 compared to 38.7 with YUV420p; the proposed method achieves 15.3 MAE under standard YUV420p, making it practical without requiring YUV444p support.

5 Discussion

TURBO’s superior MAE and good-pixel recovery stem from its monotonically increasing luminance, which distributes depth information across all three YUV channels. This allows depth to be encoded into color for side-by-side export, enabling RGB and depth to be stored in a single MP4 after multi-camera point cloud fusion. In contrast, HSV suffers from hue wrapping ($0^\circ \equiv 179^\circ$), resulting in large reconstruction errors (MAE 50.9, RMSE 252.3 on Kinect), while LAB’s gamut clipping during the CIELab-to-RGB conversion causes depth ambiguities on complex scenes (MAE 81.4, Good 72.9% on Kinect).

Despite these advantages, several limitations remain. Zone boundaries currently rely on per-scene depth histograms and are not adapted online. The far-field Zone 3 jitter of 3.9 ± 1.6 units may impact applications requiring sub-millimeter accuracy. Additionally, the k-D tree nearest-neighbor search remains the main computational bottleneck, accounting for approximately 40% of total processing time.

6 Conclusion

We presented a histogram-based zone-adaptive depth-to-color encoding method that explicitly addresses YUV420p chroma sub-sampling limitations for accurate depth reconstruction. Using a uniform gamma of $\gamma = 0.45$ per zone and a TURBO colormap LUT with $k = 5$ nearest-neighbor decoding, the method achieves an averaged MAE of 15.3 depth units across datasets (7.8 desk, 11.1 room, 27.0 Kinect), a temporal jitter of 2.6 ± 1.1 units, and a 45% and 61% MAE reduction compared to HSV and LAB encoding, respectively. These results enable practical side-by-side volumetric video compression of RGB and depth streams using standard consumer hardware and decoder pipelines.

Références

- [1] H. Schwarz, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, K. Müller, H. Rhee, G. Tech, M. Winken, et T. Wiegand. Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(10) :1935–1953, Octobre 2013.
- [2] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. M. Tourapis, et V. Zakharchenko. Emerging MPEG Standards for Point Cloud Compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1) :133–148, Mars 2019.
- [3] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, et Y. Sheikh. Neural volumes : Learning dynamic renderable volumes from images. *ACM*

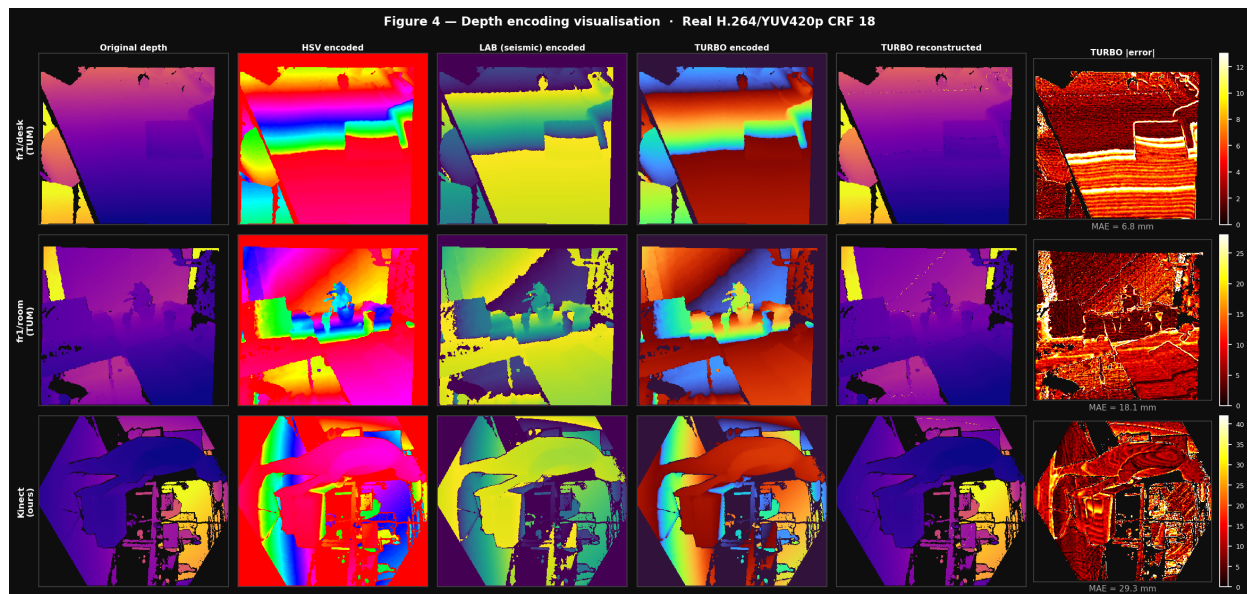


FIGURE 3 – Visual comparison across three datasets. Rows : *fr1/desk*, *fr1/room*, *Kinect*. Columns : (a) original depth, (b–d) encoded (HSV/LAB/TURBO), (e) TURBO reconstructed, (f) absolute error. Error maps (hot colormap) show TURBO’s localized errors vs. HSV’s hue-wrapping and LAB’s gamut-clipping ambiguity.

- Transactions on Graphics (TOG)*, 38(4) :1–14, Juillet 2019.
- [4] Philipp Merkle, Karsten Müller, Detlev Marpe, et Thomas Wiegand. Depth intra coding for 3d video based on geometric primitives. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3) :570–582, 2015.
 - [5] Wei Liu, Xiaogang Chen, Jie Yang, et Qiang Wu. Robust color guided depth map restoration. *IEEE Transactions on Image Processing*, 26(1) :315–327, 2016.
 - [6] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, et Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4) :1–13, 2015.
 - [7] Amritha Premkumar et Christian Herglotz. Adaptive resolution and chroma subsampling for energy-efficient video coding. *arXiv*, 2026. arXiv :2602.06100.
 - [8] Mozilla Bugzilla. Video with yuv444 chroma subsampling can’t be played back. https://bugzilla.mozilla.org/show_bug.cgi?id=1368063, 2018. Accessed 2026.
 - [9] Turbo, an improved rainbow colormap for visualization. Google Research Blog, 2019.
 - [10] F. Crameri, G.E. Shephard, et P.J. Heron. The misuse of colour in science communication. *Nature Communications*, 11 :5444, 2020.
 - [11] T. Wiegand, G. Sullivan, G. Bjøntegaard, et A. Luthra. Overview of the h.264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7) :560–576, 2003.
 - [12] Ching-Yu Wu, Po-Chyi Su, Long-Wang Huang, et Chia-Yang Chiou. Constant frame quality control for h.264/avc. *APSIPA Transactions on Signal and Information Processing*, 2(1) :e3, 2013.
 - [13] Ffmpeg compression parameters : Crf explained. *videocompress.com blog*, 2025.
 - [14] J.L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9) :509–517, 1975.

A Framework for Event-Level Representation of Wearable Healthcare Time Series

Wenda Liu*, Fabien Picarougne, Guillaume Raschia, Vincent Ricordel
Polytech Nantes, Nantes Université
LS2N, UMR 6004, F-44000, Nantes, France
wenda.liu@univ-nantes.fr

2026-03-04

Abstract

Wearable healthcare devices such as smartwatches collect physiological signals and healthcare data, producing large volumes of multivariate time-series data. Extracting meaningful event-level patterns from these time-series and producing further downstream outputs remains challenging due to long temporal dependencies and limited annotations. We propose a modular framework for processing healthcare time-series data in 3 stages : an embedding stage that translates time-series data into embedding representations, a processing stage that discovers event patterns and aggregates temporal information, and an output stage that supports downstream tasks such as prediction and retrieval. Current experiments focus on event-level representation and time-series-text alignment in the embedding stage, with the goal of providing semantic pattern discovery for downstream applications.

Keywords

Wearable healthcare; time-series representation; event-level embedding; multimodal alignment; physiological signals

1 Introduction

Wearable healthcare devices including smartwatches, are increasingly used to collect physiological signals such as heart rate, blood pressure, and other health metrics. These devices produce large volumes of multivariate time-series data that reflect physiological dynamics over long periods of time [1].

However, processing collected healthcare time-series data are challenging. Healthcare time-series often exhibit long temporal dependencies and complex patterns that are difficult to represent using traditional signal processing methods [2]. Moreover, annotations for healthcare events are usually rare, especially considering the privacy and ethical concerns of patients, making it difficult to capture signal patterns with semantic descriptions of events [3] (i.e.,

textual descriptions that provide interpretable meaning for physiological patterns).

To address these challenges, we propose a modular framework for processing wearable healthcare time-series data. The framework contains three stages : embedding, processing, and output generation. This modular design allows different components to evolve independently, thus enabling flexible choices of analysis methods.

While the framework covers the end-to-end processing of healthcare time-series data, from raw signals to final outputs such as prediction and retrieval, the current focus of our experiments is on the embedding stage. The embedding stage aims to learn *event-level patterns* and align time-series data with textual descriptions, where an *event-level pattern* refers to a localized temporal structure in a time-series signal that corresponds to a meaningful physiological or behavioral change within a short time window. This method captures healthcare event patterns and provides a semantic interpretation of physiological signals, enabling a range of downstream tasks including event detection, classification, and forecasting.

2 Framework for Time-Series Processing

The proposed framework processes healthcare time-series data through three main stages : embedding, processing, and output generation. Figure 1 illustrates the overall architecture.

2.1 Embedding Stage

In the embedding stage, raw time-series data are segmented into patches and mapped into an embedding space. Let a multivariate time series be defined as

$$X = (x_1, x_2, \dots, x_T), \quad x_t \in \mathbb{R}^d, \quad 1 \leq t \leq T.$$

The sequence is divided into patches $\{X_i\}_{i=1}^N$, where each patch represents a local temporal pattern. A time-series encoder f_θ maps each patch into an embedding vector [4]

$$z_{time}^{(i)} = f_\theta(X_i).$$

*Acknowledgements : This research was funded by the French National Research Agency (ANR) through the project RENEW (ANR-24-THCS-0002-01).

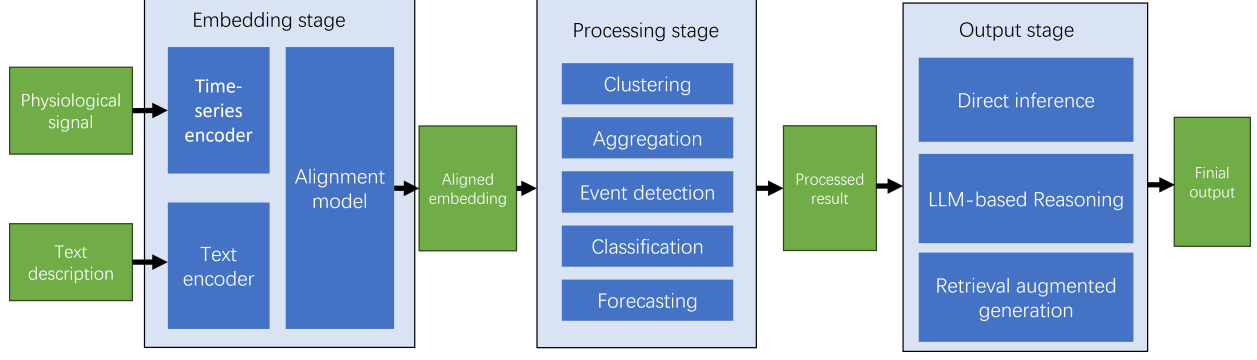


FIGURE 1 – Overview of the proposed framework for processing healthcare time-series data.

In addition, textual descriptions of time-series data as labels or events can be encoded using a text encoder g_ϕ , such as a pre-trained language model,

$$z_{text}^{(j)} = g_\phi(W_j).$$

Here, W_i denotes the textual description corresponding to the time-series patch X_i . The parameters θ and ϕ represent the learnable parameters of the time-series encoder f_θ and the text encoder g_ϕ , respectively.

To link labels and time-series patterns and obtain semantic representations, we align the two modalities in a shared embedding space using a contrastive learning objective. The similarity between a time-series embedding and a text embedding is computed using cosine similarity [5]

$$s_{ij} = \frac{z_{time}^{(i)} \cdot z_{text}^{(j)}}{\|z_{time}^{(i)}\| \|z_{text}^{(j)}\|}.$$

We adopt a contrastive objective similar to the InfoNCE [6] loss to align time-series patches and textual descriptions to have high similarity :

$$\mathcal{L}_{ts \rightarrow text} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^N \exp(s_{ij}/\tau)},$$

where τ is a temperature parameter. A symmetric objective from text to time-series is also computed, and the final loss is defined as the average of the two directions [7]. This objective encourages paired time-series and textual descriptions to be close in the embedding space while pushing mismatched pairs apart.

2.2 Processing Stage

The processing stage operates on the learned embeddings to identify meaningful structures in the time-series data. Several processing operations can be applied to the learned embeddings, from simple to more complex methods :

- clustering analysis to discover recurring patterns across patients
- temporal aggregation of embeddings to capture long-term dynamics

- event detection [8]
- time-series classification [9]
- time-series forecasting [2]

By operating on embeddings, the framework provides a generic and flexible processing mechanism that can be adapted to different datasets and tasks.

2.3 Output Stage

The output stage produces results for downstream applications. The final output can be in many forms, such as directly provided predictions from processing stage, or retrieval-based outputs that provide interpretable summaries using pre-trained language models.

In retrieval-based output methods, the embeddings and processed data can be retrieved and interpreted in a database, and the final output can be a prediction of a health event, a reasoning for user questions or a summary of time-series patterns [10].

The modular design of the framework allows the components to evolve independently, so that the output stage can be adapted to different applications such as health monitoring, event detection, and personalized healthcare recommendations.

3 Experiment Design

To validate the proposed framework, we design experiments focusing on two main aspects : dynamic time-series patch segmentation and time-series-text alignment.

First, we investigate dynamic time-series patch segmentation, which aims to provide more semantic segmentation of physiological time-series data. This experiment leverages self-supervised learning methods [11], combining contrastive learning to encourage intra-patch similarity and inter-patch separation, and reconstruction tasks to capture intra-patch variability, to identify similarities within a time-series patch while keeping temporal patterns information in single embedding of a patch.

Once segment-level embeddings are obtained, aggregation strategies ,such as mean pooling or attention-based pooling are used to construct higher-level sequence representations when required for downstream tasks.

Second, we study the alignment between time-series embeddings and textual descriptions. This experiment combines synthetic time-series–text pairs generated from mathematical functions with weakly paired time-series data from real datasets to train an alignment model that associates time-series patterns with semantic descriptions of events.

The model is trained using a contrastive alignment objective [7] that encourages embeddings of matched time-series and text pairs to be close in the shared space, while mismatched pairs are pushed apart.

By combining dynamic patch segmentation and time-series–text alignment, the experiments aim to link time-series patches to semantic textual descriptions by leveraging self-supervised learning methods to obtain event-level representations and align them with textual descriptions to provide semantic interpretation of physiological patterns, thereby validating the effectiveness of the embedding stage of the proposed framework.

4 Conclusion

This paper presents a modular framework for processing healthcare time-series data from wearable devices and adapting to various downstream tasks. The framework proposes a pipeline with three stages : embedding, processing, and output, which enables flexible integration of different methods for physiological signal analysis.

Future work will focus on implementing the full framework on real-world wearable healthcare datasets and conducting a comprehensive evaluation in practical health monitoring scenarios.

References

- [1] Matthew Smuck, Charles A Odonkor, Jonathan K Wilt, Nicolas Schmidt, et Michael A Swiernik. The emerging clinical role of wearables : factors for successful implementation in healthcare. *NPJ digital medicine*, 4(1) :45, 2021.
- [2] 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.
- [3] Rayan Krishnan, Pranav Rajpurkar, et Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12) :1346–1352, 2022.
- [4] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, et Jayant Kalagnanam. A time series is worth 64 words : Long-term forecasting with transformers. Dans *International Conference on Learning Representations*, 2023.
- [5] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et Qingsong Wen. Time-LLM : Time series forecasting by reprogramming large language models. Dans *International Conference on Learning Representations (ICLR)*, 2024.
- [6] Aaron van den Oord, Yazhe Li, et Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv :1807.03748*, 2018.
- [7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. Dans *International conference on machine learning*, pages 8748–8763. PmlR, 2021.
- [8] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, et Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. Dans *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2828–2837, 2019.
- [9] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, et François Petitjean. Inception-time : Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 2020.
- [10] Rúben Almeida, Hugo Sousa, Luís F Cunha, Nuno Guimarães, Ricardo Campos, et Alípio Jorge. Physio : an llm-based physiotherapy advisor. Dans *European Conference on Information Retrieval*, pages 189–193. Springer, 2024.
- [11] Sana Tonekaboni et al. Unsupervised representation learning for time series with temporal neighborhood coding. Dans *International Conference on Learning Representations (ICLR)*, 2021.

Compression et Diffusion de Nuages de Points Dynamiques Issus d'un Capteur Mobile

Mohamad El Dika^{1,2}, Jalal Possik¹, Petra Bilane¹, Jad Nassar³, Roger Achkar³, Charles Yaacoub¹

¹ICL, Junia, Université Catholique de Lille, LITL, F-59000 Lille, France

²American University of Science and Technology (AUST), Beyrouth, Liban

³TICKET Lab, Université Antonine, Hadat-Baabda, Liban

Résumé

Les nuages de points constituent une représentation tridimensionnelle dense utilisée dans des domaines comme la robotique mobile, la cartographie 3D et les véhicules autonomes. Lorsqu'ils sont acquis dans des environnements dynamiques par des capteurs embarqués, ils génèrent des volumes de données importants devant être transmis ou traités en temps réel, ce qui pose des défis majeurs pour les systèmes robotiques mobiles aux ressources de communication, de calcul et de stockage limitées. Cette communication étudie la compression et la diffusion de nuages de points dynamiques dans un contexte robotique simulé. Plusieurs méthodes de compression sont évaluées : une approche géométrique avec MPEG G-PCC, ainsi que des méthodes issues de l'image et de la vidéo, notamment PNG, JPEG, H.264 et H.265, en passant par une carte de profondeur intermédiaire. L'évaluation analyse le compromis entre taux de compression et qualité de reconstruction du nuage de points.

Mots clefs

Nuages de points dynamiques, compression géométrique, carte de profondeur, diffusion de données 3D, robotique mobile, ROS2.

1 Introduction

Les nuages de points représentent une collection de points tridimensionnels décrivant la géométrie d'un objet ou d'un environnement. Chaque point est défini par ses coordonnées spatiales dans un repère global et peut être accompagné d'informations supplémentaires telles que la couleur ou l'intensité du signal. Lorsque ces points sont agrégés, ils permettent de reconstruire une représentation détaillée d'une scène tridimensionnelle.

Cette représentation est aujourd'hui largement utilisée dans des applications telles que la cartographie tridimensionnelle, l'inspection industrielle ou encore la navigation des véhicules autonomes. Les capteurs LiDAR modernes peuvent produire plusieurs millions de points par seconde, ce qui génère des volumes de données particulièrement importants. Dans le cas de scènes dynamiques, ces données doivent souvent être transmises

en continu vers un nœud central chargé du traitement, du rendu ou du stockage.

Les robots mobiles, notamment les drones et les plateformes robotiques compactes, disposent cependant de ressources limitées en termes de puissance de calcul et de capacité de stockage, et même de bande passante réseau. La transmission brute de nuages de points volumineux peut ainsi entraîner une surcharge du réseau et augmenter significativement la latence des systèmes.

Dans ce contexte, la compression des nuages de points apparaît comme une solution essentielle afin de réduire la taille des données transmises tout en conservant une représentation suffisamment fidèle de l'environnement. Cette étude vise à analyser différentes stratégies de compression appliquées au streaming de nuages de points dynamiques issus d'un capteur mobile. Les performances des méthodes étudiées sont évaluées en fonction du taux de compression obtenu et de la qualité géométrique du nuage de points reconstruit.

2 Revue de littérature

Les systèmes robotiques mobiles sont souvent contraints par des limitations matérielles liées à la consommation énergétique, à la capacité de calcul et à la bande passante disponible [1]. Ces contraintes ont conduit à de nombreuses recherches visant à réduire la quantité de données générées par les capteurs embarqués ou à améliorer l'efficacité de leur transmission [2].

Dans le domaine de la compression des nuages de points, plusieurs approches ont été proposées. Le groupe MPEG a notamment introduit deux standards dédiés à ce type de données. Le standard *Geometry-based Point Cloud Compression* (G-PCC) compresse directement la géométrie tridimensionnelle du nuage de points en exploitant les redondances spatiales présentes dans la structure des données, alors que l'approche *Video-based Point Cloud Compression* (V-PCC) consiste à projeter les nuages de points dans des images bidimensionnelles avant d'appliquer des techniques de compression vidéo traditionnelles [3, 4].

Certains travaux ont exploré l'utilisation directe d'algorithmes de compression d'image ou de vidéo pour traiter les nuages de points. Dans ces approches, les données tridimensionnelles sont transformées en

représentations bidimensionnelles telles que des cartes de profondeur ou bien des séquences d'images [5, 6]. Cette transformation permet d'utiliser des codecs largement optimisés comme JPEG, H.264 ou H.265. Toutefois, cette projection peut introduire une perte d'information géométrique avant même l'étape de compression [7, 8].

3 Méthodologie

Afin d'évaluer les performances des différentes méthodes de compression dans des conditions contrôlées, un environnement de simulation a été développé à l'aide du middleware ROS2 et du simulateur Gazebo. Plusieurs scènes virtuelles présentant différents niveaux de complexité géométrique ont été construites (Figure 1). Certaines scènes contiennent principalement des surfaces planes, tandis que d'autres intègrent des objets aux formes plus complexes tels que des arêtes marquées ou des surfaces courbes.

Un robot mobile équipé d'un capteur LiDAR simulé se déplace dans ces environnements et génère des nuages de points dynamiques.

La Figure 2 illustre l'architecture du pipeline de traitement implémenté dans ROS2 pour l'évaluation des différentes techniques de compression. Le nuage de points généré par le capteur LiDAR constitue la donnée de référence utilisée pour l'évaluation. À partir de ce nuage de points initial, deux types de traitement sont appliqués selon la méthode de compression considérée.

Dans le premier cas, le nuage de points est directement compressé à l'aide de l'encodeur GPCC. Cette méthode traite directement la géométrie tridimensionnelle des données sans transformation intermédiaire. Après compression et transmission, les données sont décodées afin de reconstruire le nuage de points.

Dans le second cas, le nuage de points est d'abord transformé en une carte de profondeur. Dans cette représentation bidimensionnelle, chaque pixel encode la distance entre le capteur et un point de la scène observée. Cette carte de profondeur peut ensuite être comprimée à l'aide d'algorithmes de compression d'image ou de vidéo tels que PNG, JPEG, H.264 ou H.265.

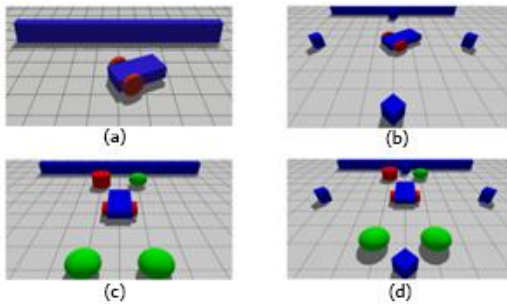


Figure 1 – Environnements de simulation : (a) mur simple, (b) environnement anguleux, (c) environnement plus lisse que (b) avec des sphères et des courbes, (d) environnement complexe.

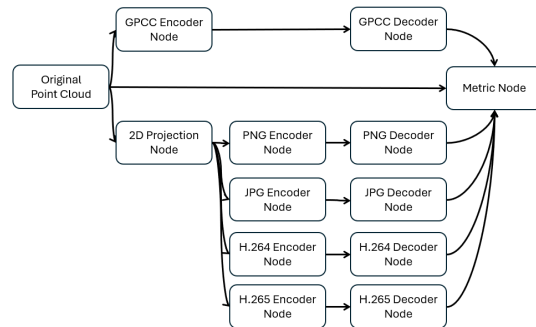


Figure 2 – Pipeline de traitement de nuage de point dans ROS2.

À l'exception du format PNG qui repose sur une compression sans perte et ne peut donc pas être ajusté, les autres méthodes permettent de contrôler le taux de compression au détriment d'une qualité variable. Le Tableau 1 présente le paramètre de qualité associé à chaque algorithme de compression.

Algorithme	Paramètre	Valeur (min → max)
GPCC	Échelle de quantification	100 → 50,000
JPG	Facteur de qualité	30 → 95
H.264	CRF	10 → 51
H.265	CRF	10 → 51

Tableau 1 – Valeurs des paramètres de qualité pour chaque algorithme de compression.

Après décompression, la carte de profondeur est convertie à nouveau en nuage de points afin de reconstruire la représentation tridimensionnelle de la scène. Ce nuage de points reconstruit est ensuite comparé au nuage de points original afin d'évaluer la qualité de la compression.

La qualité du nuage de points reconstruit est évaluée à l'aide du Peak Signal-to-Noise Ratio (PSNR). Cette métrique mesure la différence entre le nuage de points original et le nuage de points obtenu après compression et décompression. La première étape consiste à calculer l'erreur quadratique moyenne entre les deux ensembles de points. Pour chaque point du nuage reconstruit, on identifie le point le plus proche dans le nuage original et on calcule la distance euclidienne entre ces deux points. L'erreur quadratique moyenne est alors obtenue en moyennant le carré de ces distances. Le PSNR est ensuite calculé à partir de cette erreur moyenne et de la taille maximale du nuage de points, généralement définie comme la diagonale de la boîte englobante contenant l'ensemble des points. Une valeur élevée de PSNR indique que la reconstruction est proche du nuage de points original, tandis qu'une valeur faible indique une dégradation importante de la géométrie.

4 Résultats et analyse

Les résultats expérimentaux sont présentés dans la Figure 3 sous forme de courbes débit-distorsion illustrant la relation entre le PSNR et le taux de compression pour les différentes méthodes considérées.

Comme indiqué précédemment, la méthode GPCC comprime directement le nuage de points, tandis que les autres méthodes reposent sur une représentation intermédiaire sous forme de carte de profondeur. Cette différence fondamentale influence significativement les performances observées.

Les résultats montrent que GPCC conserve la meilleure qualité de reconstruction parmi les méthodes étudiées, au prix d'un taux de compression relativement faible. Cette performance s'explique par le fait que l'algorithme exploite directement la structure géométrique du nuage de points et préserve ainsi plus efficacement les relations spatiales entre les points.

Les formats PNG et JPEG sont appliqués directement aux cartes de profondeur issues de la projection du nuage de points. PNG, étant une méthode de compression sans perte, préserve fidèlement la carte de profondeur ; la dégradation de la qualité observée provient donc principalement de la transformation initiale du nuage de points en carte de profondeur, et non du processus de compression lui-même. JPEG introduit une perte d'information liée à la quantification, sans pour autant offrir de gains significatifs en termes de taux de compression.

Les codecs vidéo H.264 et H.265 atteignent les taux de compression les plus élevés parmi les méthodes étudiées. Cette efficacité s'explique par l'utilisation de techniques avancées de prédiction temporelle et de codage entropique. Toutefois, la qualité géométrique des nuages de points reconstruits reste inférieure à celle obtenue avec GPCC. Les artefacts induits par la compression vidéo, combinés à l'étape de projection en carte de profondeur, peuvent en effet altérer la précision géométrique de la scène reconstruite.

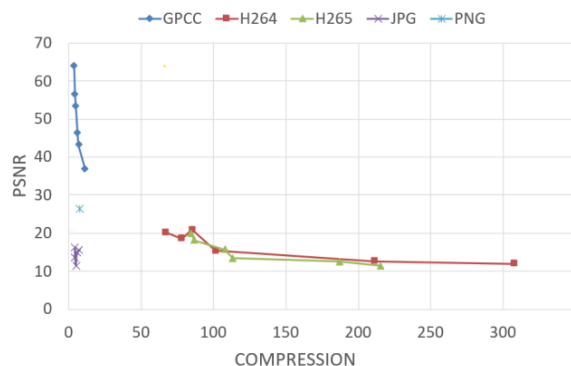


Figure 3 – Courbes de débit-distorsion.

Ces résultats mettent en évidence un compromis important entre efficacité de compression et fidélité géométrique. Les méthodes spécifiquement conçues pour les nuages de points privilégient la précision de la reconstruction, tandis que les techniques issues de la compression d'image et de vidéo permettent de réduire davantage la quantité de données transmises.

5 Conclusion

Cette communication a présenté une étude comparative de différentes techniques de compression appliquées à la diffusion de nuages de points dynamiques issus d'un capteur mobile. Les expériences ont été réalisées dans un environnement de simulation ROS2 permettant d'évaluer plusieurs méthodes dans des conditions identiques.

Les résultats montrent que la compression géométrique GPCC offre la meilleure qualité de reconstruction grâce à son traitement direct de la structure tridimensionnelle des données. Les méthodes basées sur la compression d'image et de vidéo permettent quant à elles d'obtenir des taux de compression plus élevés, mais au prix d'une dégradation plus importante de la géométrie du nuage de points.

Ces observations soulignent l'importance de choisir la méthode de compression en fonction des contraintes de l'application. Les applications nécessitant une représentation précise de l'environnement privilégieront des méthodes géométriques, tandis que les scénarios où la bande passante réseau constitue la contrainte principale pourront tirer parti des techniques de compression vidéo.

Des travaux futurs pourront explorer des approches hybrides combinant compression géométrique et compression vidéo, ainsi que l'utilisation de méthodes d'apprentissage automatique pour optimiser la compression des nuages de points dans des environnements dynamiques. Par ailleurs, des méthodes spécifiquement adaptées au codage et à la compression des cartes de profondeur pourraient également être envisagées.

Références

- [1] Cao, Y., Wang, Y., & Chen, H. (2025). Real-time LiDAR point cloud compression and transmission for resource-constrained robots. arXiv preprint arXiv:2502.06123. DOI:10.48550/arXiv.2502.06123
- [2] Gao, Y., Wenqi W., Ning C., Weilin R., and Quanyi Ye. (2025). "Real-Time Mobile Transmission and Rendering of UAV LiDAR Massive Point Cloud," IET Radar, Sonar & Navigation: e70095. DOI:10.1049/rsn2.70095
- [3] Schwarz, S., Preda, M., Baroncini, V., Madhukar B., Cesar, P., Chou, P. A., (2019). Emerging MPEG standards for point cloud compression. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 9(1), 133–148. DOI:10.1109/JETCAS.2018.2885981

- [4] Graziosi D, Nakagami O, Kuma S, Zaghetto A, Suzuki T, Tabatabai A. (2020). An overview of ongoing point cloud compression standardization activities: video-based (V-PCC) and geometry-based (G-PCC). *APSIPA Transactions on Signal and Information Processing*. 2020;9:e13. DOI:10.1017/ATSIP.2020.12
- [5] Tohidi, F., Paul, M., Ulhaq, A., Chakraborty, S. (2024). Improved video-based point cloud compression via segmentation. *Sensors*, 24(13), 4285. DOI:10.3390/s24134285
- [6] Dumić, E., Bjelopera, A., & Nüchter, A. (2022). Dynamic point cloud compression based on projections, surface reconstruction and video compression. *Sensors*, 22(1), 197. DOI:10.3390/s22010197
- [7] Nardo, F., Peressoni, D., Testolina, P., Giordani, M., & Zanella, A. (2022). Point cloud compression for efficient data broadcasting: A performance comparison. *arXiv preprint arXiv:2202.00719*. DOI:10.48550/arXiv.2202.00719
- [8] Akhtar, A., Gao, W., Li, L., Li, Z., Jia, W., & Liu, S. (2022). Video-based point cloud compression artifact removal. *IEEE Transactions on Multimedia*, 24, 2456–2469. DOI:10.1109/TMM.2021.3090148

Débruitage d'images secrètes à partir d'images partagées compressées en JPEG

V. Noyé

P. Puteaux

W. Puech

LIRMM, Univ Montpellier, CNRS, Montpellier, France

{valentin.noyé, pauline.puteaux, william.puech}@lirmm.fr

Résumé

Les systèmes de partage d'images secrètes (Secret Image Sharing – SIS), basés sur le partage de secret (Secret Sharing – SS), de Shamir sont utilisés pour protéger le contenu visuel sur les réseaux sociaux et sur le cloud. Cependant, les images partagées sont généralement soumises à du bruit lors de leur transmission ou de leur stockage. Dans ce cas, l'image secrète n'est pas parfaitement reconstruite. Dans cet article, nous proposons une analyse expérimentale du débruitage d'images secrètes quand une ou plusieurs images partagées sont sujettes à du bruit, en mettant particulièrement l'accent sur la compression JPEG avec différents facteurs de qualité (QF). Les résultats expérimentaux obtenus sur une base complète d'images valident les performances des stratégies de débruitage basées sur le filtrage médian ou l'utilisation d'un réseau neuronal.

Mots clefs

Sécurité multimédia, partage d'images secrètes, images partagées bruitées, débruitage, compression JPEG.

1 Introduction

En 1979, le système de partage de secret (*Secret Sharing* – SS) basé sur les polynômes de Shamir est devenu un tournant dans le domaine de la cryptographie [1]. Dans ce système, un secret est partagé de manière parfaitement sécurisée entre n participants, dont k peuvent le reconstruire. Formellement, on définit donc un système à paramètres (k, n) avec un seuil k pour lequel tout sous-ensemble de moins de k participants ne peut révéler aucune information sur le secret. Durant ces dernières années, avec le développement des technologies liées aux données multimédias, le partage d'images secrètes (*Secret Image Sharing* – SIS), tel qu'il a été introduit pour la première fois par Thien et Lin [2], est apparu à partir du SS comme un moyen de partager des images entre différents participants.

Cependant, les applications habituelles du SIS sur les réseaux sociaux dans des scénarios d'utilisation réels soumettent les images partagées à des distorsions, dont certaines proviennent généralement de chaînes de traitements d'images avec pertes, tels que la compression JPEG (avec différents facteurs de qualité QF). Dans ce cas-ci, l'opération de quantification garantit une forte réduction du stockage, mais a pour coût de modifier les valeurs des pixels

d'une image partagée. Ce bruit se propage alors à travers le processus d'interpolation polynomiale, rendant l'image secrète reconstruite très sensible à la moindre erreur.

Aujourd'hui, de nombreuses recherches sur le bruit dans les images partagées se concentrent sur les capacités correctives du SIS grâce à des codes correcteurs intégrés [3, 4] ou à des capacités correctives post-reconstruction [5].

Dans cet article, nous présentons et mettons en œuvre des méthodes de débruitage appliquées à des images compressées au format JPEG selon différents QF. Contrairement aux méthodes de correction intégrées au système, le débruitage est effectué après la reconstruction. Ainsi, il n'entraîne pas la sécurité cryptographique et ne nécessite pas d'augmenter la taille des images partagées.

2 Méthodologie de débruitage

Comme illustré dans la figure 1, nous considérons d'abord une image I qui est partagée entre n participants en plusieurs images I_{x_i} de même taille selon un système de SIS. Ces images partagées I_{x_i} peuvent être soumises à du bruit, tel que la compression ou la corruption.

Au cours de la phase de reconstruction, k images partagées sont combinées, parmi lesquelles un sous-ensemble de k^* d'entre elles est bruité. Dans ces circonstances, le bruit présent dans les images partagées bruitées $I_{x_u}^*$ se propage à travers la reconstruction de l'image secrète, ce qui donne une image secrète bruitée I^* . Enfin, une partie de notre travail se concentre sur le débruitage de l'image secrète reconstruite et bruitée I^* afin d'obtenir une image $\tilde{I}$ aussi proche que possible de l'image secrète originale I .

2.1 Partage d'images secrètes de Shamir

En 1979, Shamir [1] a proposé un système de SS selon un seuil (k, n) , énoncé comme suit. Soit p un nombre premier, nous travaillons dans le corps fini $\mathbb{F}_p$. Un secret $s \in \mathbb{F}_p$ est encodé comme terme constant d'un polynôme aléatoire de degré $k - 1$:

$$f(x) = s + a_1x + a_2x^2 + \dots + a_{k-1}x^{k-1} \in \mathbb{F}_p[x], \quad (1)$$

où les coefficients $a_1, \dots, a_{k-1}$ sont choisis aléatoirement et uniformément dans $\mathbb{F}_p$. Ensuite, n points d'évaluation publics, distincts et non nuls $x_1, \dots, x_n \in \mathbb{F}_p$ sont attribués aux participants. Le participant i reçoit donc la part (x_i, y_i) avec $y_i = f(x_i)$.

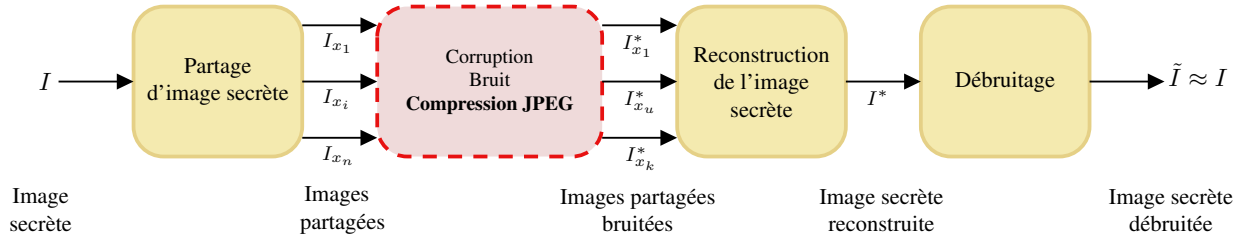


FIGURE 1 – Vue d'ensemble du SIS suivie du débruitage d'une image secrète bruitée.

La reconstruction à partir de n'importe quelles k parts est réalisée par interpolation polynomiale. Étant donné (x_j, y_j) pour $j = 1, \dots, k$, f peut être récupéré de manière unique avec le secret $s = f(0)$. De manière équivalente :

$$s = f(0) = \sum_{j=1}^k y_j \ell_j(0), \quad \ell_j(0) = \prod_{\substack{m=1 \\ m \neq j}}^k \frac{x_m}{x_m - x_j}, \quad (2)$$

où $\ell_j(x)$ sont les polynômes de base de Lagrange. Le système est parfaitement sécurisé au sens de la théorie de l'information, car tout ensemble de parts au plus $k-1$ ne fournit aucune information sur le secret s .

Le système de SS de Shamir a été étendu à des systèmes de SIS en traitant chaque valeur de pixel $p_{w,h}$ d'une image en niveaux de gris I comme un secret. Un polynôme aléatoire est choisi pour chaque secret et ses évaluations en n points distincts produisent n images partagées.

2.2 Débruitage d'une image secrète reconstruite bruitée

Dans cette section, nous détaillons les approches de débruitage visant à atténuer le bruit dans l'image secrète reconstruite après qu'elle ait été soumise au bruit provenant d'une ou plusieurs images partagées bruitées.

Dans un premier scénario, un filtre médian est appliqué comme étape de post-traitement après la reconstruction de l'image secrète. Un filtre médian a également été utilisé dans l'approche décrite par Bertojo et Puech [5]. Pour chaque pixel de l'image secrète reconstruite, un voisinage local est pris en compte. Les valeurs de niveaux de gris des pixels voisins, y compris le pixel actuel, sont triées par ordre croissant, et la valeur du pixel actuel est remplacée par la valeur médiane de niveau de gris de cet ensemble. Le filtrage médian réduit les valeurs de bruit extrêmes ou les valeurs aberrantes introduites lors de la reconstruction de l'image secrète lorsqu'au moins une image partagée est bruitée à l'aide de la compression JPEG.

Par ailleurs, les méthodes de débruitage spécifiques au format JPEG basées sur l'apprentissage profond, telles que FBCNN basé sur un réseau neuronal convolutionnel [6] ou SwinIR basé sur un transformeur Swin [7], ne sont pas directement adaptées à l'application considérée. En effet, dans notre approche, l'image secrète est reconstruite en combinant plusieurs images partagées bruitées qui ont subi

une compression JPEG, et la distribution du bruit qui en résulte ne suit pas nécessairement la distribution standard des artefacts JPEG associée à un QF fixe. Néanmoins, le processus de reconstruction induit un motif de bruit structuré qui peut être appris et compensé après la reconstruction. Ce phénomène a déjà été observé par Bertojo et Puech [5], bien que l'analyse se soit limitée à un QF égal à 100.

Ainsi, dans un deuxième scénario, un réseau neuronal de débruitage basé sur DnCNN [8] est entraîné pour résoudre ce problème. Deux configurations d'entraînement et d'inférence sont envisagées. Dans la première configuration, un seul modèle DnCNN est entraîné pour débruiter des images secrètes reconstruites sans aucune information préalable sur le QF utilisé lors de la compression JPEG des images partagées. Dans la seconde configuration, un modèle DnCNN spécialisé est entraîné pour chaque QF afin d'apprendre spécifiquement les distributions de bruit de compression JPEG du QF concerné. Cette spécialisation devrait améliorer les performances de débruitage lorsque le QF, extrait des métadonnées JPEG de l'image, est connu au moment de l'inférence.

3 Résultats expérimentaux

Toutes les expériences ont été réalisées avec un modèle de bruit correspondant à une compression JPEG standard avec différents facteurs de qualité (QF) allant de 100 à 95, en particulier, puis 90, 75 et 50, afin d'observer de manière exhaustive leurs effets sur l'image secrète reconstruite. Les expériences sont menées sur la base BOWS2, composée d'images en niveaux de gris [9], et en utilisant la méthode de SIS de Shamir de paramètres $(k, n) = (3, 255)$. Dans ces expériences, nous considérons les deux cas où $k^* = 1$ et $k^* = 3$ images partagées sont bruitées. Dans le cas où $k^* = 3$, nous supposons que toutes les images partagées sont bruitées avec le même QF. La figure 2 illustre comment l'image secrète de la figure 2.a est partagée entre plusieurs participants, parmi lesquels 3 images partagées (figures 2.b à 2.d) permettent de reconstruire l'image secrète. Nous prenons 100 images reconstruites à partir de l'ensemble de données BOWS2 [9] qui ont été partagées 100 fois selon un système de seuil $(3, 255)$, en utilisant $k = 3$ images partagées sélectionnées de manière aléatoire et uniforme pour la reconstruction de l'image secrète lorsque $k^* = 1$ et $k^* = 3$ images partagées sont bruitées, pour un total de 10 000 reconstructions bruitées. La figure 3 montre

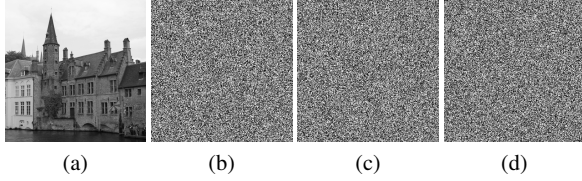


FIGURE 2 – Partage de l’image secrète n°26 issue de la base de données BOWS2 [9] en utilisant la méthode de SIS de Shamir avec les paramètres $(3, n)$: a) L’image secrète, b-d) Trois images partagées.

l’évolution du PSNR sur les 10 000 reconstructions effectuées pour les images partagées avec $k^* = 1$ et $k^* = 3$ images partagées bruitées. Les deux graphiques montrent que lorsque le QF diminue, même légèrement, le PSNR se dégrade rapidement. On observe que la combinaison de $k^* = 3$ images partagées bruitées réduit l’écart-type par rapport à une seule image partagée ($k^* = 1$) (par exemple, 3,48 dB avec $k^* = 1$ et 2,09 dB avec $k^* = 3$ lorsque QF = 100).

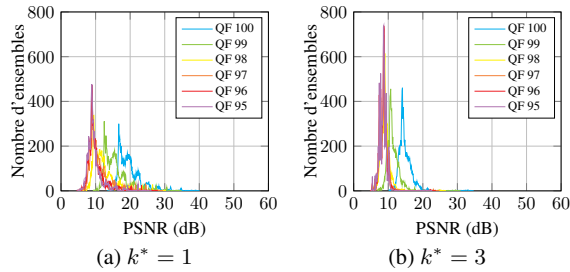


FIGURE 3 – Distribution du PSNR de 10 000 images secrètes reconstruites à partir de la base de données BOWS2 [9] (100 images avec 100 combinaisons différentes d’images partagées) : avec a) $k^* = 1$, b) $k^* = 3$ image(s) partagée(s) bruitée(s).

Les approches de débruitage abordées dans la section 2.2 sont testées et comparées aux résultats de la reconstruction de l’image secrète partagée avec $k^* = 1$ et $k^* = 3$. Nous utilisons une architecture DnCNN à 17 couches similaire à celle mise en œuvre par Zhang *et al.* [8], qui est entraînée avec une taille de lot (*batch size*) de 16 composé d’images en niveaux de gris provenant de l’ensemble de données BOWS2 [9], comprenant 8 900 images d’entraînement et 1 000 images de validation, laissant alors 100 images sur lesquelles effectuer nos tests. Dans une première partie, un modèle général est entraîné pendant 100 époques en transmettant des images secrètes ainsi que leurs images secrètes reconstruites avec bruit, en supposant que le bruit est introduit sur $k^* = 1$ ou $k^* = 3$ images partagées avec un QF allant de 100 à 95, puis 90, 75, 50 (soit deux entraînements au total). Dans une deuxième partie, le modèle est entraîné pendant 50 époques en transmettant uniquement des images d’un QF spécifié afin de le spécialiser dans le débruitage d’images secrètes reconstruites lorsque le QF est connu et lorsque $k^* = 1$ ou $k^* = 3$ (soit un total de

deux entraînements par QF, c’est-à-dire 18).

Toutes ces approches de débruitage sont illustrées dans la figure 4 et montrent l’évolution de la qualité des méthodes présentées sur une image en comparant leur PSNR avec l’image originale de la figure 2.a. Nous montrons que ces méthodes de débruitage sont efficaces, avec un PSNR de 35,45 dB pour le débruitage basé sur un DnCNN général, et un PSNR de 41,03 dB lorsqu’un DnCNN spécialisé est utilisé. Nous constatons qu’elles ont des difficultés à corriger le bruit près des bords, en particulier avec le filtre médian, qui a tendance à lisser l’image et donne un PSNR de 29,48 dB.

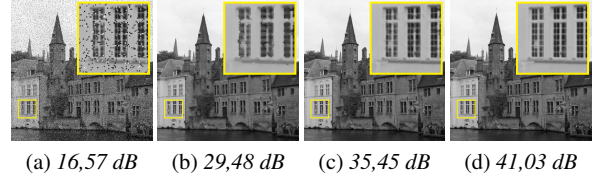


FIGURE 4 – Débruitage de l’image secrète reconstruite et bruitée de la figure 2.a et leur PSNR correspondant : a) Image secrète reconstruite et bruitée, b) Image débruitée par filtre médian, c) Image débruitée par un DnCNN général [8], d) Image débruitée par un DnCNN spécialisé.

Dans la figure 5.a, avec $k^* = 1$, nous évaluons l’évolution du PSNR en fonction de différents QF. Tout d’abord, nous remarquons que la qualité visuelle de l’image secrète reconstruite (avant débruitage) diminue rapidement entre QF = 100 et QF = 95, passant de 19,89 dB à 9,81 dB, pour atteindre ensuite une asymptote proche de 8,50 dB à QF = 50. Lorsqu’on applique un filtre médian, le PSNR dans cet intervalle augmente dans des proportions équivalentes au PSNR de la reconstruction bruitée, passant de 32,70 dB à 13,44 dB pour des QF compris entre 100 et 95, pour converger vers environ 11,09 dB pour QF = 50. Le DnCNN général donne de meilleurs résultats globaux que l’approche du filtre médian, avec une augmentation du PSNR allant de 36,16 dB (QF = 100) à 19,21 dB (QF = 95) et une asymptote proche de 12,94 dB à QF = 50. Lorsqu’il est spécialisé, nous obtenons à nouveau de meilleurs résultats pour un QF élevé, entre 42,22 dB (QF = 100) et 19,53 dB (QF = 95), et atteignons 12,91 dB pour un QF = 50. Cela prouve que pour un QF plus faible, le DnCNN spécialisé n’est pas plus performant que le DnCNN général. Dans la figure 5.b, pour $k^* = 3$, nous observons des tendances similaires à celles attendues suivant l’ajout de bruit sur trois images partagées, ce qui réduit le PSNR. Une observation intéressante avec chaque méthode est qu’elles atteignent un PSNR asymptotique similaire à celui obtenu avec $k^* = 1$.

Enfin, la figure 6 illustre que les performances de débruitage dépendent à la fois du nombre d’images partagées bruitées k^* et du QF JPEG. Pour des QF faibles, la distribution du PSNR produite par le DnCNN général ressemble étroitement à celle du DnCNN spécialisé pour $k^* = 1$ et $k^* = 3$. En revanche, pour des QF plus élevés, il existe une

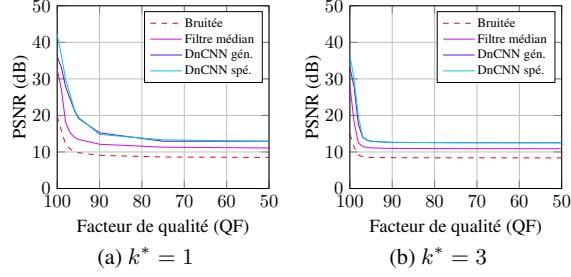


FIGURE 5 – PSNR de la reconstruction et du débruitage de 10 000 images secrètes issues de la base de données BOWS2 [9] (100 images avec 100 combinaisons différentes d’images partagées) en fonction du QF JPEG : avec a) $k^* = 1$, b) $k^* = 3$ image(s) partagée(s) bruitée(s).

séparation claire entre la distribution des images bruitées et les distributions après débruitage avec les trois méthodes, ce qui indique une amélioration générale plus importante.

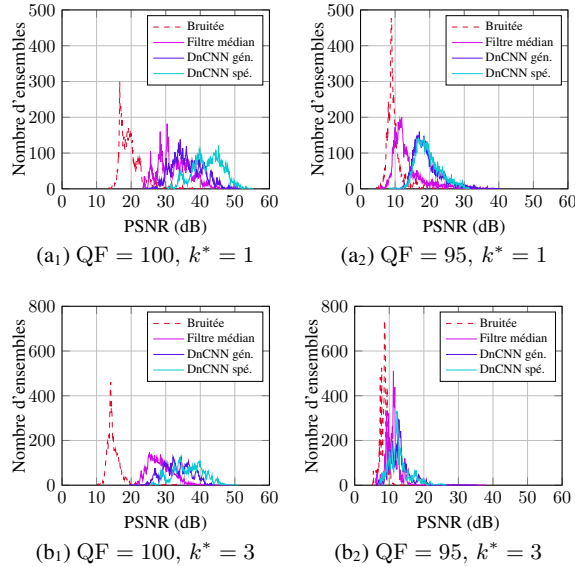


FIGURE 6 – Distribution du PSNR de 10 000 images secrètes reconstruites à partir de la base de données BOWS2 [9] (100 images avec 100 combinaisons différentes d’images partagées) par rapport à leurs versions débruitées avec $k^* = 1$ et $k^* = 3$ image(s) partagée(s) bruitée(s) : a) QF = 100, b) QF = 95.

4 Conclusion

Dans cet article, nous avons étudié expérimentalement les performances de différentes stratégies de débruitage d’images secrètes lorsque celles-ci sont reconstruites à partir d’une ou plusieurs images partagées bruitées, en focalisant notre attention sur la compression JPEG et ses différents facteurs de qualité (QF). En particulier, nous avons étudié le filtre médian, qui s’est avéré donner de bons résultats mais a tendance à lisser l’image, ainsi qu’un

un DnCNN général [8] qui confère de meilleurs résultats. Lorsqu’il est adapté à un QF spécifique, le DnCNN spécialisé présente d’excellents résultats de débruitage pour un QF élevé et se rapproche des performances du modèle de débruitage général pour un QF plus faible. Les perspective de ces travaux comprennent une étude approfondie de la propagation de l’erreur, l’exploration de la sélection optimale d’images partagées afin de réduire l’erreur dans les images secrètes reconstruites, ou l’utilisation de codes correcteurs d’erreurs.

Remerciements

Ce travail a été soutenu par une subvention du gouvernement français gérée par l’Agence nationale de la recherche dans le cadre du programme France 2030, référence ANR-22-PECY-0011.

Références

- [1] A. Shamir. How to share a secret. *Communications of the ACM*, 22(11) :612–613, 1979.
- [2] C.C. Thien et J.C. Lin. Secret image sharing. *Computers & Graphics*, 26(5) :765–770, 2002.
- [3] X. Yan, Y. Lu, Y. Chen, C. Lu, B. Zhu, et Q. Liao. Secret image sharing based on error-correcting codes. Dans *IEEE International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pages 86–89. IEEE, 2017.
- [4] Y. Sun, Y. Lu, X. Yan, L. Liu, et L. Li. Robust secret image sharing scheme against noise in shadow images. *IEEE Access*, 9 :23284–23300, 2021.
- [5] L. Bertojo et W. Puech. Correction of secret images reconstructed from noised shared images. Dans *IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2022.
- [6] J. Jiang, K. Zhang, et R. Timofte. Towards flexible blind JPEG artifacts removal. Dans *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4977–4986, 2021.
- [7] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, et R. Timofte. SwinIR : Image restoration using Swin transformer. Dans *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021.
- [8] K. Zhang, W. Zuo, Y. Chen, D. Meng, et L. Zhang. Beyond a Gaussian denoiser : Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7) :3142–3155, 2017.
- [9] P. Bas et T. Furon. BOWS-2 (Break Our Watermarking System), 2007. Image dataset (10,000 grayscale images), Accessed : 2026-01.

Quand le mouvement émerge des co-activations : modélisation probabiliste des dépendances spatio-temporelles pour les flux de caméras événementielles

Wessim Omezzine¹

Lionel Fillatre¹

¹ Laboratoire I3S

Université Côte d’Azur, CNRS

{omezzine, lionel.fillatre}@i3s.unice.fr

Résumé

L’information issue des caméras événementielles est portée par l’organisation spatio-temporelle des événements asynchrones plutôt que par des images d’intensité. Pourtant, la distribution jointe de ces événements est rarement modélisée explicitement. Nous proposons un cadre distributionnel dans lequel les événements discrétisés sont représentés comme des variables binaires et leurs co-activations sont modélisées au moyen d’une famille exponentielle intégrant des interactions d’ordre élevé. Afin d’identifier un sous-ensemble d’interactions qui est réalisable numériquement, nous développons une procédure d’apprentissage hiérarchique non supervisée de dictionnaires, guidée par des critères informationnels, conduisant à un ensemble parcimonieux et interprétable d’atomes spatio-temporels dont l’importance est quantifiée. Sur des données réelles issues de caméras événementielles, des interactions d’ordre élevé émergent avec une organisation géométrique dans l’espace-temps cohérente avec la dynamique du mouvement. En grande dimension, une contrainte de localité est introduite afin de maîtriser la croissance combinatoire des interactions candidates.

Mots clefs

Caméras événementielles, famille exponentielle, apprentissage hiérarchique, dépendances spatio-temporelles.

1 Introduction

Les caméras événementielles [1] sont des capteurs bio-inspirés qui enregistrent de manière asynchrone les variations locales de luminosité, contrairement aux caméras classiques qui échantillonnent des images à fréquence fixe. Leur très haute résolution temporelle et leur large dynamique [2] les rendent adaptées à la capture de phénomènes dynamiques rapides.

Dans un flux événementiel, l’information visuelle n’est pas portée par des intensités statiques, mais par l’organisation spatio-temporelle des événements. Les structures pertinentes — contours, trajectoires — émergent de co-activations structurées dans l’espace et le temps [3]. Caractériser ces dépendances est donc essentiel pour comprendre

la structure statistique d’une séquence événementielle.

Au-delà de l’analyse des dépendances locales, la question centrale est celle de la représentation efficace du signal événementiel. Un flux asynchrone constitue déjà une forme de codage parcimonieux de l’information visuelle ; néanmoins, la structure jointe qui gouverne les co-activations spatio-temporelles reste à expliciter pour obtenir une description réellement compacte. Les approches contemporaines [4, 5, 6] d’apprentissage apprennent des représentations internes puissantes, mais ces dernières encapsulent les dépendances de manière implicite dans des paramètres de grande dimension. Une modélisation explicite de la loi jointe offre au contraire un contrôle direct sur la complexité de la représentation, facilite son interprétation, et ouvre la voie à des schémas de compression ou de génération cohérents avec la structure statistique du signal.

Les premiers modèles probabilistes dédiés aux caméras événementielles se sont principalement concentrés sur des comportements d’ordre un, en modélisant les pixels indépendamment [7], ce qui permet d’estimer des statistiques marginales mais laisse de côté la structure jointe. Des approches plus récentes introduisent des relations locales entre événements, notamment via des graphes ou des réseaux neuronaux sur graphes [5, 8], où les dépendances sont généralement limitées à des interactions d’ordre deux induites par des proximités spatio-temporelles. D’autres travaux construisent des motifs temporels hiérarchiquement à l’aide de réseaux impulsifs [9], mais sélectionnent ces motifs en fonction de leurs performances en classification, conduisant à une organisation orientée tâche. Dans ce travail, nous adoptons une perspective distributionnelle explicite. Les événements discrétisés sont représentés comme des variables binaires, et leur loi jointe est décrite par une famille exponentielle intégrant des interactions d’ordre élevé. Afin de maîtriser la croissance combinatoire du nombre d’interactions possibles, nous proposons un apprentissage hiérarchique de dictionnaires dans lequel chaque atome correspond à un motif de co-activation spatio-temporelle. La sélection des atomes est guidée par un critère informationnel intrinsèque, permettant d’identifier un dictionnaire parcimonieux et interprétable. Nous montrons expérimentalement, sur le jeu de données DVS-

Gesture [10], que cette construction fait émerger des structures spatio-temporelles d'ordre élevé cohérentes avec la dynamique du mouvement, sans supervision ni hypothèse explicite sur la trajectoire.

Le reste de l'article est organisé comme suit. La Section 2 introduit le cadre probabiliste et la représentation des co-activations au moyen d'une famille exponentielle. La Section 3 décrit la procédure d'apprentissage hiérarchique du dictionnaire. La Section 4 présente l'évaluation expérimentale sur le jeu de données DVSGesture et analyse les structures spatio-temporelles apprises. Enfin, la Section 5 conclut et discute les perspectives.

2 Cadre probabiliste

2.1 Représentation binaire

Une caméra événementielle produit un flux d'événements (x_i, y_i, t_i, p_i) , où (x_i, y_i) désigne la position spatiale, t_i l'instant et $p_i \in \{-1, +1\}$ la polarité. Nous nous restreignons aux événements positifs et notons $e_i = (x_i, y_i, t_i)$. Le flux continu est discrétisé selon des résolutions spatiales $(\Delta x, \Delta y)$ et temporelle Δt , induisant une grille voxelisée sur $[1, W] \times [1, H] \times [0, T_{seq}]$. On définit $N_x = \lceil \frac{W}{\Delta x} \rceil$, $N_y = \lceil \frac{H}{\Delta y} \rceil$, $N_t = \lceil \frac{T_{seq}}{\Delta t} \rceil$, et l'on associe à chaque voxel une variable binaire indiquant la présence d'au moins un événement [4]. On obtient ainsi un tenseur $\mathbf{v} \in \{0, 1\}^{N_x \times N_y \times N_t}$, où $v_{x,y,t} = 1$ si le voxel correspondant contient au moins un événement. Étant donné un ensemble de M séquences voxelisées $\{\mathbf{v}^{(m)}\}_{m=1}^M$, celles-ci induisent une distribution empirique $\hat{p}(\mathbf{v}) = \frac{1}{M} \sum_{m=1}^M \mathbb{I}[\mathbf{v}^{(m)} = \mathbf{v}]$. Bien que définie sur un espace de taille 2^d avec $d = N_x N_y N_t$, la distribution $\hat{p}$ possède un support extrêmement parcimonieux. Le problème consiste alors à modéliser les dépendances statistiques dominantes de $\hat{p}$ sans explorer l'ensemble combinatoire complet des configurations possibles.

2.2 Famille exponentielle et atomes

Afin de modéliser les dépendances statistiques entre voxels, nous vectorisons le tenseur binaire $\mathbf{v}$ et l'identifions à $\mathbf{v} \in \{0, 1\}^d$, où $d = N_x N_y N_t$ et $[d] = \{1, \dots, d\}$. Les dépendances sont décrites au moyen d'*interactions*, définies comme des sous-ensembles $S \subseteq [d]$. À chaque interaction est associée la fonction indicatrice $\phi_S(\mathbf{v}) = \prod_{i \in S} v_i$, qui vaut 1 si et seulement si tous les voxels indexés par S sont simultanément actifs. Ces interactions correspondent à des motifs de co-activation spatio-temporelle interprétables (structures spatiales, répétitions temporelles ou chaînes spatio-temporelles). Pour un ensemble d'interactions $D \subseteq 2^{[d]}$, nous considérons le modèle de famille exponentielle

$$q(\mathbf{v}; \theta, D) = \exp \left(\sum_{S \in D} \theta_S \phi_S(\mathbf{v}) - \psi(\theta, D) \right), \quad (1)$$

où $\theta = \{\theta_S\}_{S \in D} \in \mathbb{R}^{|D|}$ est un vecteur de paramètres réels, et $\psi(\theta, D)$ assure la normalisation. Lorsque $D =$

$2^{[d]}$, le modèle est universel [11], mais le nombre d'interactions croît exponentiellement avec d . Nous cherchons donc à sélectionner un sous-ensemble parcimonieux $D \subset 2^{[d]}$, appelé *dictionnaire*, tel que le modèle restreint q_D approxime fidèlement la distribution empirique $\hat{p}$. Chaque interaction $S \in D$ est appelée *atome*.

3 Apprentissage du dictionnaire

Nous adoptons le cadre d'apprentissage hiérarchique de dictionnaires pour données binaires proposé dans [12]. Le problème consiste à approximer la distribution empirique $\hat{p}$ par un modèle exponentiel restreint q_D :

$$\arg \min_{D \subseteq \mathcal{A}} \text{KL}(\hat{p} \| q_D), \quad (2)$$

où $\mathcal{A}$ désigne une classe d'interactions admissibles.

Afin de sélectionner les interactions de manière progressive, nous utilisons la *refined information* (RI), définie pour une interaction candidate S relativement à un dictionnaire D par

$$\text{RI}_{D,S}(p) = \text{KL}(p \| q_D) - \text{KL}(p \| q_{D \cup \{S\}}), \quad (3)$$

qui mesure la réduction de divergence obtenue en ajoutant le motif/l'interaction S . Les interactions sont organisées hiérarchiquement par ordre croissant $|S|$. À l'ordre k , seules les interactions de cardinalité k sont considérées. Cette construction conduit à une décomposition par couches de la divergence KL, permettant d'interpréter la contribution informationnelle de chaque atome. L'apprentissage procède couche par couche : la couche k contient les interactions telles que $|S| = k$. Les atomes d'ordre k sont générés par fusion d'atomes informatifs d'ordre $k-1$: S est admissible s'il existe S_1, S_2 sélectionnés précédemment tels que $|S_1 \cup S_2| = k$. Cette contrainte hiérarchique limite la croissance combinatoire et favorise la composition progressive de motifs locaux vers des structures globales. Chaque atome correspond à une interaction $S \subseteq [d]$, interprétable comme un motif de co-activation spatio-temporelle. Pour l'analyse expérimentale, nous distinguons a posteriori trois types d'atomes : **spatiaux**, **temporels** et **spatio-temporels**. Un atome purement spatial correspond à un motif instantané ; un atome purement temporel correspond à une persistance locale du même pixel ; un atome spatio-temporel correspond à une chaîne cohérente de co-activations. Il est important de noter que ces atomes encodent des co-activations jointes et non des mouvements explicites ; l'interprétation du mouvement émerge de leur organisation géométrique dans l'espace-temps.

La Figure 1 illustre un atome spatio-temporel correspondant à une trajectoire anti-horaire. Afin d'obtenir une représentation continue en espace, chaque voxel (x_v, y_v, t_v) est remplacé par une impulsion spatiale lissée $G_\sigma(x - x_v, y - y_v)$. On définit alors une densité $\rho(x, y) = \sum_v G_\sigma(x - x_v, y - y_v)$ et une surface de temps moyen $\tau(x, y) = \frac{\sum_v t_v G_\sigma(x - x_v, y - y_v)}{\rho(x, y)}$. Le flou gaussien est purement spatial et permet d'obtenir une surface régulière τ . Le champ

vectorel affiché correspond à $-\nabla\tau(x, y)$, qui indique la direction locale d’augmentation du temps. Il ne s’agit pas d’un flot optique, mais d’un champ induit par la structure spatio-temporelle de l’atome. La Figure 1 met en évidence l’extraction du mouvement à partir de l’atome.

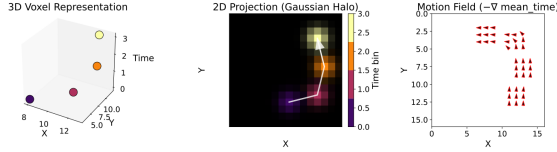


FIGURE 1 – *Illustration d’un atome spatio-temporel et de l’extraction du champ de mouvement. Gauche : représentation 3D des voxels (x, y, t) . Centre : projection 2D après flou gaussien spatial, la couleur encode le temps moyen local. Droite : champ vectoriel $-\nabla\tau(x, y)$.*

4 Expérimentation

Nous évaluons la méthode sur le jeu de données DVSGesture [10], en considérant les rotations de la main droite dans les directions horaire (CW) et antihoraire (CCW). Ces deux classes présentent une apparence spatiale similaire mais diffèrent dans leur organisation temporelle, ce qui permet d’évaluer si les interactions apprises capturent la direction du mouvement. Nous conservons uniquement les événements de polarité positive, appliquons un sous-échantillonnage spatial et un centrage, puis voxelisons chaque séquence sous la forme d’un tenseur binaire $32 \times 32 \times 10$.

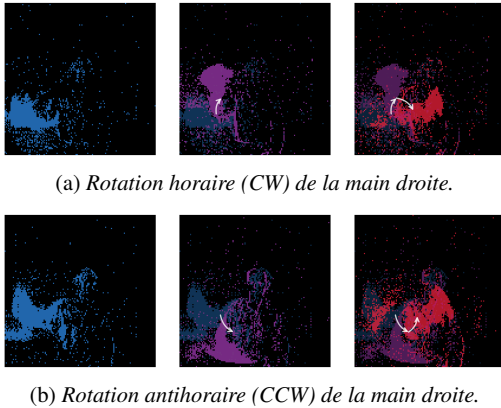


FIGURE 2 – *Exemples de mouvements dans le jeu de données DVSGesture. Chaque ligne présente trois instants successifs avec un effet de flou temporel illustrant la direction du mouvement.*

Dans DVSGesture, les classes de rotation horaire (CW) et antihoraire (CCW) présentent des projections spatiales très similaires. La structure spatiale instantanée seule ne suffit donc pas à distinguer le mouvement. Ce cadre permet d’évaluer si l’apprentissage hiérarchique d’interactions

capture la direction comme propriété distributionnelle.

La dimension ambiante ($d = 32 \times 32 \times 10$) rend la génération exhaustive des interactions non réalisable numériquement au-delà de très faibles ordres. Nous imposons donc une contrainte de localité spatio-temporelle (rayon spatial $r = 4$ et écart temporel $\Delta t = 2$) limitant la fusion des atomes à des voisinages cohérents. Cette contrainte est indépendante des classes. C’est une forme de biais qui va induire des atomes avec une proximité spatio-temporelle.

La Figure 3 montre l’évolution hiérarchique des atomes dominants pour les deux classes. Aux premières couches (par exemple ordre 3), les atomes sont compacts et indépendants de la classe : ils correspondent à des motifs locaux associés aux contours en mouvement. Ces dépendances de bas ordre encodent « ce qui est présent » mais pas encore « comment cela se déplace ». Lorsque l’ordre augmente, les atomes forment progressivement des chaînes spatio-temporelles plus étendues, alignées avec des fragments de trajectoires. Aux couches supérieures (jusqu’à l’ordre 11), ils deviennent spécifiques à la classe, avec des orientations espace-temps opposées pour CW et CCW.

La Figure 4 illustre la construction progressive d’un atome associé au mouvement. Aux couches inférieures (ordre 2–4), le support de l’atome reste spatialement compact et temporellement court : il capture des co-activations locales correspondant à des fragments de contour en déplacement. Ces structures reflètent des corrélations immédiates entre voisinages spatio-temporels proches. Lorsque l’ordre augmente (couches 6–8), la fusion hiérarchique assemble ces motifs locaux en chaînes cohérentes le long de la trajectoire. Le support s’étend simultanément dans l’espace et dans le temps, traduisant une dépendance jointe entre événements séparés mais statistiquement corrélés par la cinématique du geste. Le champ directionnel estimé à partir du support de l’atome met en évidence une orientation dominante qui devient stable aux ordres élevés. Cette stabilisation traduit le passage d’une corrélation locale à une structure géométrique globale : la direction du mouvement émerge comme propriété collective de la dépendance jointe. Ce phénomène est particulièrement significatif car aucune variable explicite de vitesse ou de trajectoire n’est introduite dans le modèle. La cohérence directionnelle résulte uniquement de la maximisation du critère informationnel, ce qui indique que la dynamique du geste constitue une composante dominante de la structure statistique du signal événementiel.

5 Conclusion

Nous avons proposé une modélisation probabiliste des flux issus de caméras événementielles fondée sur une famille exponentielle à interactions d’ordre élevé. L’apprentissage hiérarchique guidé par un critère informationnel construit un dictionnaire d’atomes spatio-temporels. Sur DVSGesture, la direction du mouvement émerge comme dépendance spatio-temporelle d’ordre élevé.

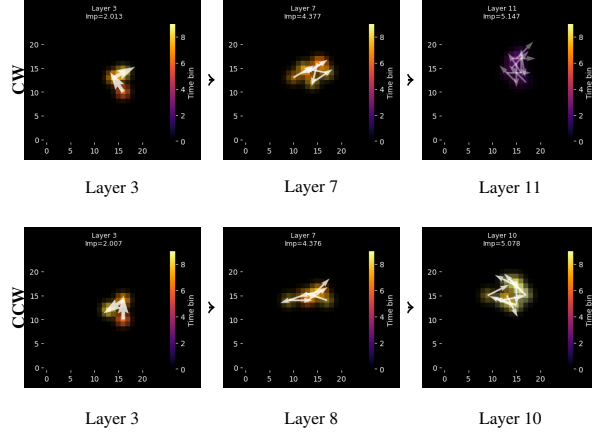


FIGURE 3 – *Évolution hiérarchique des atomes spatio-temporels pour les rotations CW et CCW. Chaque ligne montre les atomes dominants à différents ordres d'interaction. Les supports espace-temps présentent des orientations opposées selon la classe, révélant l'émergence de la direction comme dépendance d'ordre élevé.*

Références

- [1] Patrick Lichtsteiner, Christoph Posch, et Tobi Delbruck. A 128×128 120 db $15 \mu s$ latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2) :566–576, 2008.
- [2] R Wes Baldwin, Ruixu Liu, Mohammed Almatrafi, Vijayan Asari, et Keigo Hirakawa. Time-ordered recent event (tore) volumes for event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2) :2519–2532, 2022.
- [3] Tobi Delbruck. Neuromorphic vision sensing and processing. Dans *2016 46th European Solid-State Device Research Conference (ESSDERC)*, pages 7–14, 2016.
- [4] Dimche Kostadinov et Davide Scaramuzza. Unsupervised feature learning for event data : Direct vs inverse problem formulation. Dans *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5981–5987. IEEE, 2021.
- [5] Thomas Mesquida, Manon Dampfhofer, Thomas Dalgaty, Pascal Vivet, Amos Sironi, et Christoph Posch. G2n2 : lightweight event stream classification with gru graph neural networks. Dans *BMVC 2023-The 34th British Machine Vision Conference*, page 660, 2023.
- [6] Rohan Ghosh, Anupam Gupta, Siyi Tang, Alcimar Soares, et Nitish Thakor. Spatiotemporal feature learning for event-based vision. *arXiv preprint arXiv :1903.06923*, 2019.
- [7] Lionel Fillatre et Marc Antonini. Uniformly minimum variance unbiased estimation for asynchronous

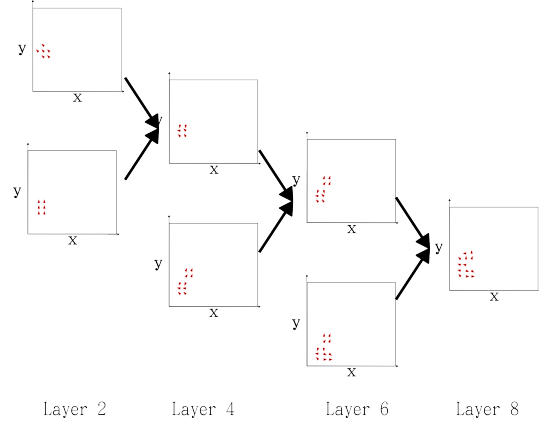


FIGURE 4 – *Formation hiérarchique d'un atome de mouvement. Atomes appris aux ordres 2, 4, 6 et 8. Les flèches rouges représentent le champ extrait $-\nabla\tau(y, x)$ à partir du support de l'atome.*

event-based cameras. Dans *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4107–4111. IEEE, 2014.

- [8] Thomas Dalgaty, Thomas Mesquida, Damien Joubert, Amos Sironi, Pascal Vivet, et Christoph Posch. Hugnet : Hemi-spherical update graph neural network applied to low-latency event-based optical flow. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3953–3962, 2023.
- [9] Hugo Bulzomi, Yuta Nakano, Rémy Bendahan, et Jean Martinet. Neuromorphic temporal pattern detection with spiking neural networks using synaptic delays. Dans *2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS)*, pages 46–50. IEEE, 2024.
- [10] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. Dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017.
- [11] Shun-ichi Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5) :1701–1711, 2001.
- [12] Wessim Omezzine et Lionel Fillatre. Hierarchical Dictionary Learning for Binary Data via Information Decomposition. Working paper, January 2026. Available on HAL : <https://hal.science/hal-05458223>, Janvier 2026.

Adaptation multi-échelle de résolution par apprentissage léger pour VVC

Kra Tchimbé KOFFI^{1,2}

Thomas AMESTOY¹

Elie MORA¹

Anissa MOKRAOUI²

¹Advanced Research Team, VITEC, Châtillon, France

²L2TI, Université Sorbonne Paris Nord, Villetaneuse, France

{landry.koffi, thomas.amestoy, elie.mora}@vitec.com, anissa.mokraoui@univ-paris13.fr

Résumé

Dans les régimes de codage contraints en débit, la préservation de la qualité des contenus haute résolution constitue un enjeu majeur des systèmes modernes de compression vidéo. Dans ce cadre, le codage fondé sur le rééchantillonnage spatial, offre une solution équilibrée en améliorant l'efficacité du compromis débit-distorsion tout en réduisant la complexité d'encodage. Cet article propose une méthode à faible complexité, basée sur l'apprentissage machine, pour l'adaptation de la résolution spatiale au niveau scène dans le standard VVC. Les décisions de résolution sont limitées aux frontières de scènes afin de capturer les variations structurelles significatives, d'assurer la cohérence temporelle et d'éviter les basculements fréquents de résolution. Deux stratégies de décision multi-échelle sont étudiées, reposant toutes deux sur des modèles légers de type MLP et exploitant exclusivement des caractéristiques issues du processus d'encodage. Les expérimentations menées sur des séquences 4K démontrent des performances comparables à la méthode de référence implémentée dans VTM, tout en offrant une stabilité décisionnelle accrue aux faibles débits.

Mots clefs

Scalabilité spatiale, codage basé sur le rééchantillonnage, prédiction d'échelle basée contenu

1 Introduction

La norme de codage Versatile Video Coding (VVC) [1] a permis des gains significatifs en efficacité de compression pour les contenus haute résolution (4K, 8K, formats immersifs 360°). Ces performances reposent sur l'intégration d'outils de codage avancés et de mécanismes d'optimisation toujours plus élaborés. Néanmoins, dans des régimes de débit faible ou très faible, la préservation de la qualité visuelle reste un défi majeur. Lorsque la résolution spatiale augmente sans augmentation proportionnelle du débit disponible, le budget binaire doit être réparti sur un nombre croissant d'échantillons. L'encodeur est alors contraint d'appliquer une quantification plus agressive, entraînant l'apparition d'artefacts visibles et une dégradation perceptible de la qualité. Le codage basé sur le rééchantillonnage spatial, qui consiste à appliquer un sous-échantillonnage

avant l'encodage suivi d'un sur-échantillonnage après le décodage, constitue une stratégie efficace dans ces régimes contraints. En réduisant le nombre de pixels effectivement codés, cette approche améliore le compromis débit-distorsion tout en diminuant la complexité computationnelle et la consommation énergétique, un point critique pour les architectures temps réel. Afin de prendre en charge ce type d'adaptation, VVC intègre nativement le mécanisme de Reference Picture Resampling (RPR) [1]. RPR autorise des changements de résolution sans recourir à un intra refresh, ce qui s'avère particulièrement pertinent dans des scénarios contraints en débit tels que le streaming adaptatif. Les avantages du codage basé sur le rééchantillonnage spatial dépendent toutefois fortement de la fréquence et de l'amplitude des changements de résolution : un sous-échantillonnage inadapté peut entraîner la perte irréversible de détails spatiaux importants, tandis que des basculements trop fréquents induisent une instabilité temporelle perceptible ainsi que des fluctuations transitoires du débit. Ces constats motivent l'exploration de stratégies alternatives de décision de résolution, compatibles avec des contraintes de complexité très faibles, tout en renforçant la cohérence temporelle.

Dans cet article, nous étendons notre approche précédente d'adaptation spatiale légère basée sur l'apprentissage [2] vers une formulation multi-échelle. Afin d'imposer une cohérence temporelle, les décisions de changement de résolution sont prises aux frontières de scènes, où les caractéristiques visuelles changent de façon significative, plutôt que de réagir à des fluctuations image par image. Les images d'une même scène partageant généralement des caractéristiques cohérentes, elles se prêtent naturellement à une décision de résolution commune. Du point de vue système, les deux stratégies proposées, Cascade Binary Classifier (CBC) et Single Multi-Class Classifier (SMC), reposent sur des approches d'apprentissage de type MLP et exploitant exclusivement des caractéristiques disponibles côté encodeur (Section 3). En s'affranchissant de tout traitement dans le domaine pixel, elles se distinguent des méthodes récentes fondées sur des réseaux neuronaux [3, 4]. Ce choix méthodologique est motivé par les contraintes architecturales et computationnelles d'un encodeur VVC temps réel implémenté sur ASIC, pour lequel l'accès au domaine pixel ainsi que les besoins supplémentaires en mémoire tampon

sont incompatibles avec les exigences matérielles. Les stratégies proposées sont comparées à la méthode d'adaptation de résolution implémentée dans VTM [5]. Conformément à la méthodologie de [6], la performance est aussi mesurée par rapport à une borne supérieure théorique de l'adaptation de résolution au niveau scène, permettant d'estimer les gains maximaux et de quantifier l'écart résiduel par rapport à la performance optimale.

La suite de cet article est organisée comme suit. La Section 2 rappelle les mécanismes de scalabilité spatiale dans VVC et les stratégies existantes de décision de résolution. La Section 3 décrit la préparation du jeu de données. La Section 4 présente les stratégies de décision fondées sur l'apprentissage machine. Enfin, la Section 5 rapporte les résultats expérimentaux obtenus selon les CTC du JVET, incluant une analyse détaillée du BD-Rate.

2 Scalabilité spatiale et de décision de résolution dans VVC

Avant VVC – Bien avant l'émergence du standard VVC, de nombreuses études ont établi de manière convergente que l'application d'un rééchantillonnage spatial avant la compression pouvait améliorer significativement les performances sous contrainte sévère en débit [7–10].

Bruckstein *et al.* [7] ont montré, avant l'émergence des standards modernes de compression vidéo, que le codage d'images JPEG fondé sur le rééchantillonnage pouvait surpasser le codage haute résolution à bas débit, tant selon des métriques objectives que du point de vue de la qualité subjective. Des études ultérieures ont formalisé ce compromis fondamental en décomposant la distorsion totale en deux composantes : (i) la distorsion due au sous-échantillonnage et (ii) la distorsion de codage. Dans le contexte de H.264, Wang *et al.* [10] ont montré que la minimisation de la distorsion globale conduit à un facteur d'échantillonnage spatial optimal, étroitement lié au débit cible.

Modèles paramétriques – Dans VVC, la scalabilité spatiale est majoritairement gérée au moyen de modèles paramétriques ou basés sur des seuils, opérant au niveau image ou GOP. Par exemple, Fu *et al.* [8] ont proposé une méthode d'accélération de la décision RPR au niveau image, reposant sur un modèle exponentiel pré-ajusté permettant de déterminer un seuil de distorsion en fonction du paramètre de quantification (QP). La distorsion induite par le rééchantillonnage est estimée via un PSNR calculé sur l'image sous-échantillonnée puis rééchantillonnée (down-up scaled), et le mécanisme RPR n'est activé que si cette distorsion reste inférieure au seuil dépendant du QP.

VTS (stratégie à seuils implémentée dans VTM) – La méthode proposée dans [5] réalise une sélection adaptative au niveau GOP parmi les facteurs de mise à l'échelle $\{\times 1.0, \times 1.25, \times 1.5, \times 2.0\}$. Elle repose sur une stratégie à seuils semblable, comparant un PSNR down-up scaled à des seuils prédéfinis, dépendant du QP de la première image du GOP. Notons cette méthode VTS qui servira de référence pour l'analyse du BD-rate en Section 5.1.

3 Préparation du jeu de données

3.1 Jeu de données et protocole d'annotation

Notre jeu de données se compose d'extraits vidéo mono-scène de deux secondes, couvrant des environnements intérieurs et extérieurs et présentant une large diversité de textures et de dynamiques de mouvement. Pour garantir une représentativité statistique, une analyse spatio-temporelle fondée sur les métriques SI (Spatial Information) et TI (Temporal Information) a été réalisée, permettant de sélectionner un sous-ensemble équilibré de segments couvrant de façon homogène les différents niveaux de complexité spatiale et temporelle. Chaque segment est encodé selon 20 débits cibles. Pour chaque couple {segment, débit}, un vecteur de 36 *features*, introduit et validé dans [2], est extrait. Chaque échantillon, représenté par son vecteur de *features*, est annoté selon le ratio maximisant le compromis *débit-distortion*. Ce protocole conduit à 2 740 échantillons.

TABLEAU 1 – Classes de débit définies par intervalles

Catégorie	Moyen	Faible	Très faible	Ultra-faible
Intervalle (Mbps)	30 – 12	11 – 5	4 – 1,5	1 – 0,5

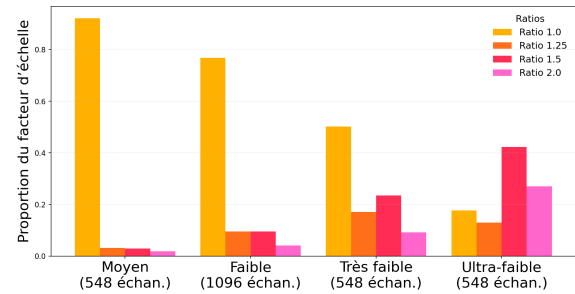


FIGURE 1 – Distribution normalisée des ratios selon les classes de débit (échan. pour échantillons).

3.2 Distribution des classes

Afin de favoriser un apprentissage équilibré, une attention particulière est portée à la distribution des étiquettes de mise à l'échelle. Une analyse pré-entraînement (Fig. 1) étudie la distribution conjointe des facteurs de scaling et des classes de débit. Les déséquilibres observés motivent l'utilisation d'une pondération de la loss dépendante des classes visant à compenser les points de fonctionnement sous-représentés et à améliorer la capacité de généralisation du modèle.

4 Stratégies de décision de résolution fondées sur l'apprentissage

Cette section présente nos stratégies pour sélectionner le ratio optimal au sein de l'ensemble des candidats $\mathcal{S}$: la *Cascade Binary Classifier* (CBC) et la *Single Multi-Class Classifier* (SMC).

4.1 Stratégie 1 : CBC (Cascade Binary Classifier)

La première stratégie reformule la sélection multi-classes en une succession de décisions binaires. À l'étape k , un classifieur binaire produit un score de confiance c_k indiquant si le facteur de scaling candidat doit être retenu. Si c_k dépasse un seuil τ_k , le facteur est sélectionné; sinon, la décision est transmise à l'étape suivante. Comme illustré Fig. 2, la cascade évalue séquentiellement des options de sous-échantillonnage de plus en plus conservatrices, en appliquant un seuil constant $\tau_k = 0.6$.

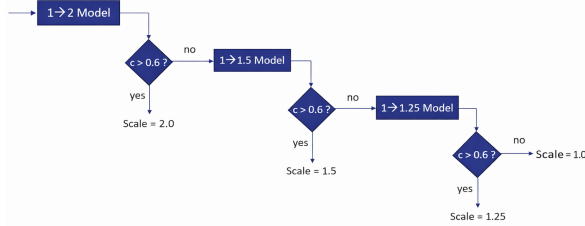


FIGURE 2 – Raffinement progressif de la décision via une cascade de classifieurs binaires.

4.2 Stratégie 2 : SMC (Single Multi-Class Classifier)

La seconde stratégie prédit le facteur de mise à l'échelle en une seule passe, au moyen d'un classifieur multi-classes avec une sortie softmax sur

$$\mathcal{R} = \{1.0, 1.25, 1.5, 2.0\}.$$

Pour un vecteur de features $\mathbf{x}$, le modèle fournit les probabilités *a posteriori* $p(r | \mathbf{x})$. Plutôt que de retenir directement $\arg \max$, une règle de post-traitement permet d'écarter les prédictions à faible confiance. Dans notre implémentation, les seuils sont fixés à $\tau_{2.0} = 0.255$ et $\tau_{\Sigma} = 0.76$. Le facteur final $\hat{r}$ est alors défini par :

$$\hat{r} = \begin{cases} 2.0, & p(2.0 | \mathbf{x}) \geq \tau_{2.0}, \\ \arg \max_{r \in \{1.5, 2.0\}} p(r | \mathbf{x}), & \sum_{r \geq 1.25} p(r | \mathbf{x}) \geq \tau_{\Sigma}, \\ \arg \max_{r \in \{1.0, 1.25\}} p(r | \mathbf{x}), & \text{sinon.} \end{cases} \quad (1)$$

4.3 Architecture des modèles

Le réseau SMC adopte une architecture en couches entièrement connectées $256 \rightarrow 128 \rightarrow 64 \rightarrow 32$, suivie d'une couche de sortie à 4 classes. La structure du CBC est composée de $128 \rightarrow 64 \rightarrow 32 \rightarrow 16$ neurones, avec des sorties binaires à chaque étape. Cela correspond à environ 52.8k paramètres entraînables pour le SMC et 46.9k pour le CBC.

5 Protocole expérimental et résultats

Les expériences sont conduites sur un ensemble représentatif de séquences 4K, incluant les séquences des CTC du JVET (H.266/VVC) des classes A1 et A2, ainsi que des contenus additionnels issus du *Common Feature Exploration* (CfE) post-VVC.

5.1 Performances BD-Rate sur les séquences CTC

Le Tableau 2 présente les performances en BD-Rate des approches CBC et SMC sur les séquences CTC 4K. Ces résultats sont comparés à une *borne supérieure* définie dans [6], correspondant à une adaptation au niveau scène par recherche exhaustive. Nos deux approches récupèrent l'essentiel des gains atteignables, en particulier aux très faibles et ultra-faibles débits, où l'adaptation de résolution apporte le bénéfice le plus marqué. En moyenne, CBC dépasse très légèrement SMC, suggérant une robustesse marginalement supérieure aux cas ambigus, grâce au raffinement séquentiel caractéristique de la cascade.

TABLEAU 2 – BD-Rate moyens (%) par rapport à l'ancre pleine résolution pour CBC et SMC, ventilé selon les classes de débit.

Catégorie	Borne sup.	CBC	SMC
	PSNR [%]	PSNR [%]	PSNR [%]
Moyen	-0.28	0.00	0.00
Faible	-0.54	-0.01	0.56
Très faible	-4.65	-3.03	-3.11
Ultra-faible	-15.56	-15.36	-15.38
Moyenne	-5.26	-4.60	-4.48

5.2 Comparaison à VTS sur CTC et CfE

Une évaluation étendue a été conduite en incluant les séquences CfE et en comparant aux performances obtenues avec la méthode VTS. Le Tableau 3 présente les résultats relatifs à l'ancre en pleine résolution, en se concentrant sur les régimes de débit faible et très faible. En moyenne, les stratégies fondées sur l'apprentissage affichent des performances comparables à celles de VTS sur ces deux régimes. À faible débit, VTS est quasi neutre en moyenne (-0.02%), mais une analyse par séquence révèle des erreurs coûteuses : sur *WomenFootball*, VTS induit un surcoût de $+4.37\%$, indiquant une décision de mise à l'échelle défavorable. À l'inverse, CBC et SMC évitent cette pénalité, témoignant de décisions plus conservatrices et robustes. Une forte dépendance au contenu est observée. Sur *CatRobot*, SMC engendre des surcoûts de $+1.31\%$ et $+3.31\%$ aux points de débit faible et très faible, respectivement. Ce comportement est cohérent avec la borne supérieure, qui indique un gain atteignable très limité (-0.16%). *CatRobot* présente en effet une grande richesse de détails et de textures fines, pour lesquels le sous-échantillonnage apporte intrinsèquement peu de bénéfices.

TABLEAU 3 – *BD-Rate (%) aux régimes de débits faible et très faible sur les séquences CTC et CjE.*

Catégorie	Borne sup.		VTS		CBC		SMC	
	Low BR	Very Low BR	Low BR	Very Low BR	Low BR	Very Low BR	Low BR	Very Low BR
Tango2	-0.75	-4.93	0.0	-4.11	0.0	-2.83	-0.13	-4.80
FoodMarket4	-3.63	-9.80	-2.87	-9.78	0.0	-8.78	0.0	-5.76
Campfire	0.0	-5.91	0.0	-0.51	0.0	-1.48	0.0	-0.51
CatRobot	0.0	-0.16	0.0	-0.15	0.0	0.06	1.31	3.31
DaylightRoad2	0.0	-0.68	0.0	0.01	0.0	-0.1	0.0	0.12
ParkRunning3	-0.06	-2.29	1.38	-1.84	0.0	-0.28	-0.06	-1.84
Crowdrun	-0.14	-2.85	0.0	-0.27	0.0	-0.76	-0.14	-2.87
DrivingPov3	0.0	-0.46	0.0	-0.27	0.0	0.0	0.0	-0.55
FashionLady	-4.42	-13.41	-3.79	-13.18	-4.42	-13.41	-0.91	-13.18
FireDance	-1.14	-3.83	0.62	-0.53	0.0	-0.85	0.0	-0.79
HallwayScene	-9.35	-3.08	0.0	-0.66	-0.7	0.0	0.0	-1.81
WomenFootball	-1.02	-5.22	4.37	-3.37	0.0	-4.18	-0.59	-5.22
Moyenne	-1.71	-4.38	-0.02	-2.89	-0.43	-2.72	-0.04	-2.83

et reste difficile à prédire. Les erreurs de SMC peuvent aussi être partiellement attribuées à la simplicité de la stratégie de post-traitement (Section 4.2), encore perfectible : une gestion plus avancée de la confiance ou un lissage décisionnel pourrait réduire ces dégradations isolées.

6 Conclusion

Cet article présente deux approches légères basées sur l'apprentissage pour l'adaptation de résolution au niveau scène. En décidant aux frontières de scènes, ces méthodes imposent la cohérence temporelle et limitent les basculements fréquents de résolution susceptibles de dégrader la qualité visuelle. CBC et SMC exploitent uniquement des features côté encodeur et des MLP compacts, évitant l'utilisation de métriques de distorsion dans le domaine pixel, peu compatibles avec les architectures matérielles. Les résultats obtenus sur les séquences CTC du JVET et CjE montrent que ces approches atteignent des performances comparables à VTS, tout en offrant un comportement plus stable aux bas débits. Au regard de la borne supérieure, ces approches récupèrent une part importante des gains atteignables aux très faibles et ultra-faibles débits. Ces observations confirment que l'adaptation au niveau scène fondée sur l'apprentissage constitue une alternative viable et pragmatique aux stratégies paramétriques classiques.

Références

- [1] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, et Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Trans. on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.
- [2] Kra-Tchimbie Koffi, Thomas Amestoy, Elie Mora, et Anissa Mokraoui. Lightweight learning of the enco-

ding scaling ratio for versatile video coding. Dans *VCIP*, 2025.

- [3] Riyu Lu, Yingwen Zhang, Hengyu Man, Meng Wang, Shiqi Wang, et Xiaopeng Fan. Learning the scale in reference picture resampling for versatile video coding. *IEEE Trans. on Multimedia*, 2025.
- [4] Shimin Huang, Cheolkon Jung, Yang Liu, et Ming Li. Cnn filter for super-resolution with rpr functionality in vvc. Dans *ICASSP*, pages 1–5. IEEE, 2023.
- [5] Junghak Nam, Sunmi Yoo, Jaehyun Lim, et Seung Hwan Kim. EE1-2.1 : Updates on RPR Encoder and Post-Filter. *JVET-AC0078-v1*, 2023. 29th Meeting of ITU-T SG16 WP3 and ISO/IEC JTC 1/SC 29, Teleconference, 11–20 January 2023.
- [6] Kra-Tchimbie Koffi, Thomas Amestoy, Elie Mora, et Anissa Mokraoui. Objective quality assessment of full-scene resampling-based coding for vvc standard. Dans *ISCAS*, 2025.
- [7] A.M. Bruckstein, M. Elad, et R. Kimmel. Down-scaling for better transform compression. *IEEE Trans. on Image Processing*, 12(9):1132–1144, 2003.
- [8] Tianliang Fu, Kai Zhang, Li Zhang, Shanshe Wang, et Siwei Ma. An efficient framework of reference picture resampling (rpr) for video coding. *IEEE Trans. on Circuits and Systems for Video Technology*, 32(10):7107–7119, 2022.
- [9] Thomas Stockhammer. Dynamic adaptive streaming over http— standards and design principles. Dans *Proceedings of the second annual ACM conference on Multimedia systems*, pages 133–144, 2011.
- [10] Ren-Jie Wang, Ming-Chen Chien, et Pao-Chi Chang. Adaptive down-sampling video coding. Dans *Multimedia on Mobile Devices*, volume 7542, pages 226–233. SPIE, 2010.

Finetuning Minimax de la règle de décision d'un réseau de neurones pour la classification de lésions cutanées avec une fonction de coût spécifique

Younes Habbal, Cyprien Gilet, Lionel Fillatre *

Université Côte d'Azur, CNRS, Laboratoire I3S

Université de technologie de Compiègne, CNRS, Laboratoire Heudiasyc

yhabbal@i3s.unice.fr, cyprien.gilet@hds.utc.fr, lionel.fillatre@i3s.unice.fr

Résumé

Dans la classification de lésions de la peau, les modèles optimisent souvent des critères tels que la précision ou le F1 score. Cependant, dans plusieurs applications réelles, toutes les classes n'ont pas les mêmes importances : détecter un cancer est plus important que détecter une lésion bénigne. Ainsi, chaque erreur devrait avoir un coût propre. De plus, les modèles de classification ont souvent des difficultés lorsque les proportions des ensembles d'entraînement ne correspondent pas à celles des ensembles de test. Dans cet article, nous utilisons le critère du Minimax Regret avec un réseau de neurones afin d'obtenir un classifieur robuste aux changements de proportions qui minimise un coût global spécialement construit pour la classification de lésions cutanées. Dans nos expériences sur le jeu de données réelles ISIC 2018, nous montrons que notre solution réduit le pire risque de classification, sans trop réduire la performance du modèle comparé à un réseau de neurones classique. Nous comparons également plusieurs architectures en tant qu'extracteurs de descripteurs informatifs ("features") pour le classifieur Minimax regret.

Mots clefs

Classification d'images, Minimax, Fonction de coût, CNN, ViT, Apprentissage profond, Extraction de descripteurs.

1 Introduction

En médecine, on privilégie les modèles précis et robustes. Pour ce faire, plusieurs types de méthodes d'apprentissage profond sont utilisées, comme les Réseaux de Neurones Convolutionnels (CNN) [1]. Classifier des images de lésions de peau peut se faire avec une multitude de méthodes, une des familles de méthodes la plus répandue est celle des CNN. Dans [2], l'auteur utilise un VGG19 pour classifier des images issues d'ISIC et Dermofit, obtenant ainsi une précision de 94,91%. D'autres méthodes basées sur les CNN les utilisent autrement comme [3] qui les emploie pour extraire des caractéristiques d'images provenant d'ISIC 2016 et 2017 avant de les associer à une des 3 classes considérées dans ces données avec 87,7% de précision moyenne. Il existe d'autres méthodes se basant pu-

rement sur de l'apprentissage profond, comme le Vision Transformer (ViT), qui exploite le modèle Transformer. Les auteurs de [4] utilisent cette méthode pour détecter des cancers sur les données d'ISIC 2019, atteignant une Area Under the Curve (AUC) de 92%.

Cependant, l'entraînement de ces méthodes nécessite des données qui sont souvent difficiles à obtenir, ce qui donne lieu à des jeux de données assez déséquilibrés. Pour y répondre, on emploie généralement des méthodes d'augmentation de données, mais celles-ci varient en fonction des données traitées et ne sont pas tout le temps facilement répliquables ou bien renseignées. De plus, des écarts de performance entre les données d'entraînement et celles de test peuvent survenir. Ceux-ci s'expliquent souvent par des différences au niveau des proportions des classes entre les deux jeux de données. Ce genre de pertes peut s'avérer problématique, surtout si une classe importante est concernée, il est donc primordial de rendre le modèle plus robuste à ces déséquilibres [5, 6]. Par ailleurs, toutes les classes n'ont pas les mêmes importances : détecter un cancer est plus important que détecter une lésion bénigne. Ainsi, il est nécessaire d'associer un coût spécifique [7] à chaque type d'erreur de classification pour que le système de classification réponde aux contraintes médicales.

Ce type de problème a été étudié en théorie de la décision statistique [5, 8] et deux solutions pour obtenir des modèles robustes se sont démarquées : minimiser le risque de Bayes maximum (approche Minimax) ou minimiser la perte de performance par rapport au classifieur optimal de Bayes (approche du regret Minimax). Cet article propose d'entraîner un classifieur, positionné en sortie d'un réseau de neurones, qui optimise le critère du regret Minimax. Cette approche peut être vue comme une forme légère d'adaptation du modèle, où seule la règle de décision est soumise à un réglage fin ("finetuning") plutôt que l'ensemble du réseau. Notre contribution principale consiste à comparer plusieurs modèles de réseau de neurones en tant qu'extracteurs de descripteurs ("features") pour notre classifieur du regret Minimax. Nous montrons que le critère du regret Minimax est moins pessimiste que celui du Minimax.

Dans la section 2, nous énonçons le problème étudié ainsi que les notations utilisées. Ensuite, la section 3 présente notre solution et nous décrivons les résultats obtenus dans

*Ce travail a été financé par la Région PACA et par l'initiative France 2030 dans le cadre du projet I-Démo PLICIA.

la section 4 avant de conclure dans la section 5.

2 Problème et notations

2.1 Classification de lésion

Commençons par introduire les données que nous allons utiliser. Elles proviennent de la tâche 3 du challenge ISIC 2018 [9, 10]. Il s’agit d’images de 7 classes de lésions différentes (voir Fig. 1). Les images sont séparées en 3 ensembles correspondants aux phases d’entraînement, de validation et de test. Le jeu de données d’entraînement contient 10015 images, celui de validation en a 1000 et celui de test en possède 1512.

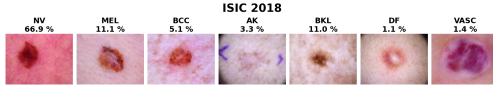


FIGURE 1 – Exemple de chaque classe d’ISIC 2018, associé à leurs proportions respectives

La classification d’images de lésions de peau est complexe car les données sont déséquilibrées et la mauvaise classification des images a des conséquences différentes en fonction des classes concernées. Nous utilisons donc une fonction de coûts construite dans [7]. Notre approche consiste à donner un score entre 1 et 10 aux sept classes du jeu de données selon 3 caractéristiques : leur traitement (son coût, sa durée et sa difficulté), leur sévérité (la lésion est-elle dangereuse ou douloureuse ?) et leur urgence (un délai de traitement est-il dangereux ?). Nous obtenons alors la matrice de coûts suivante :

$$C = \begin{bmatrix} 5.4 & 17.6 & 17.1 & 17.0 & 18.0 & 17.2 & 17.4 \\ 19.4 & 7.6 & 19.1 & 19.0 & 20.0 & 19.2 & 19.4 \\ 3.4 & 3.6 & 1.1 & 3.0 & 4.0 & 3.2 & 3.4 \\ 1.4 & 1.6 & 1.1 & 1.0 & 2.0 & 1.2 & 1.4 \\ 30.4 & 30.6 & 30.1 & 30.0 & 11.0 & 30.2 & 30.4 \\ 5.4 & 5.6 & 5.1 & 5.0 & 6.0 & 1.2 & 5.4 \\ 8.4 & 8.6 & 8.1 & 8.0 & 9.0 & 8.2 & 2.4 \end{bmatrix}. \quad (1)$$

Le coût résultant de la classification d’une image de classe k dans la classe l est égal au coefficient $C(k, l) = C_{kl}$.

2.2 Minimax

Soit $\Delta = \{\delta : \mathcal{X} \rightarrow \mathcal{Y}\}$ l’ensemble contenant tous les classifieurs possibles où $\mathcal{X}$ est l’espace d’entrée et $\mathcal{Y} = \{1, \dots, K\}$ les étiquettes des $K = 7$ classes. Soit $\mathbb{S}$ l’ensemble de tous les a priori (c’est-à-dire toutes les proportions de classes possibles) :

$$\mathbb{S} \stackrel{\text{def}}{=} \left\{ \pi \in [0, 1]^K : \sum_{k \in \mathcal{Y}} \pi_k = 1 \right\} \quad (2)$$

Le risque moyen d’un classifieur δ associé aux proportions $\pi \in \mathbb{S}$ et à la matrice de coût C est défini par

$$\hat{r}_C(\pi, \delta) = \sum_{k \in \mathcal{Y}} \pi_k \hat{R}_{k,C}(\delta), \quad (3)$$

où $\hat{R}_k(\delta)$ est le risque de classe associé à la classe k :

$$\hat{R}_{k,C}(\delta) = \sum_{l \in \mathcal{Y}} C_{kl} \hat{\mathbb{P}}(\delta(X) = l \mid Y = k), \quad (4)$$

avec $\hat{\mathbb{P}}(\delta(X) = l \mid Y = k)$ la probabilité empirique (estimée avec les données d’entraînement) que δ choisisse la classe l quand la vérité terrain est la classe k . Le plus petit risque moyen, appelé risque de Bayes, est donné par

$$V_C(\pi) \stackrel{\text{def}}{=} \min_{\delta \in \Delta} \hat{r}_C(\pi, \delta). \quad (5)$$

Le risque de Bayes convient à notre problème de classification de lésions de peau car il s’adapte bien à n’importe quelle fonction de coût C . Nous cherchons à résoudre le problème des changements de proportions (“prior probability shift”). Nous devons donc choisir une mesure adaptée afin de quantifier la robustesse des classifieurs à ces changements. Nous utilisons donc la moyenne du regret (aussi appelée déviation) $D_C(\pi, \delta)$ qui représente la différence entre le risque moyen d’un classifieur δ et le risque de Bayes pour un a priori π donné :

$$D_C(\pi, \delta) \stackrel{\text{def}}{=} \hat{r}_C(\pi, \delta) - V_C(\pi). \quad (6)$$

Nous décidons d’utiliser un “Minimax Regret Classifier (MRC)” afin d’avoir un classifieur robuste aux changements de proportions :

$$\delta^{MRC} = \underset{\delta \in \Delta}{\operatorname{argmin}} \max_{\pi \in \mathbb{S}} D_C(\pi, \delta). \quad (7)$$

Cette méthode minimise le maximum de sa déviation sur l’ensemble des a priori $\mathbb{S}$. Cela permet de diminuer les pertes de performances sur des a priori inhabituels. Nous aurions pu choisir un classifieur Minimax classique δ^M , car il est également robuste aux changements de proportions étant donné qu’il minimise le risque moyen sur $\mathbb{S}$:

$$\delta^M = \underset{\delta \in \Delta}{\operatorname{argmin}} \max_{\pi \in \mathbb{S}} \hat{r}_C(\pi, \delta). \quad (8)$$

Cependant, ce critère peut-être trop conservateur ou pessimiste, comme nous le verrons dans nos expériences.

3 Solution proposée

3.1 Démarche générale

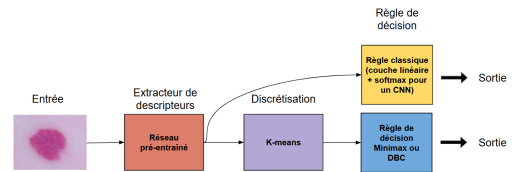


FIGURE 2 – Schéma explicatif de notre solution

Nous nous inspirons de [8] où un réseau de neurones est utilisé en tant qu’extracteur de descripteurs. Le réseau de

neurones pré-entraîné est considéré sans sa couche de classification finale afin d'extraire une représentation vectorielle de chaque image. Après l'extraction, comme illustré dans la Figure 2, il y a 2 voies possibles. La première voie évalue le réseau extracteur en utilisant sa règle de décision classique (pour un CNN, une couche dense suivie d'un softmax par exemple). La deuxième consiste à utiliser un classifieur DBC, DMC et DMRC présentés dans la sous-section suivante. Ces classifieurs nécessitent des entrées discrètes. Nous réalisons donc une discrétisation des descripteurs à l'aide de l'algorithme K-means ce qui revient à partitionner l'espace des descripteurs en K régions.

3.2 Discrétisation

Le problème du Minimax Regret Classifier (7) est le calcul du risque Bayésien $V_C(\pi)$ car il faut estimer des probabilités des descripteurs dans un espace de grande dimension. Comme cela est difficile dans le cas continu, nous choisissons de réaliser une discrétisation de l'espace des descripteurs $\mathcal{X}$ en T sous-ensembles distincts $\{\Omega_1, \dots, \Omega_T\}$ avec K-means. Ces sous-ensembles sont tels que $\cup_{t=1}^T \Omega_t = \mathcal{X}$. Cela définit l'application $\phi : \mathcal{X} \mapsto \mathcal{T} = \{1, \dots, T\}$ qui satisfait $\phi(x^{(i)}) = t$ si et seulement si $x^{(i)} \in \Omega_t$. L'ensemble d'entraînement est constitué des couples $(x^{(i)}, y^{(i)})$ où $x^{(i)} \in \mathcal{X}$ est un descripteur extrait et $y^{(i)} \in \mathcal{Y}$ est son étiquette. La discrétisation nous permet d'estimer $\hat{p}_{kt}$, la probabilité d'observer une donnée de classe k dans l'ensemble Ω_t , à l'aide des données d'entraînement :

$$\hat{p}_{kt} = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{\phi(x^{(i)})=t\}}, \quad (9)$$

avec $\mathcal{I}_k = \{i \in \mathcal{I} : y^{(i)} = k\}$ et $n_k = |\mathcal{I}_k|$ le nombre d'instances de $\mathcal{I}_k$.

D'après [8], le classifieur discret de Bayes (DBC) $\delta_{\pi, C}^B = \argmin_{\delta \in \Delta} \hat{r}(\pi, \delta)$ est donné par $\delta_{\pi, C}^B : x \in \mathcal{X} \mapsto \argmin_{l \in \mathcal{Y}} f_l(\pi, x)$, où pour chaque classe $l \in \mathcal{Y}$,

$$f_l(\pi, x) = \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} C_{kl} \pi_k \hat{p}_{kt} \mathbb{1}_{\{\phi(x)=t\}}. \quad (10)$$

Le Discrete Minimax Regret Classifiers (DMRC) et le Discrete Minimax Classifier (DMC), sont des DBC particuliers construits respectivement avec les a priori $\tilde{\pi} \in \mathbb{S}$ et $\bar{\pi} \in \mathbb{S}$. D'après [11], le DMRC est donné par :

$$\delta_{\tilde{\pi}, C}^B : x \mapsto \argmin_{l \in \mathcal{Y}} \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} C_{kl} \tilde{\pi}_k \hat{p}_{kt} \mathbb{1}_{\{\phi(x)=t\}}, \quad (11)$$

où $\tilde{\pi} = \argmax_{\pi \in \mathbb{S}} V_{\tilde{C}}(\pi)$ avec $\tilde{C}_{kl} = C_{kl} - C_{kk}$. La valeur de l'a priori $\tilde{\pi}$ du DMRC est calculé avec l'algorithme décrit dans [8]. Le DMC et son a priori $\bar{\pi} \in \mathbb{S}$ sont calculés de façon similaire.

4 Expériences

Nous allons maintenant présenter les résultats de nos expériences. Pour étudier plusieurs architectures utilisant notre

méthode, nous avons choisi un Vision Transformer [12] (un type de modèle assez récent et performant) et deux CNNs (EfficientNet-B0 qui représente un bon compromis entre le nombre de paramètres et la performance ainsi que ResNet-50 qui est un modèle très utilisé comme référence). Cette étude permet donc de comparer les représentations locales des CNN aux représentations globales des ViT. Les 3 architectures sont pré-entraînées. Après l'extraction, les règles de décisions classiques (couche linéaire + softmax) sont entraînées sur les descripteurs obtenus. Quant aux DBC, DMC et DMRC, nous réalisons une discrétisation avec Kmeans comportant 205 centroïdes afin d'apprendre les a priori optimaux pour chaque méthode (nous avons essayé plusieurs nombres de centroïdes ; le nombre retenu est assez stable et minimise correctement les risques). Pour tester la robustesse des modèles, nous générons 1000 a priori différents que nous utilisons pour comparer les 4 règles de décisions différentes (DBC, DMC, DMRC et le CNN/ViT) sur chacune des 3 architectures. Pour chaque a priori testé, nous créons une base de test avec ses proportions puis nous évaluons les classifieurs.

	AK	BCC	BKL	DF	MEL	NV	VASC
DBC	0.001773	0.030142	0.00591	0.003546	0.131206	0.792553	0.03487
DMC_eff	0.161839	0.232398	0.00000	0.000000	0.605764	0.000000	0.00000
DMRC_eff	0.162279	0.198713	0.00000	0.000000	0.495728	0.143281	0.00000
DMC_res	0.161963	0.232471	0.00000	0.000000	0.605566	0.000000	0.00000
DMRC_res	0.039258	0.165843	0.00000	0.000000	0.431908	0.362990	0.00000
DMC_vit	0.161810	0.232380	0.00000	0.000000	0.605810	0.000000	0.00000
DMRC_vit	0.104081	0.197346	0.00000	0.000000	0.550266	0.148307	0.00000

FIGURE 3 – A priori utilisés dans DBC, DMC et DMRC.

Les a priori calculés et utilisés pour les méthodes sont indiqués dans la Fig. 3 : le DBC utilise les proportions du jeu de données ISIC 2018 alors que DMC et DMRC donnent plus d'importance à la classe la plus dangereuse (MEL : la classe des mélanomes). Les classifieurs Minimax négligent certaines classes moins importantes mais le DMRC donne tout de même un peu d'importance à la classe majoritaire NV, même si elle n'est pas dangereuse.

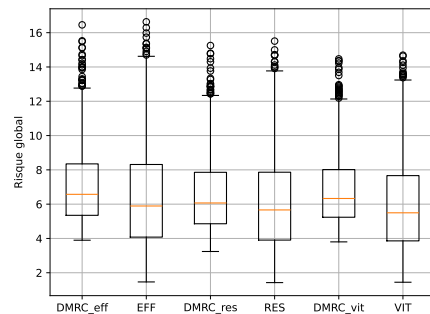


FIGURE 4 – Boxplot du risque global $\hat{r}_C$.

Le risque moyen des modèles est présenté sur la Fig. 4. Les risques moyens des méthodes classiques sont plus bas que ceux des DMRC mais leurs variances et leurs risques maximum, leur pire risque, sont plus élevés.

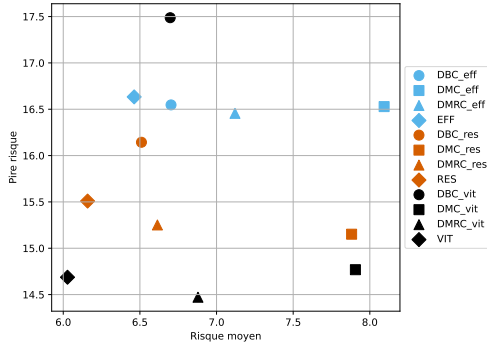


FIGURE 5 – Pire risque en fonction du risque moyen.

Cela se confirme dans la Fig. 5 où nous pouvons voir que les pires risques des méthodes Minimax sont en général plus bas que ceux des méthodes classiques, au prix d’une augmentation du risque moyen. Ensuite, la Fig. 6 représente la déviation des méthodes en fonction de la proportion de la classe la plus dangereuse π_{MEL} . Peu importe

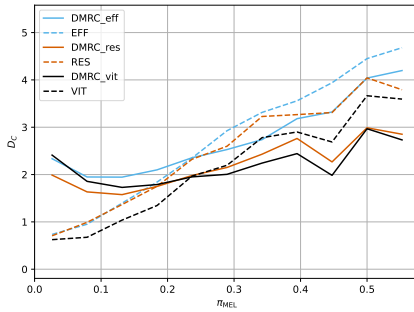


FIGURE 6 – D_C en fonction de π_{MEL} .

l’extracteur, nous constatons que lorsque π_{MEL} est faible, les déviations de DMRC sont supérieures à celles des modèles classiques. Cependant, lorsque π_{MEL} est grand, cette tendance s’inverse, ce qui est en accord avec nos observations précédentes : les cas "hors-normes" sont mieux gérés par les méthodes Minimax.

5 Conclusion

Nous avons comparé des règles de décisions robustes aux changements de proportions des classes qui minimisent un risque d’erreur adapté aux lésions cutanées. Bien que les modèles classiques aient de meilleures performances sur des proportions fixes, ils sont vulnérables à des proportions inhabituelles, ce qui peut être résolu par l’emploi de méthodes Minimax comme DMRC. De nombreux modèles

peuvent être utilisés pour extraire les descripteurs utilisés par ces méthodes, mais il semblerait que le ViT soit le plus adapté sur les données ISIC 2018.

Références

- [1] I. Mienye, T. Swart, G. Obaido, M. Jordan, et P. Ilono. Deep convolutional neural networks in medical image analysis : A review. *Information*, 16 :195, 03 2025.
- [2] M. Gupta, R. Kumar, Nandan, A.K. Pradhan, et A.J. Obaid. Skin disease detection using neural networks. Dans *2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, pages 1–4, 2024.
- [3] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, et C. Wang. Fusing fine-tuned deep features for skin lesion classification. *Computerized Medical Imaging and Graphics*, 71 :19–29, 2019.
- [4] A. Garcia, J. Zhou, G. Pinero-Crespo, T. Beachkofsky, et X. Huang. Clinical application of vision transformers for melanoma classification : A multi-dataset evaluation study. *Cancers*, 17(21), 2025.
- [5] J. O. Berger. *Statistical decision theory and Bayesian analysis ; 2nd ed.* Springer Series in Statistics. Springer, New York, 1985.
- [6] R. Alaiz-Rodríguez, A. Guerrero-Curieses, et J. Cid-Sueiro. Minimax regret classifier for imprecise class distributions. *Journal of Machine Learning Research*, 8 :103–130, Jan 2007.
- [7] Y. Habbal, C. Gilet, et L. Fillatre. Skin lesion classification combining minimax regret classifier and deep learning with task-adapted costs. Soumis à ICIP 2026, 2026.
- [8] C. Gilet, M. Guyomard, S. Barbosa, et L. Fillatre. Multiclass minimax learning for deep neural networks. Dans *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 1265–1269, 2023.
- [9] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, et A. Halpern. Skin lesion analysis toward melanoma detection 2018 : A challenge hosted by the international skin imaging collaboration (ISIC), 02 2019.
- [10] P. Tschandl, C. Rosendahl, et H. Kittler. The ham10000 dataset : A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 08 2018.
- [11] C. Gilet. *Discrete Minimax Classifier for Personalized Diagnosis in Medicine*. Thèse de doctorat, Université Côte d’Azur, 2021.
- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, et N. Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale, 2021.

Investigation des effets de dégradations appliquées aux entrées des générateurs image-vers-image

Ahmed IBRAHIM

Abderrezzaq SENDJASNI
CNRS, XLIM UMR 7252, Université de Poitiers, France

Mohamed-Chaker LARABI

Résumé

Ce travail porte sur l'évaluation de la qualité des contenus visuels générés par intelligence artificielle (Artificial Intelligence Generated Content - AIGC), un domaine dans lequel les métriques classiques telles que le PSNR et le SSIM montrent aujourd'hui leurs limites. L'étude s'intéresse à une problématique encore peu explorée : l'impact des dégradations présentes dans les images d'entrée sur les performances des modèles image-vers-image (I2I). Afin de pallier le manque de données adaptées à cette analyse, nous avons construit une nouvelle base de données nommée UP-AIGIC (University of Poitiers - AI Generated Image Content). Cet ensemble comprend 109 800 images générées à partir de neuf modèles génératifs (GAN et modèles de diffusion), déclinées selon neuf styles artistiques et plusieurs types de dégradations (bruit, flou, contraste, compression JPEG, etc.). Cette base est utilisée pour analyser la dérive sémantique et la qualité des images générées à l'aide de métriques récentes telles que CLIP, DINOv1 et FID. Les résultats obtenus permettent de mieux comprendre l'influence des dégradations d'entrée sur le comportement des modèles génératifs et sur la fidélité des contenus produits.

Mots clefs

IA générative, modèles de diffusion, GAN, génération image-vers-image, évaluation de qualité d'image, métriques sémantiques

1 Introduction

L'introduction d'artefacts dans les images et les vidéos, due aux différents systèmes de traitement du signal, a conduit au développement du domaine de l'évaluation de la qualité perceptuelle des contenus visuels. De nombreux travaux ont été consacrés à la conception de métriques objectives capables d'estimer cette qualité tout en tenant compte de la perception humaine [1].

Avec l'émergence récente de l'intelligence artificielle générative, la question de l'évaluation devient encore plus cruciale. Les modèles génératifs modernes, tels que les GAN ou les modèles de diffusion [2, 3], sont aujourd'hui capables de produire des images d'un réalisme impressionnant.

Cette évolution s'accompagne d'une forte croissance de la production de contenus visuels générés par intelligence artificielle, communément appelés AIGC (AI-Generated Content), dans de nombreux domaines tels que les médias, le divertissement ou la création artistique.

Cependant, l'évaluation de la qualité de ces contenus demeure un défi majeur. Les métriques objectives traditionnelles, telles que PSNR ou SSIM, initialement conçues pour l'évaluation de la compression ou de la restauration d'images, ne permettent pas toujours de capturer correctement les caractéristiques perceptuelles ou sémantiques propres aux images générées par IA [4].

Dans ce contexte, notre travail vise à explorer une problématique encore peu étudiée : l'impact des dégradations appliquées aux images d'entrée sur le comportement des modèles image-vers-image. En effet, dans de nombreuses applications réelles, les images d'entrée peuvent contenir différents types d'altérations (bruit, flou, compression), susceptibles d'influencer la qualité et la cohérence sémantique des images générées. Pour mener cette étude, nous avons construit une nouvelle base de données intitulée UP-AIGIC, spécifiquement dédiée à l'analyse des effets des dégradations d'entrée sur les modèles génératifs image-vers-image. Cette base comprend un large ensemble d'images générées à partir de plusieurs architectures génératives, combinées à différents styles visuels et niveaux de dégradation.

Les performances des modèles sont ensuite analysées à l'aide de métriques récentes basées sur l'apprentissage profond, telles que CLIP, DINOv1 et FID, permettant d'évaluer à la fois la qualité perceptuelle et la cohérence sémantique des images produites.

2 Les données AIGC

L'évaluation des contenus générés par IA demeure une tâche complexe, notamment en raison de l'absence de standards clairement établis pour mesurer leur qualité [4]. Contrairement aux images naturelles, les contenus générés peuvent être évalués selon plusieurs dimensions : qualité perceptuelle, cohérence sémantique, fidélité au prompt ou encore esthétique visuelle.

Par ailleurs, l'analyse des bases de données existantes constitue une étape essentielle pour comprendre les capa-

cités et les limitations des modèles génératifs actuels. Dans ce cadre, nous avons examiné les principales bases de données AIGC recensées dans [4], ainsi que les publications associées, afin d'identifier les lacunes existantes.

Plusieurs dimensions doivent être considérées lors de l'évaluation des contenus générés par IA. D'une part, il est nécessaire d'analyser la qualité technique des images produites, en veillant notamment à l'absence d'artefacts ou de dégradations susceptibles d'affecter la netteté ou la lisibilité de l'image. D'autre part, des critères plus subjectifs, tels que l'esthétique de l'image et son adéquation avec les préférences perceptuelles humaines, doivent également être pris en compte. De plus, il est important d'évaluer la capacité des modèles génératifs à interpréter correctement l'instruction fournie dans le prompt et à produire un contenu visuel cohérent avec cette instruction. Cela inclut également la génération correcte de structures complexes telles que les visages, les mains ou les membres, qui constituent souvent des éléments sensibles pour les modèles génératifs. Enfin, certaines considérations éthiques doivent également être prises en compte, notamment l'absence de contenus sensibles ou inappropriés dans les images générées.

À notre connaissance, la majorité des ensembles de données existants se concentrent principalement sur la génération texte-vers-image (Text-to-Image, T2I). Une exception notable est la base PKU-AIGQA-4k [5], qui aborde la problématique de la génération image-vers-image (Image-to-Image, I2I). Toutefois, bien que les bases de données les plus récentes couvrent désormais une diversité croissante d'images générées, un déséquilibre persiste entre les contenus produits par les modèles de diffusion [6] et ceux générés par les modèles GAN [7], ces derniers étant généralement moins représentés.

3 Construction de UP-AIGIC



FIGURE 1 – Exemples des premiers essais avec le modèle commercial (Artguru). Plusieurs types de dégradations sont appliquées sur l'image fournie en entrée du modèle.

Compte tenu du nombre limité d'images générées à partir de modèles image-vers-image, nous avons entrepris de construire un nouvel ensemble de données afin d'étudier

l'impact des dégradations appliquées aux images d'entrée sur le comportement des modèles génératifs (cf. figure 1).

Des premières expérimentations réalisées avec un modèle commercial (Artguru) ont montré que certaines dégradations appliquées aux images d'entrée peuvent entraîner une dérive sémantique dans les images générées. Comme illustré dans la figure 1, certaines altérations telles que le flou gaussien ou le flou de mouvement peuvent conduire le modèle à produire des objets différents de ceux présents dans l'image de référence. À l'inverse, d'autres dégradations préservent l'objet principal tout en modifiant certaines caractéristiques visuelles.

Ces observations ont motivé la construction d'un nouvel ensemble de données intitulé UP-AIGIC (University of Poitiers – AI Generated Image Content), dédié à l'analyse systématique de l'influence des dégradations d'entrée sur les modèles génératifs image-vers-image, comme illustré sur la figure (2).

La construction de la base UP-AIGIC s'effectue selon les étapes suivantes :

3.1 Étape 1 : Génération des images dégradées

Trente images issue de la base RAISE [8] couvrant différentes catégories visuelles (personnes, bâtiments et scènes naturelles) ont été recadrées recadrées à une résolution de 512x512 pixels. Cinq types de dégradations ont ensuite été appliqués

- Bruit gaussien
- Flou gaussien
- Flou de mouvement
- Modification du contraste
- Compression JPEG

Chaque dégradation a été appliquée selon 12 niveaux d'intensité, produisant un total de : 1800 images et 30 originales (clean).

3.2 Étape 2 : Génération des images par modèles I2I

Les images obtenues ont ensuite été utilisées comme entrées pour neuf modèles image-vers-image (I2I) incluant deux architectures basées sur GAN () et 7 modèles de diffusion (AniagineXL, ControlNet, StableDiffusion [2] 1.5 et XL, OmniConsistency [9], Flux-dev [10], InstructPix2Pix [11])

Pour chaque image, neuf styles artistiques ont été générés : Abstrait, Baroque, Ghibli, Van Gogh, Ukiyoe, Pixar, Seinen, Science-fiction et styles des années 70.

Au total, la base UP-AIGIC comprend 109 800 images générées, dont 108 000 images produites à partir d'images dégradées et 1 800 images générées à partir d'images non dégradées. La figure (2) présente la distribution des images selon les styles artistiques et les modèles génératifs utilisés.

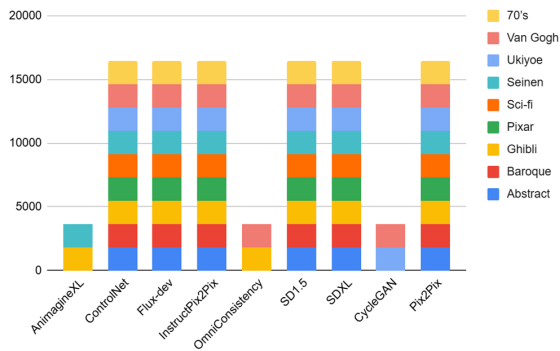


FIGURE 2 – Répartition des données de la base UP-AIGIC par rapport aux styles et aux modèles utilisés avec quelques exemples d’images présentes dans UP-AIGIC.

4 Performance des métriques objectives

Afin d’évaluer l’impact des dégradations sur les images générées, nous analysons dans cette section le comportement de plusieurs métriques objectives couramment utilisées pour l’évaluation des contenus générés par IA. Nous considérons en particulier CLIP [12], DINOv1 [13] et FID [14], qui permettent d’étudier respectivement la cohérence sémantique, la similarité des représentations visuelles et la qualité statistique des images générées. En raison des contraintes d’espace, les résultats présentés se concentrent sur un seul modèle génératif (Flux1-dev) et deux types de dégradations (le flou gaussien et la modification de contraste), sélectionnés afin d’illustrer de manière représentative les principales observations de cette étude.

À partir des résultats obtenus avec CLIP CosSim (cf. première ligne de la figure 3), nous observons que la valeur du score pour le flou gaussien varie légèrement entre les différents styles et aussi en fonction de l’intensité de la dégradation. Cependant, ces variations semblent irrégulières. Pour la modification de contraste, nous observons une tendance à la hausse modérée jusqu’au niveau 0, suivie d’une diminution lorsque le contraste devient plus intense.

En revanche, les résultats obtenus avec DINOv1 (cf. deuxième ligne de la figure 3) mettent en évidence un comportement différent. Les scores présentent peu de variations entre les styles, mais montrent des différences plus marquées entre les niveaux de dégradation, et surtout entre les différents types de dégradations.

Les résultats obtenus avec FID (cf. troisième ligne de la figure 3) montrent que la métrique ne met pas en évidence de variations significatives entre les différents niveaux de dégradation pour un même type de dégradation. En revanche, des différences plus marquées apparaissent entre les styles artistiques, même lorsque le type de dégradation

reste identique.

Ces observations suggèrent que les métriques reposant sur des représentations profondes capturent des aspects complémentaires des images générées. Alors que CLIP apparaît relativement peu sensible aux dégradations appliquées, DINO et FID révèlent des variations plus significatives en fonction du type de dégradation et du style artistique.

5 Conclusion

Ce travail introduit UP-AIGIC, une base de données dédiée à l’analyse de l’impact des dégradations des images d’entrée sur les modèles génératifs image-vers-image. Les résultats montrent que certaines dégradations peuvent entraîner une dérive sémantique importante, conduisant les modèles à générer des contenus sensiblement différents des images de référence. Nos observations révèlent également que, dans certains cas, des écarts apparaissent déjà lorsque les images d’entrée ne sont pas dégradées, ce qui se traduit par des scores CLIP et DINOv1 relativement faibles pour des modèles comme Stable Diffusion 1.5 et ControlNet.

L’analyse comparative des métriques met en évidence des comportements distincts : alors que CLIP apparaît relativement stable face aux variations de dégradation, DINOv1 et FID se montrent plus sensibles aux niveaux de dégradation et aux styles artistiques. Ces résultats soulignent la nécessité de recourir à des approches d’évaluation multidimensionnelles pour caractériser plus finement la qualité et la robustesse des contenus générés par IA.

Une suite logique à ce travail consisterait à identifier les caractéristiques permettant de déterminer si des images générées par IA, proviennent des sources dégradées. Cela ouvrirait des applications en détection d’artefacts IA ainsi qu’en traçabilité des données d’entraînement.

Références

- [1] Pedram Mohammadi, Abbas Ebrahimi-Moghadam, et Shahram Shirani. Subjective and objective quality assessment of image : A survey. *Majlesi Journal of Electrical Engineering*, 9, 06 2014.
- [2] Stable diffusion : Genai for image content. Dernière visite en mars 2026.
- [3] Nanobanana : Genai for image content. Dernière visite en mars 2026.
- [4] Abhijay Ghildyal, Yuanhan Chen, Saman Zadtootaghaj, Nabajeet Barman, et Alan C. Bovik. Quality prediction of ai generated images and videos : Emerging trends and opportunities. *ArXiv*, abs/2410.08534, 2024.
- [5] Jiquan Yuan, Jihe Li, Fanyi Yang, Xinyan Cao, Jinming Che, Jinlong Lin, et Xixin Cao. Pku-aigiga-4k : A perceptual quality assessment database for both text-to-image and image-to-image ai-generated images. Dans *Proceedings of the IEEE/CVF Interna-*

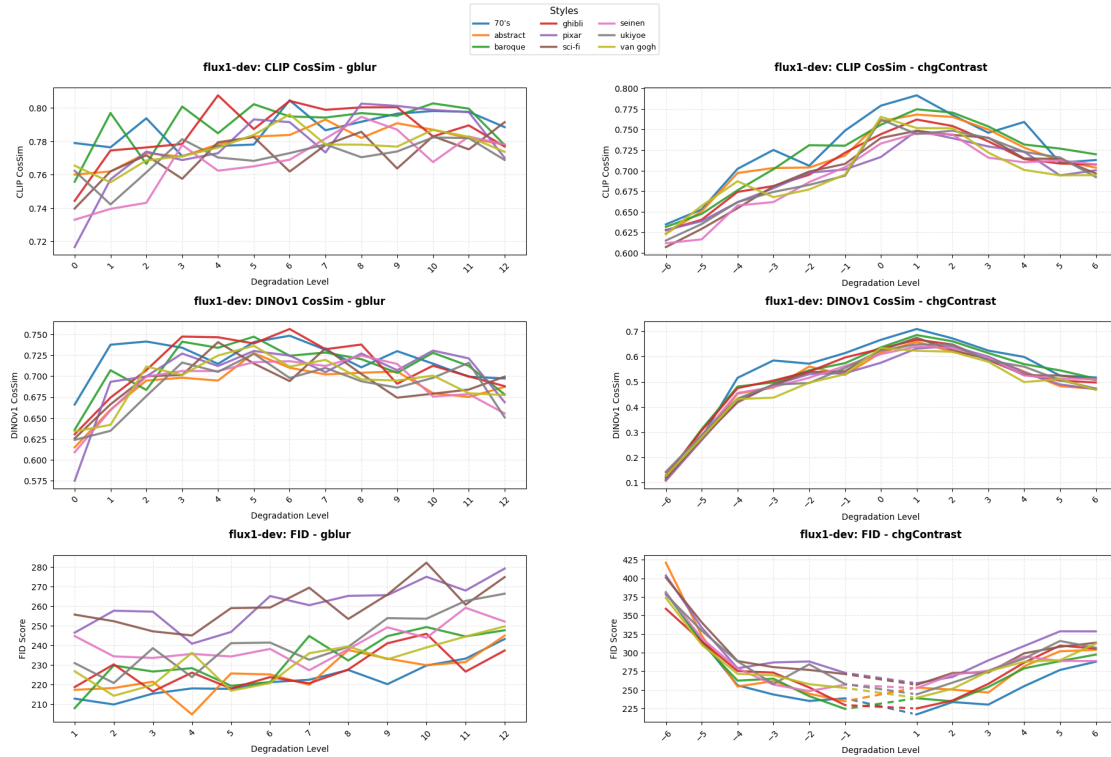


FIGURE 3 – Scores des métriques CLIP CosSim, DINOv1 CosSim et FID respectivement. Les colonnes représentent la variation des scores des styles artistiques en fonction des différents niveaux de dégradation : Flou Gaussien à gauche et Changement de Contraste à droite. Le niveau de dégradation 0 représente le cas des images générées sans dégradation. Les échelles des axes des ordonnées ne sont pas uniformes afin de mieux visualiser les tendances.

- tional Conference on Computer Vision (ICCV) Workshops, pages 3331–3340, October 2025.
- [6] Jonathan Ho, Ajay Jain, et P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.
 - [7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, et Yoshua Bengio. Generative adversarial networks. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7, 2021.
 - [8] Raise dataset, 2015. D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, G. Boato, RAISE – A Raw Images Dataset for Digital Image Forensics, ACM Multimedia Systems, Portland, Oregon.
 - [9] Yiren Song, Cheng Liu, et Mike Zheng Shou. Omniconsistency : Learning style-agnostic consistency from paired stylization data. Dans *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
 - [10] Flux-dev. Dernière visite en mars 2026.
 - [11] Tim Brooks, Aleksander Holynski, et Alexei A. Efros. Instructpix2pix : Learning to follow image editing instructions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2022.
 - [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, et Ilya Sutskever. Learning transferable visual models from natural language supervision. Dans *International Conference on Machine Learning*, 2021.
 - [13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, et Armand Joulin. Emerging properties in self-supervised vision transformers. Dans *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
 - [14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, et Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Dans *Neural Information Processing Systems*, 2017.

Rééclairage de visage et harmonisation des teintes dans l'espace latent pour un faceswapping réaliste basé sur les modèles de diffusion

Thu Hien Le[†]

Christophe Charrier[†]

Emmanuel Giguet[†]

Maxime Bérubé[‡]

[†] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen ; France

[‡] Université du Québec à Trois-Rivières, GRSF ; Canada

{thu-hien.le, christophe.charrier, emmanuel.giguet}@unicaen.fr

maxime.berube2@uqtr.ca

Résumé

La génération de deepfakes par faceswapping reste une tâche complexe nécessitant un équilibre entre préservation de l'identité, cohérence des attributs et photoréalisme. Dans cet article, nous proposons un cadre de génération en trois étapes, sans entraînement, qui améliore le réalisme en alignant explicitement l'éclairage et l'apparence de la peau avant la synthèse par diffusion. Notre approche affine la cohérence photométrique et la teinte des visages tout en préservant la structure faciale et en s'intégrant parfaitement à un modèle de faceswapping par diffusion standard. Les expériences menées sur la base CelebAMask-HQ démontrent des améliorations significatives tant au niveau du réalisme visuel que sur la préservation des attributs, atteignant un score FID de 7,16 par rapport à la méthode de référence. La méthode proposée offre une solution efficace et robuste pour un échange de visage réaliste sous diverses conditions d'éclairage et d'apparence, sans nécessiter d'entraînement supplémentaire du modèle.

Mots clefs

Échange de visage, modèles de diffusion, espace latent.

1 Introduction

L'échange de visage, ou *face swapping* consiste à transférer l'identité d'un visage source vers une image cible tout en préservant les attributs propres à cette dernière, tels que la pose, l'expression, l'illumination et la géométrie faciale, dans un rendu photoréaliste. C'est une technologie essentielle pour le divertissement, l'augmentation de données et la réalité virtuelle [1, 2, 3, 4]. Les méthodes pionnières basées sur les réseaux antagonistes génératifs (*Generative Adversarial Networks*, GAN) ont prouvé leur efficacité mais elles souffrent d'instabilité et d'une généralisation limitée [5, 2, 6]. Récemment, les modèles de diffusion se sont imposés par leur stabilité et leur haute fidélité de synthèse [3, 7, 4].

Malgré ces avancées, les systèmes basés sur la diffusion font face à des défis persistants. D'une part, les artefacts perceptuels, tels que les incohérences d'éclairage, les couleurs de peau dépareillées ou les fusions imparfaites, res-

sent fréquents. D'autre part, l'amélioration de la préservation des attributs nécessite souvent un réentraînement coûteux des modèles de grande taille.

Pour remédier à ces problèmes, nous proposons un cadre de faceswapping sans réentraînement, reposant sur une stratégie de prétraitement modulaire. Notre approche décompose le processus en trois étapes, alignant explicitement l'éclairage et l'apparence de la peau avant la synthèse. En intégrant un module dédié au rééclairage des visages et un module dédié à l'harmonisation des couleurs de peau à un système de transfert par diffusion préexistant, notre méthode améliore significativement le réalisme et la cohérence des attributs, sans nécessiter de réentraînement du modèle.

2 Travaux connexes

2.1 Méthodes de Face swapping

Pour un résultat qualitatif, il est nécessaire de dissocier explicitement les composantes identitaires des autres attributs visuels. Alors que les premières approches reposaient surtout sur des modèles génératifs adversariaux, tels que [1] et SimSwap [2], les modèles de diffusion ont redéfini les standards de qualité en synthèse d'images.

Dans ce travail, nous nous appuyons sur REFace [4], un système d'inpainting conditionnel fondé sur des modèles de diffusion latents (LDM). Le conditionnement fusionne les caractéristiques d'identité de la source avec les attributs de la cible — pose, expression et géométrie — garantissant un transfert contrôlé de l'identité tout en préservant la cohérence structurelle et photométrique de l'image cible. Bien qu'offrant une fidélité de pointe, REFace peut introduire des artefacts sous un éclairage non contraint.

2.2 Ré-éclairage et cohérence de l'éclairage

La cohérence photométrique est cruciale pour un transfert réaliste. Les méthodes de ré-éclairage décomposent souvent les images en composantes intrinsèques : albedo, ombrage et éclairage. L'intensité du pixel à la position x d'une image est représentée selon l'Éq. (1).

$$L_o(x, \omega_o) = a(x) \cdot s(x) \quad (1)$$

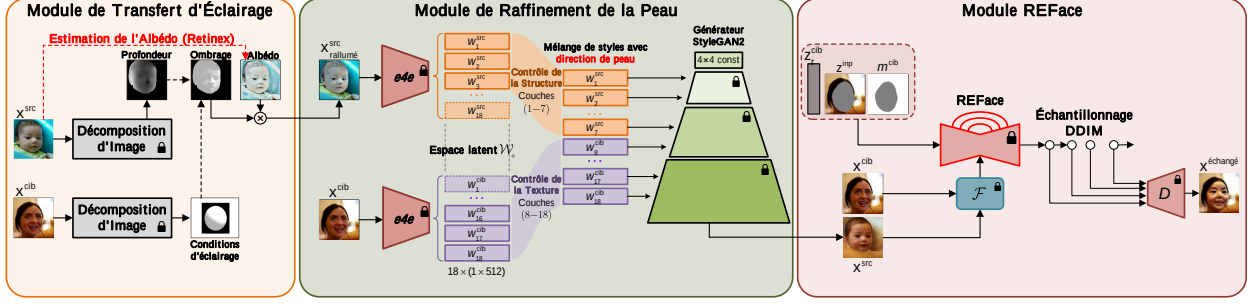


FIGURE 1 – Vue d’ensemble du pipeline de transfert de visage en trois étapes : le transfert d’éclairage, le raffinement de la peau dans l’espace latent $\mathcal{W}_+$ de StyleGAN2, et la synthèse finale en s’appuyant sur l’architecture REFace [4].

où $a(x)$ et $s(x)$ désignent respectivement l’albédo et l’ombrage du pixel. Si le cadre d’apprentissage profond de [8] permet d’estimer la géométrie et les ombres pour un ré-éclairage physiquement plausible, il souffre néanmoins de « fuites d’éclairage » dans la carte d’albédo.

2.3 Manipulation d’images basée sur l’inversion de GAN

L’inversion de GAN permet de projeter des images dans l’espace latent d’un modèle génératif pré-entraîné, facilitant ainsi l’édition sémantiques par le biais de déplacements directionnels [9]. Des architectures d’encodeurs tels que *e4e* [10], couplées à des techniques de mélange de styles [11], offrent un contrôle granulaire permettant de dissocier les attributs structuraux des détails fins. Au-delà des méthodes basées sur les encodeurs, plusieurs chercheurs ont exploré l’édition directe de l’espace latent. InterfaceGAN [12], bien qu’utilisant des SVM linéaires pour identifier les frontières d’attributs binaires (ex. : le genre), s’avère moins efficace pour les attributs continus et non étiquetés tels que la teinte de la peau.

3 Méthodologie

Le pipeline proposé, nommé PCFace (*Photometric Consistency Face-swapping*), décompose le processus de transfert de visage en trois étapes distinctes afin de garantir une cohérence photométrique optimale : le ré-éclairage de visage, l’harmonisation des teintes de peau dans l’espace latent, puis la synthèse haute fidélité (voir Fig. 1). Le *Module de Transfert d’Éclairage* atténue d’abord les incohérences d’éclairage entre la source I_{src} et la cible I_{cib} . Ensuite, le *Module de Raffinement de la Peau* manipule les caractéristiques de texture dans l’espace désentrelacé $\mathcal{W}_+$ pour un transfert d’identité perceptuel, tout en garantissant une diffusion naturelle de la teinte de la peau cible. Enfin, le transfert de visage s’appuie sur le *backbone* REFace [4] dont les poids sont maintenus fixes pour préserver les capacités en généralisation du modèle pré-entraîné.

3.1 Module de transfert d’éclairage

Inspiré par [8], ce module estime l’albédo et la carte de profondeur de l’image source, puis extrait l’éclairage cible

pour calculer l’ombrage selon la géométrie source. Afin d’affiner la décomposition intrinsèque et d’éviter les fuites d’éclairage, nous nous appuyons sur la théorie Retinex [15] pour estimer l’albédo. Nos expériences montrent que ce procédé réduit significativement les erreurs de ré-éclairage et améliore la cohérence photométrique entre le visage généré et la cible.

3.2 Module de raffinement de la peau

Ce module vise à transférer le style colorimétrique restreint à la peau de l’image cible $x_{I_{extcib}}$ vers la source I_{src} tout en préservant l’identité de cette dernière. Les deux images sont projetées dans l’espace $\mathcal{W}_+$ via un encodeur *e4e* pré-entraîné (bloc central, Fig. 1). Nous effectuons ensuite un mélange de styles [11] en combinant les couches d’identité structurelle (L1–L7) de la source avec les couches de texture (L8–L18) de la cible. Pour affiner ce transfert malgré l’absence d’étiquettes binaires pour la peau, nous utilisons un score d’approximation : des coefficients de texture sont extraits de 20 000 paires image-latent via un modèle morphable 3D (3DMM) [16]. Un partitionnement par K-moyennes sur ces coefficients permet de définir une « direction de peau » $\vec{n}$ reliant les centroïdes des clusters. En projetant l’écart entre les codes latents cible et source $z_{src} - z_{cib}$ sur $\vec{n}$, nous obtenons le code latent édité le plus proche de la source. Nos expériences confirment que cette manipulation selon la direction de peau surpasse l’encodage standard en termes de fidélité colorimétrique.

3.3 Module REFace sans optimisation

Le pipeline s’achève par l’utilisation du modèle de diffusion pré-entraîné REFace [4]. Les images sources modifiées sont directement injectées dans le processus de diffusion inverse, sans adaptation des paramètres du réseau principal ni de l’adaptateur $\mathcal{F}$. Cette stratégie évite la complexité et les risques d’instabilité liés au fine-tuning, tout en préservant les informations d’éclairage et de la teinte de la peau. Malgré le décalage de domaine (*domain shift*) induit, les résultats expérimentaux ne montrent aucun impact négatif sur les performances. Cette stratégie conserve les fortes propriétés génératrices et la stabilité du modèle de base, tout en alignant les données source et cible dans une

TABLEAU 1 – Comparaison quantitative sur le jeu de données CelebAMask-HQ.

Méthode	Similarité de Caractéristiques d'Identité ↑	Récupération d'Identité ↑		FID ↓	Pose ↓	Expression ↓	SSIM ↑
		Top-1 (%)	Top-5 (%)				
E4s [13]	0,47	80,5	89,0	14,503	4,09	1,251	0,635
FaceDancer [14]	0,28	11,5	22,6	7,920	3,892	0,95	0,835
DiffFace [7]	0,59	93,0	97,2	8,890	3,631	0,98	0,660
REFace [4]	0,632	96,3	98,6	7,435	3,184	0,95	0,743
PCFace	0,498	81,6	90,7	7,160	3,089	0,94	0,767

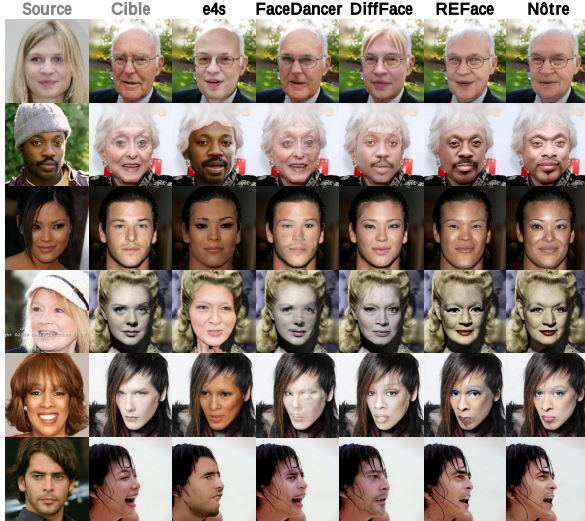


FIGURE 2 – Comparaison qualitative entre PCFace et les méthodes existantes.

représentation plus compatible, garantissant un face swap précis et robuste.

4 Expérimentations

4.1 Contexte expérimental

Bases de Données En suivant l’approche de référence [4], nous évaluons notre méthode sur le jeu de données CelebAMask-HQ [17]. Nous utilisons pour cela la partition de test fixe de 1 000 images haute résolution (1024×1024), telle que définie dans [4], afin d’évaluer la capacité de généralisation de notre pipeline sans aucun fine-tuning du modèle.

Métriques Pour évaluer les performances du transfert de visage sous plusieurs angles, nous employons diverses métriques. Pour la préservation de l’identité, le score de similarité de caractéristiques ID est calculé en comparant les vecteurs d’identité de la face source et du résultat généré via ArcFace [18]. De plus, les scores de récupération d’identité Top-1 (précision de rang 1) et Top-5 (précision de rang 5) sont utilisés. Pour évaluer le réalisme et la qualité des visages synthétisés, nous utilisons la distance de Fréchet (FID) [19]. Enfin, nous mesurons les erreurs de

pose et d’expression [20], ainsi que l’indice de similarité structurelle SSIM, afin d’évaluer la préservation des attributs cibles et la fidélité globale des résultats.

Détails d’Implémentation Notre méthode est implémentée avec le framework PyTorch. L’ensemble des expérimentations est réalisé sur des GPU Quadro RTX 6000 (24 Go) hébergés sur les serveurs du CRIANN¹. La taille des images d’entrée est fixée à 512×512 pixels. Les résultats obtenus avec PCFace sont comparés avec ceux obtenus avec E4s [13], FaceDancer [14], DiffFace [7] et l’approche initiale REFace [4].

4.2 Résultats

Nos résultats d’évaluation sont présentés dans le Tableau 1. Afin d’en faciliter la lecture, pour chaque métrique, la meilleure et la deuxième meilleure valeur sont respectivement surlignées en **bleu** et en **vert**. Notre approche modulaire obtient les meilleures performances sur plusieurs métriques : un FID de **7,160**, une erreur de pose de **3,089** et une erreur d’expression de **0,94**. Ces résultats indiquent que PCFace améliore significativement la qualité perceptuelle et la cohérence visuelle des images finales.

Les résultats qualitatifs (Fig. 2) confirment en outre les analyses quantitatives. Si REFace conserve la meilleure préservation de l’identité, il peine à harmoniser l’identité source avec l’environnement cible. Les résultats générés par E4s et DiffFace présentent des artefacts de bordure marqués sur des régions faciales, tandis que FaceDancer échoue à maintenir l’identité malgré un SSIM élevé. En revanche, en acceptant une légère réduction de la similarité d’identité au profit d’un réalisme accru, notre système génère des images plus naturelles et cohérentes grâce à une gestion supérieure de l’éclairage et de la teinte de la peau.

Conclusion

Nous avons introduit un nouveau pipeline de transfert de visage en trois étapes qui améliore le réalisme visuel sans nécessiter de réentraînement du modèle. En combinant le prétraitement par transfert d’éclairage et raffinement de la teinte de la peau avec une architecture de diffusion REFace, nous atteignons un FID de 7,16. Les résultats expérimentaux confirment que notre approche modulaire parvient à relever efficacement les défis liés à l’illumination, offrant

1. **CRIANN** Le Centre Régional Informatique et d’Applications Numériques de Normandie

ainsi une solution robuste pour un face swap de haute fidélité.

Références

- [1] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, et Fang Wen. Faceshifter : Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv :1912.13457*, 2019.
- [2] Renwang Chen, Xuanhong Chen, Bingbing Ni, et Yanhao Ge. Simswap : An efficient framework for high fidelity face swapping. Dans *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020.
- [3] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, et Jiwen Lu. Diffswap : High-fidelity and controllable face swapping via 3d-aware masked diffusion. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023.
- [4] Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, et Muhammad Haris Khan. Realistic and efficient face swapping : A unified approach with diffusion models. Dans *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1062–1071. IEEE, 2025.
- [5] Yuval Nirkin, Yosi Keller, et Tal Hassner. Fsgan : Subject agnostic face swapping and reenactment. Dans *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019.
- [6] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, et Zhenan Sun. One shot face swapping on megapixels. Dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4834–4844, 2021.
- [7] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, et KwangHee Lee. Diffface : Diffusion-based face swapping with facial guidance. *Pattern Recognition*, 163 :111451, 2025.
- [8] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, et Xiaoming Liu. Face relighting with geometrically consistent shadows. Dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4217–4226, 2022.
- [9] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, et Sylvain Paris. Ganspace : Discovering interpretable gan controls. *Advances in neural information processing systems*, 33 :9841–9850, 2020.
- [10] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, et Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4) :1–14, 2021.
- [11] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, et Daniel Cohen-Or. Encoding in style : a stylegan encoder for image-to-image translation. Dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.
- [12] Yujun Shen, Jinjin Gu, Xiaoou Tang, et Bolei Zhou. Interpreting the latent space of gans for semantic face editing. Dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.
- [13] Maomao Li, Ge Yuan, Cairong Wang, Zhian Liu, Yong Zhang, Yongwei Nie, Jue Wang, et Dong Xu. E4s : Fine-grained face swapping via editing with regional gan inversion. *arXiv preprint arXiv :2310.15081*, 2023.
- [14] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, et Cristofer Englund. Facedancer : Pose-and occlusion-aware high fidelity face swapping. Dans *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023.
- [15] Edwin H Land et John J McCann. Lightness and retinex theory. *Journal of the Optical society of America*, 61(1) :1–11, 1971.
- [16] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, et Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning : From single image to image set. Dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.
- [17] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, et Ping Luo. Maskgan : Towards diverse and interactive facial image manipulation. Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [18] Jiankang Deng, Jia Guo, Niannan Xue, et Stefanos Zafeiriou. Arcface : Additive angular margin loss for deep face recognition. Dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
- [19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, et Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [20] Nataniel Ruiz, Eunji Chong, et James M Rehg. Fine-grained head pose estimation without keypoints. Dans *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018.

CAFe-GS : Densification orientée compacité par guidage fréquentiel pour le 3D Gaussian Splatting

L.P. Huar^{*†} G. Sandri[†] N. Sabater[†] C. Guillemot^{*} P. Hellier^{*}

^{*}Centre INRIA de l'Université de Rennes, Rennes [†]InterDigital, Video Lab, Rennes

prénom.nom@inria.fr prénom.nom@interdigital.com

Résumé

Le 3D Gaussian Splatting (3DGS) représente une scène par un modèle de gaussiennes 3D et permet la synthèse de vues nouvelles en temps réel. Un élément clé du pipeline d'optimisation est le contrôle adaptatif de densité (Adaptive Density Control, ADC), qui sélectionne les gaussiennes à densifier (c.-à-d. dupliquer) afin d'augmenter la capacité du modèle. Dans 3DGS, cette densification tend à sélectionner des zones déjà bien reconstruites, entraînant des duplications redondantes et un contrôle limité du compromis qualité-compacité (fidélité vs. nombre de primitives). Nous proposons CAFe-GS, une stratégie de densification orientée compacité, qui utilise (i) un score d'erreur par gaussienne, obtenu en redistribuant l'erreur de rendu vers les gaussiennes, et (ii) un guidage fréquentiel favorisant la densification dans les régions riches en détails tout en pénalisant les zones lisses. Un paramètre de seuil unique permet un contrôle interprétable sur le compromis. CAFe-GS atteint une qualité comparable en PSNR à 3DGS avec 2–4× moins de gaussiennes, et jusqu'à 12–15× moins avec une baisse de qualité contrôlée.

Mots clefs

3D Gaussian Splatting · Compacité · Densification

1 Introduction

Le 3D Gaussian Splatting (3DGS) [1] s'est imposé comme une méthode de référence pour la synthèse de vues nouvelles en temps réel à partir d'images multi-vues. Il représente une scène par un ensemble de gaussiennes 3D anisotropes optimisées de bout en bout via un rendu différentiable, et produit des reconstructions photoréalistes en temps réel. Grâce à son excellent compromis qualité-performance, 3DGS a été largement adopté pour la reconstruction et le rendu interactif de scènes complexes.

Un mécanisme central du pipeline d'optimisation de 3DGS est l'Adaptive Density Control (ADC), qui densifie (c.-à-d. dupliquer) périodiquement certaines gaussiennes sélectionnées pour compenser une initialisation parcimonieuse et augmenter la capacité du modèle. Dans la formulation originale, le critère de sélection repose sur la norme du gradient positionnel moyen (typiquement évaluée toutes les $T=100$ itérations). En pratique, ce déclencheur pro-

voque fréquemment des duplications redondantes dans des régions déjà bien reconstruites (p. ex. autour de contours ou d'occlusions), ce qui dégrade la compacité sans gain de qualité et augmente inutilement les coûts de mémoire, stockage et rendu. De plus, le seuil sur le gradient est difficilement interprétable, ce qui complique le contrôle du compromis entre taille du modèle et qualité de reconstruction.

Plusieurs travaux visent à rendre la densification plus efficace tout en conservant la fidélité. 3DGS-EDC [2] propose une sélection *edge-aware* et un *Long-Axis Split* pour limiter la redondance, avec des mécanismes de stabilisation du processus. Mini-Splatting [3] s'appuie sur une réorganisation géométrique des primitives (p. ex. *blur splitting* et réinitialisation en profondeur) suivie d'une simplification pour réduire le nombre final de gaussiennes, tandis que Taming 3DGS [4] formalise un entraînement sous budget via une densification guidée par score et un calendrier constructif. Côté apparence, Color-Cued [5] exploite les gradients du coefficient SH d'ordre zéro pour mieux cibler les variations de couleur. BoGausS [6] combine des mises à jour *precision-aware*, un *Sparse Adam* non biaisé et des stratégies *split/prune* inspirées du compromis débit-distorsion. Reduced-3DGS [7] vise surtout la compaction mémoire (pruning, adaptation du nombre de bandes SH, quantification). Enfin, Bulò *et al.* [8] remplacent le déclencheur par gradient par un critère *pixel-error driven* en redistribuant une erreur de rendu vers les primitives via leurs poids de compositing, et proposent en plus un contrôle de croissance et des correctifs liés au clonage.

Nous proposons CAFe-GS, une stratégie de densification orientée compacité qui redéfinit le mécanisme d'ADC afin de réduire le nombre de primitives tout en préservant la qualité de reconstruction. Au lieu d'un critère basé sur le gradient positionnel, CAFe-GS s'appuie sur un score par gaussienne issu de l'erreur de reconstruction, modulé par des indices fréquentiels locaux. Cette formulation concentre la densification sur les détails mal reconstruits et limite les duplications redondantes dans les régions déjà bien modélisées. Un unique paramètre de seuil τ permet de contrôler explicitement le compromis qualité-compacité : une valeur faible entraîne une densification accrue, tandis qu'une valeur élevée conduit à un modèle plus compact.

2 Méthode

2.1 Vue d'ensemble

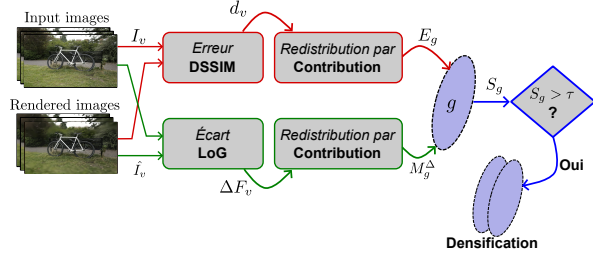


FIGURE 1 – **Vue d'ensemble de CAFe-GS.** La distortion entre I_v et $\hat{I}_v$ est redistribuée vers les gaussiennes via leur contribution $C(g, u)$. En parallèle, l'écart fréquentiel entre les images d'entrée et de rendu est également reprojété vers les gaussiennes. Les deux signaux sont combinés en un score S_g . Les gaussiennes avec $S_g \geq \tau$ sont densifiées. **Couleurs :** rouge : distortion, vert : fréquence, bleu : décision.

Comme illustré en Fig. 1, CAFe-GS vise à allouer la capacité du modèle là où elle améliore réellement la qualité perceptuelle du rendu. D'une part, nous remplaçons le critère basé sur le gradient positionnel par un signal directement aligné avec l'objectif d'optimisation : l'erreur de reconstruction. En redistribuant la distortion des images de rendu vers les gaussiennes selon leur contribution, nous estimons pour chaque primitive un score reflétant son impact perceptuel sur l'erreur finale.

D'autre part, nous introduisons un signal fréquentiel fondé sur l'observation que les détails fins (contours, structures minces, variations rapides) requièrent davantage de capacité que les régions lisses. En comparant les réponses fréquentielles locales des images d'entrée et rendues, nous identifions les zones où le rendu manque de détails.

Ces deux informations sont combinées en un score de densification S_g associé à chaque gaussienne. Lors des étapes d'ADC, seules les primitives telles que $S_g \geq \tau$ sont dupliquées. La densification est ainsi guidée conjointement par un critère perceptuel et par la complexité locale, concentrant la capacité sur les régions mal reconstruites tout en limitant les duplications redondantes.

Contribution effective. Nous notons $C(g, u)$ la contribution effective de la gaussienne g au pixel u sous *front-to-back* alpha compositing :

$$C(g, u) = \alpha_g(u) \prod_{h <_g} (1 - \alpha_h(u)), \quad (1)$$

où $\alpha_g(u)$ est l'opacité projetée de g et $<_g$ l'ordre en profondeur. Ce terme correspond au poids de compositing de g , intégrant sa transmittance accumulée le long du rayon.

2.2 Métrique d'erreur par gaussienne

Nous remplaçons le déclencheur basé sur le gradient positionnel par un critère directement aligné avec la perte d'entraînement. L'objectif est d'estimer, pour chaque gaussienne, sa contribution effective à l'erreur perceptuelle de reconstruction. Notre formulation s'inspire fortement de la redistribution d'erreur de Bulò *et al.* [8], ici adaptée avec une mesure de distortion DSSIM.

Pour une vue d'entraînement v , nous calculons une carte de distortion pixel à pixel entre l'image d'entrée I_v et le rendu $\hat{I}_v$:

$$d_v(u) = \text{DSSIM}(I_v, \hat{I}_v)(u), \quad (2)$$

où DSSIM correspond à $1 - \text{SSIM}$. Cette distortion est ensuite redistribuée aux gaussiennes selon leur contribution. En utilisant le poids $C(g, u)$ (Eq. 1), nous définissons l'erreur attribuée à g pour la vue v :

$$e_{g,v} = \sum_{u \in \Omega_v} C(g, u) d_v(u). \quad (3)$$

Afin d'éviter qu'une erreur importante, visible seulement sur un sous-ensemble de vues, ne soit diluée par une moyenne sur l'ensemble des vues d'entraînement, nous agrégeons les contributions par un maximum :

$$E_g = \max_v e_{g,v}. \quad (4)$$

Le score E_g quantifie ainsi l'impact perceptuel de chaque gaussienne sur l'erreur de reconstruction. Les primitives présentant un E_g élevé sont considérées comme responsables de défauts significatifs et sont candidates à la densification.

2.3 Guidance fréquentielle

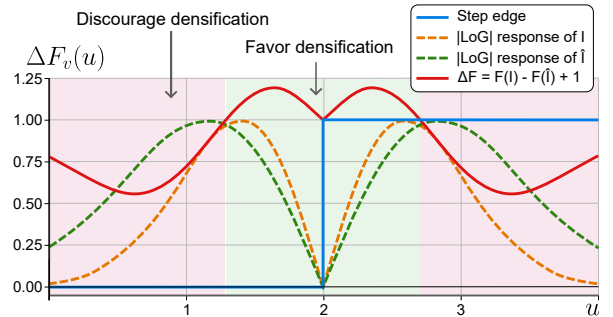


FIGURE 2 – **Guidance par écart fréquentiel sur un bord en escalier 1D.** Bleu : signal d'entrée. Orange/vert : $|\text{LoG}|$ de l'entrée I et du rendu $\hat{I}$. Rouge : écart fréquentiel. Les zones où $\Delta F > 1$ favorisent la densification (détails manquants), tandis que $\Delta F \leq 1$ indiquent des régions déjà suffisamment reconstruites.

En complément du terme d'erreur, nous introduisons un signal mesurant le déficit local de détails dans le rendu.

L'objectif est d'identifier les régions où la représentation manque de hautes fréquences, indépendamment du niveau absolu de distorsion. Pour chaque vue v , nous calculons une carte de magnitude fréquentielle à partir de la luminance de l'image d'entrée en appliquant un filtre *Laplacian-of-Gaussian* (LoG, $\sigma_{LoG}=1$) :

$$F(I_v)(u) = |(LoG_{\sigma_{LoG}} * I_v)|(u). \quad (5)$$

La même opération est appliquée au rendu $\hat{I}_v$. Les cartes $F(I_v)$ et $F(\hat{I}_v)$ sont ensuite normalisées indépendamment par vue dans l'intervalle $[0, 1]$. Nous définissons alors un écart fréquentiel :

$$\Delta F_v(u) = F(I_v)(u) - F(\hat{I}_v)(u) + 1, \quad (6)$$

tel que $\Delta F_v(u) = 1$ indique une correspondance fréquentielle, $\Delta F_v(u) > 1$ révèle un manque de détails dans le rendu, et $\Delta F_v(u) < 1$ correspond à des régions déjà suffisamment détaillées (voir Fig. 2).

Comme pour l'erreur, l'écart fréquentiel est redistribué via $C(g, u)$ et agrégé par un maximum :

$$M_g^\Delta = \max_v \sum_{u \in \Omega_v} C(g, u) \Delta F_v(u), \quad (7)$$

afin de ne pas atténuer un déficit marqué sur un sous-ensemble de vues.

2.4 Critère de densification

Le score final de densification est défini par

$$S_g = E_g \cdot M_g^\Delta, \quad (8)$$

qui combine contribution à l'erreur perceptuelle et déficit fréquentiel local. À chaque étape d'ADC, les gaussiennes telles que $S_g \geq \tau$ sont sélectionnées pour duplication. Le seuil τ contrôle directement le compromis qualité-compacité : un τ élevé rend la sélection plus stricte (modèle plus compact), tandis qu'un τ faible augmente la capacité et la qualité du modèle.

Soient $s_g = (s_{g,x}, s_{g,y}, s_{g,z})$ les paramètres de taille de la gaussienne g . En utilisant le même seuil s_{th} que 3DGS, l'opération appliquée est :

- **Clone** si $s_{g,i} \leq s_{th}$ pour tout i : copie des attributs et léger décalage le long de l'axe dominant.
- **Split** sinon : copie des attributs, réduction des échelles ($s_{g,i} \leftarrow \frac{1}{2}s_{g,i}$) et léger décalage.

3 Expériences

3.1 Protocole expérimental

Nous suivons le protocole standard de 3DGS sur 13 scènes issues de Mip-NeRF 360 [9], Tanks and Temples [10] et Deep Blending [11], en réservant une image sur huit pour le test. La qualité de reconstruction est évaluée par PSNR, SSIM et LPIPS, et la compacité par le nombre total de gaussiennes (abscisse sur les courbes débit-distorsion).

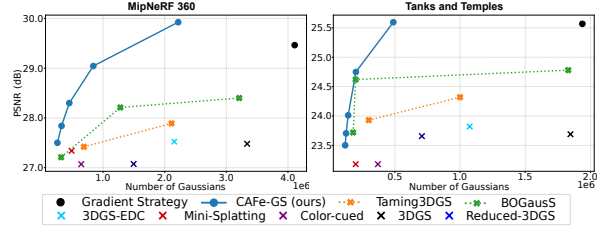


FIGURE 3 – *Compromis débit-distorsion sur Mip-NeRF 360 [9] et Tanks and Temples [10]. PSNR en fonction du nombre de gaussiennes pour CAFE-GS et la stratégie de référence. CAFE-GS atteint une fidélité comparable ou supérieure avec significativement moins de gaussiennes sur l'ensemble des jeux de données.*

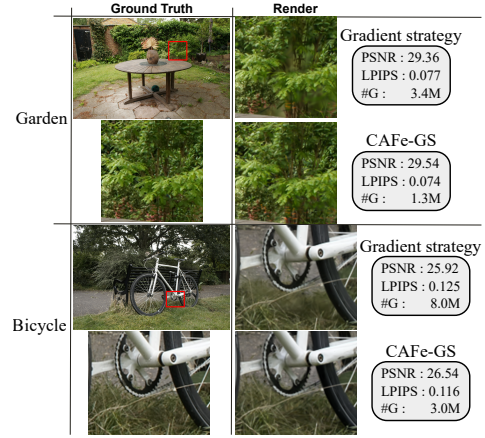


FIGURE 4 – *Comparaison qualitative sur garden et bicycle [9]. Image de référence, rendu avec la densification par gradient, puis rendu CAFE-GS au réglage de meilleure qualité ($\tau = 10$). Les agrandissements montrent une meilleure restitution des détails fins avec significativement moins de gaussiennes.*

Les expériences sont réalisées sous PyTorch avec `gsplat` v1.5.3 [12] sur un GPU NVIDIA L40S, en conservant l'ensemble des hyperparamètres de 3DGS (optimiseur, calendrier, rendu différentiable, perte $L_1 + DSSIM$). Les poses de caméra et le modèle 3DGS sont initialisés via COLMAP [13] comme fournis dans les datasets. Après un warm-up de 500 itérations, la densification est appliquée toutes les 100 itérations jusqu'à 15k itérations, l'entraînement se poursuivant jusqu'à 30k. Nous évaluons cinq valeurs du seuil $\tau \in \{10, 25, 50, 75, 100\}$.

La comparaison principale est effectuée avec la stratégie de densification par gradient de 3DGS [1] dans le même cadre `gsplat`, avec des hyperparamètres identiques. À titre indicatif, nous reportons également des résultats publiés pour *Taming 3DGS* [4], *BoGausS* [6], *3DGS-EDC* [2], *Mini-Splatting* [3], *Color-Cued* [5] et *Reduced-3DGS* [7].

3.2 Résultats

Comme illustré par les courbes dans la Fig. 3, CAFe-GS induit une progression régulière du compromis qualité-compacité. Lorsque τ augmente, le nombre de gaussiennes diminue (modèle plus compact) au prix d'une baisse de PSNR. Inversement, un τ plus faible augmente la capacité et la fidélité. Ainsi, $\tau = 100$ correspond au régime le plus compact, tandis que $\tau = 10$ maximise la qualité. À PSNR comparable (Fig. 3), CAFe-GS réduit le nombre de gaussiennes d'environ 2–4× par rapport à la densification par gradient sur l'ensemble des jeux de données. Dans un régime fortement orienté compacité, la réduction atteint jusqu'à 12–15×, pour une baisse contrôlée d'environ 2 à 2.5 dB.

Les comparaisons qualitatives de la Fig. 4 confirment ces tendances : la densification guidée par la fréquence restitue plus fidèlement les hautes fréquences et réduit le flou résiduel, notamment dans les détails fins comme les buissons de la scène *garden* [9], et l'herbe dans *bicycle* [9].

4 Conclusion

Nous avons introduit une stratégie de densification orientée compacité pour le *3D Gaussian Splatting* [1]. CAFe-GS remplace le critère basé sur le gradient positionnel par un score d'erreur par gaussienne, modulé par une guidance fréquentielle locale, afin d'allouer la capacité là où elle améliore réellement la qualité perceptuelle du rendu. Un unique seuil τ permet de contrôler explicitement le compromis qualité-compacité.

Expérimentalement, CAFe-GS atteint une qualité comparable (PSNR) avec environ 2 à 4× moins de gaussiennes, et jusqu'à 12 à 15× moins dans les régimes plus fortement contraints.

Annexe

Ces travaux font également l'objet d'une soumission à *International Conference on Pattern Recognition* 2026. Des détails méthodologiques supplémentaires, ainsi que des résultats complémentaires, notamment des études d'ablation et des analyses approfondies, sont disponibles dans la version complète soumise à cette conférence.

Références

- [1] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, et George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.
- [2] Xiaobin Deng, Changyu Diao, Min Li, Ruohan Yu, et Duanqing Xu. Efficient density control for 3d gaussian splatting. *arXiv preprint arXiv :2411.10133*, 2024.
- [3] Guangchi Fang et Bing Wang. Mini-splatting : Representing scenes with a constrained number of gaussians. Dans Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, et Gül Varol, éditeurs, *Computer Vision – ECCV 2024*, pages 165–181, Cham, 2024. Springer Nature Switzerland.
- [4] Saswat Subhajyoti Mallick, Rahul Goel, Bernhard Kerbl, Markus Steinberger, Francisco Vicente Carrasco, et Fernando De La Torre. Taming 3dgs : High-quality radiance fields with limited resources. Dans *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery.
- [5] Sieun Kim, Kyungjin Lee, et Youngki Lee. Color-coded efficient densification method for 3d gaussian splatting. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 775–783, June 2024.
- [6] Stéphane Pateux, Matthieu Gendrin, Luce Morin, Théo Ladune, et Xiaoran Jiang. Bogauss : Better optimized gaussian splatting. Dans *2025 33rd European Signal Processing Conference (EUSIPCO)*, pages 765–769, 2025.
- [7] Panagiotis Papantonakis, Georgios Kopanas, Bernhard Kerbl, Alexandre Lanvin, et George Drettakis. Reducing the memory footprint of 3d gaussian splatting. *Proc. ACM Comput. Graph. Interact. Tech.*, 7(1), Mai 2024.
- [8] Samuel Rota Bulò, Lorenzo Porzi, et Peter Kontschieder. Revising densification in gaussian splatting. Dans Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, et Gül Varol, éditeurs, *Computer Vision – ECCV 2024*, pages 347–362, Cham, 2025. Springer Nature Switzerland.
- [9] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, et Pratul P. Srinivasan. Mip-nerf : A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021.
- [10] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, et Vladlen Koltun. Tanks and temples : Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.
- [11] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, et Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Trans. Graph.*, 37(6), Décembre 2018.
- [12] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, et Angjoo Kanazawa. gsplat : An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34) :1–17, 2025.
- [13] Johannes Lutz Schönberger et Jan-Michael Frahm. Structure-from-motion revisited. Dans *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Détection de la double compression des vidéos HEVC par analyse du domaine compressé et attention neuronale

Syamantak Sarkar^{†‡,*}, Christophe Charrier[†], Emmanuel Giguet[†], Kiran Raja[‡]

[†] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

[‡] Norwegian University of Science and Technology (NTNU), Norvège

5 mars 2026

Résumé

La détection de la double compression constitue un enjeu majeur en analyse forensique vidéo car les opérations de falsification, de transcodage ou de manipulation malveillante de contenus s'accompagnent fréquemment de ré-encodages successifs laissant des traces exploitables pour leur identification. Bien que largement étudiée pour les images et les codecs antérieurs, l'identification de la re-compression HEVC reste difficile en raison de la partition flexible des blocs, des structures de prédiction hiérarchiques et de l'utilisation des B-frames. Les méthodes existantes reposent soit sur des descripteurs du domaine compressé, soit sur des modèles neuronaux "boîtes noires", ce qui limite l'interprétabilité et la valeur probatoire.

Dans cet article, nous proposons un cadre forensique hybride guidé par l'attention neuronale pour la détection de la double compression des vidéos HEVC. L'idée principale est d'exploiter l'instabilité temporelle de l'attention comme indicateur des décisions de ré-encodage et de la corréler aux anomalies de prédiction liées au désalignement du GOP. Des caractéristiques du flux HEVC sont combinées avec la dynamique des cartes d'attention extraites image par image. L'évolution temporelle de l'attention est quantifiée à l'aide de descripteurs *CAM-change* basés sur le *rank-biased overlap* (RBO). La représentation hybride obtenue est modélisée par un Bi-LSTM afin de capturer les motifs périodiques associés à la double compression. Les résultats expérimentaux montrent une amélioration significative des performances de détection par rapport aux méthodes existantes.

1 Introduction

Avec la généralisation des contenus vidéos haute résolution et le développement des usages mobiles, la conception de techniques de compression toujours plus optimisée est devenue un enjeu majeur. Les techniques mises en œuvre se sont sophistiquées, rendant de fait plus difficiles les opérations d'investigation numérique lorsque la question de l'authenticité du contenu se pose [1].

TABLE 1 – Comparaison des méthodes de détection de double compression.

Méthode	Domaine	HEVC	B	Guidage
Wang et al.	Compressé	✗	✗	✗
Stamm et al.	Compressé	✗	✗	✗
Vázquez-Padín	Compressé	✓	Partiel	✗
Yu et al.	Compressé	✓	Limité	✗
He et al.	CNN	✓	✓	Implicite
Furushita et al.	Compressé	✓	✓	✗
Proposé	Hybride	✓	✓	Explicite

La détection fiable et probante de contenus altérés est aujourd'hui essentielle avec la démocratisation des outils d'édition vidéo qui entraîne la génération et la diffusion massives de contenus malveillants. Parmi les techniques d'identification classique, la détection de la *double compression* est un indice de manipulation possible. Les vidéos issues directement de la caméra étant généralement encodées une seule fois, la présence d'une double compression révèle un post-traitement : la vidéo a été décodée, éventuellement modifiée, puis ré-encodée [1, 2].

La nouvelle norme de compression HEVC (H.265) permet une très forte réduction du débit tout en conservant une très haute qualité visuelle, comparé à son prédécesseur H.264. Cette amélioration résulte d'une partition adaptative des blocs, de modes de prédiction spatiale et temporelle améliorés, ainsi que d'une hiérarchie de B-frames généralisée [3]. Ce format complexe pose des défis pour les techniques d'investigation. Mais en cas de réencodage, le processus entraîne des incohérences de prédiction et l'apparition de *Relocated I-frames* (RI-frames) [4] qui sont autant de traces exploitables.

La table 1 résume les principales approches de l'état de l'art pour la détection de la double compression des vidéos. Pour HEVC, aucune ne relie explicitement les caractéristiques du flux à la dynamique de l'attention neuronale en présence de B-frames. Les méthodes existantes exploitent soit les descripteurs du flux compressé [1, 5], soit des modèles d'apprentissage profond appliqués aux images reconstruites [2, 6]. Les premières offrent une bonne robustesse mais peu d'interprétabilité spatiale, tandis que les secondes ignorent les traces syntaxiques du codec et se comportent souvent comme des boîtes noires, ce qui réduit leur valeur probatoire.

*Ce travail est actuellement soumis à IWBF 2026. La notification d'acceptation est prévue en mars 2026.

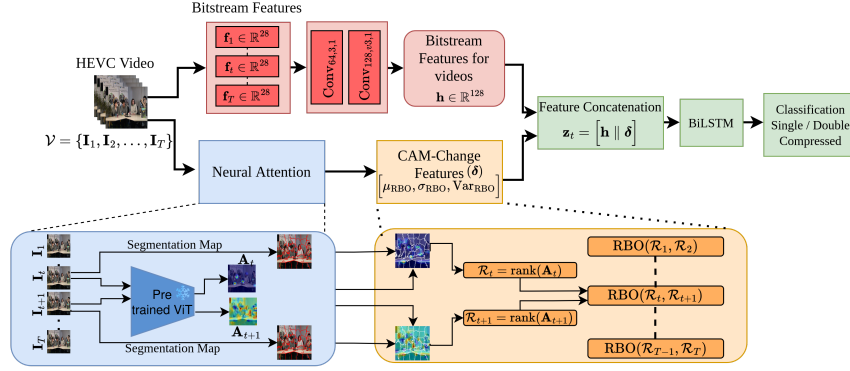


FIGURE 1 – Méthode de détection de la double compression HEVC : après extraction des caractéristiques du domaine compressé et calcul des cartes d’attention, la dynamique de l’attention est quantifiée par des descripteurs CAM-change, combinée avec les caractéristiques du flux, puis classée via un Bi-LSTM.

Des travaux récents montrent que les artefacts de recompression induisent une instabilité mesurable des cartes d’attention neuronale pour les images doublement compressées [7]. Nous faisons l’hypothèse que la double compression induit une *instabilité temporelle de l’attention neuronale*, liée au désalignement du GOP et aux RI-frames. Pour exploiter cette propriété, nous proposons un cadre forensique hybride combinant les caractéristiques HEVC et la dynamique des cartes d’attention. L’évolution temporelle de l’attention est quantifiée à l’aide du *rank-biased overlap* (RBO), produisant des descripteurs *CAM-change* fusionnés avec les caractéristiques du flux binaire et modélisés par un Bi-LSTM.

Contributions :

- une représentation hybride reliant caractéristiques HEVC et dynamique de l’attention neuronale ;
- des descripteurs *CAM-change* basés sur le RBO pour quantifier l’instabilité temporelle ;
- une modélisation temporelle par Bi-LSTM adaptée aux B-frames ;
- une amélioration des performances avec interprétabilité spatiale.

2 Méthode proposée

L’objectif est de détecter la double compression HEVC en modélisant conjointement les traces du flux binaire et la dynamique de l’attention neuronale. Le ré-encodage induit des irrégularités syntaxiques ainsi qu’une instabilité temporelle de l’attention. Une séquence vidéo décodée est notée

$$\mathcal{V} = \{I_1, \dots, I_T\}, \quad (1)$$

où I_t désigne la t -ième image. La méthode combine l’extraction de caractéristiques HEVC, l’analyse des cartes d’attention, le calcul des descripteurs *CAM-change* et une modélisation temporelle pour la classification.

2.1 Extraction des caractéristiques du flux binaire

Suivant [1], des descripteurs d’encodage sont extraits image par image à partir du flux HEVC à l’aide du décodeur HM 16.25 [8]. Chaque image est divisée en blocs 4×4 et représentée par un vecteur $\mathbf{f}_t \in \mathbb{R}^{28}$, comprenant le type d’image (I, P, B), les statistiques des unités de codage (CU), la distribution des modes (Intra, Skip, Merge), les catégories de vecteurs de mouvement ainsi que les directions de prédiction intra en luminance et chrominance. La séquence de caractéristiques est notée $\mathcal{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_T\}$.

Afin de capturer les dépendances temporelles locales, cette séquence est traitée par deux convolutions unidimensionnelles :

$$\mathbf{h}_t^{(1)} = \text{Conv}_{64}(\mathbf{f}_t) ; \mathbf{h}_t^{(2)} = \text{Conv}_{128}(\mathbf{h}_t^{(1)}). \quad (2)$$

produisant une représentation compacte $\mathbf{h} \in \mathbb{R}^{128}$ résumant les motifs d’encodage sensibles aux RI-frames et aux incohérences de prédiction.

2.2 Analyse de l’attention neuronale

Chaque image décodée I_t est analysée par un classifieur Vision Transformer (ViT) [9]. Une carte d’attention est obtenue par Grad-CAM [10] :

$$\mathbf{A}_t \in \mathbb{R}^{H \times W}. \quad (3)$$

Contrairement à une utilisation purement visuelle, ces cartes sont considérées comme des signaux dynamiques permettant d’évaluer la stabilité temporelle de l’attention.

2.3 Descripteurs CAM-change via RBO

Chaque carte $\mathbf{A}_t$ est segmentée en K régions spatiales fixes à l’aide de SLIC [11]. Les régions sont ensuite ordonnées selon leur intensité moyenne, ce qui définit le classement $\mathcal{R}_t = \text{rank}(\mathbf{A}_t)$.

La similarité entre deux instants consécutifs est mesurée par le *rank-biased overlap* (RBO) [12] :

$$\text{RBO}(\mathcal{R}_t, \mathcal{R}_{t+1}) = (1-p) \sum_{d=1}^K \frac{|\mathcal{R}_t^{(d)} \cap \mathcal{R}_{t+1}^{(d)}|}{d} p^{d-1}. \quad (4)$$

Trois descripteurs sont extraits pour quantifier l'instabilité temporelle de l'attention, plus marquée en cas de double compression : $\delta = [\mu_{\text{RBO}}, \sigma_{\text{RBO}}, \text{Var}_{\text{RBO}}]$.

2.4 Fusion des caractéristiques et classification

Les caractéristiques du flux binaire et les descripteurs *CAM-change* sont fusionnés par concaténation $\mathbf{z} = [\mathbf{h} \parallel \delta]$. Le vecteur résultant est modélisé par un réseau BiLSTM [13], $\hat{y} = \text{BiLSTM}(\mathbf{z})$, suivi d'une couche entièrement connectée produisant la décision au niveau de la vidéo. Le Bi-LSTM capture les motifs périodiques associés au désalignement du GOP et aux RI-frames.

3 Protocole expérimental

Cette section présente les données, les configurations d'encodage et le protocole d'évaluation utilisés pour valider la méthode proposée. Le protocole suit celui de [1] afin d'assurer la reproductibilité et une comparaison équitable.

3.1 Base de données

Les expériences sont réalisées sur 43 séquences vidéo YUV non compressées issues de la base publique Xiph.org [14], couvrant trois résolutions (720p, 1080p et 4K) et des contenus variés. Seules les 500 premières images de chaque séquence sont utilisées afin de conserver les caractéristiques temporelles tout en limitant la complexité de calcul.

Pour améliorer la robustesse, trois variantes sont générées : **Original (O)**, **CLAHE (C)** et **Blur (B)**. Ces variantes sont considérées comme une unité unique lors de la division apprentissage/test afin d'éviter toute fuite de contenu.

3.2 Configurations d'encodage

Les vidéos sont compressées en HEVC avec l'encodeur $\times 265$ en GOP fermé, sans pyramide de B-frames ni *open-GOP*.

Compression simple : débits {1000, 3000, 5000} kbps, tailles de GOP {9, 25, 70} et nombres de B-frames {3, 5, 7}, soit 81 vidéos par séquence source.

Double compression : un premier encodage en I/P uniquement (GOP {12, 30}), suivi d'un second encodage avec les mêmes paramètres que la compression simple. Un sous-ensemble équilibré de 81 vidéos doublement compressées est sélectionné.

3.3 Mesures d'évaluation

Les performances sont évaluées à l'aide du TPR, du TNR, de l'accuracy (ACC) et de l'aire sous la courbe ROC (AUC). Les descripteurs *CAM-change* sont également analysés via leur moyenne, écart-type et variance.

3.4 Protocole d'apprentissage

Une validation croisée en 10 plis est adoptée : 32 unités pour l'apprentissage, 4 pour la validation et 4 pour le test, en garantissant que les variantes d'une même séquence restent dans le même pli. Le modèle hybride est entraîné avec l'optimiseur Adam et un *early stopping*, en utilisant les mêmes hyperparamètres que [1].

4 Résultats et discussion

Cette section évalue l'efficacité de la méthode proposée pour la détection de la double compression HEVC et analyse l'apport de la dynamique de l'attention neuronale.

4.1 Performance de détection

La table 2 présente les performances en termes de TPR, TNR, accuracy (ACC) et écart-type (SD). La méthode proposée obtient la meilleure accuracy globale (86.43%).

TABLE 2 – Performance de détection de la double compression (%).

Méthode	TNR	TPR	ACC	SD
Vazquez <i>et al.</i>	77.71	45.12	61.42	1.81
He <i>et al.</i>	33.51	82.07	57.79	0.60
Furushita <i>et al.</i>	78.07	82.05	80.06	1.16
Méthode proposée	85.90	86.96	86.43	2.05

Par rapport aux approches basées uniquement sur le flux binaire ou sur l'apprentissage profond appliqué aux images reconstruites, la méthode hybride améliore significativement la détection tout en conservant un équilibre entre les classes. Un gain d'environ 6% est observé par rapport à [1]. Cette amélioration provient de la combinaison des caractéristiques HEVC et de la dynamique de l'attention, qui permet de capturer des artefacts spatiaux non visibles dans les seules statistiques syntaxiques. Les performances restent stables pour différentes résolutions, débits et structures de GOP, montrant la robustesse du modèle.

4.2 Visualisation de l'espace des caractéristiques

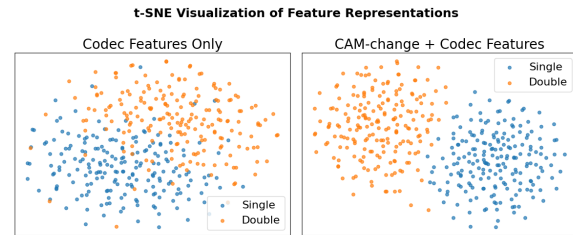


FIGURE 2 – Visualisation t-SNE des caractéristiques. La représentation hybride (droite) montre une meilleure séparation entre vidéos simplement et doublement compressées que le flux binaire seul (gauche).

La figure 2 présente une visualisation t-SNE des représentations. Les caractéristiques issues uniquement du flux bi-

TABLE 3 – Étude d'ablation (%).

Méthode	TNR	TPR	ACC	SD
CAM-change seul	71.84	73.26	72.55	2.41
Flux binaire seul	77.62	78.19	77.91	1.63
Fusion sans Bi-LSTM	82.05	83.11	82.58	1.97
Fusion + Bi-LSTM	85.90	86.96	86.43	2.05

naire montrent un chevauchement notable entre les classes. En revanche, la représentation hybride combinant flux HEVC et descripteurs *CAM-change* produit des clusters mieux séparés, confirmant le caractère complémentaire des deux types d'informations pour la détection de la double compression.

4.3 Étude d'ablation

La table 3 évalue la contribution des différents modules. Les caractéristiques du flux binaire ainsi les descripteurs *CAM-change* fournissent isolément une base solide mais insuffisante. La modélisation temporelle par Bi-LSTM permet de capturer les motifs périodiques liés au désalignement du GOP et conduit aux meilleures performances. Ces résultats confirment l'intérêt de la représentation hybride pour la détection robuste de la double compression HEVC.

5 Conclusion et perspectives

Cet article propose une méthode forensique hybride guidée par l'attention neuronale pour la détection de la double compression des vidéos HEVC comportant des B-frames. La méthode combine des caractéristiques du flux binaire avec la dynamique temporelle des cartes d'attention issues de l'analyse CAM. L'instabilité de l'attention est quantifiée par des descripteurs basés sur le RBO et modélisée temporellement par un Bi-LSTM, permettant d'identifier des anomalies à la fois dans le flux HEVC et dans les images décodées. Les résultats expérimentaux montrent une amélioration significative par rapport aux approches reposant uniquement sur les statistiques du codec ou sur l'apprentissage profond dans le domaine spatial, en particulier dans des scénarios comportant des B-frames et des désalignements de GOP.

Cette étude porte sur la double compression HEVC dans des conditions contrôlées. L'évaluation sur des bases de données forensiques réelles constitue une perspective importante. Les travaux futurs porteront sur la double compression inter-codec (e.g., H.264/AVC vers HEVC) ainsi que sur la détection en cas de paramètres d'encodage variables, afin d'étendre l'applicabilité à des scénarios forensiques plus complexes et réalistes.

Références

[1] Y. Furushita, T. Yasuda, and H. Kiya, "Detection of double compression in HEVC videos containing B-frames using encoding feature sequences," *Journal of Imaging*, vol. 11, no. 7, p. 211, 2025.

[2] P. He, H. Li, H. Wang, S. Wang, X. Jiang, and R. Zhang, "Frame-wise detection of double hev

compression by learning deep spatio-temporal representations in compression domain," *IEEE Transactions on Multimedia*, vol. 23, pp. 3179–3192, 2021.

- [3] V. Sze, M. Budagavi, and G. J. Sullivan, Eds., *High Efficiency Video Coding (HEVC) : Algorithms and Architectures*. Springer International Publishing, 2014. [Online]. Available : <https://link.springer.com/book/10.1007/978-3-319-06895-4>
- [4] A. C. Gallagher and T. Chen, "Detecting gop structure in mpeg video," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 566–578, 2012.
- [5] X. Yu, Y. Yang, R. Ni, and Y. Zhao, "Detection of double compression in HEVC videos using coding unit and transform unit statistics," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, 2019.
- [6] Y. Wu, W. AbdAlmageed, and P. Natarajan, "Mantranet : Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9543–9552.
- [7] S. Sarkar, S. N. George, and K. Raja, "Detecting double compression in images using difference of class activation maps," in *Proceedings of the IEEE International Workshop on Biometrics and Forensics (IWBF)*, 2025.
- [8] Fraunhofer Heinrich Hertz Institute, "Hm software (hm-16.25) – HEVC test model," <https://vcgit.hhi.fraunhofer.de/jvet/HM/-/tree/HM-16.25>.
- [9] A. Dosovitskiy, "An image is worth 16x16 words : Transformers for image recognition at scale," *arXiv preprint arXiv :2010.11929*, 2020.
- [10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam : Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [11] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [12] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Transactions on Information Systems (TOIS)*, vol. 28, no. 4, pp. 1–38, 2010.
- [13] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2. IEEE, 1997, pp. 1228–1231.
- [14] "Xiph.org video test media," <https://media.xiph.org/video/derf/>.

Egocentric Collision Perceptual Envelopes: Comparing Dynamic and Static Conditions in Immersive visualization.

Hussein AMMAR, Toinon VIGIER, Pierre LEBRANCHU, Patrick LECALLET
LS2N Laboratoire des Sciences du Numériques de Nantes, Equipe IPI (Image Perception Interaction)
hussein.ammар@etu.univ-nantes.fr, toinon.vigier@univ-nantes.fr, pierre.LEBRANCHU@chu-nantes.fr, patrick.lecallet@univ-nantes.fr.

Abstract

Perception of potential collisions with approaching individuals is crucial for understanding safe navigation in crowds. The collision perception in this study was examined in a monitor-based virtual pedestrian environment where participants needed to assess whether an oncoming pedestrian would hit them. There were two viewing conditions (dynamic and static) and pedestrians could walk at one of two speeds (1.0 m/s, 1.6 m/s). Collision thresholds were estimated and used to create collision envelopes that define the perceptual boundary between colliding and non-colliding trajectories. Results indicate that collision envelope shape depends on experimental condition and pedestrian speed.

Key Words

Collision perception, Virtual environment, Collision Envelope.

1 Introduction

The ability to judge whether approaching individuals are on a collision course is required for a safe navigation in crowded environments. This skill is critical especially for people with visual impairments, who often face daily challenges avoiding person-to-person collisions in places such as shopping malls and transportation stations [1]. Impaired collision detection not only increases the risk of accidents but also reduces independence and quality of life.

Several solutions have been proposed to reduce collision risk. For example, peripheral visual field expansion devices like prism glasses [2] [3], which can improve obstacle detection by enlarging the wearer's visual field. However, such interventions are evaluated in static conditions, without reproducing the complex motion cues and dynamic interactions of real walking environments. This imposes ecological limitations on their validity.

To solve this problem, and to measure mobility performance more realistically, physical obstacle courses have been used [4] [5]. These courses typically involve navigating around static obstacles and are often part of orientation and mobility (O&M) tests. Yet, in practice, the implementation of courses with moving obstacles is difficult: they require precise control of trajectories and timing, involve a risk of physical collision, and are

constrained by the available space and courses configuration.

To overcome these challenges, virtual environments (VEs) provide a safe, controlled and flexible platform for testing collision detection. They allow precise manipulation of pedestrian paths, speeds, and approach angles without physical risks. In this context, the concept of 2D egocentric collision envelope (ECE) was introduced by Hwang et al [6], a spatial boundary around an observer within which an approaching pedestrian is likely to be perceived as colliding. Using a large screen-based virtual simulation of forward walking, they showed that both approaching angle of the pedestrian and the walking speed affect the size of the ECE. In later work, Hwang et al [7] developed a head-mounted display (HMD) based VR system evaluating collision detection performance in immersive environments. This approach improves ecological validity but introduces the problem of cybersickness, a form of visually induced motion sickness caused by sensory conflicts during simulated self-motion [8] [9]. Hwang et al. [7] [6] shows that speed strongly influence collision perception, where faster pedestrian's approach speed associated with larger safety margins. In other hand, the work completed by Bjorn Jorge et al. [10] showed that time to collision (TTC) predictions biased by self-motion. Also, Yan et al. [11] in their study verified that forward self-motion leads to TTC over estimation compared with equivalent object-motion displays.

One possible compromise is to use a static, screen-based display that replays pedestrian trajectories recorded during dynamic walking simulations. If static replay produces similar perceptual judgments to dynamic viewing, it could serve as more comfortable alternative for repeated testing, especially in visually impaired populations.

In the present work, we extended the ECE framework by comparing:

1. Dynamic condition: simulated forwarded self-motion with pedestrians approaching from different angles toward different destinations.
2. Static condition: fixed viewpoint with pedestrians replaying the exact trajectories recorded in the dynamic condition.

We investigate whether static replay produces similar collision perception outcomes to the dynamic condition. If successful, this approach could help to reduce user

discomfort while maintaining the ecological validity of collision detection tasks in virtual environments.

2 Methodology

2.1 Experimental Setup

The experiment used a screen-based virtual environment developed with Unity 3D game engine and ran on Windows 11 PC (Intel i7, 32G RAM) with Nvidia GeForce RTX-4080 super GPU. The walking simulation was displayed on a 65" TV.

Figure 1 shows that subjects were sitting at a distance of 50 cm from the TV (65"), with these conditions the field of view of the participant covers 104° horizontally and 81.5° vertically. To maintain the correct angular size of the pedestrians and environment and eliminate any perspective distortion caused by the angular size mismatch between rendering and display condition, the field of view of the Unity 3D's rendering camera was matched with the FOV of the TV from the viewing position.



Figure 1-The experimental setup used during the test. A participant is seated 50 cm from a 65-inch screen displaying the virtual environment developed in Unity.

During the test, participants observed a virtual pedestrian approaching from a defined initial bearing angle toward a specific destination as shown in Figure 2.



Figure 2-The experimental screen shown during a trial in the screen-based virtual environment. The pedestrian moving from an initial bearing angle to pre-defined destination.

The virtual camera in the Unity 3D environment was positioned at a height of 150 cm, approximating the eye level of a standing pedestrian in order to reproduce a realistic walking perspective on the screen. Participants adjusted the seat height so that their eye level aligned with the center of the display. This alignment ensured that the viewpoint in the real world corresponded to the virtual camera position, minimizing perspective distortions between the rendered environment and the display.

2.2 Experimental design

The experiment followed a within-subject factorial design with two independent factors:

1. Viewing motion condition:
 - a. Dynamic condition: simulated forward self-motion.
 - b. Static condition: fixed viewpoint with pedestrian motion replayed from the dynamic trajectories.

2. Pedestrian motion speed:

Two walking speeds were tested for the virtual pedestrians:

- a. 1.0 m/s
- b. 1.6 m/s

These values correspond to typical human walking speeds and were used to examine how pedestrian motion speed influences collision perception.

The dynamic condition followed the collision-judgment protocol introduced by Hwang et al. [6], in which the virtual camera simulates forward self-motion toward a planned collision point (PCP) while pedestrians approach the observer along predefined trajectories. In the present study, as shown in Figure 3, pedestrians appeared from nine initial bearing angles ($\pm 45^\circ$, $\pm 30^\circ$, $\pm 20^\circ$, $\pm 10^\circ$, and 0°) and walked toward one of five possible destination offsets relative to the PCP. These offsets, illustrated in Figure 4, determine whether the pedestrian passes in front of the observer, behind the observer, or directly intersects the observer's path.

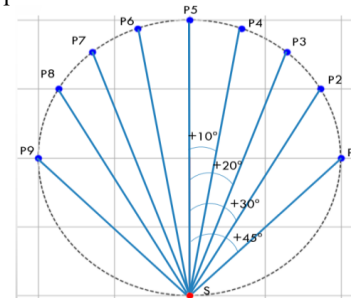


Figure 3-Virtual pedestrian approach angles defined relative to the vertical issued from the initial position of the subject S in the dynamic condition.

In the static condition, the observer viewpoint remained fixed while pedestrians followed trajectories that were pre-computed from the dynamic condition. These trajectories correspond to those that would be observed from the moving reference frame of the subject in the dynamic scenario, Figure 5 shows an example of the relative trajectories followed by the pedestrian in the static condition. By replaying these trajectories with a fixed viewpoint, the experiment preserved the same geometric motion patterns while removing the observer's self-motion.

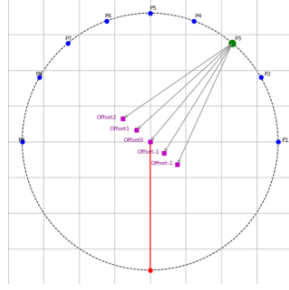


Figure 4-Five offset destinations used in the experiment, centered on the planned collision point.

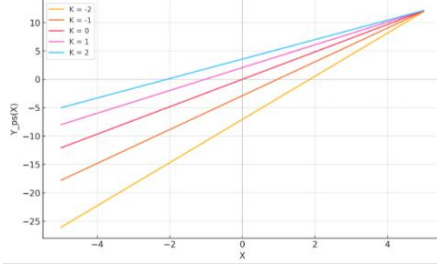


Figure 5-The trajectories followed by the pedestrian in the static condition coming from the bearing angle 30°, these trajectories are replayed from those observed from the moving reference frame of the subject in the dynamic scenario

As in the original protocol, only one pedestrian appeared in each trial. Participants observed the approaching pedestrian and were asked to report whether the trajectory would lead to a collision. The response interface appeared shortly before the predicted collision point, prompting participants to indicate their judgment using the keyboard as shown in the Figure 6.

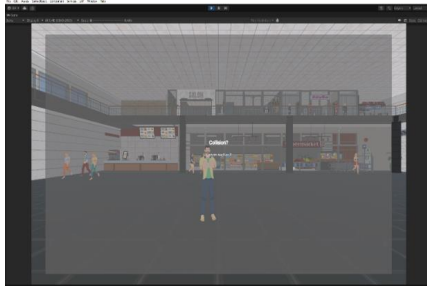


Figure 6-The UI panel prompts the participant with the question "Collision?" to indicate the moment of decision, asking whether the approaching pedestrian is perceived as colliding or not.

2.3 Participants

16 subjects with normal or corrected to normal vision participated in the test. The sample included 8 male and 8 female participants. Based on age grouping, 7 participants were classified as young adults and 9 as older adults. Participants' ages ranged from 23 to 65 years. All participants

gave written informed consent and the study followed ethical guidelines for psychophysical research.

3 Data Analysis

Responses were organized by pedestrian bearing angle and trajectory offset, and pooled across participants to compute collision response proportions. Collision thresholds were estimated by fitting psychometric functions relating perceived collision probability to offset, with the threshold defined at the 50% point separating colliding from non-colliding trajectories. To improve robustness, a bootstrap resampling procedure was applied by repeatedly sampling the data with replacement and re-estimating thresholds, producing a distribution of values whose mean was taken as the final estimate. These thresholds were then used to construct egocentric collision envelopes following the methodology of Hwang et al. [6], and 95% confidence intervals were derived from the bootstrap distributions.

4 Results

Collision envelopes were estimated for the two experimental conditions (dynamic and static) and for the two walking speeds (1.0 m/s and 1.6 m/s). For each condition and speed, the thresholds obtained for all pedestrian bearing angles were used to construct the collision envelope representing the perceptual boundary between trajectories judged as colliding and those judged as safely passing.

For the static condition, the thresholds corresponding to both front and rear passing trajectories were first gathered together. An ellipsoidal fitting method was then applied to the full set of thresholds in order to estimate a global collision envelope. The fitted ellipse represents the region around the observer within which a pedestrian trajectory is perceived as potentially colliding.

For the dynamic condition, a slightly different procedure was used. Because the presence of simulated self-motion introduces asymmetries between front and rear collision judgments, the thresholds were first separated into two groups corresponding to front passing trajectories and rear passing trajectories. An ellipsoidal fitting was then performed independently for each group. The two resulting ellipses were subsequently combined to obtain the final collision envelope describing the perceptual collision boundary for the dynamic condition.

Figures 7 and 8 illustrate the collision envelopes for the four tested configurations, enabling direct comparison across viewing conditions and walking speeds. Dynamic envelopes are larger than static ones, and envelope size increases with pedestrian speed, indicating that both simulated self-motion and motion speed enlarge perceived safety margins. This effect was confirmed statistically using a univariate ANOVA with viewing condition as a within-subject factor, revealing significantly larger thresholds in dynamic compared to static conditions.

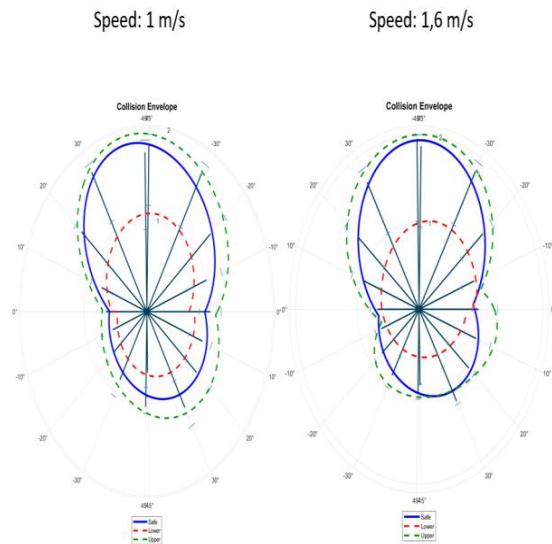


Figure 7-Collision Envelopes for the 'Dynamic condition' with two walking speeds 1.0 and 1.6 m/s

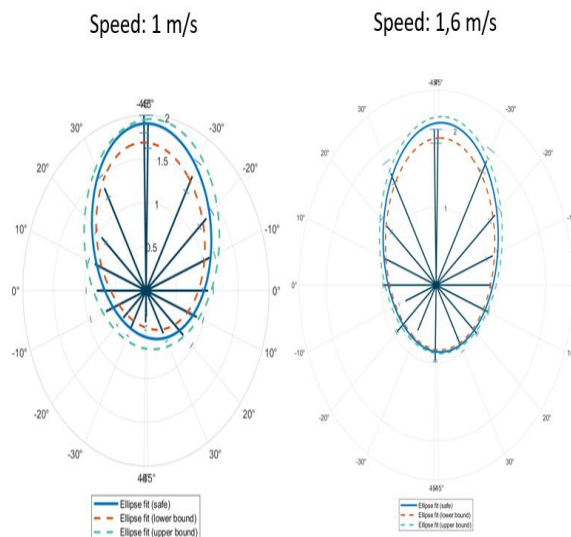


Figure 8-Collision Envelopes for the 'Static condition' with two walking speeds 1.0 and 1.6 m/s

5 Conclusion

This study investigated how egocentric collision perceptual envelopes vary across different viewing and motion conditions in a virtual pedestrian environment. By estimating collision thresholds across approach angles and walking speeds, we compared envelope geometry between dynamic and static viewing conditions. The results show that envelope shape depends on viewing condition and pedestrian walking speed, highlighting the influence of observer motion on collision envelopes and safety distances. Further research will explore these impacts with immersive virtual environments.

6 References

- [1] K. Turano, D. Geruschat, J. Stabl et R. Massof, «Perceived Visual Ability for Independent Mobility in Persons with Retinitis Pigmentosa,» *Invest Ophthalmol Vis Sci*, vol. 40, pp. 865-877, 1999.
- [2] P. Eli, «Prentice Award Lecture: Peripheral prisms for visual field expansion: A translational journey,» *Optom Vis Sci*, 2020.
- [3] C. Qiu, J.-H. Jung, M. Tuccar-Burak, L. Spano, R. Goldstein et E. Peli, «Measuring Pedestrian Collision Detection With Peripheral Field Loss and the Impact of Peripheral Prisms,» *Transl Vis Sci Technol*, vol. 7, pp. 1-10, 2018.
- [4] G. P. Spoong, J. E. Lovie-Kitchin et B. Brown, «Does mobility performance of visually impaired adults improve immediately after orientation and mobility training?,» *Optom Vis Sci*, vol. 78, pp. 657-666, 2001.
- [5] A. Bowers, G. Luo, N. Rensing et E. Peli, «Evaluation of a prototype Minified Augmented-View device for patients with impaired night vision,» *Ophthalmic Physiol Opt*, vol. 24, n° 14, pp. 296-312, 2004.
- [6] A. Hwang, J. Jung, A. Bowers et E. Peli, «Egocentric Boundaries on Distinguishing Colliding and Non-Colliding Pedestrians while Walking in a Virtual Environment,» *IS&T Int. Symp. Electron. Imaging*, vol. 36, p. 2141–2148, 2024.
- [7] A. Hwang, E. Peli et J.-H. Jung, «Development of Virtual Reality Walking Collision Detection Test on Head-mounted display,» *Proceedings of SPIE, The international Society for Optical Engineering*, vol. 12449, March 2023.
- [8] S. Weech, S. Kenny et M. Barnett-Cowan, «Presence and Cybersickness in Virtual Reality Are Negatively Related: A Review,» *Front Psychol*, vol. 10, 2019.
- [9] L. Laviola, «A discussion of cybersickness in virtual environments,» *ACM SIGCHI Bull*, vol. 32, n° 11, pp. 47-56, 2000.
- [10] B. Jorges et L. Harris, «The impact of visually simulated self-motion on predicting object motion,» *PLOS ONE*, 2024.
- [11] J.-J. Yan, B. Lorv, H. Li et H.-J. Sun, «Visual Processing of Impending Collision: Differential Processing of Object Motion and Self-motion,» *Journal of Vision*, vol. 12, n° 19, 2012.

Évaluation de la Qualité d'Expérience et de l'impact environnemental de systèmes immersifs

Emerik Aji, Amar Tious, Toinon Vigier

Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000

Nantes, France

emerik.aji@etu.univ-nantes.fr

Résumé

Les expériences immersives voient récemment leur popularité grandir, au fil de la démocratisation des appareils personnels de visionnage en réalité virtuelle (RV) tout comme le nombre grandissant d'expositions immersives grand public. Cela entraîne des enjeux écologiques importants, ces contenus étant particulièrement énergivores en rendu et en affichage, et incite au développement de méthodes d'optimisation pour réduire leur impact environnemental tout en préservant une bonne expérience pour les spectateurs. Dans cet article, nous nous intéressons à identifier, à travers une revue de l'état de l'art, des leviers technologiques prometteurs pour la diminution de la consommation énergétique. Nous questionnons ensuite les méthodologies pertinentes pour évaluer l'impact de ces leviers sur la qualité visuelle et la qualité d'expérience des spectateurs, dans un protocole expérimental comprenant différentes configurations de visionnage. Nous concluons sur l'effet que pourront avoir les résultats de cette expérience sur la création immersive artistique.

Mots clefs

Réalité Virtuelle, Environnements immersifs, Rendu 3D, Sobriété énergétique, Qualité d'Expérience

1 Introduction

L'art immersif est une forme d'art digital définie par sa nature multisensorielle, qui plonge le spectateur dans le monde virtuel de l'œuvre [1]. Cette immersion est principalement achevée par les moyens d'affichages utilisés, parmi lesquels les salles immersives avec projection sur les surfaces entourant le spectateur, les écrans immersifs haute résolution occupant une grande partie du champ de vision, et les casques de réalité virtuelle sont les plus populaires. L'ajout de l'immersion à l'art digital altère les comportements des spectateurs en leur donnant plus de liberté lors de l'exploration de l'œuvre, leur permettant de se déplacer et de regarder autour d'eux. Huang [2] observe que cela conduit les spectateurs à participer activement à l'interprétation de l'œuvre et au processus de création d'art en devenant "les protagonistes de l'œuvre d'art". De plus, un niveau d'immersion sensorielle plus élevé peut conduire à des changements d'états émotionnels du spectateur [3, 4]

et contribuer à la création de l'état de présence tel que décrit par Slater et al. [5], la sensation d'être dans le monde virtuel. Au-delà des créations immersives purement artistiques, ce travail s'intéresse à l'ensemble des expériences immersives dans le domaine des industries culturelles et créatives.

Nous pouvons observer que les systèmes d'affichage ont des caractéristiques spécifiques à prendre en compte pour les expériences immersives, notamment le nombre de spectateurs simultanés et le niveau d'immersion. Par exemple, les casques de réalité virtuelle fournissent plus d'immersion [6] mais peuvent causer malaise et inconfort lors d'une utilisation prolongée [3]. Par ailleurs, bien que les casques de réalité virtuelle soient individuels, il est possible de placer les individus dans une même expérience immersive partagée avec des avatars, ou même leur apparence réelle comme le permet VR2Gather [7].

Le contenu visuel d'une expérience immersive peut largement varier. Dans ce travail, nous portons notre attention sur rendu des contenus 3D pour les économies d'énergie possibles. En effet, les calculs complexes effectués par les cartes graphiques pour l'affichage et les tâches intensives comme les simulations physiques sont les plus largement corrélés à la consommation énergétique totale [8]. Considérant la demande de qualité suffisante pour non seulement préserver l'intention de l'œuvre mais également créer l'immersion, ainsi que la variabilité élevée des contenus 3D et moyens d'affichage immersifs, nous souhaitons donc répondre à la question suivante : Que pouvons-nous faire pour maximiser la sobriété énergétique sans affecter la qualité d'expérience des systèmes immersifs ?

2 État de l'art

2.1 Optimisations énergétiques de l'affichage

Une large partie de l'état de l'art en terme d'optimisation de contenu pour l'économie d'énergie est centrée sur l'utilisation d'écrans OLED. Puisque ces écrans fonctionnent sans rétroéclairage avec des LEDs émettant directement la couleur et luminance voulue, leur consommation dépend du contenu affiché [9], et sera en général inférieure à celle d'un écran LCD [10]. Cela permet par exemple

de développer des méthodes pour rendre les images plus sobres énergétiquement en modifiant dynamiquement la luminance [11]. Dans la plupart des cas, nous ne pouvons pas appliquer ces optimisations aux expériences immersives, les dispositifs d'affichage actuels n'étant pas ou plus basés sur la technologies OLED. Les casques de réalité virtuelle grand public modernes tels que le Meta Quest 3 ou le Steam Frame sortant prochainement sont équipés de dalles LCD, tandis que les projecteurs nécessitent un rétroéclairage.

2.2 Optimisations énergétiques du rendu 3D

Afin d'obtenir des économies d'énergie, nous nous penchons particulièrement sur les méthodes d'optimisation du rendu, qui voient des évolutions rapides à travers le domaine du jeu vidéo, très demandeur de telles optimisations graphiques. Celles-ci tournent autour de deux méthodes principales qui fonctionnent en tandem : l'upsampling ou super-resolution (SR), et la génération de frames.

Premièrement, la super-resolution est une méthode d'augmentation de la qualité visuelle qui consiste à augmenter la résolution et le niveau de détails d'une image basse résolution, comme le décrivent Anwar et al. [12]. Spécifiquement au rendu 3D, l'objectif est de produire un rendu à basse résolution qui consomme peu d'énergie, et d'utiliser des méthodes SR pour obtenir la pleine résolution désirée. Pour pouvoir effectuer cette conversion avec un coût de calcul plus bas que le rendu direct à la résolution voulue, la plupart des modèles dépendent soit d'algorithmes spécialisés, soit de réseaux neuronaux profonds. En effet, les algorithmes temporels, qui exploitent les frames précédentes pour augmenter la résolution de la frame actuelle, sont longtemps restés compétitifs face aux méthodes d'augmentation en deep learning. Les modèles plus récents et spécialisés aux jeux vidéos comme celui de Xiao et al. [13] font utilisation d'informations présentes dans les moteurs de jeu comme les vecteurs de mouvement et profondeur pour améliorer la qualité des rendus. Ce sont les modèles commerciaux, respectivement Deep Learning Super Sampling (DLSS) pour NVIDIA [14] et FidelityFX Super Resolution Upscaler (FSR) pour AMD [15], qui ont démocratisé l'utilisation de la super-resolution [16] et continuent aujourd'hui à recevoir des améliorations les plaçant à la frontière de ce domaine.

La génération de frames est beaucoup plus récente, s'intégrant dans la même lignée des technologies de super-resolution en se basant sur les mêmes algorithmes et réseaux neuronaux. Là où la super-resolution augmente les performances et baisse la quantité d'énergie consommée en réduisant la qualité initiale du rendu, la génération de frame génère des frames intermédiaires pour diviser par deux le nombre de rendus nécessaires. Il s'agit là encore de méthodes surtout implémentées par les modèles commerciaux, DLSS, FSR Frame Generation [17] et plus récemment Intel XeSS 2 [18]. La dernière version de DLSS, avec les cartes graphiques de la série RTX 50, permet notamment de générer jusqu'à trois frames intermédiaires pour

diviser le nombre de rendus nécessaires par seconde par quatre [19]. Il existe cependant des modèles spécialisés comme celui de Yang et al. [20] pour l'utilisation de la super-resolution et génération de frames sur plateformes mobiles, tels que les casques de réalité virtuelle.

Ces modèles sont paramétrables, et selon le niveau de gain en performances demandé, l'impact sur la qualité du rendu de la super-résolution peut être relativement peu perceptible. Sen et Bhushan [21] comparent des rendu natifs avec des rendus obtenu en mode Performance de DLSS et FSR (soit à 50% de la résolution réelle), et obtiennent un indice de similarité structurelle (SSIM) supérieur à 85% dans tous les cas. Par ailleurs, ils observent que le réseau neuronal de DLSS produit une image plus proche d'un rendu natif que l'algorithme temporel de FSR.

De plus, les gains en consommation énergétique des ces méthodes d'optimisations du rendu sont significatifs, comme le montre Berg [22] en comparant les gains en énergie de DLSS, FSR et XeSS dans plusieurs jeux ; Le gain maximal obtenu étant de 433% frames par joule avec DLSS dans le jeu Cyberpunk 2077, soit quatre fois plus d'images obtenues pour la même quantité d'énergie consommée.

Nous avons cependant observé un manque de recherche sur l'impact des modèles d'optimisation du rendu 3D sur la qualité visuelle subjective, particulièrement pour la génération de frame en tant que technologie récente et pratiquement unique aux jeux vidéos et à la réalité virtuelle. Nous n'avons trouvé aucun article mettant en relation la qualité visuelle et la consommation énergétique avec ces méthodes d'optimisation.

2.3 Méthodes d'évaluation

L'évaluation subjective des expériences immersives va dépendre du type de contenu évalué et repose ainsi sur une variété de méthodologies selon les dimensions expérimentales étudiées. Les recommandations ITU-T P.919 [23] définissent les bonnes pratiques pour l'évaluation subjective de contenus 360° sur des séquences vidéo d'une dizaine de secondes, avec les méthodes ACR (Absolute Category Rating) et DCR (Degradation Category Rating) [24]. Ces recommandations ne permettent pas de s'intéresser à la réception globale, sur un temps long, d'un contenu immersif artistique ou culturel. Elles ne s'appliquent pas non plus aux contenus plus immersifs comme les vidéos volumétriques, ne prenant pas en compte les degrés de liberté supplémentaires ou la calibration des casques de réalité virtuelle [25]. D'autres méthodologies subjectives, elles non standardisées, peuvent être utilisées dans ce contexte, telles que l'évaluation des émotions positives et négatives par questionnaire [26, 4], la mémorisation du contenu [4, 27], ou des questions ouvertes sur les ressentis du participant. [3]. Afin de mieux définir l'évaluation subjective et objective des expériences immersives, Zhang [28] compare les différents moyens d'en mesurer la qualité à travers les axes de la "validité, fiabilité, sensibilité, non-intrusivité et nouveauté". Les résultats de cette comparaison montrent

que les mesures neuro-psycho-physiologiques comme l'activité électrodermale sont les plus équilibrées sur ces différents axes. L'oculométrie est également pertinente, car la création de l'immersion dans ces contenus affecte la perception humaine et donc le processus d'attention visuelle [29]. L'évaluation de la qualité d'expérience peut ainsi être basée sur les dimensions perceptives de haut-niveau comme l'engagement avec l'expérience immersive, et donc se faire à travers des mesures physiologiques ou attentionnelles, qui ont pour autre avantage de laisser le participant observer l'œuvre sans tâche assignée ni distractions.

3 Méthodologie

3.1 Implémentation

Nous cherchons dans cette étude à évaluer l'influence des optimisations de rendu, la super-resolution et la génération de frames, sur la qualité d'expérience et la consommation énergétique d'œuvres immersives. Nous nous intéressons donc à leur implémentation dans deux moteurs de jeu largement utilisés dans le développement d'applications XR, Unreal Engine et Unity. Le travail de Perez et al. [30] comparant ces deux moteurs de jeu dans des cas exemples nous montre que leur consommation énergétique peut différer selon les nécessités du contenu. De plus, nous souhaitons observer comment les différences d'implémentation de DLSS, FSR et XeSS affectent les performances ; Unity n'est notamment à ce jour pas compatible avec les technologies de génération de frames.

3.2 Mesures et évaluation

A partir d'un contenu référence, nous allons appliquer les différentes options d'optimisation de rendu et suivre une procédure d'évaluation subjective de la qualité tout en mesurant la consommation énergétique à partir des capteurs système. Nous avons observé dans l'état de l'art qu'il n'existe pas de méthode standardisée pour l'évaluation subjective d'une expérience immersive qui soit applicable à notre expérience, un problème déjà mis en avant par Tious et al. [25] pour la qualité d'expérience des nuages de points dynamiques. Nous allons donc construire une méthodologie adaptée aux expériences immersives qui couplera mesures subjectives standardisées de la qualité visuelle, évaluations subjectives de la qualité d'expérience et de la réception artistique à travers des dimensions de haut-niveau comme les émotions ou la présence, et mesures physiologiques, oculométriques et comportementales durant le visionnage des œuvres. Nous nous attacherons également à comparer différents dispositifs d'affichage immersif dont des casques de réalité virtuelle, une salle immersive comprenant des projecteurs pour les murs et le sol, et un grand écran immersif haute résolution consistant d'un mur d'images de 6x4 écrans LCD.

4 Conclusion et travaux futurs

Les expériences immersives sont particulièrement énergivores en raison de la puissance de calcul nécessaire au

rendu et à l'affichage des contenus 3D. Afin que le processus de création et d'exposition des expériences immersives culturelles puisse entrer dans un cadre de sobriété énergétique, nous établissons un protocole expérimental pour évaluer différents moyens d'optimisation sur les économies d'énergie qu'ils permettent tout en préservant une qualité d'expérience élevée. Dans l'ensemble, notre objectif est de contribuer au développement de la chaîne de création d'expériences immersives en servant de recommandation pour des expériences immersives écoresponsables. Nous espérons présenter nos premiers résultats lors de la conférence en mai, afin de mettre en avant le protocole expérimental que nous aurons développé et les différentes méthodologies d'évaluation que nous aurons utilisées.

Références

- [1] Yanling Qi et Shijun Zhang. Research on the Characteristics of Immersive Installation Art from the Perspective of Digital Media. *Frontiers in Art Research*, 6(9), Septembre 2024.
- [2] Nanyan Huang. Immersive Exhibition : Its Theoretical Development, Its Audiences and the Re-discovery of Modern Art Exhibition. *Communications in Humanities Research*, 3 :309–315, Mai 2023.
- [3] Adam Greenfeld, Artur Lugmayr, et Wesley Lamont. Comparative Reality : Measuring User Experience and Emotion in Immersive Virtual Environments. Dans *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 204–209, Décembre 2018.
- [4] Mila Bujic, Mikko Salminen, et Juho Hamari. Effects of Immersive Media on Emotion and Memory : An Experiment Comparing Article, 360-video, and Virtual Reality. *International Journal of Human-Computer Studies*, 179 :103–118, Novembre 2023.
- [5] Mel Slater et Sylvia Wilbur. A Framework for Immersive Virtual Environments (FIVE) : Speculations on the Role of Presence in Virtual Environments. *Presence : Teleoperators and Virtual Environments*, 6(6) :603–616, Décembre 1997.
- [6] Kyle Johnsen et Benjamin Lok. An Evaluation of Immersive Displays for Virtual Human Experiences. Dans *2008 IEEE Virtual Reality Conference*, pages 133–136, Mars 2008.
- [7] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, et Pablo Cesar. VR2Gather : A Collaborative, Social Virtual Reality System for Adaptive, Multiparty Real-Time Communication. *IEEE MultiMedia*, 30(2) :48–59, Avril 2023.
- [8] Jyothsna Priya Kattakinda, Sachin Dhananjaya, Samuel Moji, Amine Nadif, Mattheus Hanna, Radu Ap-san, et Ivano Malavolta. Towards Understanding the Energy Consumption of Virtual Reality Applications in Gaming, Education, and Entertainment. Dans *2024*

- IEEE/ACM 8th International Workshop on Games and Software Engineering (GAS)*, GAS '24, pages 50–57, New York, NY, USA, Avril 2024. Association for Computing Machinery.
- [9] Claire-Hélène Demarty, Laurent Blondé, et Olivier Le Meur. Display Power Modeling for Energy Consumption Control. pages 760–764, Octobre 2023.
 - [10] Janisley O. De Sousa, Rafael M. Ribeiro, Bruno De M. Pinheiro, et Jussif J. A. Arnez. Power Consumption Analysis in LCD and AMOLED Display Technologies for Mobile Devices. pages 1–2, Janvier 2025.
 - [11] Emmanuel Sampaio, Claire-Hélène Demarty, et Olivier Le Meur. How to make images less power-hungry : An objective benchmark study. Dans *Proceedings of the Second International ACM Green Multimedia Systems Workshop*, GMSys '24, pages 1–7, New York, NY, USA, Avril 2024. Association for Computing Machinery.
 - [12] Saeed Anwar, Salman Khan, et Nick Barnes. A Deep Journey into Super-resolution : A Survey. *ACM Comput. Surv.*, 53(3) :1–34, Mai 2020.
 - [13] Lei Xiao, Salah Nouri, Matt Chapman, Alexander Fix, Douglas Lanman, et Anton Kaplanyan. Neural supersampling for real-time rendering. *ACM Trans. Graph.*, 39(4) :1–12, Août 2020.
 - [14] NVIDIA DLSS. <https://developer.nvidia.com/rtx/dlss>.
 - [15] AMD FSR Upscaling - AMD GPUOpen. <https://gpuopen.com/amd-fsr-upscaling/>.
 - [16] Alexander Watson. Deep Learning Techniques for Super-Resolution in Video Games. Mémoire de master, Bournemouth University, Department of Computing and Informatics, Décembre 2020.
 - [17] AMD FSR Frame Generation - AMD GPUOpen. <https://gpuopen.com/amd-fsr-framegeneration/>.
 - [18] Intel Xe Super Sampling 2 for Developers. <https://www.intel.com/content/www/us/en/developer/topic-technology/gamedev/xess2.html>.
 - [19] Aman Sirdesai, Vivekanand Limbikai, Pooja Kalyani, Samarth Joshi, et Sandeep N. Kugali. A Concise Overview of DLSS-Based Frame Generation. *ISJEM Journal*, 04(06), Juin 2025.
 - [20] Sipeng Yang, Qingchuan Zhu, Junhao Zhuge, Qiang Qiu, Chen Li, Yuzhong Yan, Huihui Xu, Ling-Qi Yan, et Xiaogang Jin. Mob-FGSR : Frame Generation and Super Resolution for Mobile Real-Time Rendering. Dans *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, pages 1–11, New York, NY, USA, Juillet 2024. Association for Computing Machinery.
 - [21] Shubhojit Sen et Bharat Bhushan. Image Quality Comparison Between Nvidias Deep Learning Super Sampling and AMDs FidelityFX Super Resolution. Dans *2024 International BIT Conference (BITCON)*, pages 1–6, Décembre 2024.
 - [22] Mikael Berg. Energy implications of upscaling and frame generation in games. Mémoire de master, Uppsala University, Department of Information Technology, 2024.
 - [23] Recommendation ITU-T P.919 : Subjective test methodologies for 360° video on head-mounted displays, Octobre 2020.
 - [24] Jesús Gutiérrez, Pablo Pérez, Marta Orduna, Ashutosh Singla, Carlos Cortés, Pramit Mazumdar, Irene Viola, Kjell Brunnström, Federica Battisti, Natalia Ciepliska, Dawid Juszka, Lucjan Janowski, Miokoaj Leszczuk, Anthony Adeyemi-Ejeye, Yaosi Hu, Zhenzhong Chen, Glenn Van Wallendael, Peter Lambert, César Díaz, John Hedlund, Omar Hamsis, Stephan Fremerey, Frank Hofmeyer, Alexander Raake, Pablo César, Marco Carli, et Narciso García. Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360° Videos : ITU-T Rec. P.919. *IEEE Transactions on Multimedia*, 24 :3087–3100, 2022.
 - [25] Amar Tious, Toinon Vigier, et Vincent Ricordel. Emerging challenges in subjective quality assessment of volumetric videos in eXtended Reality. Dans *2025 17th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–7, Septembre 2025.
 - [26] Valentijn T. Visch, Ed S. Tan, et Dylan Molenaar. The emotional and cognitive effect of immersion in film viewing. *Cognition and Emotion*, 24(8) :1439–1445, Décembre 2010.
 - [27] Jakki O. Bailey, J. Bailenson, A. S. Won, J. Flora, et K. Armel. Presence and Memory : Immersive Virtual Reality Effects on Cued Recall. Dans *Proceedings of the International Society for Presence Research Annual Conference*, 2012.
 - [28] Chenyan Zhang. The Why, What, and How of Immersive Experience. *IEEE Access*, 8 :90878–90888, 2020.
 - [29] Huiyu Duan, Xilei Zhu, Yuxin Zhu, Xiongkuo Min, et Guangtao Zhai. A Quick Review of Human Perception in Immersive Media. *IEEE Open Journal on Immersive Displays*, 1 :41–50, 2024.
 - [30] Carlos Pérez, Javier Verón, Francisca Pérez, M Ángeles Moraga, Coral Calero, et Carlos Cetina. Towards green game software engineering : A comparative analysis of energy consumption between the widespread unity and unreal video game engines. *Information and Software Technology*, 191 :107991, Mars 2026.

From Vision to Perception : A Pipeline for Assessing Walkability in Urban Environments

Abderrahim Menadi¹ Vincent Tourre² Jean-Philippe Rivière³ Myriam Servières²

¹ Lab-STICC, CNRS, UMR 6285, Université de Bretagne Occidentale (UBO), Brest, France

² Nantes Université, ENSA Nantes, École Centrale Nantes, CNRS, AAU-CRENAU, UMR 1563, Nantes, France

³ Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France

abderrahim.menadi@univ-brest.fr vincent.tourre@ec-nantes.fr
jean-philippe.riviere@univ-nantes.fr myriam.servieres@ec-nantes.fr

Abstract

Perceived walkability partly depends on cues that pedestrians can recognize while walking. This paper presents a pipeline that combines mobile eye-tracking and semantic segmentation to study how pedestrians visually experience urban environments while walking. Nine participants walked along a predefined route in Nantes, France, while wearing Tobii Pro Glasses 3. After the walk, they rated the perceived walkability of route segments and briefly commented on their experience. Scene videos were processed with a semantic segmentation model trained on the Mapillary Vistas dataset, allowing each fixation to be associated with an urban semantic class such as sidewalk, vegetation, road, facade, or vehicle. Preliminary results show that segments rated as more walkable attracted more visual attention to sidewalks and vegetation, whereas lower-rated segments were associated with more attention to facades and roads. Higher-rated segments also showed higher saccade rates, suggesting more active visual exploration. These first results support the value of combining gaze and scene semantics to better understand perceived walkability in real urban conditions.

Keywords

Walkability, urban perception, mobile eye-tracking, semantic segmentation, visual attention.

1 Introduction

The way people experience urban environments, particularly regarding sustainability, quality of life, and public health, has become increasingly important [1]. One key element of this relationship is how people perceive their ability to "walk" in an urban area, which influences comfort outdoors, physical activity levels, and the overall experience of daily life in the city [2, 3]. Traditionally, urban planners have described walkability using attributes of the built environment, including connectivity, land use diversity, street design, and pedestrian infrastructure [4, 5, 6]. These indicators are useful in describing the built environment, but do not fully capture how a place is actually experienced while walking. Two streets may appear identical in their urban structure, yet feel significantly different to pedestrians in terms of openness, safety, visual appeal, and other factors.

To enhance understanding of walkability, several studies have used eye tracking to determine how pedestrians vi-

ually interact with their surroundings in real time [7, 2, 8, 9]. In parallel, computer vision methods such as semantic segmentation now allow large-scale identification of elements like sidewalks, roads, facades, vegetation, and vehicles at the pixel level [10, 11, 12, 13, 14, 15, 16]. These approaches rely on large-scale urban scene datasets such as Mapillary Vistas and Cityscapes [17, 18]. However, eye-tracking studies often require manual definition of Areas of Interest (AOI), which limits scalability, while image-based studies are often restricted to static views and do not capture the dynamic first-person experience of navigating a city or town while walking.

More recently, several studies have begun to link perceptual and computational methods by using behavioral observation, deep learning, and urban experience analysis [19, 20]. In this paper, we continue in that direction by proposing a fixation-to-semantics pipeline that connects mobile eye-tracking data with semantic segmentation of first-person video. Our aim is to study how visual attention is distributed across urban elements and how this distribution relates to the perception of walkability. More specifically, our contributions are : (1) a reproducible workflow for mapping fixations to urban semantic classes in realistic walking conditions, and (2) preliminary evidence demonstrating that the way people distribute their visual attention differs significantly when they walk in areas considered more or less walkable than others. In this sense, this study also serves as a methodological bridge between subjective experiences of urban life and automated scene understanding.

2 Method

2.1 Pilot study

Nine participants walked individually along a 450m predefined route in central Nantes, France, during the daytime in summer conditions 1. The path included various urban environments, ranging from areas with different levels of greenery (pedestrian zones) to more mineral settings (sidewalks, etc.). During the walk, participants wore Tobii Pro Glasses 3, which recorded both the video scenes at 25 frames per second (fps) and participants' eye movements/gaze at 100 times per second (Hz).

After completing the walk, the participants divided the route into segments, assigned a walkability rating to each segment, and provided a brief explanation for each rating. This study has been approved by the ethics committee of



FIGURE 1 – *Experimental Walking Path - Start : 16 Bd de l'Estuaire, End : 16 Rue Louis Blanc, Nantes.*

Nantes University.

2.2 Processing pipeline

The workflow is summarized in figure 2. First, valid fixations were identified in the Tobii eye-tracking data and synchronized with the video frames using timestamps. The video frames were then processed with a semantic segmentation model (Mask2Former) trained on the Mapillary Vistas dataset to generate pixel-level masks. Each fixation was then mapped to the semantic segmentation of the corresponding frame, allowing the fixation to be assigned to one of the following classes : sidewalk, vegetation, facade, road, or vehicle.

Following the mapping step, we calculate the fixation time (duration) and count for each semantic class, for each route segment, and for each participant. In addition, we calculated the "attention share," which was the percentage of total fixation time spent in each semantic class. We also measured *saccade rate* as an additional indicator of visual exploration.

2.3 Comparison strategy

To identify shared tendencies among participants, we compared the segment each participant rated highest with the segment each participant rated lowest. This comparison produced two grouped data sets : one for the best-rated segments and one for the worst-rated segments. We then analyzed the semantic attention profiles and saccade dynamics for both data sets. Finally, we considered participants' qualitative comments as supporting context to help explain the patterns identified in their visual behavior.

3 Results and Discussion

The analysis reveals consistent differences between the two groups of segments; both were viewed differently, as expected from earlier studies on visual focus and how people perceive the urban environment while walking [7, 2, 9, 19]. In the highest-rated segments, participants devoted more fixation time to sidewalks and vegetation. In the lowest-rated segments, attention was more strongly directed to-



FIGURE 2 – *Overview of the processing pipeline linking mobile eye-tracking, semantic segmentation, and segment-level walkability ratings.*

ward facades and roads. These results suggest that a more visually pleasant walking environment leads people to focus on components that support walking, such as vegetation and sidewalks, rather than on more rigid or constrained urban features.

Figure 3 shows the difference in the mean share of fixations on each object category for the two types of street segments. Participants' attention is primarily focused on the sidewalk and vegetation in the most favorably rated streets, while it is mainly on the façade and road in the least favorably rated ones. These trends are consistent with participants' verbal comments about their experiences. They frequently mentioned greenery, openness, and comfort as reasons for liking certain areas, and cited monotony, mineral surfaces, enclosure, or traffic exposure as reasons for disliking other areas.

A second results concern visual exploration. As shown in Figure 4, Best-rated segments showed higher saccade rates overall in comparison to lower-rated segments, suggesting that pleasant environments facilitate more active and distributed visual exploration of the scene. In contrast, less comfortable sections were associated with less exploratory gaze patterns than best-rated sections, suggesting that they encourage visually focused attention on a smaller set of dominant elements instead of visually exploring multiple features of the environment.

A third view of the results is presented in the difference plot

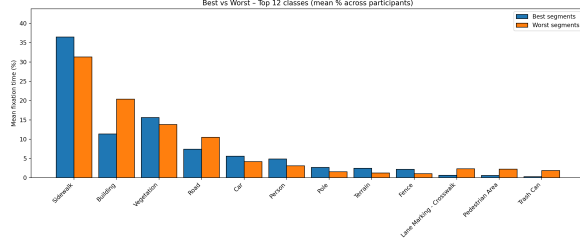


FIGURE 3 – Mean fixation share by semantic class in the best- and worst-rated segments across participants.

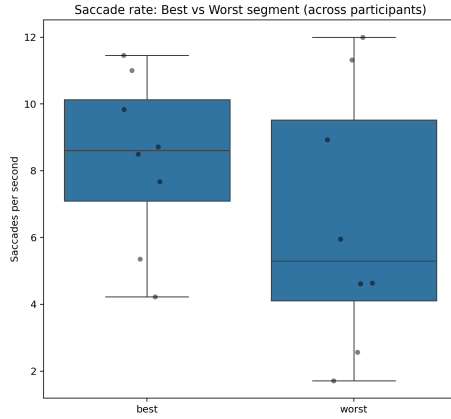


FIGURE 4 – Comparison of saccade rates between the best- and worst-rated segments across participants.

5. Positive values on this plot indicate classes with a greater proportion of time spent fixating in the top-rated segments. Conversely, negative values indicate classes more frequently fixated in the bottom-rated segments. The difference plot provides a compact visual comparison of the pedestrian-friendly aspects of the environment with those that are more restricted or less so, and helps summarize the general tendency identified among all participants.

Taken together, these patterns suggest that perceived walkability arises not only from the physical composition of the environment (the "scene") but also from the way pedestrians visually navigate it while in motion. These findings are significant for two reasons. First, they suggest that perceived walkability is reflected not only in what people report, but also in measurable ways individuals visually interact with the environment. Second, these findings highlight the value of an automated fixation-to-semantics approach for studying urban perception without manually defining AOI.

Beyond the new methodology, this study has implications for urban design practitioners, especially given the substantial evidence that the presence of green spaces, pedestrian-friendly features, and urban environments that are easy to visually understand and navigate can improve both the overall walking experience and the perception of

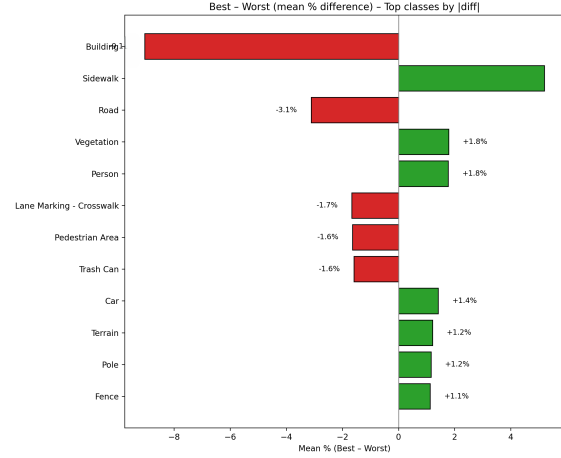


FIGURE 5 – Difference in fixation share between best- and worst-rated segments. Positive (resp. negative) values indicate greater attention in the best (resp. worse) segments.

an area's environmental quality. If confirmed on larger datasets, the proposed methodology could help identify which spatial features consistently support positive pedestrian experience. In particular, the contrast observed between attention directed toward sidewalks and vegetation in positively rated segments, and toward facades and roads in negatively rated segments, suggests that perceived walkability may depend not only on the presence of infrastructure, but also on how visually accessible and engaging pedestrian-oriented elements are during walking.

3.1 Limitations and Perspectives

This study is a pilot investigation based on a limited number of participants and a single urban route. The reported results should therefore be interpreted as preliminary tendencies rather than broadly generalizable findings. In addition, the present analysis is descriptive and does not yet include inferential statistical testing. Future work should extend the protocol to a larger and more diverse participant group, include multiple urban contexts, and formally test the robustness of the observed relationships between semantic attention patterns, gaze dynamics, and perceived walkability.

4 Conclusion

This paper presents a processing pipeline that integrates mobile eye-tracking and semantic segmentation to study perceived walkability in real urban environments. This method enables the association of gaze fixations with specific urban elements without manual annotation, enhancing reproducibility and scalability.

Preliminary results indicate that more walkable segments are associated with greater attention to sidewalks and vegetation, as well as higher visual exploration, while less walkable segments draw more attention to facades and roads.

These findings support the idea that gaze-based measures can provide valuable behavioral insights into how pedestrians experience the city.

More broadly, these results align with the growing literature advocating for stronger integration of perceptual, behavioral, and computational approaches in walkability assessment. Future research should include more participants, a greater variety of urban routes, and statistical analyses to test the robustness of these trends.

Acknowledgments

This work was supported by the FAISTOS cluster (Fostering Artificial Intelligence in Industrial Systems and Technologies) at Nantes University.

Références

- [1] Jonas De Vos, Katrin Lättman, Anna-Lena van der Vlugt, Janina Welsch, et Noriko Otsuka. Determinants and effects of perceived walkability : A literature review, conceptual model and research agenda. 43(2) :303–324, 2023.
- [2] James Simpson, Kevin Thwaites, et Megan Freeth. Understanding visual engagement with urban street edges along non-pedestrianised and pedestrianised streets using mobile eye-tracking. 11 :4251, 2019.
- [3] Reid Ewing et Susan Handy. Measuring the unmeasurable : Urban design qualities related to walkability. 14(1) :65–84, 2009.
- [4] Ester Cerin, Brian E. Saelens, James F. Sallis, et Lawrence D. Frank. Neighborhood environment walkability scale : Validity and development of a short form. 2006.
- [5] Lin Lin et Anne Vernez Moudon. Objective versus subjective measures of the built environment : Which capture associations with walking ? 2010.
- [6] Fernando Fonseca, George Papageorgiou, Simona Tondelli, Paulo Ribeiro, Elisa Conticelli, Mona Jabbari, et Rui Ramos. Perceived walkability and respective urban determinants : Insights from bologna and porto. 14 :9089, 2022.
- [7] J. C. Simpson, M. Freeth, K. Simpson, et K. Thwaites. Visual engagement with urban street edges : Insights using mobile eye-tracking. 12(3) :259–278, 2018.
- [8] Justin B. Hollander, Alexandra Purdy, Andrew Wiley, Veronica Foster, Robert J. K. Jacob, Holly A. Taylor, et Tad T. Brunyé. Seeing the city : Using eye-tracking technology to explore cognitive responses to the built environment. 2018.
- [9] Congying Fang, Riken Homma, Tianfu Qiu, et Qikang Zhong. The mediating role of visual behavior in perceived walkability evaluation. 2024.
- [10] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, et Piotr Dollár. Panoptic segmentation. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9404–9413, June 2019.
- [11] Hao Zhou, Shenjing He, Yuyang Cai, Miao Wang, et Shiliang Su. Social inequalities in neighborhood visual walkability using street view imagery and deep learning. page 101605, 2019.
- [12] Xuan He et Sylvia He. Using open data and deep learning to explore walkability in shenzhen, china. page 103696, 2023.
- [13] Shohei Nagata, Tomoki Nakaya, Tomoya Hanibuchi, Shiho Amagasa, Hiroyuki Kikuchi, et Shigeru Inoue. Objective scoring of streetscape walkability with semantic segmentation of google street view images. page 102428, 2020.
- [14] Bon Woo Koo, Subhrajit Guhathakurta, et Nisha Botchwey. Walkability factors and walking behaviors : A big data approach using street view images. 2021.
- [15] Xiaoran Huang, Li Zeng, Hanxiong Liang, Daoyong Li, Xin Yang, et Bo Zhang. Comprehensive walkability assessment of urban pedestrian environments using big data and deep learning. 14 :26993, 2024.
- [16] Koichi Ito, Yuhao Kang, Ye Zhang, Fan Zhang, et Filip Biljecki. Understanding urban perception with visual data : A systematic review. 152 :105169, 2024.
- [17] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, et Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. Dans *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009, 2017.
- [18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, et Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.
- [19] Arash Tavakoli, Isabella P. Douglas, Hae Young Noh, Jackelyn Hwang, et Sarah L. Billington. Psycho-behavioral responses to urban scenes : An exploration through eye-tracking. 156 :105568, 2025.
- [20] Yuexin Wang et Yu Zhang. Time perception and street visual elements during walking : An eye-tracking study. 24(3) :2133–2153, 2025.

Fusion de YOLO à multi-échelles pour la détection d’objets à tailles très variable

A. Vedel^{1,4}, J. Pasquet^{2,3}, R. Thevenoux¹, J. Cortet⁴, C. Villenave¹

¹ ELISOL Environnement, France

{alvin.vedel, romain.thevenoux, cecile.villenave}@elisol-environnement.fr

² AMIS, Université Paul-Valéry Montpellier 3, France

³ TETIS, INRAE, AgroParisTech, CIRAD, CNRS, Univ. Montpellier, France

⁴ Univ Paul Valéry Montpellier 3, Univ. Montpellier, EPHE, CNRS, IRD, CEFU UMR 5175, F34000, Montpellier, France

{jerome.pasquet, jerome.cortet}@univ-montp3.fr

Résumé

La détection de nématodes sur plaque microscopique présente plusieurs défis, notamment par la taille conséquente des images mais également par la grande variabilité de taille des objets à détecter. Le redimensionnement peut rendre impossible la détection des plus petits objets. De plus, les modèles de détection basés sur les convolutions présentent une limite au niveau du champ réceptif rendant la détection des plus grands nématodes impossible d’un bloc. Pour résoudre ce problème, nous nous comparons à la littérature et proposons l’utilisation de modèle de détection à différentes échelles ainsi que plusieurs méthodes d’agrégation de leurs prédictions mettant en avant un gain du macro- F_1 .

Mots clefs

Analyse d’images, détection d’objets, images microscopiques, écologie des sols.

1 Introduction

Les nématodes sont des vers microscopiques présents en abondance dans le sol. Ils peuvent être utilisés comme bio-indicateurs : ils renseignent par leur présence, leur absence et leur diversité sur le fonctionnement biologique d’un écosystème. Caractériser l’abondance en nématodes d’un échantillon est une étape clé pour évaluer l’état de santé biologique du sol par l’étude de la nématofaune. Bien qu’un écart de comptage entre la détection et la vérité terrain puisse être toléré, la diversité nématologique doit être conservée. Certains des plus grands nématodes par exemple sont des prédateurs qui traduisent la stabilité du milieu, il est donc essentiel de bien les détecter. La difficulté de cette tâche réside principalement dans la grande variabilité de taille entre les nématodes avec un facteur 200 entre l’aires des petits et celles des grands. Certains nématodes peuvent mesurer moins d’une centaine de pixels, d’autres dépassent le millier et dépassent les limites du champ réceptif des modèles de détection convolutifs. Il est impossible de redimensionner l’image sous peine de

perdre en qualité de détection sur les plus petits nématodes car certains patterns discriminant font moins de 8×8 pixels. Pour répondre à cette problématique, nous commencerons par présenter et discuter les méthodes existantes dans la littérature, puis nous exposerons notre approche, détaillerons nos expériences, et enfin nous analyserons les résultats obtenus.

2 Etat de l’art

La détection d’objet est un paradigme du Deep Learning très populaire et répandu. Elle consiste à prédire un ensemble de boîtes englobantes pour une image donnée. La famille la plus populaire d’algorithme de détection est celle des YOLO (You Only Look Once) [1], prise pour ses performances et sa rapidité. Le modèle que nous avons choisi d’utiliser est YOLOv5 [2]. C’est un modèle convolutif constitué d’un extracteur CSPDarknet-53, un Spatial Pyramid Pooling - Fast (SPPF), un Path Aggregation Network (PANet) et trois têtes de prédiction à trois résolutions spatiales différentes : 1/8ème, 1/16ème et 1/32ème.

Pour pouvoir traiter des images haute résolution malgré les limites matérielles, le Slicing Aided Hyper Inference (SAHI) a été mis en place et consiste à inférer l’image par morceau, [3] explore des méthodes d’agrégation des prédictions par SAHI sur des images panoramiques. Leur travail consiste principalement à prédire de petits objets avec précision. De grands objets en haute résolution forment une double peine pour les modèle convolutifs travaillant au niveau local, ils deviennent d’autant plus incapable de cerner l’objet s’il est proche ou au-delà du champ réceptif.

Les limites des articles actuels reposent sur leurs spécificités : beaucoup font état d’améliorations pour la détection des petits objets comme [4] en inférant également des parties de l’image ou des travaux plus proches de [5] qui apportent une modification architecturale pour mieux prédire les petits objets. Il s’agit du cas inverse au notre où ce sont les grands objets qui sont problématiques, réduire la qualité des images pour se rapprocher de leur situation n’est pas envisageable car la qualité des images serait trop dégradée par un redimensionnement, rendant la détection

des plus petits impossible. Certains travaux comme [6] explorent empiriquement la relation entre taille de l'image et taille des objets détectés par YOLO. En utilisant un unique YOLO, ils estiment pour un redimensionnement donné la plage de taille d'objet détectés par le modèle. Ils sélectionnent ensuite deux redimensionnements permettant de couvrir, par l'union de leurs plages, toute la plage des objets réels. Nous proposons d'aller plus loin que cette démarche en intégrant le concept du SAHI et sans se restreindre sur le nombre de modèles utilisés, d'autant plus que nous utilisons un modèle entraîné à chaque résolution là où ils réutilisent le même.

3 Données et métriques

Les données consistent en une centaine d'images acquises au microscope automatique, chaque photo est obtenue par mosaïcage d'une centaine d'images. La taille finale de ces photos est de 24000 pixels de haut pour 16000 de large. Une séparation entre entraînement, validation et test a été réalisée avec 66% images utilisées pour l'entraînement, 17% pour la validation et 17% pour le test, soit 80, 20 et 20 respectivement. Pour entraîner les modèles, des sous-images de différentes tailles ont été extraites sur les images complètes avant d'être redimensionnées. Le jeu de validation et de test servent chacun leur tour de jeu de calibration pour la sélection paramétrisée des prédictions entre modèles afin d'apporter une validation croisée des résultats. Le point le plus important concerne la distribution de taille des nématodes, comme on peut le voir en figure 1. La plupart sont de petits nématodes inférieurs à 350 pixels sur chaque dimension mais il existe certains extrêmes, principalement dans le jeu d'entraînement qui est plus représentatif, pouvant dépasser les 1000 pixels.

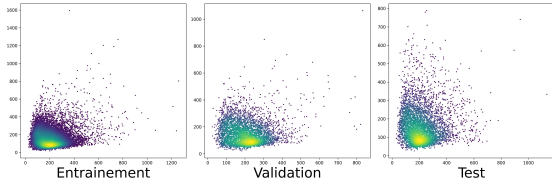


FIGURE 1 – Distribution de la taille en hauteur / largeur des nématodes sur chaque jeu de données

Pour évaluer notre méthode nous utiliserons un macro- F_1 par tranche de tailles défini comme :

$$\text{Macro-}F_1 = \frac{1}{C} \sum_{c=1}^C F_{1c} = \frac{1}{C} \sum_{c=1}^C 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (1)$$

avec C l'ensemble des tranches de tailles. Le détail de la séparation par taille est fourni dans le tableau 1, les tailles sont celles de la diagonale de la boîte. Les tranches de tailles ont été adaptées pour ne pas contenir trop peu de nématodes mais on voit bien l'asymétrie du problème.

Tailles	0–200	200–300	300–400	400–550	>550
Entraînement	4344	8739	4323	1107	363
Validation	853	1997	1080	265	57
Test	984	2099	1204	516	179

TABLEAU 1 – Nombre de nématodes par tranche de taille.

4 Méthodes d'aggrégation

Afin de résoudre le problème de champ réceptif limitant pour les plus grands organismes et de conserver une qualité d'image suffisante à la détection des plus petits, nous avons mis en place une méthode qui combine le SAHI et l'apprentissage de modèle à différentes échelles ainsi que la fusion de leurs prédictions. Concrètement, notre méthode consiste à entraîner N YOLOv5 sur différentes base de données constituées en extrayant des sous-images de tailles variables avant de les redimensionner en 1024×1024 . Nous avons extrait des sous-images à résolution réelle (1:1), en 2048 (avant redimensionnement) (1:2), 3072 (1:3), 4096 (1:4), 5120 (1:5), 6144 (1:6) ainsi que 768 (1:0, 75). Les modèles sont entraînés indépendamment puis vient la phase d'aggrégation dans la quelle nous proposons 4 variantes :

- Fusion naïve : il s'agit de la méthode de référence qui pourrait être mise en place en première approximation et que les méthodes développées permettent d'améliorer. Elle consiste à récupérer les prédictions de chaque modèle et les considérer comme un ensemble unique, appliquer un NMS et filtrer sur la confiance.
- Sélection experte : cette seconde méthode proposée est une première amélioration de la méthode naïve qui prend en compte les spécificités de chaque modèle. En se basant sur la taille des objets détectés par chaque modèle, il est possible de définir des plages de taille pour chacun et de sélectionner seulement une partie des modèles qui couvrirait l'ensemble des objets possibles. Dans les résultats présentés, c'est la combinaison des modèles 1:1, 1:2 et 1:5 qui est utilisé avec des plages de détection de $[0, 300]$, $[0, 1200]$, $[700, 10000]$ respectivement. Ces trois modèles correspondent à nos besoins primaires : détection des plus petits, détection des plus grands et un bon détecteur sur les tranches intermédiaires.
- Optimisation génétique : à l'aide d'un algorithme génétique comme celui développé par [7], il est possible d'optimiser directement la métrique d'intérêt en trouvant les meilleurs hyper-paramètres. Nous proposons la recherche des hyper-paramètres suivants : une taille minimum et maximum de boîte prédite afin que les prédictions restent dans la tranche "facile" pour le modèle, un seuil de confiance et un seuil d'IoU, le tout pour chaque échelle. Les prédictions de chaque modèles sont

alors filtrées avant d’être agrégées par un IoU également trouvé par l’algorithme génétique. Dans toutes les expériences l’optimisation s’est faite avec une population de 1000 individus et les paramètres de références utilisées par [7] : F_1 , $F_2 = 0,25$, F_3 , $F_4 = 0,2$, $CR = 0,8$, jusqu’à la convergence sur un jeu de validation.

- Filtrage conditionnel par forêt aléatoire : cette dernière proposition met en avant une sélection plus centrée sur les caractéristiques des boîtes avec un aspect moins global que l’algorithme génétique. En effet, il est possible d’apprendre une forêt aléatoire d’arbres de décision qui décide de conserver ou non une boîte pour procéder à l’aggrégation. La calibration de réseaux neuronaux par forêts aléatoires a été explorée par [8] et les travaux de [9] montrent réellement le potentiel en terme de calibration de telles méthodes. Le modèle de Machine Learning recevra en entrée les dimensions de la boîte prédite ainsi que sa confiance, le modèle à l’origine de la prédiction et les mêmes informations pour les N boîtes les plus proches de chaque modèle avec leurs IoU respectifs. Dans ce contexte-ci, on définit une boîte comme valide si elle correspond à une vérité terrain, plusieurs boîtes peuvent être considérées bonne via la même vérité terrain c’est pourquoi un NMS est ensuite appliqué.

5 Expériences et résultats

Le modèle de référence auquel on se comparera est un YOLOv5 entraîné à la résolution 1:2, le YOLO à l’échelle 1:1 n’étant pas capable de prendre les plus grands nématodes du jeu d’entraînement sur une seule sous-image. De plus, les performances varient drastiquement selon l’échelle employée et celui à la résolution 1:2 présente le meilleur F_1 -macro. Nous comparons l’utilisation d’un seul modèle avec la méthode de [6] utilisant un modèle à deux redimensionnements ainsi que nos différentes méthodes d’aggrégation proposées, à la fois dans la configuration de [6] avec un modèle à N redimensionnements et N modèles à leurs échelles respectives. Les optimisations post-entraînement sont cross-validées via le jeu de validation et de test en

les inversants, seule la moyenne est renseignée dans les tableaux de résultats. Les résultats calculés au seuil d’IoU 0.75 sont présentés dans le tableau 2 et ceux au seuil d’IoU 0.9 dans le tableau 3.

5.1 Résultats au seuil 0.75

Comme on peut le constater dans le tableau 2 affichant les F_1 au seuil d’IoU de 0.75, l’ajout de modèles de façon naïve a tendance à ajouter des erreurs qui ne sont pas compensées par les nouvelles détections. Un contrôle de ces multiples détecteurs par de simples règles expertes reposant sur la taille des objets limite déjà grandement la chute de performance bien que ne permette pas d’obtenir un gain significatif sur le meilleur modèle unique. L’algorithme génétique n’apporte pas de gains et fait davantage chuter les performances qu’une fusion naïve, on pourrait s’attendre à retrouver des valeurs équivalentes à la sélection experte mais il est très sensible à l’overfit d’autant plus considérant la faible quantité de données de validation. De plus, il opère sur des règles très globales qui manquent de contexte local. A l’inverse, la forêt aléatoire utilise des informations de prédictions voisines pour décider de conserver une boîte et cela semble être une information très utile au vu du gain significatif que la méthode apporte sur le modèle unique, de plus cette méthode d’aggrégation permet d’exprimer directement une relation entre la taille et la confiance d’une boîte ce qui n’est pas permis par des seuils fixes. Ce gain est également présent dans la méthode de détection de [6] ce qui démontre l’intérêt de la méthode. Au seuil d’IoU 0.75, les résultats entre l’approche de [6] utilisant un seul modèle à différents redimensionnements sont relativement équivalents avec l’utilisation de différents modèles entraînés sur un redimensionnement spécifique.

5.2 Résultats au seuil 0.9

L’évaluation des méthodes d’aggrégation est assez similaire au seuil d’IoU de 0.9 avec la fusion naïve qui fait chuter les performances, la sélection experte qui arrive à maintenir un score plus proche du meilleur modèle unique, l’optimisation génétique qui fait chuter les performances, et la forêt aléatoire qui permet un gain encore plus marqué à la fois via la méthode de [6] et la notre. Cependant, on

Modèles	<200	200–300	300–400	400–550	>550	Macro- F_1
Résolution 1:2	0,817	0,926	0,943	0,939	0,905	0,906
Fusion naïve	0,763	0,916	0,940	0,927	0,885	0,886
Sélection experte	0,819	0,924	0,943	0,939	0,909	0,907
Optimisation génétique	0,805	0,918	0,94	0,932	0,897	0,893
Forêt aléatoire	0,885	0,952	0,969	0,954	0,921	0,936
[6] original (2 modèles)	0,790	0,913	0,927	0,916	0,819	0,873
[6] + fusion naïve (tous modèles)	0,763	0,916	0,940	0,927	0,885	0,886
[6] + Sélection experte	0,797	0,920	0,943	0,939	0,91	0,902
[6] + Optimisation génétique	0,778	0,911	0,928	0,926	0,882	0,886
[6] + Forêt aléatoire	0,893	0,948	0,965	0,947	0,934	0,937

TABLEAU 2 – F_1 par tranches de tailles et macro- F_1 au seuil d’IoU 0.75

Modèles	<200	200–300	300–400	400–550	>550	Macro- F_1
Résolution 1:2	0,606	0,704	0,701	0,682	0,662	0,671
Fusion naïve	0,512	0,655	0,667	0,643	0,658	0,627
Sélection experte	0,596	0,697	0,701	0,681	0,647	0,664
Optimisation génétique	0,477	0,595	0,625	0,646	0,657	0,6
Forêt aléatoire	0,694	0,81	0,833	0,804	0,755	0,779
[6] original (2 modèles)	0,514	0,617	0,592	0,547	0,481	0,55
[6] + fusion naïve (tous modèles)	0,549	0,659	0,666	0,606	0,624	0,621
[6] + Sélection experte	0,549	0,667	0,7	0,681	0,666	0,652
[6] + Optimisation génétique	0,507	0,623	0,598	0,599	0,631	0,592
[6] + Forêt aléatoire	0,739	0,817	0,823	0,753	0,674	0,761

TABLEAU 3 – F_1 par tranches de tailles et macro- F_1 au seuil d'IoU 0.9

peut noter un écart plus conséquent au seuil d'IoU 0.9 en faveur de l'agrégation des N modèles plutôt que l'utilisation d'un seul modèle à N redimensionnements. En effet, peu importe la méthode d'agrégation il semble préférable d'utiliser des modèles entraînés spécifiquement sur leur redimensionnement car ils permettent de mieux respecter les dimensions de la vérité terrain. Le fait d'être entraînés directement sur un redimensionnement garantit une meilleur correspondance entre prédictions et vérités terrains, c'est particulièrement vrai lors de l'agrégation par forêt aléatoire.

6 Conclusion

Au regard des résultats, il est évident que l'intégration de plusieurs modèles de détection à différentes échelles n'est pas une tâche triviale et peut engendrer une chute des performances globales. L'utilisation d'une forêt aléatoire intégrant les informations de boîtes environnantes permet d'obtenir un gain de **0.108** du macro- F_1 au seuil d'IoU de 0.9. Les méthodes plus simples reposant sur des filtres de tailles ou de confiance par modèle peinent à être équivalentes au meilleur modèle individuel, ce qui sous-entend que la règle qui régit la sélection optimale des boîtes est assez complexes. Considérant les excellentes performances de l'agrégation par forêt aléatoire utilisant uniquement les informations des boîtes les plus proches de chaque modèle, il semble prometteur de s'orienter vers des méthodes tirant parti d'une information plus globale bien que les performances actuelles sont déjà largement satisfaisantes pour notre besoin, surtout si l'on regarde l'augmentation du macro- F_1 dans les plages extrêmes de tailles avec +0.088 sur les petits nématodes et +0.093 sur les grands nématodes au seuil d'IoU 0.9.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2025-AD011016362 attribuée par GENCI.

Références

- [1] Joseph Redmon, Santosh Divvala, Ross Girshick, et Ali Farhadi. You only look once : Unified, real-time object detection. Dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.
- [2] Glenn Jocher. Ultralytics yolov5, 2020.
- [3] Yelyzaveta Zhabska Andrii Yaroshenko Oleksii Bychkov, Kateryna Merkulova. Enhancing object detection and classification in high resolution images using sahi algorithm and modern neural networks. *Information Technology and Implementation*, 2024.
- [4] Robert Laganière Xingyi Yang, Yonghuai Wang. A scale-aware yolo model for pedestrian detection. *International Symposium on Visual Computing*, 2020.
- [5] Lei Qian Shilong Zhou, Haijin Zhou. A multi-scale small object detection algorithm sma-yolo for uav remote sensing images. *Scientific Reports*, 2025.
- [6] Daisuke Minami et Kiyoshi Nishikawa. Yolo for high resolution images without retraining. *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024.
- [7] Ji Qiang et Chad E. Mitchell. A unified differential evolution algorithm for global optimization. *IEEE Transactions on Evolutionary Computation*, 2014.
- [8] Dhruv Nigam. Forecal : Random forest-based calibration for dnns. Dans *Proceedings of the 8th International Conference on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD)*, CODS-COMAD Dec '24, page 44–51. ACM, Décembre 2024.
- [9] Mohammad Hossein Shaker et Eyke Hüllermeier. Random forest calibration. *Knowledge-Based Systems*, 328 :114143, 2025.

Hand Movement Tracking for Motor Pattern Identification in Learners with Trisomy 21 : Preliminary Results

Caio Franco Porto¹, Deise Santana Maia², Marian Oliveira¹

¹Universidade Estadual do Sudoeste da Bahia, Centro Saber Down, PPGLin, Vitória da Conquista, Brazil

²Université de Lille, CNRS, Centrale Lille, UMR 9189 CRISTAL, F-59000 Lille, France
caiofranco246@gmail.com, deise.santanamaia@univ-lille.fr, marian.oliveira@uesb.edu.br

Abstract

The objective of this work is to present the results of an experiment as a first step in the development of an algorithm aimed at tracking hand movements during writing. Using metrics such as stroke size and number, stroke duration, and average writing speed, we aim to collect kinematic data of hand movements extracted from videos during writing to determine motor patterns of the population with Trisomy 21 (T21). We tested a preliminary version of our algorithm with two participants with T21. The results of the first experiment show the algorithm's potential for defining motor patterns. Our long term goal is to combine technological and educational approaches to promote inclusive practices adapted to the specific needs of people with T21, which can contribute to the development of fine motor coordination and, consequently, improve the teaching-learning process in this population.

Keywords

Writing, Motor skills, Down Syndrome, Trisomy 21

1 Introduction

This work presents the results of the first experiment conducted as a step in the development and validation of an algorithm designed to quantify precise kinematic data of handwriting, such as writing speed, pause duration, number of pauses, and stroke count, from video recordings of individuals, focusing specifically on those with Trisomy 21 (T21). Our hypothesis is that an algorithm such as the one being developed can determine the motor patterns in the handwriting development of people with T21. Our proposed methodology goes beyond the traditional handwriting analysis based on digitizing tablets, which directly capture spatiotemporal characteristics, by extracting comparable kinematic parameters from standard video footage [1]. This approach is particularly valuable given the difficulties associated with direct interaction with digitizing tablets for certain populations, offering a less invasive and more adaptable data collection method. Furthermore, by estimating kinematic data from videos, this method can circumvent the limitations of specialized hardware, allowing for broader applicability and accessibility in various

research environments, including clinical-educational settings. The application of kinematic writing parameter analysis has been shown to provide quantitative measures for understanding deficits in motor skills, which are frequently observed in neurodevelopmental disorders [2]. This approach facilitates the analysis of subtle variations in motor control that may not be discernible through observational methods alone. This is particularly relevant for conditions such as T21, where the development of fine motor coordination and handwriting proficiency may exhibit unique characteristics [3], since this population, due to their phenotype, face motor challenges that impact writing development [4] beyond those inherent to the relationship between purely linguistic aspects and those related to how writing materializes them. As previously mentioned, traditional methods often rely on specialized digitizing tablets, such as the Wacom Intuos 4, to record writing trajectories and pen pressure. However, these devices can pose challenges for individuals with intellectual disabilities or fine motor impairments, requiring alternative data acquisition techniques that are less invasive and more accessible [5]. Given that learning difficulties, which encompass challenges with basic skills such as writing, affect approximately 10% of children globally, identifying deficits in motor skills is crucial [2].

2 Methodology

Our methodology is based on computer vision for extracting kinematic metrics. In this way, we seek to replicate the accuracy of standard devices for this analysis. The proposed experiment was developed using the MediaPipe Hands library [6], a solution that has proven to be highly accurate for hand and finger tracking. To simulate the behavior of a digital pen without the need for one, a detection logic based on the proximity between the distal phalanges (bones at the fingertips) was implemented. The system identifies the coordinates of the hand landmarks (exemplified in Figure 1), specifically the tip of the thumb and the tip of the index finger. The "writing" state is activated when the Euclidean distance between these two points is less than a pre-established threshold, simulating the pinching movement used when holding a pen. The same state is deac-

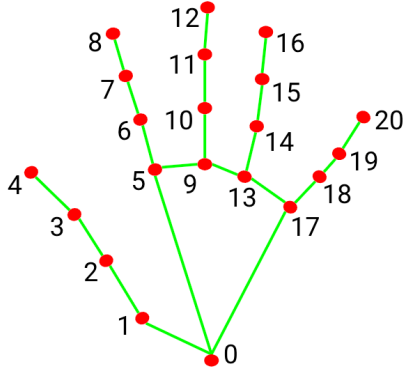


FIGURE 1 – Landmarks created by MediaPipe Hands. Image adapted from the MediaPipe Hands documentation [7].

tivated when it detects distance from the aforementioned phalanges.

The processing flow occurs frame by frame, where the captured images are converted from BGR color space to RGB for processing by the model. After inference, the midpoints between the thumb and index finger are calculated and stored in a list of tracks, allowing for the visual and mathematical reconstruction of the writing trajectory. Figure 2 illustrates what is mapped with the algorithm under development. The green line shows the path taken when writing, while the violet dots show the pauses that occurred. In the hand, the green dots illustrate the locations (fingertips, wrist, etc.) picked up by MediaPipe, and the red lines show the connections between these points, almost like the human skeleton. For image capture, a non-invasive configuration was chosen, using mobile device cameras and webcams positioned in a top-down view. The test environment was standardized with controlled lighting to ensure stability in detecting points of interest and to avoid distortions or vibrations in the tracking model. The extraction of kinematic features seeks to illustrate motor performance through quantitative metrics. Our implementation calculates and records the parameters either in real time or as a post-processing step.

The average writing speed is calculated by dividing the sum of the stroke sizes by the total duration of active writing. The time is calculated incrementally, adding the duration of each sample to the average, using Python’s own timing functions. A pause is defined as an interval where the spatial variation of the coordinates (dx and dy) is less than a given threshold (currently defined between 2 and 4 pixels) for a period greater than 0.5 seconds. The system identifies the stagnation of movement and triggers a time counter, which is interrupted as soon as writing resumes. Additionally, while this counter is interrupted, a second one is triggered to identify the moments in which the hand is moving in between the pauses. The length of each stroke is obtained by summing the Euclidean distances between



FIGURE 2 – Illustration of the algorithm in the testing phase performed by the paper’s first author. The fingers’ trajectory are illustrated in green and the positions in which a pause occurred by violet dots.

all consecutive points that form the digital stroke. A new stroke is counted whenever there is a discontinuity in the writing movement.

3 Preliminary results and discussion

The first experiment was conducted with two children with T21 at the Saber Down Research Center, *Universidade Estadual do Sudoeste da Bahia* (UESB), in Brazil. Both children are currently in the literacy acquisition stage and are receiving support from the research center, where they participate in cognitive, motor, and educational stimulation activities. Sample video frames of both participants are shown in Figure 3. We collected nine videos from Participant 1 and two videos from Participant 2. In each of those videos, the participants wrote the isolated letter A or short words such as *vaca*, *letra* and *saber*, which in Portuguese mean *cow*, *letter* and *to know*, respectively.

For Participant 1, the number of pauses ranged from 3 to 19, depending on the word being written, whereas for Participant 2 it ranged from 6 to 12 pauses. In most cases, the pauses corresponded to those that could be observed with the naked eye. The duration of the pauses was also consistent with what we observed through visual inspection : from 0.35 to 1.52 seconds for Participant 1, and from 0.63 to 0.91 seconds for Participant 2. Although these preliminary results do not yet allow us to identify global handwriting patterns for the participants, we observe that the pause durations are close to those reported in [8], which establishes an average pause duration of approximately 0.4 seconds for a group of eight 3rd- and 5th-grade students, and remain below the 2-second threshold reported in the literature.

Although it is in the early stages of development, the algorithm shows promising results in detecting indicators such as fine motor coordination, attention, inhibitory control,

visuospatial processing, and motor processing. The experiments successfully detected pauses during writing and mapped a high number of pauses in complex sections of graphemes. Visualizing the "skeleton" of the hand along with the generated stroke allowed for the visualization of the fluidity of the movement. The data collected suggest that it is possible to quantify fine motor coordination through variations in the linearity of the strokes and the consistency of writing speed.

Regarding the algorithm's shortcomings, we may highlight its inability to detect two hands at the same time, to work with varying camera angles, and to detect moments when the participant's hand is not touching the paper. In addition, camera stability proved to be a critical factor, since we observed that movements in the recording device can introduce noise into the kinematic data, requiring the equipment to remain static during the execution of the writing task. These shortcomings will be the target of improvements in the next stages of the algorithm development. We also aim to develop machine learning models more specialized in handwriting detection. Moreover, we observed that the accuracy of pause detection is sensitive to the distance between the hand and the camera. Tests indicated that the pixel variation considered as "rest" (pause) depends on the depth of the image, suggesting the need for rigorous standardization of the focal length.

Although the literature establishes digitizing tablets as the standard for kinematic analysis, the video approach has demonstrated feasibility for extracting the same classes of metrics (spatial and temporal). Unlike digitizing tablets, which capture pen pressure and absolute coordinates with high frequency, the proposed experiment offers the advantage of portability and the use of conventional writing instruments (pencil and paper), reducing cognitive load and being less invasive for the participant.

The metrics extracted via video, such as stroke speed and fluency, show promise for comparative studies. However, the inability to detect pressure, along with difficulties in precisely identifying when the pen loses contact with the paper, remain key limitations of the video capture method.

4 Conclusion

This study presented preliminary results from an experiment within the scope of a larger project aimed at developing an algorithm for the kinematic tracking of handwriting and the subsequent determination of motor patterns in individuals with Trisomy 21 (T21). The initial stage of the algorithm was validated by the experimental results presented here, proving the tool to be a low-cost and accessible alternative for collecting kinematic data.

Future work should focus on refining the detection of the beginning and end of the writing process (pen-paper contact), possibly through advanced machine learning techniques or depth calibration, alongside creating a more intuitive execution environment. Additionally, we plan to expand the database with a larger cohort of participants to

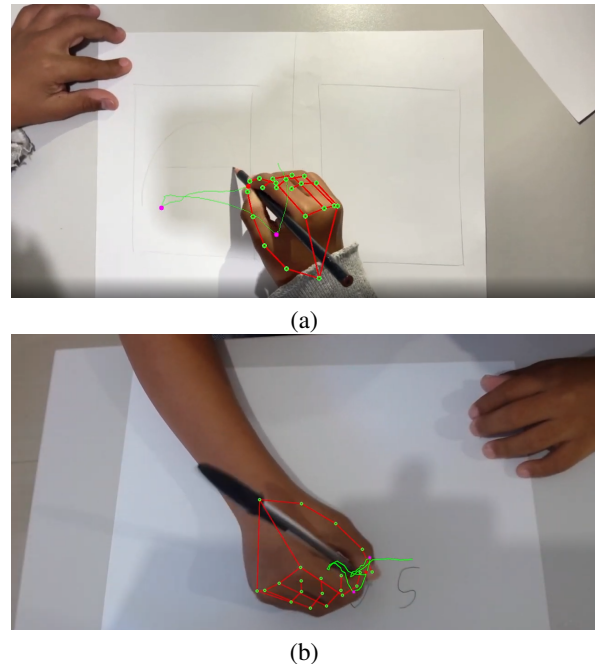


FIGURE 3 – Sample video frames of Participant 1 (a) and of Participant 2 (b). Data were acquired by the center Saber Down with approval from the Ethical Committee of UESB (protocol number 048533012.6.0000.0055). All rights reserved.

establish motor profiles with greater statistical robustness, while also addressing current limitations related to occlusion and the detection of multiple hands.

Acknowledgments

This work was partially supported by the CAPES-COFECUB program (project number : PE3335662P), funded by the French Ministry for Europe and Foreign Affairs, the French Ministry for Higher Education and CAPES - Brazil. Research conducted within the context of the International Associated Laboratory SAVOIR 21 funded by the Université de Lille.

References

- [1] Jérémy Danna, Frédéric Puyjarinet, and Caroline Jolly. Tools and methods for diagnosing developmental dysgraphia in the digital age : a state of the art. *Children*, 10(12) :1925, 2023.
- [2] Jayakanth Kunthoth, Somaya Al-Maadeed, Suchithra Kunthoth, Younes Akbari, and Moutaz Saleh. Automated systems for diagnosis of dysgraphia in children : a survey and novel framework : J. kunthoth et al. *International Journal on Document Analysis and Recognition (IJDAR)*, 27(4) :707–735, 2024.
- [3] Raphaela Tsao, Moy Eloise, Jean-Luc Velay, Nicolas Carvalho, and Carole Tardif. Handwriting in children

- and adults with down syndrome : Developmental delay or specific features? *American journal on intellectual and developmental disabilities*, 122(4) :342–353, 2017.
- [4] Marian Oliveira, Vera Pacheco, Luana Pereira-Souza, Alaine Leite Gama, Gilsara Madeira de Souza, and Lucrecia de Aquino Santos. Entrecruzamento fala - escrita e os desafios para a alfabetização do estudante com síndrome de down. In *Nuances da linguagem em uso : a síndrome de Down em foco*, pages 270–294. Campina Grande : EDUEPB, 2021.
 - [5] Oliver Schabos, Knut Hoffmann, Bjoern Enzi, Georg Juckel, and Paraskevi Mavrogiorgou. Kinematic analysis of handwriting movements in individuals with intellectual disabilities with and without obsessive compulsive symptoms. *Psychopathology*, 52(6) :346–357, 2020.
 - [6] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands : On-device real-time hand tracking. *arXiv preprint arXiv :2006.10214*, 2020.
 - [7] Media Pipe Hands. <https://mediapipe.readthedocs.io/en/latest/solutions/hands.html>.
 - [8] Mariona Pascual, Olga Soler, and Naymé Salas. In a split second : Handwriting pauses in typical and struggling writers. *Frontiers in psychology*, 13 :1052264, 2023.

Spatially-Grounded Gaussian-Prior Attention for Handwritten Mathematical Expression Recognition

Ibtissem Hadj Ali
Harold Mouchère

Laboratoire des Sciences du Numérique de Nantes (LS2N)
Nantes Université, École Centrale Nantes, CNRS
UMR 6004, F-44000 Nantes, France
`{ibtissem.hajali,harold.mouchere}@univ-nantes.fr`

Abstract

Handwritten Mathematical Expression Recognition (HMER) aims to convert handwritten mathematical formulas into structured markup languages such as $\text{\LaTeX}$. Recent encoder-decoder architectures based on Transformers have significantly improved recognition accuracy by modeling long-range dependencies through attention mechanisms. However, most existing models rely mainly on content-based attention and do not explicitly enforce spatial alignment between decoded symbols and image regions, which may lead to attention drift when recognizing complex mathematical layouts.

In this paper, we introduce a spatially grounded Transformer architecture integrating a *Latent Gaussian Prior (LGP)* into the cross-attention mechanism. The Gaussian prior acts as a probabilistic spatial anchor guiding the decoder toward coherent symbol regions. Additionally, we propose a *Geometry-Aware Encoder* that preserves spatial topology by modeling relative geometric relationships between feature grid cells before flattening the feature map. Experiments on CROHME benchmarks demonstrate that introducing geometric inductive bias improves recognition accuracy and stabilizes attention alignment.

Keywords: Handwritten Mathematical Expression Recognition, Transformer, Spatial Attention, Gaussian Prior, Geometry-Aware Encoding

1 Introduction

Handwritten Mathematical Expression Recognition (HMER) converts images of handwritten formulas into structured markup languages such as $\text{\LaTeX}$. Unlike classical OCR tasks, mathematical notation conveys meaning through spatial relationships including superscripts, subscripts, fractions and radicals.

Transformer architectures [1] have recently improved recognition performance. Models such as BTTR [6] and CoMER [7] achieve strong results on CROHME datasets. However, these models rely mainly on content similarity

and lack explicit spatial constraints, which may cause attention drift in dense expressions.

To address this issue, we introduce a **Latent Gaussian Prior (LGP)** that constrains cross-attention using a probabilistic spatial distribution predicted during decoding. The Gaussian prior acts as a spatial anchor guiding the attention mechanism toward coherent symbol regions.

2 Related Work

The CROHME competitions have played a central role in advancing handwritten mathematical expression recognition by providing standardized datasets and evaluation protocols [2]. The latest edition, CROHME 2023 [3], introduced new datasets and evaluation settings for both online and offline recognition tasks.

Early approaches relied on grammar-based parsing and symbol segmentation. Later methods introduced deep learning architectures combining convolutional encoders with attention-based decoders. Watch-Attend-Parse [4] and DenseWAP [5] demonstrated the effectiveness of CNN-RNN architectures for HMER.

More recently, Transformer-based approaches have become dominant. BTTR [6] introduced bidirectional Transformer training, while CoMER [7] proposed coverage modeling to improve attention alignment. Counting-Aware Network (CAN) [8] further improved decoding with auxiliary supervision.

3 Proposed Method

3.1 Overview

Given an input handwritten expression image X , the goal of HMER is to predict a sequence of LaTeX tokens

$$Y = \{y_1, y_2, \dots, y_T\}$$

Our architecture follows an encoder-decoder Transformer framework consisting of three components:

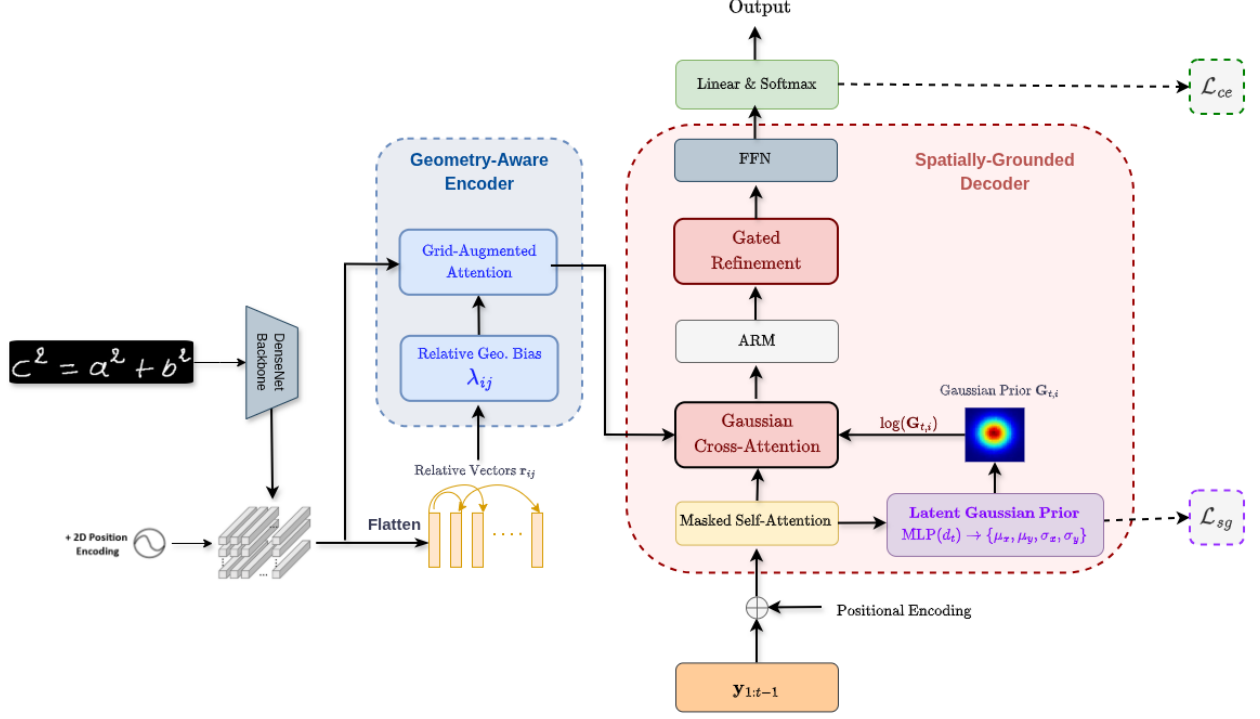


Figure 1: Overview of the proposed spatially grounded Transformer architecture for handwritten mathematical expression recognition.

- DenseNet backbone extracting visual feature maps
- Geometry-Aware Encoder modeling spatial relationships
- Transformer decoder with Gaussian-guided cross-attention

3.2 Geometry-Aware Encoder

The DenseNet backbone extracts a feature map

$$V \in \mathbb{R}^{h \times w \times d}.$$

To preserve spatial topology, we introduce a relative geometry representation

$$r_{ij} = [\log |x_i - x_j|, \log |y_i - y_j|]^T$$

which is injected into the attention energy

$$E_{ij} = \frac{Q_i K_j^T}{\sqrt{d_k}} + \lambda_{ij}.$$

This mechanism allows the encoder to capture spatial relationships between visual tokens.

3.3 Gaussian-Guided Cross Attention

During decoding, the hidden state predicts Gaussian parameters

$$(\mu_x, \mu_y, \sigma_x, \sigma_y) = MLP(d_t)$$

The spatial distribution is defined as

$$G_{t,i} = \exp \left(-\frac{(x_i - \mu_x)^2}{2\sigma_x^2} - \frac{(y_i - \mu_y)^2}{2\sigma_y^2} \right)$$

The Gaussian prior is injected into cross-attention

$$E_{t,i} = \frac{Q_t K_i^T}{\sqrt{d_k}} + \log(G_{t,i})$$

combining content similarity and spatial likelihood.

4 Experiments

4.1 Datasets

We evaluate our method on the CROHME benchmarks. The CROHME 2016 dataset contains 8,836 training expressions and 1,147 test samples. CROHME 2019 provides 1,199 test expressions and has become a standard benchmark for recent Transformer-based models. The latest CROHME 2023 dataset introduces additional handwritten equations and improved annotations [3].

Performance is evaluated using the **Expression Recognition Rate (ExpRate)**, which measures the percentage of expressions recognized without any symbol error.

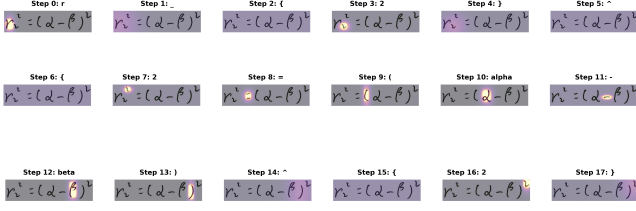


Figure 2: Visualization of cross-attention maps during decoding.

4.2 Implementation Details

Our model is implemented in PyTorch. The encoder uses a DenseNet backbone with three dense blocks. The Transformer decoder contains three layers with eight attention heads and a hidden dimension of $d_{model} = 256$. The Gaussian parameters are predicted using a two-layer MLP.

Training uses stochastic gradient descent with momentum 0.9 and weight decay 10^{-4} . Data augmentation includes random scaling of input images.

4.3 Comparison with State-of-the-Art

Table 1: Comparison with state-of-the-art methods (ExpRate %).

Method	2016	2019
DenseWAP [5]	47.50	-
CAN [8]	56.15	55.96
CoMER [7]	59.81	62.97
Ours	62.87	63.56

The results show that incorporating spatial grounding improves recognition accuracy compared with previous Transformer-based models.

4.4 Ablation Study

Table 2: Ablation study on CROHME 2023.

Model	ExpRate
Baseline (CoMER)	60.12
+ Geometry Encoder	61.32
+ Gaussian Prior	61.89
Full Model	64.69

The ablation results confirm that both the geometry-aware encoder and the Gaussian prior contribute to improved performance.

4.5 Qualitative Analysis

Figure 2 shows attention maps produced during decoding. The Gaussian prior generates localized attention patterns aligned with handwritten symbols.

5 Conclusion

We proposed a spatially grounded Transformer architecture for handwritten mathematical expression recognition. By integrating a Latent Gaussian Prior and a Geometry-Aware Encoder, the model improves recognition accuracy and stabilizes attention alignment.

References

- [1] Vaswani, A. et al.: Attention Is All You Need. NeurIPS (2017)
- [2] Mouchère, H. et al.: CROHME competitions. IJDAR (2016)
- [3] Xie, Y., Mouchère, H. et al.: ICDAR 2023 CROHME competition. ICDAR (2023)
- [4] Zhang, J. et al.: Watch, Attend and Parse. Pattern Recognition (2017)
- [5] Zhang, J. et al.: Multi-scale attention with dense encoder. ICPR (2018)
- [6] Zhao, W. et al.: Bidirectionally trained transformer for HMER. ICDAR (2021)
- [7] Zhao, W., Gao, L.: CoMER. ECCV (2022)
- [8] Li, B. et al.: Counting-aware network. ECCV (2022)
- [9] Xie, Y., Zanibbi, R., Mouchère, H.: Graph-based modeling for HMER. arXiv (2024)
- [10] Truong, T. N. et al.: Survey on HMER. Pattern Recognition (2024)

Sensibilité des représentations latentes aux distorsions : implications pour l'évaluation de la qualité d'image

Abderrezzaq Sendjasni et Mohamed-Chaker Larabi
CNRS, Univ. Poitiers, XLIM, UMR 7252, France

{abderrezzaq.sendjasni, chaker.larabi}@univ-poitiers.fr

Résumé

Des représentations visuelles capables de capturer les distorsions sont essentielles pour l'évaluation perceptuelle de la qualité d'image (IQA). Dans cet article, nous menons une analyse complète d'encodeurs visuels de pointe, y compris CLIP, DINO, ConvNeXt, EfficientNet et ResNet, soumis à des distorsions courantes telles que le flou gaussien, le flou de mouvement, les artefacts de compression et le bruit gaussien. Nous analysons la divergence des caractéristiques latentes ainsi qu'une analyse statistique basée sur l'ANOVA. Nos résultats révèlent comment les choix de l'architecture, les méthodes d'apprentissage et la diversité des données façonnent la capacité d'un modèle à encoder les distorsions dans son espace latent. Ces résultats établissent un lien entre l'apprentissage de représentations et la modélisation de la qualité perceptuelle, offrant de nouvelles perspectives pour le développement d'encodeurs résilients aux distorsions pour l'IQA.

Mots clefs

Robustesse de l'espace latent, Évaluation de la qualité d'image, analyse d'image.

1 Introduction

L'avènement des réseaux de neurones profonds et des transformers [1, 2] a considérablement fait progresser la vision par ordinateur. Malgré leurs performances remarquables, ces modèles peuvent se comporter de manière imprévisible en présence de distorsions réelles (bruit, flou, compression). Parallèlement, la communauté d'évaluation de la qualité visuelle (IQA) étudie comment ces distorsions affectent la perception humaine, mais peu d'études analysent comment les représentations internes des encodeurs visuels réagissent à ces distorsions.

Les encodeurs visuels sont de plus en plus utilisés en IQA ; cependant, leur robustesse aux distorsions, souhaitable pour la détection d'objets, la segmentation et la classification, peut nuire à l'évaluation de la qualité : un modèle trop robuste peut négliger des dégradations subtiles et ne pas refléter la baisse de qualité perçue. Cela soulève une question fondamentale : **dans quelle mesure un encodeur doit-il être invariant aux distorsions pour l'IQA, et quand doit-il être sensible ?**

Pour y répondre, nous évaluons des encodeurs variés

(ResNet-50 [3], EfficientNet [4], ConvNeXt [5], DINO [6], CLIP [7]) sous quatre distorsions (bruit, flou gaussien, flou de mouvement, compression JPEG). Nous analysons l'évolution de l'espace latent via les distances euclidienne et de Hellinger entre embeddings pristins et dégradés, et leur corrélation avec l'intensité des distorsions. Voici les contributions : (i) cadre systématique pour sonder la sensibilité des encodeurs, (ii) benchmark empirique de robustesse latente, (iii) discussion sur les implications pour la conception de pipelines IQA conscients des distorsions.

2 Étude Proposée

Nous analysons le comportement de l'espace latent d'encodeurs visuels modernes (CNN, ViT, multimodaux) face à des distorsions courantes. Des métriques de distance sont utilisées pour quantifier l'altération des représentations internes afin de révéler les compromis entre robustesse et sensibilité (Fig. 1).

L'étude utilise l'ensemble Raise-1k [8] (images pristines haute résolution). Des patches de 224×224 pixels sont extraits pour garantir la compatibilité avec les encodeurs. Seuls les patches visuellement informatifs sont conservés, via l'indice d'information spatiale (SI) [9] :

$$SI(p_i) = \max_{k \in \{x, y\}} \{\sigma(G_k(p_i))\}, \quad (1)$$

où G_x, G_y sont les gradients de Sobel et σ l'écart-type. Les patches sous le 25e percentile SI sont écartés pour éviter les zones pauvres en information.

Les patches sélectionnés sont ensuite soumis à des dégradations contrôlées (voir Sec. 2.1). Pour chaque patch original p_i et sa version dégradée $\tilde{p}_{i,j}$, un encodeur m produit des embeddings $e_{i,0}^m$ et $e_{i,j}^m$. La comparaison de ces paires par des métriques de distance révèle la perturbation de l'espace latent, mesurant ainsi la robustesse ou la sensibilité.

2.1 Encodeurs Visuels et Types de Distorsion

Types de Distorsion Sélectionnés. Nous appliquons quatre distorsions courantes [10] directement sur les patches pour analyser leur impact localisé :

- **Flou Gaussien** : Convolution avec un noyau gaussien 2D. Écart type $\sigma_b \in \{0.8, 1.4, 2.0, 2.6, 3.2, 3.8, 4.4, 5.0, 5.6, 6.2, 6.8, 7.4\}$.
- **Flou de Mouvement** : Convolution avec un noyau linéaire. Longueurs $L \in$

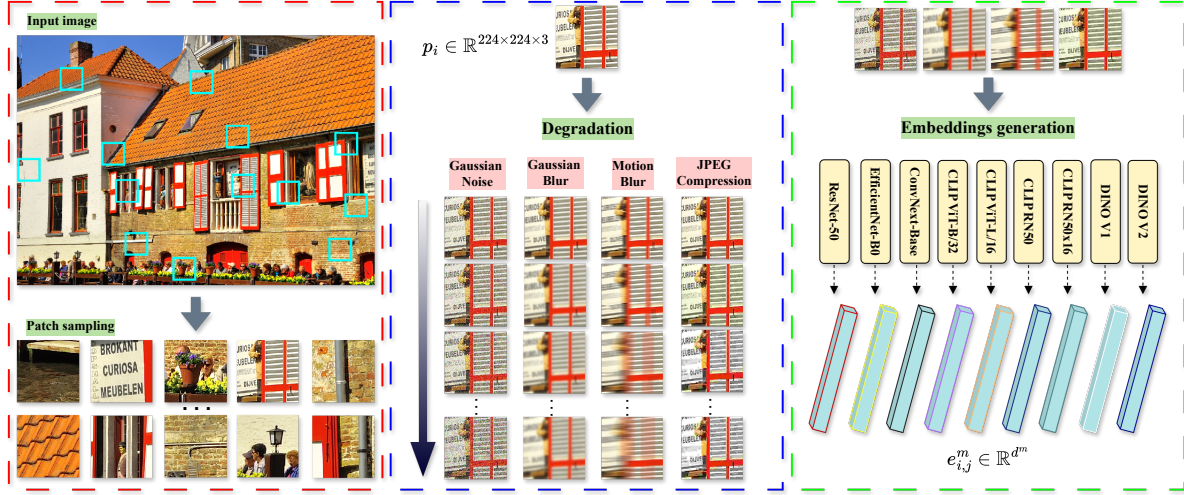


FIGURE 1 – Cadre d’analyse : des patches extraits d’images pristes sont dégradés (4 types de distorsion), puis traités par divers encodeurs. Les embeddings $e_{i,j}^m \in \mathbb{R}^{d^m}$ (patch i , distorsion j , modèle m) servent à quantifier les changements latents.

{5, 8, 11, 15, 20, 25, 30, 35, 40, 45, 50, 55} pixels à angle fixe.

- **Compression JPEG** : Facteurs de qualité $Q \in \{80, 60, 40, 30, 20, 15, 10, 7, 5, 3, 2, 1\}$.
- **Bruit Gaussien** : Ajout de bruit gaussien avec écarts types $\sigma \in \{0.001, 0.002, 0.004, 0.008, 0.012, 0.02, 0.03, 0.04, 0.06, 0.08, 0.12, 0.15\}$.

Encodeurs Visuels Sélectionnés. Nous évaluons des encodeurs pré-entraînés couvrant les architectures clés, en extrayant les vecteurs de leur avant-dernière couche. ResNet-50 [3] sert de baseline CNN. EfficientNet-B0 [4] est optimisé pour l’efficacité. ConvNeXt-Base [5] modernise les CNN avec des principes inspirés des ViT. DINO (V1 et V2) [6] est une méthode auto-supervisée produisant des caractéristiques sémantiques riches. CLIP [7] (ViT-B/32, ViT-L/14, ViT-RN50, ViT-RN50x16) est un modèle multimodal entraîné sur de larges paires image-texte, offrant robustesse et généralisation.

3 Résultats et Discussion

3.1 Configuration Expérimentale

Ensemble de données : Dans cette étude, nous utilisons l’ensemble de données RAISE-1k [8]. Il se compose de 1 000 images RAW haute résolution et non compressées capturées avec des appareils photo reflex numériques.

Métriques d’évaluation : Nous quantifions les changements induits par la distorsion dans l’espace latent à l’aide de deux métriques. La distance euclidienne mesure le décalage géométrique entre les embeddings pristes et dégradés, tandis que la distance de Hellinger (sur des vecteurs normalisés par softmax) évalue leur divergence probabiliste. De plus, nous utilisons la taille d’effet Eta-carré (η^2) [11] issue d’une ANOVA pour déterminer la proportion de

variance des caractéristiques attribuable à chaque distorsion, indiquant ainsi la sensibilité globale du modèle.

3.2 Évaluation de la Divergence

Pour évaluer la sensibilité des encodeurs visuels, nous avons analysé la divergence entre les embeddings d’images pristes et dégradées pour les distorsions considérées. Deux métriques de divergence sont utilisées, à l’aide des distances euclidienne et de Hellinger, pour capturer à la fois les changements géométriques et probabilistes dans l’espace d’embedding. Une divergence plus élevée indique une plus grande sensibilité à la distorsion appliquée, ce qui peut refléter une faible robustesse à la dégradation et une meilleure capacité à capturer les nuances des distorsions introduites. Les courbes de divergence correspondantes sont présentées dans la Fig. 2.

Pour tous les types et niveaux de sévérité de distorsion, les modèles DINO (v1 et v2) présentent la plus grande sensibilité en termes de distance euclidienne, avec une divergence de caractéristiques significativement plus élevée que les autres encodeurs visuels. Ils sont suivis par ResNet-50, qui montre également des augmentations notables du décalage des caractéristiques sous distorsion. En revanche, des modèles tels que EfficientNet-B0, ConvNeXt-Base et CLIP ViT-L/14 démontrent des embeddings plus stables, maintenant une divergence euclidienne relativement faible et constante à travers les distorsions. En analysant la distance de Hellinger, une tendance similaire est observée : DINO (V1 et V2) montrent à nouveau la divergence la plus élevée, même sous des niveaux de distorsion légers, suggérant une sensibilité accrue aux perturbations. ResNet-50, EfficientNet-B0 et CLIP ViT-L/14 suivent avec une divergence modérée, tandis que les modèles restants présentent une distance de Hellinger minimale et presque constante à travers les types et intensités de distorsion. Ces résultats indiquent que, tandis que les représentations basées

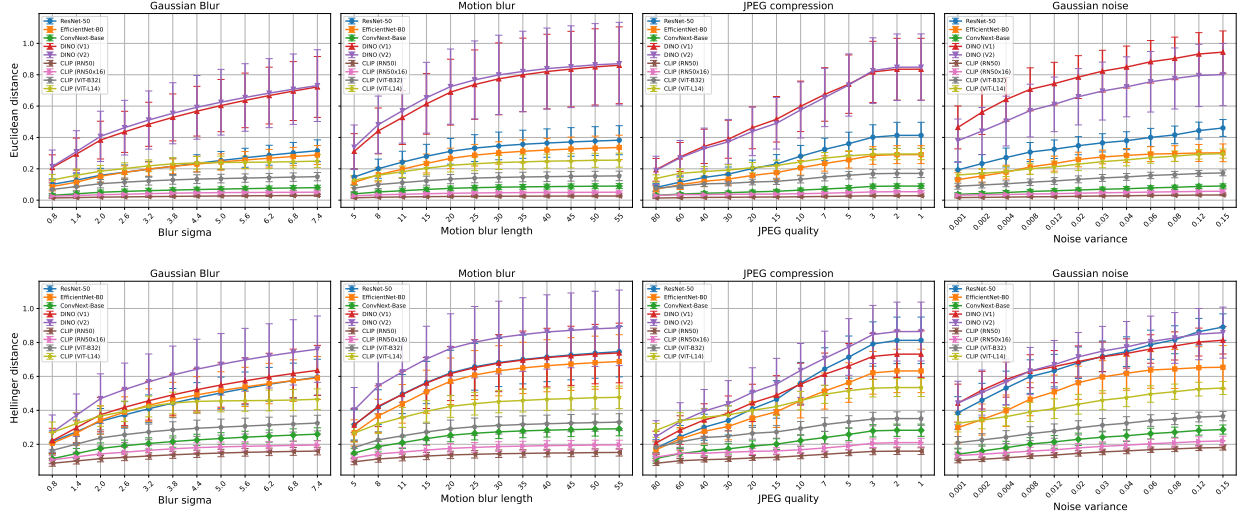


FIGURE 2 – Analyse de la divergence (haut) distances euclidienne et (bas) de Hellinger. Chaque sous-graphique illustre la divergence par rapport aux embeddings originaux en fonction des niveaux de distorsion, avec des barres d'erreur représentant l'écart type.

sur DINO sont très sensibles aux dégradations de la qualité d'image, d'autres architectures maintiennent des espaces de caractéristiques plus invariants sous distorsion. Ces tendances de divergence présentent des similitudes notables avec les principes de l'IQA perceptuelle. Particulièrement dans les métriques avec référence, les caractéristiques en IQA qui montrent une sensibilité élevée aux distorsions de bas niveau correspondent fréquemment aux perceptions humaines de la dégradation de la qualité. Il est suggéré que les embeddings DINO encodent des indices visuels riches et fins, similaires à ceux utilisés dans l'estimation de la qualité perceptuelle, comme l'indique la forte divergence observée dans leurs représentations, en particulier sous des distorsions légères. Cependant, pour des tâches qui nécessitent une invariance aux distorsions, cette sensibilité élevée pourrait également indiquer un manque de robustesse. En revanche, les modèles qui maintiennent une faible divergence à travers les niveaux de distorsion, comme CLIP ViT-L/14 et EfficientNet-B0, peuvent capturer des représentations plus invariantes aux distorsions, répondant mieux aux exigences de l'IQA sans référence, qui privilégie la stabilité et la robustesse face à la variabilité naturelle des images. Ainsi, ce comportement de divergence fournit des informations sur la façon dont différents encodeurs pourraient être plus appropriés pour divers paradigmes d'IQA en reflétant un compromis entre la stabilité représentationnelle et la sensibilité perceptuelle.

3.3 Analyse Statistique

Pour mieux comprendre la sensibilité des encodeurs visuels à diverses distorsions, nous avons mené une analyse ANOVA basée sur la différence absolue moyenne (MAD) calculée par dimension de caractéristique dans l'espace latent. Pour chaque encodeur, la taille d'effet (η^2) est calculée.

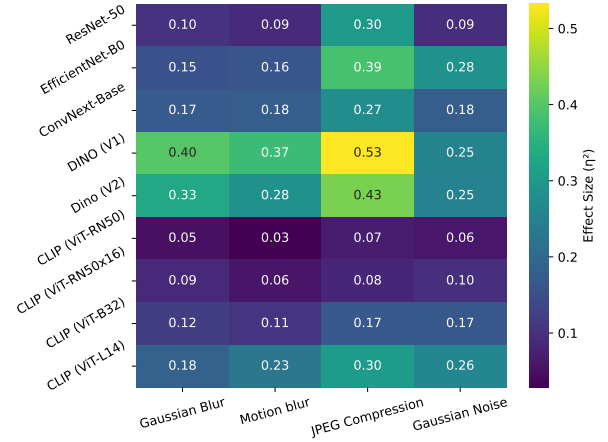


FIGURE 3 – Carte de chaleur de la taille d'effet (η^2) montrant l'impact de différents types de distorsion sur divers modèles.

lée pour évaluer comment et avec quelle force chaque distorsion perturbe la représentation interne du modèle. Cette analyse vise à quantifier la discriminabilité des distorsions dans les espaces latents et à mieux comprendre leurs implications pour des tâches telles que l'évaluation de la qualité. Une carte de chaleur représentant la taille d'effet des modèles sélectionnés par rapport à chaque distorsion est fournie dans la Fig. 3

Globalement, tous les modèles ont montré une sensibilité statistiquement significative aux distorsions, cependant l'ampleur de l'effet (telle que capturée par η^2) a varié considérablement. Les architectures comme DINO (V1) et DINO (V2) ont démontré des valeurs η^2 élevées pour toutes les distorsions, indiquant un fort déplacement interne des caractéristiques. En revanche, les variantes CLIP

basées sur ResNet ont systématiquement montré des tailles d'effet notablement plus faibles, contrastant fortement avec les modèles CLIP basés sur ViT ou les encodeurs DINO, qui avaient des valeurs η^2 supérieures à 0.16, certains modèles atteignant des valeurs aussi élevées que 0.53, comme observé pour DINO (V1) sur JPEG.

Cette divergence peut être attribuée à deux facteurs. Premièrement, la nature architecturale des CNN comme RN50 leur confère naturellement une certaine résilience aux perturbations locales et haute fréquence, en raison de leur utilisation de filtres convolutifs localisés et d'opérations de pooling. Deuxièmement, ces variantes CLIP ont été entraînées en utilisant un objectif d'apprentissage contrastif sur des ensembles de données d'images-texte bruyants à l'échelle d'Internet, tels que YFCC100M [12]. Ces données de pré-entraînement contiennent inévitablement une proportion élevée d'images déjà compressées ou dégradées, guidant implicitement le modèle à se concentrer sur des caractéristiques sémantiques robustes et de haut niveau, invariantes au bruit ou aux artefacts de bas niveau. Cela contraste avec des modèles comme les DINO, qui ont été entraînés avec des objectifs auto-supervisés mettant l'accent sur des motifs fins. De plus, les valeurs élevées de η^2 obtenues pour DINO soutiennent l'observation sur la divergence des embeddings et confirme que l'ampleur de la divergence est significative.

Ces résultats indiquent que la robustesse aux distorsions dans l'espace latent n'est pas seulement une fonction de l'architecture mais aussi de la stratégie d'entraînement et de la diversité des données. Le fait que les variantes CLIP basées sur ViT ne présentent pas la même résilience JPEG que les variantes RN50 soutient davantage ceci, car les transformers manquent de biais inductifs comme la localité, les rendant plus sensibles aux perturbations de la structure spatiale causées par JPEG.

Du point de vue de l'IQA, cette analyse établit un parallèle clair : les modèles qui sont entraînés sur des données sémantiquement riches et diversifiées en termes de distorsions, et encouragés à se concentrer sur des invariants de haut niveau, tendent à montrer une plus grande robustesse aux artefacts perceptuels, un trait primordial pour une prédiction fiable de la qualité consciente des distorsions. Ainsi, comprendre les profils de robustesse latente, comme nous le faisons ici avec la MAD par dimension d'embedding et l'ANOVA, fournit une perspective précieuse sur l'adéquation des modèles pour l'IQA et les tâches perceptuelles plus largement.

4 Conclusion

Nous avons introduit un cadre d'analyse fin, basé sur une analyse dimensionnelle, pour évaluer la robustesse des encodeurs visuels face aux distorsions courantes, révélant des différences clés entre les modèles dans la manière dont les perturbations se manifestent dans leurs espaces latents. Nos analyses de divergence et de la taille d'effet basée sur l'ANOVA ont systématiquement classé les modèles basés

sur les transformers, en particulier DINO (V1 et V2) et CLIP (ViT-L/14), comme les plus sensibles, tandis que les modèles basés sur la convolution ont présenté des représentations plus stables et plus robustes aux distorsions. Il est important de noter que nos résultats font écho aux principes de l'IQA, où la pertinence perceptuelle des distorsions varie à la fois avec le contenu et le type de distorsion. Ce travail offre une nouvelle voie pour étudier la robustesse des caractéristiques et établit un pont entre l'apprentissage de représentations et la modélisation de la qualité perceptuelle, ouvrant la voie au développement de systèmes de vision sensibles aux distorsions.

Références

- [1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al.. An image is worth 16x16 words : Transformers for image recognition at scale. *ICLR*, 2021.
- [2] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, et H. Jegou. Training data-efficient image transformers & distillation through attention. Dans *38th International Conference on Machine Learning*, volume 139, pages 10347–10357. PMLR, 2021.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et Jian Sun. Deep residual learning for image recognition. Dans *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.
- [4] PMLR. *EfficientNet : Rethinking model scaling for convolutional neural networks*, 2019.
- [5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, et Saining Xie. A convnet for the 2020s. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, et Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv :2104.14294*, 2021.
- [7] Alec Radford, Jong Wook Kim, Christopher Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al.. Learning transferable visual models from natural language supervision. Dans *International Conference on Machine Learning (ICML)*. PMLR, 2021.
- [8] D. Dang-Nguyen, C. Pasquini, V. Conotter, et G. Boato. Raise : A raw images dataset for digital image forensics. Dans *ACM International Workshop on Multimedia Security and Content Protection*, pages 219–224, 2015.
- [9] ITUITUR Rec. P. 910 subjective video quality assessment method for multimedia applications. *ITU : Geneva, Switzerland*, 2018.
- [10] Zhou Wang et Alan Conrad Bovik. *Modern image quality assessment*. Thèse de doctorat, Springer, 2006.
- [11] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd édition, 1988.
- [12] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, et Li-Jia Li. Yfcc100m : The new data in multimedia research. *Communications of the ACM*, 59(2) :64–73, 2016.

Synchronicity Between Gestures and Speech in People with Trisomy 21 : Preliminary Results and Perspectives

Laura Chaves Miranda¹, Gia-Thuy Pham², Deise Santana Maia¹, Marian Oliveira³, Vera Pacheco³

¹Université de Lille, CNRS, Centrale Lille, UMR 9189 CRISAL, F-59000 Lille, France

² CENOVA, Marcq-en-Barœul, France

³Universidade Estadual do Sudoeste da Bahia, Centro Saber Down, PPGLin, Vitória da Conquista, Brazil
laura.chaves-miranda.etu@univ-lille.fr, phamgiathuy@gmail.com,
deise.santanamaia@univ-lille.fr, marian.oliveira@uesb.edu.br, vera.pacheco@uesb.edu.br

Abstract

Given the broad consensus in the literature that speech and gestures are closely related in humans, we aim to investigate this relationship in people with Trisomy 21 (T21), one of the most common chromosomal alterations worldwide, which is associated to delays in language and motor development. Using a deep-learning based model for body landmark detection, we analyzed two videos of young women with T21 and observed a notable level of synchronization between movements of different body landmarks, as well between moments of accelerated body movements and higher vocal pitch. Further analysis with a larger and more diverse data is required to validate these preliminary observations.

Keywords

Gestures, Speech, Down Syndrome, Trisomy 21

1 Introduction

The alignment between human gestures and speech has been a subject of study for decades among researchers in a variety of fields, including linguistics, cognitive and neuroscience [1]. The general hypothesis is that humans coordinate their gestures in accordance with the words and sentences being pronounced, as well as with their intonation and voice volume. Though this hypothesis has been globally accepted among scientists for a long time, one of the first works to objectively quantify the synchronicity between gestures and speech was published by Pouw and Dixon [2]. In [2], the authors measure the synchronization between the moment of maximum effort in a gesture (*i.e.*, the peak velocity and acceleration of hand gestures) with the moment of peak voice pitch (F0) by computing the time gap between those two moments. In their experiments with four male right-handed adults, the authors observed that such synchronicity is indeed present, with the moments of peak acceleration of their right hand closely leading or following the moments of peak pitch. Given that such studies are commonly performed on participants

with typical development, in this work, we aim to extend the analysis of the synchronicity of gestures and speech to people with Trisomy 21 (T21), also known as Down syndrome. T21 is among the most common chromosomal alterations in the world, affecting around one in every one thousand births. This chromosomal alteration is the cause of a series of specificities in people with T21, including motor, cognitive and language delay, and modifications in important structures for speech production, such as small oral cavity, protruding tongue, teeth eruption disorders, and hearing impairment. Studies on the speech intonation and voice pitch of people with T21 [3] and on their gestures and words acquisition [4] are present in the literature, but their synchronization hasn't been tackled so far, to the best of our knowledge. Considering the aforementioned specificities commonly present among people with T21, our long term goal is to provide answers to the following questions :

- Are the speech and gesture synchronicity patterns observed among people with typical development also observable among the T21 population ?
- If not, is there a specific pattern in the alignment of gestures and speech among people with T21 (*e.g.*, a consistently larger or smaller gap between the peak pitch and acceleration moments) ?

As a first step toward answering those questions, we provide in this paper preliminary results on the quantification of speech and gesture synchronicity from the videos of two people with T21. We first revise the gesture-speech quantification metric proposed in [2], followed by an adaptation of the proposed metric to the case where gestures are not annotated. We then describe our data, our preliminary observations and the perspectives of the present research.

This study is part of a larger project whose aim is to better understand the difficulties and capabilities of people with T21, through the precise quantification of their body movements using computer vision, as well as through speech and auditory perception analysis. As a long-term goal, we hope that such understanding will lead to the improvement of speech and psychomotricity therapy aimed specifically at the T21 population.

2 Gesture-Speech Synchronicity

In [2], the authors tackle the problem of quantifying the alignment between the moments of peak pitch during the speech with the moments of maximum effort of a gesture, which correspond to the moments of highest acceleration or velocity of the body movements. More precisely, their experiments consisted in recording the voice and the right-hand movements of the participants while describing a cartoon they had just watched. The temporal data were then segmented into different types of gestures (beat, iconic and pointing), and, from each gesture segment, the moments of highest pitch and of highest hand velocity and acceleration were extracted. The final metric consisted in computing the difference in timing (D) between those two moments. The authors reported the distribution of the values of D , which showed a clear tendency toward the moments of high pitch happening just before or just after rapid and accelerated hand movements, indicating an alignment between those moments.

3 Methodology

Inspired by the metric D introduced in [2], we make the following adaptations in order to consider all the upper body parts movement on non-annotated video data : though the most noticeable body movements during speech might be performed by the hands, we consider that the head and the shoulders also play a role in the gesture and speech synchronization. Unlike [2], we consider not only the movements of the participant’s dominant hand, but also their non-dominant hand, as well as their wrists, elbows, shoulders and nose. Figure 1 shows four frames from both participants’ videos, in which the detected landmarks are shown in green.

Instead of analyzing segmented gestures individually, as done in [2], we search for the moments of highest maximum effort, which are computed as the peak acceleration moments among the ones higher than the third quartile. In other words, we consider only 25% of the highest maximum effort moments. Similarly, we only take into account the 25% highest pitch moments. Then, for every peak acceleration moment, we compute D as the difference between its timestamp and the closest peak pitch.

In [2], the kinematic data extracted from the participant’s dominant hand were obtained through a glove equipped with a high temporal resolution motion sensor (240 Hz), while that the kinematic data used in the present paper were estimated from videos of 30Hz using the deep-learning based MediaPipe library [5]. Though estimating body movements from videos using deep-learning models is less precise than using specialized motion sensors, the former method has been validated in [6] as a sound approach for quantifying gesture-speech synchrony, with the advantage of being an affordable method when compared to expensive motion-tracking systems.

All data analysis was performed in Python using the MediaPipe [5] and Parselmouth [7] libraries for landmark and

pitch extraction, respectively. The Parselmouth library is a Python interface for the Praat open-source software [8], widely used for speech analysis by researchers in linguistics. The scripts for computing our proposed gesture-speech analysis are available in our GitHub repository.

4 Data

Preliminary experiments were performed on videos of two young women with T21 who have been accompanied for several years by the *Saber Down* research center, at the *Universidade Estadual do Sudoeste da Bahia* (UESB), Brazil, which provide an environment for linguistic and cognitive stimulation, as well as reading and writing practice of Portuguese and foreign languages. In both videos, the participants 1 and 2 recorded themselves with the aim to invite their peers to participate in a seminar organized by *Saber Down*. In the video of Participant 1, of approximately 14 seconds, the participant’s face, arms and hands are visible along the whole duration of the video. In the video of Participant 2, of approximately 19 seconds, the hands and elbows are not always visible and, hence, only the head and shoulder movements are considered in our analysis.

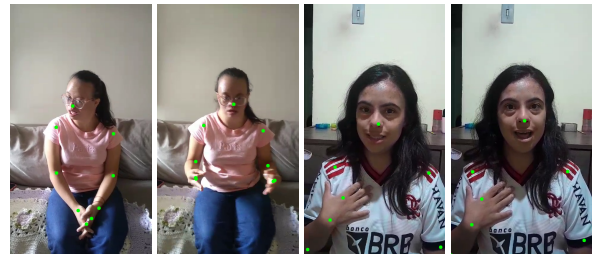


FIGURE 1 – From left to right : two frames from the video of Participant 1 and Participant 2, respectively, with the detected landmarks shown as green dots. Data were acquired by the center *Saber Down* with approval from the Ethical Committee of UESB (protocol number CAAE04853012.6.0000.0055). All rights reserved.

5 Experimental setup

The following parameters have been set up in the extraction of peak pitch and acceleration from our data :

Audio sampling rate : 1 sample per 33.3 milliseconds in order to match the video temporal resolution of 30 frames per second.

Speech frequency interval : given that both participants are female, we chose a F0 frequency range of 100-500 Hz.

Peak pitch, acceleration and velocity : as mentioned previously, we considered only 25% of the maximum effort moments (*i.e.* moments of highest velocity and acceleration), as well as the 25% highest F0 values. To avoid having several velocity, acceleration and pitch peaks concentrated around a same gesture, we consider that those peaks should be spaced by at least 10 sample points (333.3 milliseconds).

6 Preliminary results

As already mentioned previously, we take into consideration all the upper body landmarks that are visible in each of the videos in order to compute the “total quantity of movement” at any given timestamp. To ensure the robustness of this technique, we first had to answer to the following question : are the movements of the head, shoulders and arms synchronized among themselves ? In the positive case, this indicates that moments of accelerated right-hand movements, for instance, generally will coincide with moments of high acceleration in other body parts.

To illustrate this, Figure 2 present the normalized accelerations curves of the left and right wrists of Participant 1 along the video duration. By analyzing those acceleration curves, we can observe the symmetry of the left and right sides of those body parts, and that the moments of peak acceleration happen approximately at the same timestamps. This observation was then confirmed through the Pearson correlation coefficients presented in Table 1 and 2, which show that the acceleration of all landmarks are highly correlated for both participants, and, therefore, the average acceleration of all landmarks is also highly correlated to each of the individual landmarks.

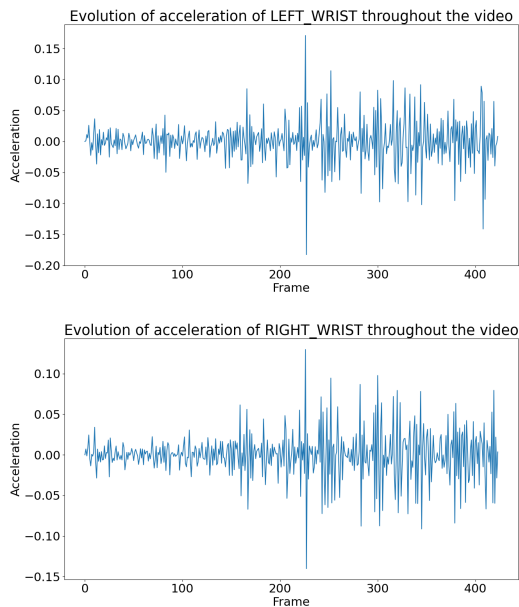


FIGURE 2 – Normalized acceleration of the right and left wrists of Participant 1.

Given the high correlation between each landmark’s acceleration with the average acceleration in both videos, we consider only the average accelerations and velocities in our speech-gesture synchrony analysis. A more in depth analysis using each of the extracted landmarks individually will be left for an extended version of the present paper. Figure 3 shows the histograms and the estimated density

functions for the metric D computed from the peak accelerations and velocities, and from the peak pitches. For both participants, we can draw a similar conclusion : the moments of “peak pitch” are either shortly preceded or shortly followed by the moments of “peak acceleration” and “peak velocity”, which demonstrates a certain synchronicity between the participants’ speech and gestures.

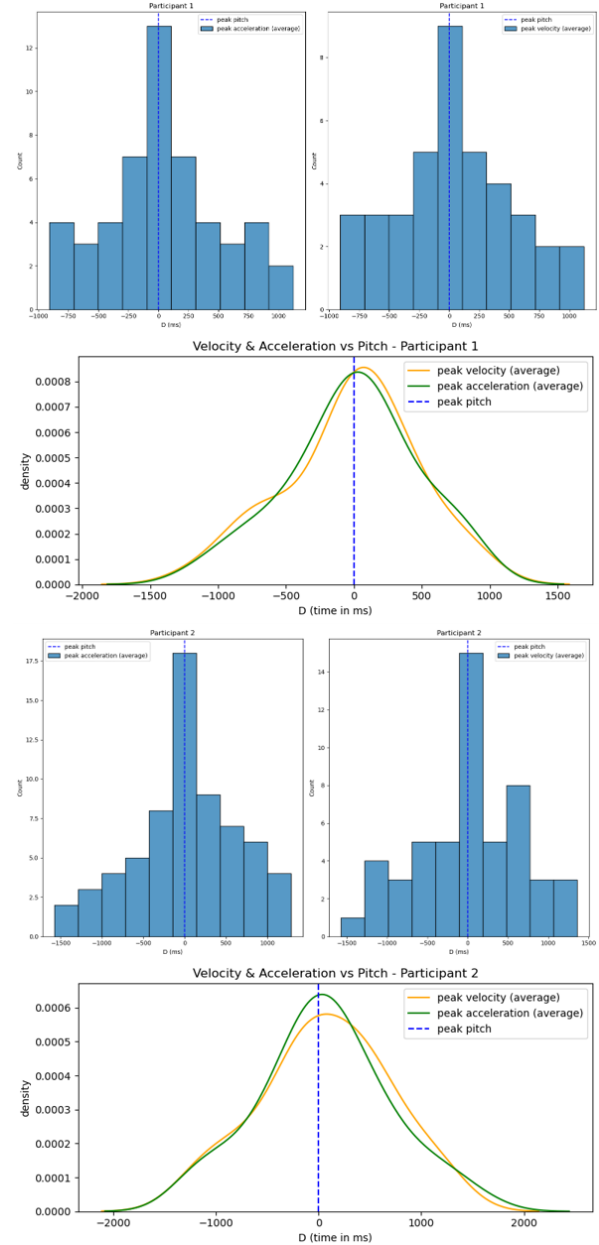


FIGURE 3 – First and third lines : histograms of timestamp differences (D) between the peak accelerations and velocities and peak pitches of participants 1 and 2, respectively. Second and fourth lines : estimated density functions based on the aforementioned histograms for participants 1 and 2, respectively.

	N	LS	RS	LE	RE	LW	RW	LI	RI	AVG
N	1.0	0.93	0.94	0.84	0.87	0.69	0.72	0.55	0.68	0.87
LS	-	1.0	0.91	0.89	0.83	0.70	0.68	0.56	0.63	0.86
RS	-	-	1.00	0.82	0.90	0.70	0.71	0.55	0.66	0.87
LE	-	-	-	1.00	0.88	0.86	0.82	0.69	0.77	0.94
RE	-	-	-	-	1.00	0.77	0.87	0.61	0.81	0.93
LW	-	-	-	-	-	1.00	0.84	0.87	0.81	0.92
RW	-	-	-	-	-	-	1.00	0.69	0.98	0.92
LI	-	-	-	-	-	-	-	1.00	0.68	0.80
RI	-	-	-	-	-	-	-	-	1.00	0.89
AVG	-	-	-	-	-	-	-	-	-	1.00

TABLE 1 – Pearson correlation coefficients between the acceleration of the landmarks of Participant 1 : nose (N), right (RS) and left (LS) shoulders, right (RE) and left (LE) elbows, right (RW) and left (LW) wrists, and right (RI) and left (LI) index fingers, and average acceleration (AVG).

	N	LS	RS	AVG
N	1.00	0.82	0.91	0.96
LS	-	1.00	0.85	0.94
RS	-	-	1.00	0.96
AVG	-	-	-	1.00

TABLE 2 – Pearson correlation coefficients between the acceleration of the landmarks of Participant 2 : nose (N), right (RS) and left (LS) shoulders, and average acceleration (AVG).

When compared to the distribution of D presented in [2], the estimated density curves for D computed from peak velocities and from peak accelerations presented in the present study follow a similar pattern : they are centered around the origin on the vertical axis, with the velocity curve slightly shifted to the right relative to the acceleration one. These preliminary results validate our video-based methodology for analyzing gesture-speech synchronicity in people with T21. However, to provide more robust evidence of this phenomenon, further experiments on larger and more diverse datasets remain to be conducted.

7 Perspectives

Our preliminary observations on videos featuring two young woman with T21 showed a certain level of synchronization between the moments of highest acceleration and velocity in upper body movements and the moments of highest voice pitch. Although these results are coherent with previous studies [2] on this topic, much remains to be done in to better describe the patterns of gesture-speech synchrony among people with T21. Our next steps include :

- The inclusion of a larger variety of participants with T21 in regard to gender and age ;
- The analysis of longer videos, from which gestures can be segmented, as done in [2] ;
- The analysis of gesture-speech synchronicity in different contexts : either during a monologue, such as the videos considered in this paper, or during interactions with other individuals, in which the participants alternate between listening and responding ;
- An analysis of the synchronicity between speech and individual body landmarks.

Acknowledgments

This work was partially supported by the CAPES-COFECUB program (project number : PE3335662P), funded by the French Ministry for Europe and Foreign Affairs, the French Ministry for Higher Education and CAPES - Brazil. Research conducted within the context of the International Associated Laboratory SAVOIR 21 funded by the Université de Lille.

Références

- [1] Anna Esposito and Antonietta M Esposito. On speech and gestures synchrony. In *Analysis of Verbal and Non-verbal Communication and Enactment. The Processing Issues : COST 2102 International Conference, Budapest, Hungary, September 7-10, 2010, Revised Selected Papers*, pages 252–272. Springer, 2011.
- [2] Wim Pouw and James A Dixon. Quantifying gesture-speech synchrony. In *the 6th gesture and speech in interaction conference*, pages 75–80. Universitaetsbibliothek Paderborn, 2019.
- [3] Marian Oliveira, Vera Pacheco, and Thaís Ferreira Brito. Expressão emocional em pessoas com síndrome de down : análise acústica da alegria e da tristeza (the emotional expression in down syndrome people : acoustic analysis of happiness and sadness). *Estudos da Língua (gem)*, 2019.
- [4] Jana M Iverson, Emiddia Longobardi, and M Cristina Caselli. Relationship between gestures and words in children with down’s syndrome and typically developing children in the early stages of communicative development. *International journal of language & communication disorders*, 38(2) :179–197, 2003.
- [5] Camillo Lugaesi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe : A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019, page 2. Long Beach, CA, 2019.
- [6] Wim Pouw, James P Trujillo, and James A Dixon. The quantification of gesture–speech synchrony : A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior research methods*, 52(2) :723–740, 2020.
- [7] Yannick Jadoul, Bill Thompson, and Bart De Boer. Introducing parselmouth : A python interface to praat. *Journal of Phonetics*, 71 :1–15, 2018.
- [8] Paul Boersma. Praat, a system for doing phonetics by computer. *Glott. Int.*, 5(9) :341–345, 2001.