Ensemble des publications

21^{e} édition du colloque CORESA

COmpression et REprésentation
des Signaux Audiovisuels

3-5 novembre 2021
Sophia Antipolis
France

Comité local d'organisation : Marc ANTONINI, Frédéric PAYAN
UCA / I3S / CNRS

Table des matières

Codage et compression

Learning sparse structured auto-encoder for image coding

Michel Barlaud
Laboratoire I3S CNRS, Cote d'Azur University
Email: michel.barlaud@i3s.unice.fr

September 27, 2021

Abstract

In recent years, deep neural networks autoencoder have been applied to different domains and achieved dramatic performance improvements over state-of-the-art classical methods such as lossy image compression [14].The performances in term of compression are superior to JPG2000. However the main issue of autoencoders is the memory requirement and computational power.
In order to cope these issues, we propose a new structured sparse learning method. We design algorithms for two constraints: the classical ℓ_1 constraint and the new $\ell_{1,1}$ constraint. Experiments results show that ℓ_1 constraint provides the best structured sparsity resulting in high reduction of memory.

1 Learning sparse sparse autoencoder with $\ell_{1,1}$ constraints

Deep neural networks have been applied recently to different domains and have shown a dramatic improvement in accuracy of image recognition [9], speech recognition [10] or natural language processing [11]. These studies relied on deep networks with millions or even billions of parameters. For instance, the original training of ResNet-50 [5] (image classification) contains 25.6M parameters and required 29 hours of processing using 8 GPUs. The recent development of DNNs, hardware accelerators like GPUs and the availability of deep learning frameworks for smartphones [8] suggest seamless transfer of DNN models trained on servers onto mobile devices. However, it turns out that the memory [12] and energy consumption [4] are still the main bottlenecks for running DNNs on such devices.

Autoencoders were introduced into the field of neural networks decades ago and their most efficient application was dimensionality reduction [7]. Autoencoders have been used for denoising different types of data [15], and lossy image coding [14] to extract relevant features. An autoencoder is a discriminative model that maps feature points to a high dimensional space to labels in a low dimensional latent space [6] and [13]. One of the main advantages of the autoencoder is the projection of the data in the low dimensional latent space. Let us recall that when a model properly learns to construct a latent space, it identifies general features which are relevant for predicting classes.
Autoencoders have the potential to address an increasing need for flexible lossy compression algorithms [14]. However, it is known that autoencoders are largely overparametrized and that in practice, relatively few network weights are actually necessary to learn accurately data features. Numerous methods have been proposed in order to remove network weights (*weight sparsification*) either on pre-trained models or during the training phase. These methods generally produce sparse weight matrices but unfortunately networks with random sparse connectivity.
In this work, we propose a new constrained approach where the constraint is directly related to the number of zero-weights.
Let Z , the latent space, $\widehat{X}$ the reconstructed data and W the weights of the neural network.

The goal is to compute the weights W of the autoencoder network minimizing the total loss which depends on both the rate loss and reconstruction loss and thus we propose to minimize the following criterion :

$$Loss(W) = \lambda\phi(Z) + \psi(\widehat{X} - X) \text{ s.t. } \|W\|_1^1 \leqslant \eta. \quad (1)$$

Where the rate loss ϕ is a function of the latent variable distribution. We use the robust Smooth ℓ_1 (Huber) Loss

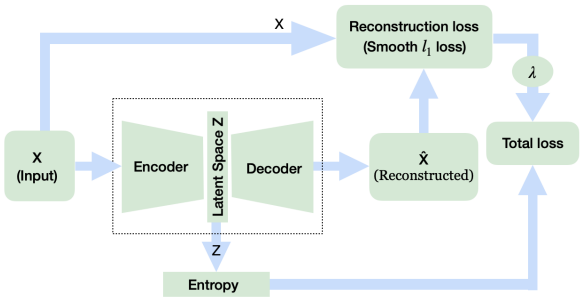


Figure 1: Autoencoder

as reconstruction loss function ψ.

The main difference of the criterion in the paper [14] is the introduction of a constraint to sparsify the neural network.
Classical Group LASSO consists in using the $\ell_{2,1}$ norm for the constraint on W but unfortunately does not induce sparsity [1]. Thus we propose in a previous works to use the $\ell_{1,1}$ constraint [2], [3] which is computed with the following algorithm: we first compute the radius t_i and then project the rows using the ℓ_1 adaptive constraint t_i (See [2] for more details):
Then we run the double descent algorithm [16] where we replace the thresholding by our $\ell_{1,1}$ projection.

Algorithm 1 Projection on the $\ell_{1,1}$ norm—$proj_{\ell_1}(V, \eta)$ is the projection on the ℓ_1-ball of radius η, $\nabla\phi(W, M_0)$ is the masked gradient with binary mask M_0, and f is the ADAM optimizer, γ is the learning rate

Input: $W*, \gamma, \eta$
for $n = 1, \ldots, N(epochs)$ **do**
  $V \leftarrow f(W, \gamma, \nabla\phi(W))$
end for
$t := proj_{\ell_1}((\|v_i\|_1)_{i=1}^d, \eta)$
for $i = 1, \ldots, d$ **do**
  $w_i := proj_{\ell_1}(v_i, t_i)$
end for
Output: W, M_0
Input: $W*$
for $n = 1, \ldots, N(epoch)$ **do**
  $W \leftarrow f(W, \gamma, \nabla\phi(W, M_0))$
end for
Output: W

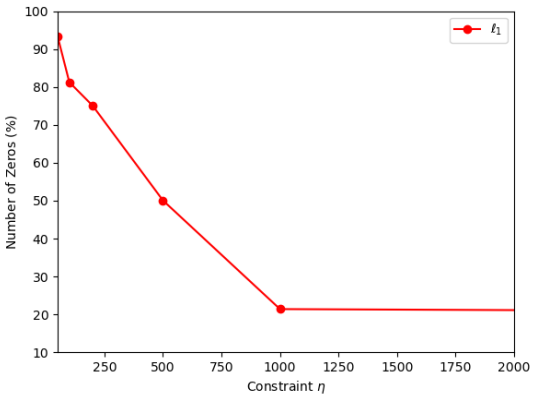


Figure 2: Number of zeros (%) as a function of η

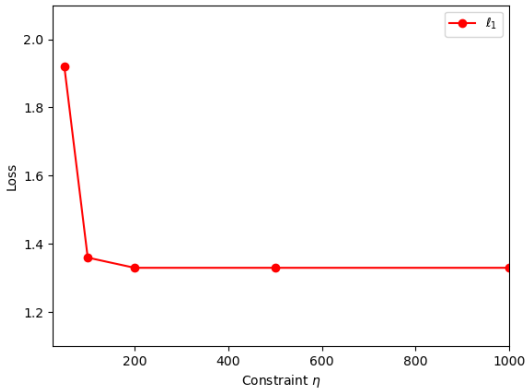


Figure 3: Loss as a function of η

2 Experimental results

We implemented our method in PyTorch. We choose as baseline Adam optimizer in PyTorch. We use the python code https://github.com/alexandru-dinu/cae implementation of the autoencoder method proposed in [14].
In this first experiment, we set $\lambda = 0$. We use the number of zeros of the weights to compute memory footprint and we show test decoded images for visual evaluation for different values of η.

The figures 4 show for different images the original image and the encoded-decoded image with a sparse autoencoder neural network.

Figures 2, 3 and 8 shows that we can sparsify the network thanks to ℓ_1 constraint thus reducing the

Figure 4: Left: original Right Coded

Figure 5: Left: original Right Coded

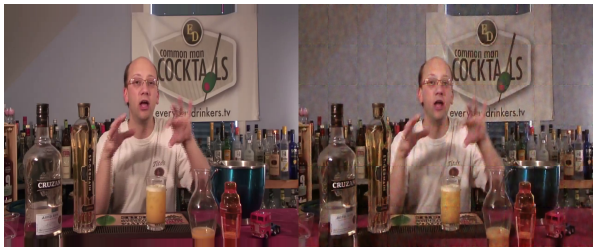


Figure 6: Left: original Right Coded

Figure 7: Left: original Right Coded

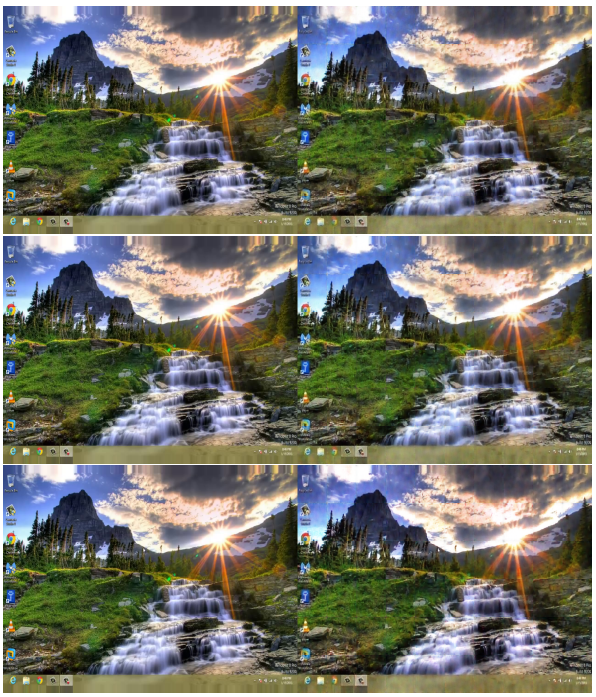


Figure 8: Left: Original image. Right decoded images: Up $\eta = 100$, Middle $\eta = 200$, Bottom $\eta = 500$

memory footprint without visual quality reduction.
Note that a constraint on (bit rate) or entropy leads to an increase of the coding loss which will then allow to improve the sparsification without additional losses.

3 Conclusion

We have proposed a framework to reduce the storage cost and the computation cost of autoencoder networks which is crucial for smart phone.
Our future work will propose a complete coding scheme and provide different autoencoders adapted to the required data rate and/or quality.
The author would thank internship Axel Gustovic for processing simulations and Dr Frederic Guyard for fruitfull discussions.

References

[1] Michel Barlaud, Antonin Chambolle, and Jean-Baptiste Caillau. Classification and feature selec-

tion using a primal-dual method and projection on structured constraints. *International Conference on Pattern Recognition, Milan*, 2020.

[2] Michel Barlaud and Frédéric Guyard. Learning sparse deep neural networks using efficient structured projections on convex constraints for green ai. *International Conference on Pattern Recognition, Milan*, 2020.

[3] Michel Barlaud and Frederic Guyard. Learning a sparse generative non-parametric supervised autoencoder. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, TORONTO , Canada*, June 2021.

[4] Song Han, Junlong Kang, Huizi Mao, Yiming Hu, Xin Li, Yubin Li, Dongliang Xie, Hong Luo, Song Yao, Yu Wang, et al. Ese: Efficient speech recognition engine with sparse lstm on fpga. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 75–84. ACM, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Geoffrey Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[7] Zemel Richard Hinton, Geoffrey. Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pages 3–10. 1994.

[8] Andrey Ignatov, Radu Timofte, William Chou, Ke Wang, Max Wu, Tim Hartley, and Luc Van Gool. AI benchmark: Running deep neural networks on android smartphones. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165, 2019.

[11] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning in natural language processing. *arXiv:1807.10854*, 2018.

[12] S Rallapalli, H Qiu, A Bency, S Karthikeyan, R Govindan, B Manjunath, and R Urgaonkar. Are very deep neural networks feasible on mobile devices. *IEEE Trans. Circ. Syst. Video Technol*, 2016.

[13] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[14] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv stat.ML /1703.00395*, 2017.

[15] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, 2010.

[16] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3597–3607. Curran Associates, Inc., 2019.

Étude comparative de l'impact d'un codage à précision variable sur des données de simulation en géosciences

Lauriane Bouard[1,3], Laurent Duval[2], Frédéric Payan[3], Christophe Preux[2], and Marc Antonini[3]

[1]IFP Energies nouvelles, Rond-point de l'échangeur de Solaize, 69360 Solaize, France
[2]IFP Energies nouvelles, 1-4 Avenue du Bois Préau, 92852 Rueil-Malmaison, France
[3]Université Côte d'Azur (UCA), CNRS, Laboratoire I3S, 06900 Sophia Antipolis, France

Résumé : *Le volume des données scientifiques produites par simulation numérique est en augmentation croissante. Cette volumétrie incite depuis peu la communauté à employer des méthodes de compression à précision variable. Si l'usage de données scientifiques comprimées peut atténuer les problèmes croissants de stockage primaire, il peut également jouer d'autres rôles à différentes étapes du workflow pour améliorer les performances des outils de calculs, limités notamment en capacité mémoire et en bande passante. Néanmoins ce type d'approche ne doit pas se faire au détriment de la qualité de la simulation. L'étude présentée ici analyse l'impact d'un codage à précision variable (obtenu par combinaison d'un compandeur, d'une transformée en ondelette et d'un codage par arbre-zéro (zerotree)) sur des données de simulation issues des géosciences. Pour valider notre approche nous le comparons à deux codeurs de référence utilisés dans différents domaines scientifiques (SZ [2], ZFP [4]) selon différentes métriques objectives corrélées à la validation subjective de la simulation.*

Mots-clés : Compression, géosciences, maillages, métriques de qualité objectives/subjectives, simulation

1 Introduction

Les supercalculateurs actuels (high performance computers ou HPC en anglais) ont de très forte capacités de calcul, dépassant les centaines de PétaFLOPS (10^{15} opérations en virgule flottante par seconde). Cela permet de simuler des phénomènes complexes avec un très grand réalisme dans de nombreux domaines (cosmologie, climatologie [1], mécanique des fluides, etc.). La contrepartie est l'explosion démesurée de la quantité de données générées, qu'il s'agisse de résultats stockés, de données transférées, ou encore pour sauvegarder et restaurer un état intermédiaire par des checkpoints. En conséquence, l'usage de codeurs à précision variable pour des données numériques simulées se généralise. Il facilite le stockage, l'empreinte mémoire et l'accélération des transferts, et par conséquent le temps d'exécution des simulations. SZ [2] et ZFP [4] (développés notamment au sein des laboratoires d'Argonne et de Lawrence Livermore) sont des codeurs à précision variable ayant fourni des résultats satisfaisants sur différents types de données, notamment en cosmologie (NYX) ou en climatologie (CESM). Par réduction de précision des données codées en nombres flottants, ces codeurs peuvent limiter la taille des données, sans pour autant impacter la qualité visuelle. Des résultats analogues ont été obtenus dans différents contextes de simulation. Ces travaux posent de manière plus générale la question suivante : quelle est la fidélité numérique nécessaire pour l'obtention de résultats précis et réalistes pour la simulation de phénomènes physiques ?

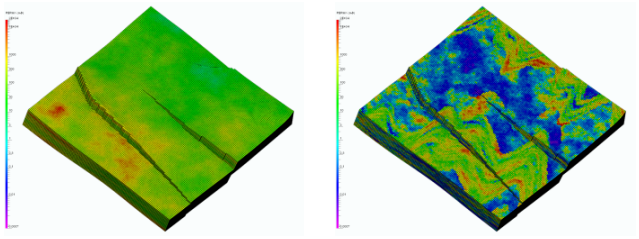


FIGURE 1 – Exemple de maillages hexahédriques en géosciences.

2 Travail proposé

En géosciences par exemple, une simulation d'écoulement s'effectue au travers de maillages hexaédriques (cf. illustrations 1) qui représentent des formations géologiques profondes. Ces simulations sont souvent lancées en grand nombre, pour des études de sensibilité. Elles requièrent une exécution en un temps raisonnable, ce qui a conduit au développement de méthodes d'upscaling, réduisant la résolution des maillages d'origine. Ce type de méthodologie est intégré de manière implicite à différentes résolutions dans le système de compression multi-échelle HexaShrink proposé en [5]. Nous complétons ici ce travail par un codage progressif des propriétés continues associées à ces maillages, notamment la perméabilité. Ce nouveau schéma de codage permet de jouer sur la précision binaire, par l'incorporation d'un codage de type zerotree (ZT) combiné à l'usage d'un compandeur (compresseur-expanseur, notés Λ et V) et d'une transformée en ondelette (9/7 DWT). Notre schéma, qui est donc à précision et à résolution variable, est comparé aux codeurs standards SZ [2] et ZFP [4]. Pour cela, nous utilisons des métriques objectives classiques, mais proposons aussi une nouvelle métrique (Λ-SNR) plus conforme à l'évaluation subjective des résultats de simulation ainsi qu'aux exigences du métier.

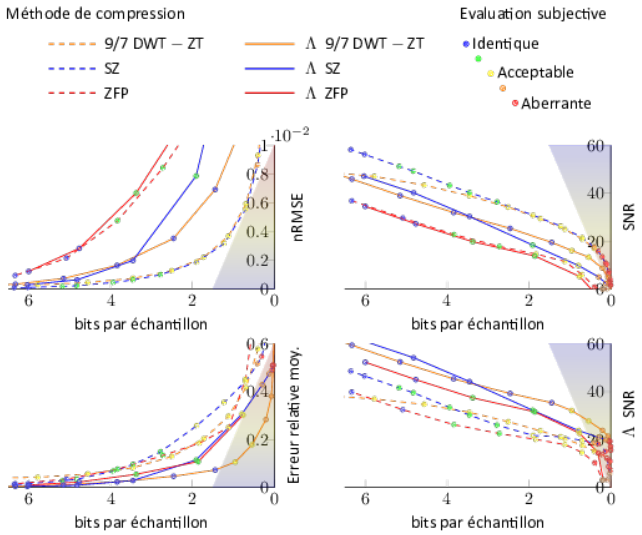


FIGURE 2 – Courbes débit/distorsion obtenues avec SZ, ZFP, et notre codeur progressif (9/7 DWT - ZT) pour des données de perméabilité décompressées à différentes précisions, avec ou sans compandeur (Λ). En ordonnée : différentes métriques objectives (le triangle dégradé symbolise l'échelle de qualité (la base bleue pour les hautes qualités, la pointe rouge pour les plus faibles). La validation subjective des simulations est indiquée par les marqueurs de couleur.

3 Résultats

Nous présentons ici les résultats de compression obtenus avec le codeur progressif proposé, ainsi que SZ et ZFP à différentes précisions. Ces trois codeurs sont utilisés avec ou sans compandeur. Les résultats sont résumés par des courbes débit/distorsion selon 4 mesures (figure 2). Indépendamment, la justesse des simulations obtenues à partir de ces données décompressées a été validée par l'attribution d'un score subjectif (échelonné en cinq valeurs, d' « identique » à « aberrant », et symbolisé par des marqueurs colorés sur les courbes). Nous constatons que :

— les métriques objectives standards ou classiques (nRMSE, SNR) reflètent mal l'évaluation subjective ;

— l'usage d'un compandeur améliore les performances objectives (erreur relative moyenne, Λ-SNR) et subjectives des trois codeurs (en traits pleins) ;

— Sur toute la gamme des résultats de simulation jugés « identiques » aux résultats de référence (marqueurs de couleur bleu), notre méthode atteint des débits inférieurs aux algorithmes SZ [2] et ZFP [4] (voir Λ-SNR). En d'autres termes, de meilleurs taux de compression peuvent être atteints pour une fidélité optimale des résultats de simulation.

4 Conclusion et perspectives

Nous montrons dans cette étude qu'il est possible de maintenir des résultats de simulation quasi-identiques à ceux obtenus à pleine précision tout en réduisant de manière notable la taille des données. Cela s'obtient en jouant sur la résolution et la précision, en exploitant conjointement la multi-résolution des ondelettes et la profondeur binaire des données numériques par l'utilisation du codage zerotree et d'un compandeur logarithmique. Il est donc possible de réduire drastiquement la quantité de données binaires à traiter, tout en conservant la pleine qualité de la simulation. De plus nous confirmons que l'utilisation d'un compandeur permet d'améliorer la compression de données scientifiques ([3]), dont la grande dynamique rend la compression difficile.

Références

[1] Allison H. Baker, Dorit M. Hammerling, Sheri A. Mickelson, Haiying Xu, Martin B. Stolpe, Phillipe Naveau, Ben Sanderson, Imme Ebert-Uphoff, Savini Samarasinghe, Francesco De Simone, Francesco Carbone, Christian N. Gencarelli, John M. Dennis, Jennifer E. Kay, and Peter Lindstrom. Evaluating lossy data compression on climate simulation data within a large ensemble. *Geosci. Model Dev.*, 9(12) :4381–4403, dec 2016.

[2] Franck Cappello, Sheng Di, Sihuan Li, Xin Liang, Ali Murat Gok, Dingwen Tao, Chun Hong Yoon, Xin-Chuan Wu, Yuri Alexeev, and Frederic T. Chong. Use cases of lossy compression for floating-point data in scientific data sets. *Int. J. High Perform. Comput. Appl.*, jul 2019.

[3] Xin Liang, Sheng Di, Dingwen Tao, Zizhong Chen, and Franck Cappello. An efficient transformation scheme for lossy data compression with point-wise relative error bound. In *2018 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, sep 2018.

[4] Peter Lindstrom. Fixed-rate compressed floating-point arrays. *IEEE Trans. Visual Comput. Graph.*, 20(12) :2674–2683, 12 2014.

[5] Jean-Luc Peyrot, Laurent Duval, Frédéric Payan, Lauriane Bouard, Lénaïc Chizat, Sébastien Schneider, and Marc Antonini. HexaShrink, an exact scalable framework for hexahedral meshes with attributes and discontinuities : multiresolution rendering and storage of geoscience models. *Computat. Geosci.*, 23 :723–743, Aug. 2019.

A SEQUENCING NOISE RESISTANT CODE MAPPING ALGORITHM FOR IMAGE STORAGE IN DNA

Melpomeni Dimopoulou, Eva Gil San Antonio, Marc Antonini
Université Côte d'Azur, CNRS, Laboratoire I3S

Abstract : *The continuous exponential increase in the generation of digital information is becoming inconsistent with the capacity and longevity limitations imposed by conventional storage devices which can't be reliable for more than 10-20 years. More precisely, 90% of the data on the internet has been only generated in the last 2 years while 80% of this information consists of "cold" data which is very rarely or never accessed but still needs to be stored in off-line back-up drives for security and compliance reasons. To ensure data reliability data centers are nowadays purging metric tons of hardware for the frequent replacement of those drives which is extremely expensive both in terms of money and energy. To handle this problem scientists have recently proposed the use of synthetic DNA as a means of digital data storage. This idea is inspired by the biological properties of the DNA molecule which contains all the necessary information for living organisms to survive, stored in a very limited volume such as the cells' nucleus. Furthermore, when stored under specific conditions, DNA can be decodable without loss of information for hundreds of years. In this work we propose an algorithm which optimally maps input quantization vectors to DNA codewords for the storage of quantized images into DNA.*

Keywords : DNA data storage, Image coding, Robust encoding, Vector Quantization.

1 Introduction

DNA data storage is a very promising yet challenging procedure as it requires the use of the delicate biological processes of DNA synthesis (writing) and sequencing (reading). More precisely DNA sequencing imposes some important restrictions in the encoding workflow as it can introduce errors in the decoded DNA sequence. In [2], there has been a first attempt to store data into DNA while also providing a study of the two main causes of this sequencing error. An additional restriction has been later included by a biological study in [7]. More precisely it has been noted that in order to reduce the sequencing error the encoding algorithms should respect the three following rules. 1) No repetitions of the same symbol more than 3 times (homopolymer rule). 2) The percentage of C and G should be lower or equal to the percentage of A and T. 3) Short repeated patterns should be avoided. In order to deal with errors, some studies in [1] and [6] suggest the use of Reed-Solomon codes in order to treat the erroneous sequences. In [5] we have proposed a new constrained fixed length algorithm for the encoding of quantized images into DNA. This work has been extended in the use of Vector Quantization (VQ) in [4]. However, respecting the rules imposed by the sequencing does not guarantee an error free decoding. This is because some sequencers like the Nanopore sequencer introduce very high error rates which can't be avoided. One can imagine the sequencing noise as the one introduced by noisy channels in telecommunications. In [3] there has been proposed an interesting algorithm for optimally assigning input values to binary words to achieve better resistance to the channel noise. Inspired by the work proposed in [3], we extend this algorithm to a quaternary representation for the mapping of input symbols to DNA codewords which aims to create an encoding that is more resistant to the sequencing noise. This algorithm uses a more sophisticated method for the mapping of input symbols to the different DNA codewords so that in case of an error the erroneous decoded symbol will be closer to the original one.

2 Proposed work

The work presented in this paper is an extension of the method proposed in [3]. The goal is finding an optimal mapping between input vectors obtained by a VQ algorithm and quaternary codewords so as to achieve resistance to sequencing errors. VQ is useful for the efficient compression of an image before it is stored into DNA to reduce the synthesis cost which can be relatively high. The purpose of the proposed algorithm is the mapping of close (in terms of Euclidean distance) quantization vectors $v_i, i \in [1, \dots, M]$ from a codebook V to codewords from a code W which have a small Hamming distance. The idea behind this mapping lies in the fact that in case of an error during sequencing and assuming that the sequencing noise rate is small enough a correct codeword will be transformed to another one which will have a small Hamming distance with the correct one. The algorithm can be very roughly described by the following parts:

For each codeword : Create a sphere $H(w_i)$ containing the B_i codewords which have a Hamming distance of 1 compared to $w_i, i \in \{1, \dots, L\}$. Define $B = max_i(B_i), i \in \{1, \dots, L\}$.

- For each input vector v_i: Find a set $S(v_i)$ of B neighboring vectors v_l which are the closest to v_i in terms of Euclidean distance $d(v_i, v_l)$.
- For each input vector : Compute the empirical function $F(v_i) = \frac{p(v_i)}{\alpha^{\beta}(v_i)}$ where $p(v_i)$ is the probability of in the input sequence, and $\alpha(v_i) = \sum_{j|v_j \in S(v_j)} d(v_j, v_i)$ with $\beta \geq 0$ a trade-off parameter Progressively perform assignment of vectors v_i to codewords w_i such that vectors with a bigger $F(v_i)$ as well as its neighboring vectors $v_l \in S(v_i)$ will be assigned to the same

sphere of codewords $H(w_i)$ whenever possible as depicted in figure 1. If this is not possible assignment is performed such that vectors v_i are assigned to codewords with a small Hamming distance from the codewords already assigned to their neighboring vectors. The algorithm for this step is relatively complicated and thus for further information readers can refer to [3].

- Optimization of the first assignment:
  - Exchange the previously mapped codewords between each pair of vectors.
  - For each exchange check if the average distortion has decreased. If true keep this change, else keep the initial state of mapping.

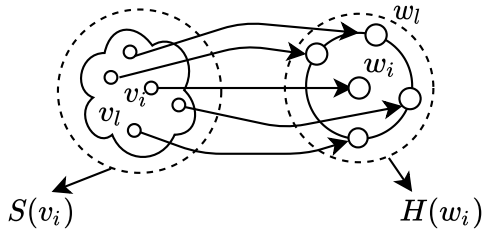


Figure 1: Assuming a vector v_i, the set $S(v_i)$ contains the B closest to v_i vectors v_l in terms of Euclidean distance. Then given a codeword w_i, the Hamming sphere $H(w_i)$ of radius 1 contains all possible codewords w_l with $i \in \{1, \ldots, B_q\}$ for which $d_H(w_i, w_l) = 1$. An optimal case of mapping would be the one where all vectors that belong to the same neighborhood $S(v_i)$ are assigned to the same sphere $H(w_i)$. However as this mapping is not possible for all the words $w_i \in \mathcal{C}^*$ we search for a solution that globally optimizes the assignment such that close codevectors are mapped to close codewords.

3 Results

For the experiments we quantized an image of 512x512 pixels using VQ with 100 vectors of length $n = 2$. We then decoded the image adding 3% of noise. This noise ratio is equal to the estimated percentage of noise added by the Nanopore sequencer. Figure 2 depicts the visual quality of the noisy decoding without optimal mapping while figure 3 represents the noisy decoding for the case where optimal mapping was used. The visual quality has improved significantly providing a gain of 3 dB in the PSNR.

Without optimal mapping PSNR=20.39 dB

With optimal mapping PSNR=23.4 dB

Figure 2: Visual result of substitution noise for the two different ways of mapping.

4 Conclusion and perspectives

In this work we proposed a new mapping algorithm for the optimal assignment of quantization vectors obtained by VQ quantization to DNA codewords. The obtained results are very promising encouraging us to further study and improve this approach.

References

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

[2] George M Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in DNA. *Science*, page 1226355, 2012.

[3] JR Boisson De Marca, NS Jayant, et al. An algorithm for assigning binary indices to the codevectors of a multi-dimensional quantizer. In *1987 IEEE International Conference on Communications (ICC'87)*, pages 1128–1132. , 1987.

[4] Melpomeni Dimopoulou and Marc Antonini. Image storage in DNA using Vector Quantization. In *EUSIPCO 2020*, Amsterdam, Netherlands, January 2021.

[5] Melpomeni Dimopoulou, Marc Antonini, Pascal Barbry, and Raja Appuswamy. A biologically constrained encoding solution for long-term storage of images onto synthetic DNA. In *EUSIPCO 2019*, 2019.

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

[7] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36, 2012.

TOWARDS ACCURATE RATE ESTIMATION FOR 3D POINT CLOUD COMPRESSION BY TSPLVQ

Amira Filali, Vincent Ricordel and Nicolas Normand

LS2N Laboratory, Nantes University, Polytech, rue Christian Pauc, 44306 Nantes, France

ABSTRACT

Point clouds are widely emerged as a promising 3D visual representation model for immersive and high quality experiences using dense point cloud. However, as they are usually made up of thousands up to billions of points, advanced techniques of data compression are essential to store and transmit this type of data. This paper improves prior work, which provided a new top-down hierarchical geometry representation based on adaptive Tree-Structured Point-Lattice Vector Quantization (TSPLVQ), to make the point cloud geometry compression more adaptive to the input point cloud characteristics. In the paper, more robust Rate-Distortion optimization process is introduced to perform efficient and accurate rate-aware splitting decisions when building and coding the tree structure. Experimental results in geometry point cloud compression, considering the tree structures and the leaves coding, show that the solution takes advantage of well-established principles that have been paramount to reach higher levels of point cloud geometry compression performance.

Index Terms— 3D point cloud, compression, rate-distortion optimization, G-PCC

1. INTRODUCTION

Among various newly acquired content types, 3D point cloud (PC) appears to be one of the most efficient representation of immersive media content thanks to the fast development of 3D scanning techniques, establishing a convergence between real and virtual realities and enabling more sophisticated experiences applications.

Characterized by geometry and multiple associated attributes, point cloud forms a spatially discrete set of points in a 3D geometric coordinate system. The development of even more precise capturing devices and the increasing requirements to realistically represent and to vividly render the 3D scenes, inevitably not only induce thousands up to billions of points, but also cause high complexity in the scattered random distribution of the spatial distribution, which brings great challenges to the storage consumption and transmission system. Thus, more advanced compression techniques are in urgent demand to make point clouds useful in practice.

The Moving Picture Experts Group (MPEG) is leading the process of seeking technologies and building an open standard for point cloud compression (PCC) [1]. The targeting standard adresses two main classes of solutions dealing with PCC: V-PCC which takes advantage of the usage of well-known 2D video technologies by projecting the point characteristics onto 2D frames, and a second class called G-PCC, for geometry-based compression of static point clouds. G-PCC is more appropriate for the context of this paper, as the G-PCC uses native 3D data structures and has large potential of improvements.

Both G-PCC and V-PCC are based on conventional models, such as octree decomposition [2], triangulated surface model, region-adaptive hierarchical transform [3], [4], and plane projection [5]. Other explorations related to the PCC relied on graphs [6], binary tree embedded with quadtree [7], or recently volumetric model [8]. To compare the proposed compression solutions, MPEG provided quality evaluation metrics for PCC leading to the selection of the point-to-point and point-to-plane as baseline metrics [9].

The rest of the paper is organized as follows. Section 3 describes briefly the adaptive TSPLVQ method. Details on the new tree rate estimation are given in Sections 3.1 and 3.2. In Section 4 we present and analyze the coding performance of the proposed approach. Finally, conclusions are drawn in section 5.

2. RELATED WORK

Actual researches in point cloud compression field can be classified as point cloud geometry compression and attributes compression [10, 11]. In the present paper, our work is mainly related to point cloud geometry compression. Many solutions for point cloud geometry compression have been explored using the octree structure to split the whole point cloud into smaller voxel volumes to better structure and represent the 3D point cloud [12] [13]. *Cohen et al.* [14, 15] extended the well-known shape adaptive Discrete Cosine Transform (SA-DCT) [16], to the voxelized 3D point clouds. Authors in [17] employed intra prediction and binary tree structure for effectively partitioning unorganized points into block structure in a lossless point cloud geometry compression approach. A

lossless intra encoder of voxelized point clouds is introduced in [18] that views the point cloud geometry as an array of bi-level images using a dyadic decomposition instead of the popular octree decomposition.
Ricordel et al. have introduced in [19, 20] a Vector Quantizer (VQ) based on truncated cubic lattices embedding in the context of classical image and video encoding according to a trade-off between rate and distortion using the Lagrangian method.
Inspired by the formulation of quantization in [19, 20], we proposed in [21] to significantly expand it by using new tools, different rate-distortion optimizations and several practical adaptations to the case of G-PCC.

The present paper builds on the work proposed in [21] to introduce a more acurate rate optimization for the purpose of efficient 3D point cloud geometry coding. Exactly, we improve the adaptive Tree-Structured Point-Lattice Vector Quantization (TSPLVQ) by using two splitting schemes (2×2×2 or 3×3×3) of a cubic Voronoï cell, and by adapting the distortion versus rate trade-off. Thus, the purpose of the present work is to reduce more the amount of data of a 3D point set while preserving as much information as possible by considering the distortion in the rendered 3D content from the decoded point cloud. In this work, we consider more accurate rate computation based on the entropic cost estimation of the tree. It is therefore better for rate-distortion optimization during the TSPLVQ procedure. Additionally, in contrast to the MPEG G-PCC reference model, the optimized TSPLVQ method produces higher resolution point clouds for lower bitrates.

3. POINT CLOUD COMPRESSION BY TSPLVQ

Our prior work in [21] is at the crossroads of static G-PCC, vector quantization of 3D data and rate-distortion optimization driven compression. The TSPLVQ approach, based on the embedding of truncated cubic lattices, permits hierarchical description of the 3-D PC through an unbalanced tree structure. The tree growing structure is achieved by using an iterative process such as, at each loop the best choice has to be done, between the node to split and according to two splitting schemes (2×2×2 versus 3×3×3) to better map the PC splitting. This choice is based on a rate-distortion criterion, the optimization is then performed locally, where the Lagrange multiplier λ controls the trade-off between rate and geometric distortion. The rate considered in [21], is either a constant cost of node splitting (8 bits if the splitting is 2×2×2, 27 bits if the splitting is 3×3×3), or a basic entropy estimation taking into account the entropy of the population in relation to node points.

3.1. Optimized estimation of the tree rate cost

Entropy coding is critical for source compression to exploit statistical redundancies. Theoretically, the entropy bound of the source symbol (e.g., splitting bitstream in our case) is closely related to its probability distribution, and accurate rate estimation plays a major role in rate-distortion optimization driven compression.
We propose to use a more accurate estimation of the entropic cost of the node splitting during the tree growing process. For each node splitting is computed, the increase in rate calculated in terms of entropic encoding cost of the node splitting, and the decrease in geometric distortion.

At this level of process, our objective is to code PC geometry stored in its 3D volumetric representation, this is referred as the *voxelization*. Considering the geometry of PC, a voxel V in the 3D representation at (i, j, k) position, corresponds also to a node (n_j) in the tree structure, is set to 1, e.g., $V(i, j, k) = 1$, if it is occupied (contains one or more PC points) and split, and $V(i, j, k) = 0$ otherwise. Each splitting scheme partitions the 3D cube associated to a node n_j either into 8 or 27 embedded smaller cubes (so children nodes of n_j). For each branch connecting n_j to its child, one bit is used, let's define it as the splitting state. Splitting state is a bitstream associated with each voxel node. Each voxel node is divided into several voxels (children) depending on the quantization scheme being considered (namely, 2×2×2 or 3×3×3) and every leaf node has in turn an associated voxel value (1 or 0). Note also that each node indicates by a position index where it lies inside its parent's splitted 3D cube (for instance when considering 2×2×2 splitting case, the position index are simply listed from 1 up to 8). If after a loop, a child node undergoes splitting (namely, its corresponding cube splitting), we update the splitting state of its parent node. The splitting state determines using 0s and 1s the children nodes that are split, e.g. a given splitting by 2×2×2, could have $splittingstate = 01001101$. So children nodes with position index 1, 4, 5, 7 are split and children nodes with position index 0, 2, 3, 6 are not split. At each level of quantization step, every splitting state has associated occurency probability which gives how likely the splitting state occurs in the tree. In other words, the occurency probability of a splitting state is the ratio between the number of time this splitting state occurs in the tree, and the number of splitted nodes. We can then approximate the actual rate of every splitting state by computing its entropy.
For instance, the rate $R(s_j)$, used to calculate λ score in equation (4) in [21], is equal to the entropy of the splitting states of the tree nodes at the $j-th$ loop in the TSPLVQ growing tree s_j.
The splitting states entropies of the tree nodes has to be updated dynamically after each loop of the TSPLVQ splitting process, by taking into account of the incrementing and

decrementing of the splitting states counters.
In order to count and to store the probabilities of all the splitting states, we use a hash table. Advantage of this strategy is that splitting state of any length can be dynamically stored in the table to avoid storing all the possible combinatories of all the splitting states (2^8 for 2×2×2 splitting and 3^{27} for 3×3×3 splitting). Due to this, we are able to run a hybrid quantization method using 2×2×2 and 3×3×3 in competition with each other and locating probability occurency of any state becomes fast due to hashing.

3.2. Final Bit-stream and attribute Compression

In the tree, each occupied leaf node corresponds exactly to one representant point which 3D location is set to the average value of the initial cloud points contained within the cube associated to the leaf. The mean colour is used to represent the colour information of all the cloud points inside the leaves cell. To further compress the corresponding reproduction vectors geometry and colour at the leaves level, we propose in this work to perform range coding by using Lempel–Ziv–Markov chain algorithm (LZMA) [22] for a lossless compression.
In other context, we could consider the voxels centers, instead of the mean points coordinates, to represent the input point cloud geometry and colour using the optimized TSPLVQ. In this case, we do not need to encode any explicit geometry information at nodes level, only build the tree structure by using recursive splittings described arithmetically in the bit stream when traversing the tree in different orders. Thus to encode the point cloud geometry : either we proceed progressively by using the scalable descriptions obtainable at the tree nodes (for instance by scanning its successive depth levels) the PC points in each leaf node are replaced by the corresponding center point, either we consider directly and only the final points positions and colours associate to the tree leaves.
For the purposes of this paper we aim to encode and decode only the point cloud thus the geometry and colour information at the leaves level are enough.

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

We used our optimized TSPLVQ, relied on Point-to-Point geometry distortion, to encode 3D point clouds. We compared our approach against the MPEG octree-based reference test model (lossy G-PCC model) [23]. All parameters in the G-PCC test model are kept unchanged for fair comparisons. We selected four static point clouds from people object dataset: $Soldier$, $LongDress$, $Loot$ and $LongDress$ suggested by MPEG-3DG group [24] as a test dataset. The performance is then measured in terms of the point-to-plane symmetric PSNR metric given in [25].

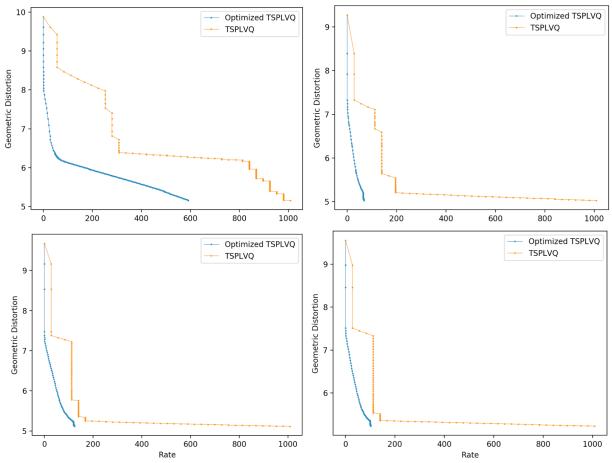


Fig. 1. RD curves showing the performance of the different steps of the optimized TSPLVQ. (first row): $Soldier$ and $RedandBlack$, (second row): $Loot$ and $LongDress$. Distortion is computed using the Point-to-Point geometric distortion [21] and the rate is the entropic cost of the splitting scheme.

4.2. Results

A selection of results is presented in Table 1. The analysis of the PSNR metric of the four point clouds, $Soldier$, $LongDress$, $Loot$ and $Redandblack$, shows that our method when considering the leaves cost only, obviously outperforms the reference model on all the testing point clouds in term of PSNR and Bitrate. Considering these 4 points clouds, the reference model has an average bitrate of 0.03 $bpov$ (bits per occupied voxel) and an average PSNR of 45.93 dB while our method has an average bitrate of 0.0002 $bpov$ and an average PSNR of 59.90 dB. It is worth observing that the entropy cost of the built trees of all point clouds, arithmetically encoded, is greater than the reference model and our method considering the leaves encoding. Moreover, we should take into consideration the colour information that also has to be transmitted, it is why we add this information (the colours) to the leaves, when we consider the leaves cost in Table 1.

We compute RD curves for each sequence of the test point clouds. We compared both rate-distortion driven TSPLVQ: the adaptive TSPLVQ [21] and the rate-adaptive TSPLVQ. The performance comparisons in term of rate and distortion between the 2 methods are reported in Fig. 1. The optimized method outperforms the previous TSPLVQ on all point clouds providing a maximal decrease in distortion, and a minimal increase in rate.
In Fig. 2, we show examples on the four selected 3D objects. These particular examples show that our method produces more points than the reference model at lower bitrates. The MPEG G-PCC model does not show details in complex regions. It also leads to halo artifacts and the resulting point

Table 1. Comparison of symmetric PSNR and Bitrate metrics results between the optimized TSPLVQ based on Point-to-Point distortion and MPEG reference model.

Point Cloud	Number of points	MPEG G-PCC			Our approach			
		Number of points	PSNR	Bitrate (bpov)	Number of points	PSNR	Leaves cost (bpov)	Tree nodes cost (bpov)
Soldier	1089091	20065	52,79	0.029	73103	58.86	0.0002	0.1473
LongDress	857966	15685	52.78	0.031	48869	59.74	0.0002	0.0418
Loot	805285	14828	52.78	0.029	47616	60.08	0.0003	0.0444
Redandblack	757691	13832	25.40	0.032	45034	60.08	0.0003	0.3203

Fig. 2. Rendering results for original point clouds (first row), compressed point clouds using MPEG lossy PCC (second row): (a), (b), (c), (d) and using our optimized TSPLVQ (third row): (e), (f) (g), (h).

cloud requiring therefore a post-processing. For instance, for the *Soldier* point cloud, the MPEG reference model cannot show details in the complex and smooth regions inducing a great loss of visual details (see Fig. 2 (a), (b), (c), (d)) while with our method using a lower bitrate for the leaves (0.0002 bpov), reasonable quality was achieved on all point clouds, as shown in 2(e), (f), (g), (h). Hence, the optimized method can preserve the details and the global look of the original point cloud.

In addition, we could obtain very close rendering to the original point cloud when we produce more points thanks to the proposed multiscale approach with low bitrate. Our visual rendering seems qualitatively very close to the original rendered point clouds compared to the reference test model. The benefits of the optimized method over the reference model in terms of both objective and subjective quality are easily observable.

5. CONCLUSION

In this paper, we introduced a rate-optimized TSPLVQ method for lossy compression of the 3D point cloud geometry. Our method takes advantage of well-established principles that have been paramount to reach higher levels of point cloud compression performance.

6. REFERENCES

[1] S. Schwar et al., "Emerging mpeg standards for point cloud compression," 2019.

[2] D. Meagher, "Geometric modeling using octree encoding," Computer graphics and image processing, vol. 19, no. 2, pp. 129–147, 1982.

[3] R. L. de Queiroz and P. A. Chou, "Compression of 3D point clouds using a region-adaptive hierarchical transform," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3947–3956, Aug. 2016.

[4] R. L. de Queiroz and P. A. Chou, "Transform coding for point clouds using a gaussian process model," IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3507–3517, July 2017.

[5] E. Lopes, J. Ascenso, C. Brites, and F. Pereira, "Adaptive plane projection for video-based point cloud coding," IEEE International Conference on Multimedia and Expo (ICME), pp. 49–54, July 2019.

[6] D. Thanou, P. A. Chou, and P. Frossard, "Graph-based compression of dynamic 3D point cloud sequences," IEEE Transactions on Image Processing, vol. 25, no. 4, pp. 1765–1778, April 2016.

[7] B. Kathariya, L. Li, Z. Li, J. Alvarez, and J. Chen, "Scalable point cloud geometry coding with binary tree embedded quadtree," IEEE International Conference on Multimedia and Expo (ICME), p. 1–6, July 2018.

[8] M. Krivokuća, P. A. Chou, and M. Koroteev, "A volumetric approach to point cloud compression–part ii: Geometry compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 2217–2229, December 2020.

[9] S. Schwarz, G. Cocher, D. Flynn, and M. Budagavi, "Common test conditions for point cloud compression," inISO/IECJTC1/SC29/WG11 MPEG output document N17766, July 2018.

[10] X. Yiqun and et al., "Rate-distortion optimized scan for point cloud color compression," Visual Communications and Image Processing (VCIP), 2017.

[11] C. Zhang, D. Florencio, and C. Loop, "Point cloud attribute compressionwith graph transform," International Conference on ImageProcessing (ICIP), pp. 2066–2070, 2014.

[12] R. Schnabel and R. Klein, "Octree-based point-cloud compression," Eurographics Symposium on Point-Based Graphics, pp. 111–120, 2006.

[13] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," International Conference on Robotics and Automation, Shanghai, 2011.

[14] R.A. Cohen, D. Tian, and V. Vetro, "Point cloud attribute compression using 3-D intra prediction and shape-adaptive transforms," Data Compression Conference (DCC), pp. 141–150, 2016.

[15] R.A. Cohen, D. Tian, and V. Vetro, "Attribute compression for sparse point clouds using graph transforms," International Conference on Image Processing (ICIP), Phoenix, AZ, USA, pp. 1374–1378, Sept. 2016.

[16] T. Sikora and B. Makai, "Shape-adaptive dct for generic coding of video," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 59–62, 1995.

[17] W. Zhu, Y. Xu, L. Li, and Z. Li, "Lossless point cloud geometry compression via binary tree partition and intra prediction," IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), Luton, pp. 1–6, Oct. 2017.

[18] E. Peixoto, "Intra-frame compression of point cloud geometry using dyadic decomposition," IEEE Signal Processing Letters, vol. 27, pp. 246–250, 2020.

[19] V. Ricordel and C. Labit, "Tree-structured lattice vector quantization," European Signal Processing Conference (EUSIPCO), Italy, Sept 1996.

[20] V. Ricordel and C. Labit, "Vector quantization by packing of embedded truncated lattices," in *Proceedings., International Conference on Image Processing, Washington, DC, USA*, 1995, vol. 3, pp. 292–295 vol.3.

[21] A. Filali, V. Ricordel, N. Normand, and W. Hamidouche, "Rate-distortion optimized tree-structured point-lattice vector quantization for compression of 3D point clouds geometry," International Conference on ImageProcessing (ICIP), sept. 2019.

[22] A. Akoguz, S. Bozkurt, A. Gozutok, G. Toprakci, M. Bogaz E. Turan, and S. Kent, "Comparison of open source compression algorithms on VHR remote sensing images for efficient storage hierarchy," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, July 2016.

[23] K. Mammou, P. A. Chou, D. Flynn, M. Krivokuća, O. Nakagami, and T. Sugio, "G-PCC codec description v1," ISO/IEC JTC1/SC29/WG11 N18015, October 2018.

[24] C. Tulvan, R. Mekuria, and Z. Li, "Draft dataset for point cloud coding (PCC)," ISO/IEC JTC1/SC29/WG11N16333, June 2016.

[25] Julien Ricard, "PCC CE 0.3 on new metrics," ISO/IEC JTC1/SC29/WG11 N18032, October 2018.

Compression des hologrammes numériques basée sur le codeur JPEG optimisé

Meha Hachani [1], Ines Bouzidi [2], Azza Ouled Zaid [2],Frédéric Dufaux [3]
[1]Institut Supérieur d'Informatique Université de Tunis El Manar, Ariana, Tunisie

[2] Laboratoire Systèmes de Communications, ENIT, Université de Tunis El Manar, Tunisie
[3]Laboratoire des Signaux et Systèmes (L2S), CNRS, Université Paris-Sud, Gif-sur-Yvette, France

Résumé : *La représentation 3D par hologrammes numériques connait un succès croissant face aux représentations stéréoscopiques classiques. La diversité des applications et des secteurs pouvant bénéficier de cette technique met en exergue les limites encore subsistantes pour le stockage et la transmission des hologrammes. Le groupe JPEG a relevé le défi en travaillant sur un nouveau standard : JPEG Pleno, permettant de compresser divers objets plénoptiques tout en préservant l'interopérabilité avec les différents formats existants. Cet article présente une solution pour le codage orienté objet des hologrammes numériques basée sur le standard JPEG.*

Mots-clés : Hologramme numérique, Codage orienté objet, JPEG, allocation binaire

1 Contexte et état de l'art

Les percées technologiques relatives aux dispositifs de capture et d'affichage d'image ont permis de migrer rapidement d'une représentation planaire vers des objets volumiques. Jusqu'à présent, la représentation stéréoscopique est la plus couramment utilisée dans différents secteurs d'activité (cinéma, applications médicales, architecture ...). Cependant, la tendance semble pencher de plus en plus en faveur des hologrammes numériques qui présentent l'avantage majeur d'offrir une représentation 3D ultra réaliste sans imposer à l'utilisateur aucun dispositif optique spécial (lunettes 3D). Ce basculement est dû principalement à l'apparition de caméras dédiées à la capture d'hologrammes, facilitant l'accès aux objets holographiques notamment les CGH (Computer Generated Holograms). L'hétérogénéité existante entre les différentes technologies et modalités de représentation 3D a rapidement nécessité la mise en place de nouveaux formats standards pour représenter efficacement cette nouvelle génération d'objets. Ce besoin s'est d'autant plus confirmé face au volume important des données à coder. En effet, les hologrammes sont des signaux à très haute fréquence pouvant être représentés par des valeurs complexes dont les caractéristiques sont incompatibles avec ceux des images pixelliques, limitant ainsi les performances des algorithmes de compression classiques. Quelques approches de compression ont déjà été proposées pour les données holographiques [1]. Actuellement, le groupe JPEG travaille sur les spécifications d'un nouveau standard JPEG Pleno prenant en charge la spécificité des contenus plénoptiques notamment les hologrammes numériques [2]. Le but de cet article est de présenter une méthode de compression des hologrammes numériques basée sur le standard JPEG et intégrant un mécanisme d'allocation binaire. L'idée est d'appliquer une version optimisée de JPEG sur la région importante, contenant l'objet d'intérêt, et d'appliquer JPEG classique sur la partie restante. La solution de codage proposée est ensuite évaluée et comparée à différentes méthodes de référence.

2 Travail proposé

Dans le cadre de ce travail, nous nous proposons d'appliquer un codeur orienté objet du type JPEG sur des hologrammes générés par ordinateur, à base de décalage de phase. L'algorithme de codage proposé intègre un processus d'allocation binaire qui sert à optimiser la performance du codage de la région d'intérêt. A la différence de JPEG classique qui génère une matrice de quantification par facteur de compression, l'algorithme utilisé construit des tables de quantification personnalisées, en se basant sur les valeurs estimées de la distorsion et du débit dans chacune des 63 localisations fréquentielles des coefficients AC. L'objectif étant de définir la table de quantification Q optimale pour un bloc de l'image sous une contrainte préfixé d'un budget binaire R^*. La table de quantification Q sélectionnée est celle qui satisfait la contrainte budgétaire : $R(Q) \leq R^*$, tout en minimisant la distorsion estimée du bloc, $D(Q)$. Le problème d'optimisation du compromis débit/distorsion est alors formulé par la minimisation de l'équation lagrangienne.

Nous avons évalué la performance du codeur sur deux représentations holographiques différentes : à distances décalées (D1/D2) et Réelle/Imaginaire (Rea/Ima). Chacune de ces représentations nécessite deux fichiers distincts pour la description d'un hologramme. Nous avons choisi de coder séparément les deux composantes holographiques en définissant le même débit binaire pour chacune. Le jeu de test utilisé est composé de 3 CGH (Venus8KS, Ball8KS et Earth8KS) choisis à partir de la base de test Interfere-II. La reconstruction des hologrammes après la procédure de compression\décompression s'est basée sur la méthode Angular Spectrum Method (ASM).

3 Résultats

Afin d'évaluer la performance du codeur, les différents hologrammes reconstruits après décompression sont comparés (tableau 1) aux résultats obtenus par le codeur HEVC en termes de Delta-PSNR (en utilisant le modèle

	BD-PSNR				
Model	**HEVC-JPEG-Optimisé**	**HEVC-JPEG**	**HEVC-JPEG2000**	**HEVC-QT-L**	**HEVC-SPIHT**
Venus8KS 8192x8192	1.15	2.81	1.23	1.27	1.24
Ball8KS 8192x8192	2.2	5.15	3.94	3.31	1.6
Earth8KS 8192x8192	2.31	4.68	2.41	2.13	0.95
Moyenne	1.88	**4.21**	2.52	2.23	**1.26**

TABLE 1 – Evaluation de la qualité de reconstruction par rapport à HEVC (en termes de BD-PSNR).

BjØntegaard), avec un débit variant entre 0.1 bpp et 1 bpp. Globalement Le codeur HEVC surpasse les différents codeurs et ce, pour la représentation (Réelle/Imaginaire) mais ceci au prix d'une complexité calculatoire exorbitante. D'après les mêmes résultats, reportés dans le tableau 1, nous pouvons constater que la performance de notre codeur est comparable à celles des codeurs à base d'ondelettes et dépasse celles des codeurs intra-bandes (JPEG2000 et Q-TL). Sur la base des images illustrées dans la figure 1, il est bien clair que le codeur proposé offre la meilleure qualité visuelle de la région sélectionnée. Ceci n'est pas surprenant du fait que notre méthode utilise une allocation binaire presque optimale sur la région sélectionnée tandis que HEVC intègre un processus d'allocation binaire qui se charge de la régulation dynamique de plusieurs paramètres de codage qui n'est pas profitable pour le codage des données statiques.

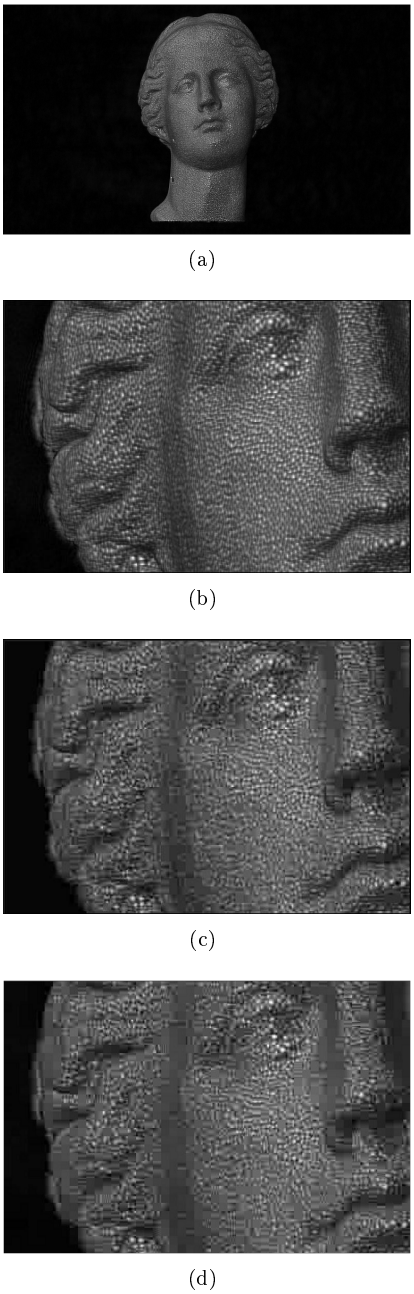
4 Conclusion et perspectives

Dans ce travail, nous avons démontré que le codeur JPEG-optimisé offre des performances en compression compétitives avec celles offertes par HEVC et SPIHT tout en bénéficiant de la rapidité et la simplicité de l'algorithme de compression JPEG. La méthode proposée est donc une alternative tout à fait valable pour la compression des hologrammes numériques et peut contribuer au développement du standard JPEG Pleno d'autant plus que la sortie du codeur JPEG-optimisé est compatible avec le flux binaire JPEG.

Références

[1] F. Dufaux, Y. Xing, B. Pesquet-Popescu, and P. Schelkens, "Compression of digital holographic data : an overview," in SPIE Proc. Applications of Digital Image Processing XXXVIII, San Diego, CA, USA, Aug. 2015.

[2] JPEG PLENO Abstract and Executive Summary, ISO/IEC JTC 1/SC 29/WG1 N6922, Sydney, Australia, 2015.



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

FIGURE 1 – Comparaison visuelle de la région d'intérêt (b), extraite du modèle décompressé Venus8KS (a), après une compression à 0.1 bps en utilisant notre méthode (c) et HEVC (d).

Représentations arithmétiques flottantes de taille réduite pour le Deep Learning

Diana Resmerita[1,2], Rodrigo Cabral Farias [1], Benoit Dupont de Dinechin [2], Lionel Fillatre [1]
[1] Université Côte d'Azur, CNRS, I3S
[2] Kalray

Résumé : *Les réseaux de neurones sont utilisés de plus en plus pour des applications embarquées telle que la détection d'obstacles pour les voitures autonomes. Les systèmes embarqués ont généralement des ressources limitées. Par conséquent, il est nécessaire de réduire la taille mémoire et les coûts énergétiques. Dans cet article, on s'intéresse à des nouveaux formats numériques à virgule flottante : posit, bfloat16 et msfp8. Ces différentes représentations arithmétiques sont utilisées pour le stockage des paramètres d'un réseau de neurones convolutifs. Cet article propose une analyse numérique sur la justesse des réseaux de neurones profonds compressés. Il montre que le format bfloat16 donne une meilleure justesse que le format classique FP16. Le format posit peut être utilisé sans trop modifier la justesse pour des réseaux de neurones convolutifs classiques comme VGG16. Globalement, msfp8 donne de mauvais résultats expérimentaux. Il n'est donc pas adapté aux types de réseaux de neurones utilisés.*

Mots-clés : Réseau de neurones profonds, Compression, Représentation Arithmétique.

1 Introduction

Les architectures de réseaux de neurones profonds (RNPs) évoluent pour résoudre de nouvelles classes de problèmes. L'exécution des modèles nécessite des ressources de calcul et de mémoire considérablement accrues. Dans le domaine de la vision par ordinateur, les principales applications des techniques d'apprentissage profond sont l'inférence pour la classification d'images, la segmentation sémantique et la détection d'objets. Les modèles de classification, de segmentation et de détection s'appuient fortement sur les couches de convolution (CONV) et les couches entièrement connectées (FC), qui représentent la majeure partie des ressources impliquées lors de l'inférence.

L'arithmétique à virgule flottante 32 bits (FP32) est traditionnellement utilisée pour l'inférence des réseaux de neurones profonds. Cependant, la représentation standard IEEE a une plage dynamique très large, bien plus grande que nécessaire pour les RNPs. Il a été observé que des économies importantes en termes d'empreinte mémoire et des augmentations de performance/efficacité peuvent être réalisées en utilisant des représentations 16 bits pour l'apprentissage [7] et des représentations 8 bits pour l'inférence avec une perte de précision acceptable [5]. Actuellement, l'approche la plus étudiée pour la compression des réseaux est la quantification des paramètres en virgule fixe [2]. Un modèle à virgule flottante peut être quantifié en un modèle à virgule fixe avec presque aucune perte de précision. Une telle approche apporte plusieurs avantages : l'empreinte mémoire est plus petite (on réduit la taille des données), le transfert est plus rapide et moins de mémoire vive et cache sont requises. Par conséquent, la consommation énergétique est réduite. En revanche, les données vont perdre en précision et les réseaux risquent d'avoir plus d'erreurs de classification.

Récemment, des nouveaux formats de données ont été introduits grâce à leur efficacité au niveau matériel. Le format bfloat16 (BF16) [6] est une représentation en FP32, tronqué en 16 bits. Il conserve les caractéristiques d'un FP32, mais ne prend en charge qu'une mantisse de 7 bits. Msfp8 [3] est un format de données proposé par Microsoft. Il représente l'équivalent du FP16, tronqué en 8 bits, avec seulement 2 bits de mantisse. Un autre type de données, appelé posit, a été introduit par Gustafson et al. [4]. Le type posit a été conçu pour remplacer les formats de type float. Un nombre posit $<n, es>$ de n bits comprend 4 éléments : 1 bit de signe, r bits de régime qui peut être vu comme une base, es bits d'exposant et f bits de mantisse, tandis qu'un nombre flottant n'a pas de régime.

2 Travail proposé

L'objectif de ce travail est la compression des paramètres pour le stockage des réseaux de neurones profonds. On s'intéresse à l'utilisation des formats numériques de taille réduite qui sont plus adaptés aux exigences des RNPs : BF16, msfp8 et posit. Par conséquent, on transforme les paramètres d'un réseau entrainé d'un format FP32 vers d'autres formats de taille réduite.

On identifie des avantages pour chacun des trois formats étudiés. Les trois représentations arithmétiques occupent moins de place en mémoire qu'un FP32. BF16 est attirant pour l'apprentissage profond parce qu'il peut représenter la même plage de valeurs que le FP32, mais la conversion avec FP32 est plus rapide. L'implémentation d'un tel type est peu coûteuse. Msfp8 est l'équivalent de bfloat16 pour FP16 et présente les mêmes avantages que bfloat16 mais avec une taille plus petite et une précision plus limitée. Les posits offrent une meilleure plage dynamique, une précision plus grande et une meilleure cohérence entre les machines tout en étant plus simples au niveau matériel. Ils sont donc plus économiques en consommation énergétique. On choisit de faire l'évaluation de posit sur 8 bits, avec un exposant es qui varie entre 0 et 3. On a remarqué que posit $<8, 0>$ et posit $<8, 1>$ sont représentables exactement en FP16. Posit $<8, 2>$ a 8 valeurs de plus grande magnitude qui ne sont pas représentables en FP16, mais en

DNNs		FP32	FP16	BF16	msfp8	posit			
						<8, 0>	<8, 1>	<8, 2>	<8, 3>
VGG16	ACC	70.6	70.6	70.8	69.7	10.2	70.8	70.5	70
	MSE	-	1.33E-12	3.58E-10	3.08E-07	8.51E-06	2.38E-07	8.02E-08	1.09E-07
ResNet50	ACC	75.7	71.3	75.5	62.8	0	27.7	73.2	66
	MSE	-	1.29E-10	3.46E-09	3.04E-06	644.0062	502.5577	29.99232	17.8428
InceptionV3	ACC	71.1	71.1	71.3	44.8	65.1	69.4	69.7	63.1
	MSE	-	6.71E-11	6.4E-09	5.64E-06	2.07E-05	1.87E-06	2.01E-06	4.49E-06
Xception	ACC	73.5	73.4	73.6	37.5	70.6	72.4	72.1	63.8
	MSE	-	7.02E-08	2.83E-08	2.49E-05	0.79345	0.16113	0.01105	0.01107
YOLO	mAP	0.41595	0.41595	0.41585	0.3022	0.4025	0.4155	0.411	0.394
	MSE	-	1.44E-11	3.67E-09	3.23E-06	1.87E-05	1.02E-06	7.10E-07	9.39E-07

TABLE 1 – Résultats pour les réseaux de classification et un réseau de détection. La conversion est appliquée à tous les paramètres.

BF16. Pour posit <8, 3>, 46 valeurs ne sont pas représentables en FP16 et 12 valeurs ne sont pas représentables en BF16. FP16 et BF16 ne peuvent pas représenter les plus petites magnitudes, et FP16 ne peut pas représenter les plus grandes magnitudes. Dans nos expérimentations, BF16 est considéré comme étant un FP32 avec les 16 derniers bits figés à 0 et le msfp8 est un FP16 avec les 8 derniers bits à 0. Les valeurs de posit8 ont été obtenues en s'appuyant sur différentes implémentations [1, 4].

3 Résultats

Les expérimentations ont été réalisées sur 13 réseaux de classification et 1 réseaux de détection d'objets. Différents critères d'évaluation ont été étudiés comme l'Accuracy (ACC) pour la classification, le Mean Average Precision (mAP) pour la détection, et le Mean Square Error (MSE). La Table 1 contient les résultats obtenus pour 4 réseaux de classification (VGG16, ResNet50, InceptionV3, Xception) qui ont des architectures différentes et le réseau de détection d'objets (YOLO). Nous avons également affiché les résultats obtenus avec FP32 et FP16 pour pouvoir comparer avec la représentation virgule flottante standard. À noter que les opérations sont faites en FP32. Pour chaque réseau et chaque format de données, 2 valeurs sont affichées. La première valeur représente l'accuracy du réseau et la deuxième est le MSE entre les paramètres au format FP32 et les mêmes paramètres au format de taille réduite.

Globalement, la compression avec BF16 donne des meilleurs résultats qu'avec FP16. Malgré une précision plus faible, elle est suffisante pour les réseaux de neurones. En revanche, pour msfp8, le manque de précision conduit à une perte importante sur la sortie de tous les réseaux testés. Les formats posit <8, 0> et posit <8, 3> ne donnent pas de bons résultats mais produisent une perte négligeable pour les réseaux de classification classique (comme VGG16) et de détection. Tandis que pour les réseaux qui contiennent des couches de normalisation (ResNet50, InceptionV3, Xception), la perte est plus importante.

4 Conclusion et perspectives

Cet article compare trois représentations arithmétiques de taille réduite en vue de diminuer le stockage en mémoire des paramètres. BF16 donne des bons résultats pour tous les types des réseaux de neurones. Posit avec 1 bit ou 2 bits d'exposant se comporte bien avec la plupart des réseaux de neurones. Pour la suite, on envisage de changer la représentation arithmétique des activations et de comparer nos résultats avec la quantification en 8 bits à virgule fixe. Nos prochains travaux s'intéresseront également à la formalisation de l'impact du codage.

Références

[1] `https://posithub.org/docs/PDS/PositEffortsSurvey.html`.

[2] Jian Cheng, Pei-song Wang, Gang Li, Qing-hao Hu, and Han-qing Lu. Recent advances in efficient computation of deep convolutional neural networks. *Frontiers of Information Technology & Electronic Engineering*, 19(1) :64–77, 2018.

[3] Eric Chung, Jeremy Fowers, Kalin Ovtcharov, Michael Papamichael, Adrian Caulfield, Todd Massengill, Ming Liu, Daniel Lo, Shlomi Alkalay, Michael Haselman, et al. Serving dnns in real time at datacenter scale with project brainwave. *IEEE Micro*, 38(2) :8–20, 2018.

[4] John L Gustafson and Isaac T Yonemoto. Beating floating point at its own game : Posit arithmetic. *Supercomputing Frontiers and Innovations*, 4(2) :71–86, 2017.

[5] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. G. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *IEEE CVPR 2018*, pages 2704–2713, 2018.

[6] David Lutz. Arm floating point 2019 : Latency, area, power. In *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, pages 97–98. IEEE, 2019.

[7] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *ICLR (Poster)*. OpenReview.net, 2018.

An Analytical Model of the End-to-End Performance for Linear Video Delivery Under Bandwidth Constraints

Anthony TRIOUX, Mohamed GHARBI, François-Xavier COUDOUX, Patrick CORLAY
UMR 8520 - IEMN, DOAE, Univ. Polytechnique Hauts-de-France, CNRS, Univ. Lille, YNCREA, Centrale Lille, F-59313 Valenciennes, France

Abstract: *Recently, Linear Video Coding and Transmission (LVCT) schemes have been proposed as an alternative to traditional video transmission schemes, the latter experiencing cliff-effect [3] in wireless error-prone environments. In this paper, we propose an analytical model to estimate/predict the end-to-end performance of SoftCast [2] video transmission, pioneer of the LVCT schemes. This model, based on the PSNR metric, accounts for channel conditions (CSNR) as well as data compression due to bandwidth constraints. We show that regardless of the available channel bandwidth, the end-to-end video quality can be accurately modeled and predicted according to the characteristics of the video and the channel conditions.*

Keywords: Wireless Video Transmission, SoftCast, Distortion Model, Bandwidth constraints, Linear Video Delivery.

1 Introduction

The basic scheme of SoftCast [2] is introduced in Fig. 1. SoftCast first takes a Group of Pictures (GoP) and uses a 3D full-frame DCT as a decorrelation transform. These DCT frames are then divided into N small rectangular blocks of transformed coefficients called *chunks*. In the SoftCast scheme, the data compression can be done after the decorrelation transform. Specifically, when the available channel bandwidth for the transmission is less than the signal bandwidth, SoftCast discards chunks with less energy. This is generally the case especially for the transmission of High Definition (HD) content as mentioned in [5, 6]. At the receiver side, these discarded chunks are replaced by null values [2].

For ease of reading, the compression level denoted hereafter CR is usually used in SoftCast and in LVCT schemes. It is defined as follows [4]:

$$\mathrm{CR} = \frac{K}{N} \tag{1}$$

where K represents the number of transmitted chunks per GoP and N the total number of chunks within a GoP. This ratio is between 0 (no information transmitted, i.e., $K = 0$) and 1 (no compression, i.e., $K = N$).

The third block at the transmitter level called Power Allocation or Scaling is used to provide error resilience. SoftCast scales the magnitude of the DCT coefficients to offer a better protection against transmission noise. Since the total transmission power available P is limited and fixed, it is distributed to all the chunks in a way that minimizes the Mean Square reconstruction Error (MSE) between transmitted and decoded chunks. This is a typical Lagrangian problem which leads to the following solution [1, 2] given by:

$$g_i = \lambda_i^{-1/4} \cdot \sqrt{\frac{P}{\sum_i \sqrt{\lambda_i}}}, \tag{2}$$

where $g_i, i = 1, 2, \ldots, N$ is the scaling factor for the i^{th} chunk, and λ_i the energy of the ith transmitted coefficient (after 3D-DCT) [2].

The Hadamard transform is then applied to the scaled chunks to provide packet loss resilience. This process transforms the chunks into *slices*. Each slice is a linear combination of all scaled-chunks.

Finally, these packets are transmitted in a pseudo-analog manner using Raw-OFDM [2], i.e., classical coding (e.g., Forward Error Correction code) and modulation stages are skipped.

In parallel, the SoftCast transmitter sends an amount of data referred as metadata. These data consist of the mean and the variance of each transmitted chunk as well as a bitmap, which indicates the positions of the discarded chunks into the GoP. Metadata are strongly protected and transmitted in a robust way (e.g., BPSK [2]) to ensure correct delivery and decoding process.

At the receiver side, if an estimation of the channel noise is available, a Linear Least Square Error (LLSE) decoder can be used to get the best estimation of the received values. Otherwise, a Zero-Forcing (ZF) decoder is used. Using the metadata, the decoded values are then reassembled to form DCT-frames, which are then passed through an inverse 3D-DCT process.

In a recent paper [8], Xiong et al. modeled the end-to-end performance of SoftCast for any channel Signal-to-Noise Ratio (CSNR, expressed in decibels).

They showed that the total distortion that affects the reconstructed video quality without data compression can be obtained from:

$$\mathrm{D}_{[\mathrm{ZF/FB}]} = \sum_{i=1}^{N} D_i = \frac{\sigma_n^2}{P} \left(\sum_{i=1}^{N} \sqrt{\lambda_i} \right)^2, \tag{3}$$

where σ_n^2 is the noise power.

Based on the following definition of the CSNR and PSNR expressed in decibels:

$$\mathrm{CSNR} = 10\log_{10}(\bar{P}/\sigma_n^2), \quad \bar{P} = P/N, \tag{4}$$

$$\mathrm{PSNR} = 10\log_{10}(255^2/\bar{D}), \quad \bar{D} = \mathrm{D}_{[\mathrm{ZF/FB}]}/N. \tag{5}$$

They showed that the expected reconstructed video quality can be finally obtained from:

$$\mathrm{PSNR}_{[\mathrm{ZF/FB}]} = c + \mathrm{CSNR} - 20\log_{10}\left(H\right), \tag{6}$$

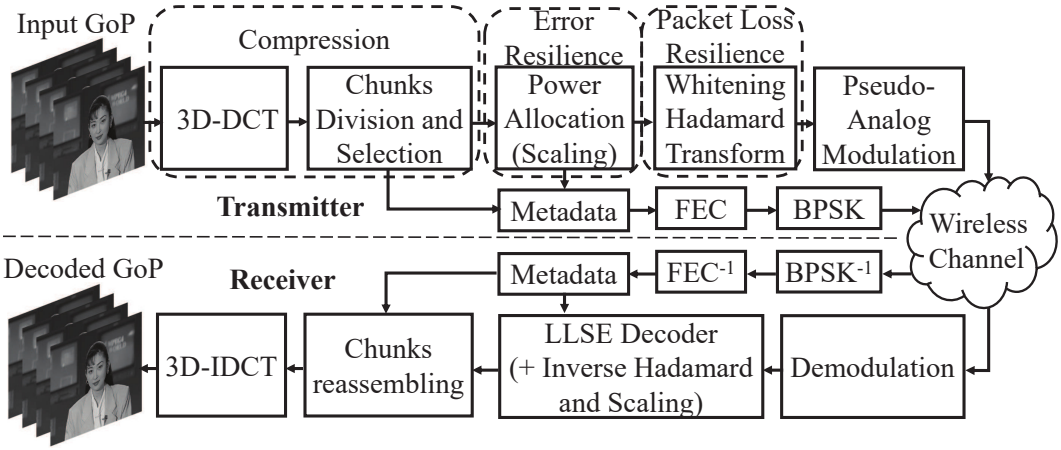


Figure 1: Block diagram of the SoftCast scheme.

where $c = 20\log_{10}(255)$ and

$$H = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\lambda_i}, \tag{7}$$

refers to the *data activity* of the video content [8]. The higher the data activity H, the lower the reconstructed PSNR, showing the importance of taking into account the characteristics of the transmitted video content in a SoftCast context. Note the linear characteristic of the $\mathrm{PSNR}_{[\mathrm{ZF/FB}]}$ that depends on the channel transmission conditions.

We note that Xiong's model relies on two assumptions: the first is that the available bandwidth of the application allows the transmission of the N elements of $\mathbf{x}$ (*i.e.*, CR=1, no compression applied). However, in practice, this is generally not the case since bandwith ressources are limited. This is especially true when considering the transmission of high resolution (HD, 4K, etc.) video formats. The second hypothesis assumes that a ZF estimator used at the receiver side. This is not valid when considering the original SoftCast scheme proposed by [2] which uses an LLSE decoder.

2 Proposed work

The first objective of this work is to consider the more realistic and general case *i.e.* only the $K \leq N$ largest energy chunks are transmitted due to bandwidth constraints.

The challenge consists in proposing a more realistic theoretical model than Xiong's model [8], i.e., addressing the weaknesses of the initial model (inaccurate prediction for bandwidth constrained applications).

Since $N-K$ chunks are discarded due to bandwidth constraints, the total distortion $\mathrm{D}_{[\mathrm{ZF/CB}]}$ now consists of two parts:

- The distortion D_i that affects each of the K transmitted coefficients x_i, given by: $D_i = E\left[(\hat{x}_i - x_i)^2\right]$. For ease of reading, let us denote the total distortion due to the transmitted coefficients $D_s = \sum_{i=1}^{K} D_i$.
- The distortion D_j due to each of the $N-K$ discarded coefficient x_j, given by: $D_j = E\left[(0 - x_j)^2\right]$.

Likewise, we denote the total distortion due to the discarded coefficients $D_d = \sum_{j=K+1}^{N} D_j$.

Therefore, the overall distortion (3) becomes:

$$\begin{aligned}\mathrm{D}_{[\mathrm{ZF/CB}]} &= D_s + D_d, \\ &= \frac{\sigma_n^2}{P}\left(\sum_{i=1}^{K}\sqrt{\lambda_i}\right)^2 + \sum_{j=K+1}^{N}\lambda_j.\end{aligned} \tag{8}$$

We note that the average transmission power in (4) becomes $\bar{P} = P/K$ as the total transmission power is here distributed over the K transmitted coefficients and in (5) $\bar{D} = \mathrm{D}_{[\mathrm{ZF/CB}]}/N$.

By inserting (8) into (5), we get:

$$\begin{aligned}\mathrm{PSNR}_{[\mathrm{ZF/CB}]} &= 10\log_{10}\left(\frac{255^2 \cdot N}{D_s + D_d}\right), \\ &= c - 10\log_{10}\left(1 + \frac{D_d}{D_s}\right) + 10\log_{10}\left(\frac{\bar{P}}{\sigma_n^2}\right) \\ &\quad - 10\log_{10}\left(\frac{1}{NK}\left(\sum_{i=1}^{K}\sqrt{\lambda_i}\right)^2\right).\end{aligned} \tag{9}$$

By analogy with (6), we identify the new data activity of the transmitted coefficients as:

$$H_t = \frac{1}{\sqrt{NK}}\sum_{i=1}^{K}\sqrt{\lambda_i}. \tag{10}$$

For ease of reading we also define E_d, the overall energy of all dropped coefficients:

$$E_d = \frac{1}{N}\sum_{j=K+1}^{N}\lambda_j. \tag{11}$$

According to these new definitions, the end-to-end video quality considering bandwidth constraints for the ZF estimator is finally given by:

$$\begin{aligned}\mathrm{PSNR}_{[\mathrm{ZF/CB}]} &= c + \mathrm{CSNR} - 20\log_{10}\left(H_t\right) \\ &\quad - 10\log_{10}\left(1 + \frac{\mathrm{CSNR}_{lin} \cdot E_d}{{H_t}^2}\right).\end{aligned} \tag{12}$$

where $\mathrm{CSNR}_{lin} = \frac{\bar{P}}{\sigma_n^2}$.

The above equation includes a new term in comparison to (6) that reflects the effect of the data compression applied. The PSNR now depends on three parameters: first, the CSNR which depends on the transmission conditions, and then the two other terms E_d and H_t that are directly related to the video data characteristics. For a given bandwidth, the higher E_d, the greater degradation. However, as E_d is multiplied by the CSNR_{lin}, the degradation becomes less noticeable at low CSNR environments.

When $K = N$, *i.e.*, CR=1, (12) and (6) are identical. In other words, the video quality scales linearly with the CSNR as stated in [8].

3 Results

To evaluate the effectiveness of the proposed model, we perform full end-to-end simulations. We create a Mixed sequence by slicing the first 128 frames of ten HD 720p sequences (*Ducks, Four People, In to tree, Johnny, Kristen and Sara, Old town, Parkjoy, Parkrun, Shields* and *Stockholm*). We use GoPs of 16 frames and divide each frame into 64 chunks as it represents the original and mostly used configuration [2]. We verified similar results for other GoP-sizes and chunk-sizes. We then consider four different compression ratio $\text{CR} = 1, 0.75, 0.5, 0.25$.

As observed in Fig. 2, the proposed model (colored lines) perfectly matches the simulation results (colored dots), regardless of the considered CR or CSNR values. When CR=1 (red color), we logically obtain the same linear characteristic as in [8].

However, Xiong's model represented with the red line [8] is no longer valid when the available channel bandwidth decreases (cyan, green and blue dots) as the data compression is not considered. In practice, it is mandatory to consider such loss since it drastically degrades the received video quality and implies non-linear characteristics at high CSNR values. This is the well-known leveling-off effect [5] that appears and implies huge decibel losses (e.g. ΔPSNR up to 20dB for the considered case). Unlike Xiong's model, ours (colored lines) perfectly predicts and models this leveling-off effect regardless of the amount of discarded chunks.

4 Conclusion and perspectives

In this paper, we present a theoretical model that can be used in the context of linear video delivery under bandwidth constraints. In contrast to Xiong's model [8], the proposed model takes into account the losses due to data compression/bandwidth constraints. Regardless of the available channel bandwidth, results show that the model accurately represents the full end-to-end performance by predicting the leveling-off [5] effect that appears when some chunks are discarded. This model can help for parameters optimization in an LVCT transmission context subject to bandwidth limitations as in [6]. It can also be used to quickly evaluate schemes without requiring extensive end-to-end simulations. Further works concern the extension of the model to different versions of SoftCast that bring additional PSNR gain (e.g., by taking into account the LLSE estimator) or the extension of the model to other objective metrics such as SSIM [7].

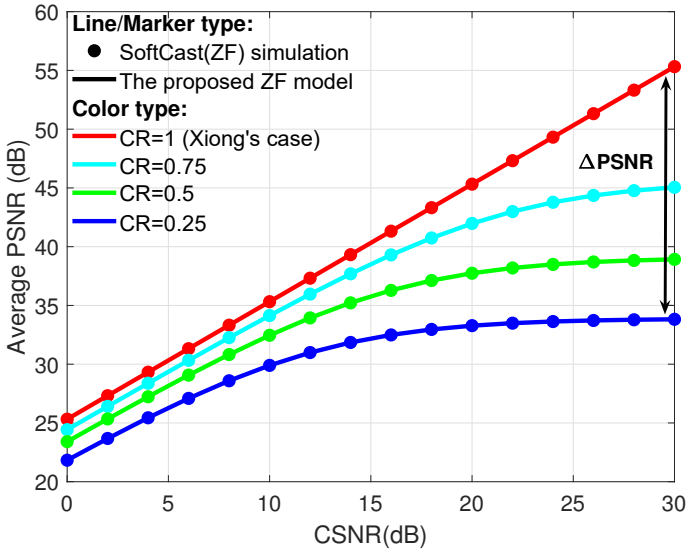


Figure 2: Average PSNR results for the proposed theoretical model (solid lines) and SoftCast simulations with ZF estimator: (dots markers) for the *Mixed HD720p* sequence. Configuration: GoP-size=16 frames, 64 chunks/frame. The colors red, cyan, green and blue represent CR=1, 0.75, 0.5 and 0.25, respectively.

References

[1] T. Fujihashi, T. Koike-Akino, T. Watanabe, and P. V. Orlik. High-Quality Soft Video Delivery With GMRF-Based Overhead Reduction. *IEEE Transactions on Multimedia*, 20(2):473–483, February 2018.

[2] Szymon Jakubczak and Dina Katabi. SoftCast: Clean-slate scalable wireless video. *MIT Technical report*, February 2011.

[3] S. Kokalj-Filipović and E. Soljanin. Suppressing the cliff effect in video reproduction quality. *Bell Labs Technical Journal*, 16(4):171–185, March 2012.

[4] Zexue Li, Hancheng Lu, and Yanglong Wu. Compressed uncoded screen content video transmission in bandwidth-constrained wireless networks. In *IEEE Int. Conf. Wireless Commun. & Signal Process. (WCSP)*, pages 1–5, November 2016.

[5] F. Liang, C. Luo, R. Xiong, W. Zeng, and F. Wu. Superimposed Modulation for Soft Video Delivery with Hidden Resources. *IEEE Trans. Circuits Systems Video Technol.*, 28(9):2345–2358, September 2018.

[6] Anthony Trioux, François-Xavier Coudoux, Patrick Corlay, and Mohamed Gharbi. Temporal information based GoP adaptation for linear video delivery schemes. *Signal Processing: Image Communication*, 82:115734, March 2020.

[7] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, April 2004.

[8] Ruiqin Xiong, Feng Wu, Jizheng Xu, Xiaopeng Fan, and al. Analysis of decorrelation transform gain for uncoded wireless image and video communication. *IEEE Trans. Image Process.*, 25(4):1820–1833, April 2016.

Voxel-based Deep Point Cloud Geometry Compression

Giuseppe Valenzise, Maurice Quach, Dat-Than Nguyen, Frédéric Dufaux
Université Paris-Saclay, CNRS, CentraleSupelec, Laboratoire des Signaux et Systèmes (L2S)

Abstract : *We present two learning-based methods for coding point clouds geometry. The two methods target lossy and lossless compression, respectively, and have in common the fact to use a voxel-based representation of geometry. This representation enables us to extend well-known architectures used for 2D image generation and compression to 3D. We show that, when the point cloud density is sufficiently high, the voxel-based approach achieves state-of-the-art performance compared to conventional octree-based methods such as MPEG G-PCC.*

Keywords: point cloud, geometry coding, voxels, auto-encoder, generative models.

1 Introduction

Due to recent advances in visual capture technology, point clouds have been recognized as a crucial data structure for 3D content. In particular, point clouds are essential for numerous applications such as virtual and mixed reality, sensing for autonomous vehicle navigation, architecture and cultural heritage, etc. Point clouds are sets of 3D points identified by their coordinates, which constitute the geometry of the point cloud. In addition, each point can be associated with attributes like colors, normals and reflectance. Point clouds can have a massive number of points, especially in high precision or large scale captures. This entails a huge storage and transmission cost. As a result, Point Cloud Compression (PCC) is fundamental in practice. The Moving Picture Experts Group (MPEG) has recently released two PCC standards [1]: Geometry-based PCC (G-PCC) and Video-based PCC (V-PCC). G-PCC approaches PCC from a 3D perspective and compresses point clouds in their native form using 3D data structures such as octrees. On the other hand, V-PCC approaches PCC from a 2D perspective, projects 3D data onto a 2D plane and makes use of video compression technology. Recently, deep point cloud compression (DPCC) methods have been proposed and shown to provide significant coding gains compared to traditional methodologies [2, 3, 4, 5].

In this paper, we focus on the compression of point cloud geometry, and we review two recently proposed learning-based methods for lossy and lossless coding. We consider the case of voxelized point clouds. Voxelization is the process that quantizes the coordinates of a point cloud to integer precision prior to the coding process and is typically applied in most codecs to discretize the geometry. We also make the implicit hypothesis that the point cloud is dense enough to exhibit local correlations among neighboring points on the voxel grid – in other terms, we assume there is not too much "empty space" between points. This enables us to employ deep neural networks with voxel-based 3D convolutions (see, e.g., [2]), which have been shown to be particularly effective in point cloud compression. On the other hand, point-based convolutions [6, 7] are also possible [8], but their performance in PCC is still lagging behind traditional hand-crafted methods such as those used in MPEG G-PCC.

When a point cloud is voxelized, its geometry can be expressed as a binary signal over the voxel grid. In particular, a voxel is consider occupied if it contains at least one point, and is non occupied otherwise. Based on this observation, learning-based methods for geometry coding typically cast decoding as a binary classification problem, see Section 2. Instead, in lossless compression an explicit, accurate estimation of the likelihood of voxel occupancy is necessary: in this case, the decoding can be interpreted as a voxel generation process, as we will see in Section 3.

2 Lossy compression

Our lossy compression scheme is inspired by the success of variational auto-encoder (VAE) methods for image compression [9, 10]. The general architecture of a VAE-based codec is illustrated in Figure 1. An input signal x (pixels for the case of 2D images, binary voxel occupancies for 3D PCs) is transformed by an *analysis* network f_a into a latent representation y and quantized into $\tilde{y}$. This is later used as input to a *synthesis* network, producing an approximated reconstruction $\tilde{x}$ of the original signal. The quantizer Q represents the main difference with respect to a conventional VAE, in which the latent space is typically continuous. Since quantization is not differentiable, several approximations have been proposed; in this work, we replace quantization noise by uniform noise during training, as initially suggested in [9]. Quantization, along with the fact that y has smaller dimensionality compared to x, contribute both to achieve compression. The quantized latent code $\tilde{y}$ is entropy coded and transmitted as bitstream. A basic version of the VAE-based codec assumes that the components of $\tilde{y}$ are i.i.d. and computes symbol probabilities for entropy coding accordingly. However, this has been shown to be suboptimal, and later versions introduce a *hyperprior* model [10], where the probability of the quantized latent variables is also modeled through a VAE. This allows the codec to capture the residual spatial dependencies among voxels.

The model is trained end-to-end using a $D + \lambda R$ loss function, for a given value of λ. The rate term includes both the bits for the latent variables $\tilde{y}$ and $\tilde{z}$, which are approximated by their differential entropy at training time. The D part is computed using the *focal loss*, a variant of the binary cross-entropy loss [11] used in classification, which has been shown to be more effective when the class distribution is strongly unbalanced (in the case of PC, most of the voxels are empty). The output of the VAE is

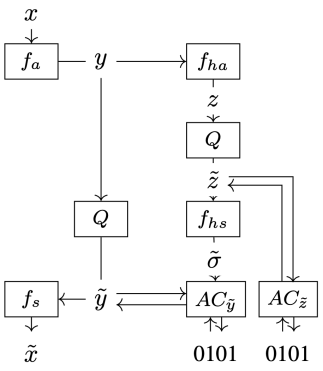


Figure 1: Scheme of a VAE-based codec with hyperprior.

Point cloud	Metric	BD-PSNR
loot	D1	5.91 / 6.99
	D2	6.87 / 6.13
redandblack	D1	5.01 / 6.48
	D2	5.93 / 5.63
longdress	D1	5.55 / 6.94
	D2	6.60 / 6.01
soldier	D1	5.57 / 6.93
	D2	6.57 / 6.04
Average	D1	5.51 / 6.83
	D2	6.50 / 5.95

Table 1: RD performance of the proposed VAE-based codec compared to G-PCC (version 10.00). We specify BD-PSNR values (dB) compared to G-PCC trisoup and G-PCC octree in each cell (trisoup BD-PSNR / octree BD-PSNR).

a set of per-voxel probabilities of occupancy, which need to be thresholded in order to provide the final binary occupancy values.

The choice of the threshold is of paramount importance for coding performance. We choose to optimize this threshold at the encoder side and transmit it as a side information into the bitstream. Our codec operates on a block-by-block basis, in such a way that all decisions are taken locally and adapted to the spatially varying density in the point cloud. Further details about the architecture of the codec are reported in [3], and the code is publicly available as a toolbox [12].

2.1 Performance evaluation

We report RD performance on a subset of four dense point clouds. Details about the training dataset, as well as the training hyperparameters, are given in [3]. We evaluate the different conditions using G-PCC trisoup and octree as baselines. The octree is the basic coding structure of G-PCC: the point cloud is recursively subdivided into octants, and only those nodes that contain at least a point are further split. On top of the basic vanilla octree, G-PCC adds a number of additional modes and optimizations, including direct coding for isolated points, planar modes and sophisticated contexts for entropy coding (see [1] for a survey). The triangle soup (trisoup) mode, in particular, adds local triangular approximations at the octree leaves, and is typically included in PC compression benchmarks (although it is not included in the released standard). The distortion metrics D1 and D2 are obtained, respectively, from symmetrized point-to-point and point-to-distance mean squared errors, which are converted to PSNR using the original PC bit depth as peak error [13].

Table 1 reports Bjontegaard Delta PSNR of the proposed scheme compared to G-PCC trisoup and octree: the coding gains are significant for all the considered point clouds, demonstrating the potential of voxel-based convolutions in the compression of dense point clouds.

3 Lossless compression

Voxel-based convolutional architectures can be successfully used also for lossless geometry coding of dense point clouds. Compared to lossy compression, the goal here is to estimate accurately the voxel occupancy probabilities for entropy coding, rather than handling class imbalance to favor a precise binary reconstruction. In the following, we present briefly the *VoxelDNN* codec we proposed in [14], whose general architecture is illustrated in Figure 2.

The basic element of the codec is the context model, which is based on an auto-regressive generative model inspired by PixelCNN [15]. Specifically, let v_i denote the binary occupancy of a voxel i. We factorize the joint distribution $p(v)$ of a block of voxels v as a product of conditional distributions $p(v_i|v_{i-1},\dots,v_1)$ over the voxel volume: $p(v) = \prod_{i=1}^{N} p(v_i|v_{i-1}, v_{i-2},\dots,v_1)$, with N the number of voxels in a block. Each term $p(v_i|v_{i-1},\dots,v_1)$ is the probability of the voxel v_i being occupied given the occupancy of all *previous* voxels. An illustration is given in Figure 2(c). We approximate $\hat{p}(v_i|v_{i-1},\dots,v_1)$ using a convolutional neural network, which we train by minimizing the binary cross-entropy $\mathbb{E}_{v\sim p(v)}\left[\sum_{i=1}^{N} -\log \hat{p}(v_i)\right]$. This is equivalent to minimize the distance between the estimated conditional distributions and the real data distribution, yielding accurate context distributions for arithmetic coding. This process can be carried out on blocks of different sizes (typically, N ranges between 8^3 and 64^3), using a rate-optimization algorithm. Since empty blocks in a point cloud do not bring any useful context, we apply an octree-based partitioning to pre-process the PC and remove the non-occupied space. Further details about VoxelDNN are available in [14].

3.1 Performance evaluation

A comparison of the bitrates of VoxelDNN and G-PCC (v. 12) for various PC categories is reported in Table 2. We observe that VoxelDNN achieves significant gains of up to 37% on dense point clouds (MVUB and 8i). On sparser point clouds the gains are smaller, but still competitive compared to G-PCC. The only exception is the PC "Arco Valentino", which has large density variations and very sparse regions where context modeling is ineffective.

Notice that a drawback of VoxelDNN is the sequential decoding of voxels, which is equivalent to a sequential sampling from the voxel occupancy distribution. As a result, the decoding times are significantly higher than G-PCC. In a follow-up work [16], we propose a partial solution consisting in breaking some dependencies to parallelize decoding, achieving execution times withing one order of magnitude from G-PCC.

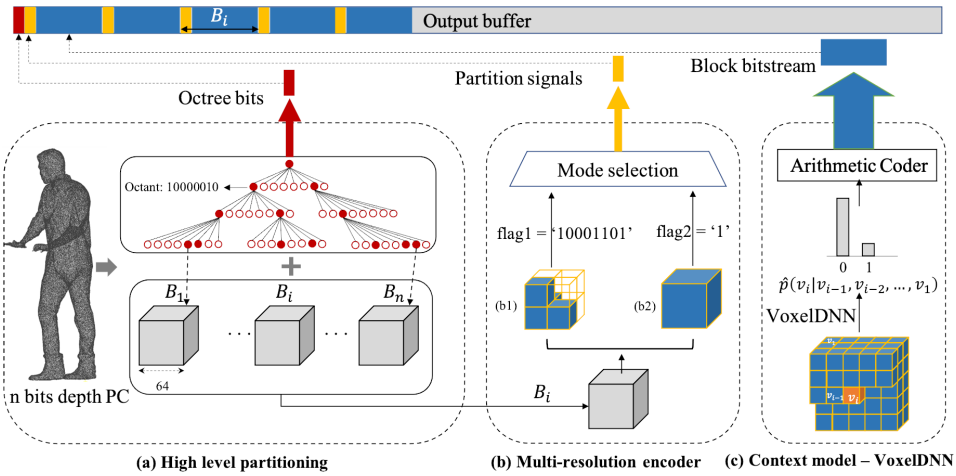


Figure 2: General architecture of the VoxelDNN codec, composed by a high-level octree partitioning part; a multi-resolution encoder; and the basic context model unit.

		G-PCC	VoxelDNN	
Dataset	Point Cloud	bpov	bpov	Gain over G-PCC
MVUB	Phil	1.1599	0.8252	-28.86%
	Ricardo	1.0673	0.7572	-29.05%
	Average	**1.1136**	**0.7912**	**-28.95%**
8i	Redandblack	1.0893	0.7003	-35.71%
	Loot	0.9524	0.6084	-36.12%
	Thaidancer	0.9990	0.6627	-33.66%
	Boxer	0.9492	0.5906	-37.78%
	Average	**0.9975**	**0.6405**	**-35.79%**
CAT1	Frog	1.8990	1.7071	-10.11%
	Arco Valentino	4.8531	4.9900	+2.82%
	Shiva	3.6716	3.5135	-4.31%
	Average	**3.4746**	**3.4035**	**-3.86%**
USP	BumbaMeuBoi	5.4068	5.066	-6.29%
	RomanOiLight	1.8604	1.6231	-12.76%
	Average	**3.6336**	**3.4855**	**-9.52%**

Table 2: Average rate in bits per occupied voxel (bpov) of proposed method and percentage reductions compared with MPEG G-PCC.

4 Conclusion and perspectives

We have described two deep learning-based architectures for point cloud geometry compression (lossy and lossless). In both cases, the use of voxel-based convolutions provides significant gains over the reference G-PCC solution for dense point clouds. We believe this advantage is given by the ability to represent the underlying geometric structure (local surfaces, objects, etc.), which is not captured by simple octree-based approaches. On the other hand, when the point cloud is sparser or scant (as for LiDAR data), voxel-based techniques break down, and other kinds of approaches are more suitable, such as point-based and graph convolutions [17]. Compression of very sparse PC is still an open challenge.

References

[1] C. Cao, M. Preda, V. Zakharchenko, E. S. Jang, and T. Zaharia, "Compression of sparse and dense dynamic point clouds—methods and standards," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1537–1558, 2021.

[2] M. Quach, G. Valenzise, and F. Dufaux, "Learning convolutional transforms for lossy point cloud geometry compression," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4320–4324, ISSN: 1522-4880.

[3] ——, "Improved deep point cloud geometry compression," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.

[4] A. F. R. Guarda, N. M. M. Rodrigues, and F. Pereira, "Deep learning-based point cloud coding: A behavior and performance study," in *2019 8th European Workshop on Visual Information Processing (EUVIP)*, pp. 34–39, ISSN: 2164-974X.

[5] J. Wang, H. Zhu, H. Liu, and Z. Ma, "Lossy point cloud geometry compression via end-to-end learning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[6] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[7] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] W. Yan, S. Liu, T. H. Li, Z. Li, G. Li *et al.*, "Deep autoencoder-based lossy geometry compression for point clouds," *arXiv preprint arXiv:1905.03691*, 2019.

[9] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *2017 5th International Conference on Learning Representations (ICLR)*.

[10] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *2018 6th International Conference on Learning Representations (ICLR)*.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, ISSN: 2380-7504.

[12] M. Quach, G. Valenzise, and F. Dufaux, "A Deep Point Cloud Geometry Coding Toolbox," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Shenzhen, China, Jul. 2021.

[13] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, "Geometric distortion metrics for point cloud compression," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 3460–3464. [Online]. Available: http://ieeexplore.ieee.org/document/8296925/

[14] D. T. Nguyen, M. Quach, G. Valenzise, and P. Duhamel, "Lossless Coding of Point Cloud Geometry using a Deep Generative Model," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[15] A. van den Oord and N. Kalchbrenner, "Pixel RNN," in *ICML*, 2016.

[16] D. T. Nguyen, M. Quach, G. Valenzise, and P. Duhamel, "Multiscale deep context modeling for lossless point cloud geometry compression," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Shenzhen (virtual), China, Jul. 2021.

[17] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Transactions On Graphics*, vol. 38, no. 5, pp. 1–12, 2019.

Analyse et réprésentation

Localisation rapide de falsifications dans des images 4K Ultra HD

Laura Bertojo, Olivier Strauss, William Puech
LIRMM, Université de Montpellier, CNRS, Montpellier, France

Résumé : *Le copié-déplacé est une technique de falsification d'image très répandue. Les méthodes de détection basées sur les points d'intérêt se sont révélées très efficaces pour identifier les zones ayant subi une attaque par copié-déplacé. Bien que ces méthodes soient rapides, la phase d'appariement des points caractéristiques a une complexité élevée. Dans cet article, une méthode avec un coût de calcul réduit par rapport aux méthodes classiques est étudiée. Celle-ci permet d'analyser des images 4K Ultra HD en un temps raisonnable. Après extraction des points d'intérêt de l'image, un nouvel algorithme de mise en correspondance est appliqué. Les différents clusters sont obtenus avec l'algorithme DBSCAN afin de générer un masque binaire. Les résultats expérimentaux montrent que la méthode est capable de détecter précisément les différentes falsifications sur de grandes images avec un temps de calcul fortement réduit.*

Mots-clés : CMFD, copié-déplacé, falsification, image, points d'intérêt, DBSCAN, SURF

1 Introduction

Chaque année, des trillions de photos numériques sont produites. Cette explosion de données a favorisé le développement des logiciels de traitement d'images tels que Gimp ou Photoshop permettant à n'importe qui d'effectuer des retouches indétectables à l'œil nu à moindre coût. De ce fait, avec ces logiciels permettant donc au grand public d'embellir des images et favorisant ainsi le partage, il est devenu difficile d'établir la fiabilité et l'authenticité des images diffusées en ligne. Si la plupart des manipulations effectuées restent anodines, certaines peuvent être effectuées à des fins malveillantes.

Dans ce contexte, il est devenu essentiel de pouvoir garantir l'authenticité d'une image mais aussi de pouvoir détecter avec précision de potentielles falsifications. Deux grands types d'approches ont été proposés pour résoudre ce problème. Les méthodes actives consistent à modifier l'image originale afin de détecter de futures modifications. Les méthodes passives, quant à elles, cherchent les falsifications uniquement à l'aide du contenu de l'image interceptée. Il existe aujourd'hui dans la littérature trois grandes catégories de falsification à savoir le copié-déplacé, le copié-collé et l'inpainting. Pour détecter ces falsifications, une première approche consiste à utiliser des incohérences statistiques de l'image afin d'identifier les zones impliquées dans une falsification. Ce type de procédure est inefficace dans le cadre de manipulations de type copié-déplacé. En effet, comme les zones falsifiées proviennent de la même image, ces zones là restent cohérentes par rapport au reste de l'image en terme de bruit, luminosité, couleurs et autres propriétés de l'image. Une approche plus efficace pour détecter un copié-déplacé consiste à chercher des similarités au sein d'une même image. Il existe dans la littérature deux types d'approches pour traiter ce type de falsification. La première consiste à découper l'image en différents blocs superposables [6, 3] tandis que la deuxième consiste à analyser que les zones caractéristiques de l'image en détectant, caractérisant et comparant les points d'intérêt de l'image [7, 14]. Si les méthodes basées sur l'utilisation de points d'intérêts restent plus rapides que celles basées sur l'utilisation de blocs superposables, la phase de mise en correspondance a une complexité quadratique en fonction du nombre de points d'intérêt.

Dans cet article, nous proposons un algorithme avec un temps de calcul réduit par rapport à la méthode [1] tout en gardant une efficacité similaire à celle de l'algorithme original permettant ainsi d'analyser des images 4K Ultra HD. Notre méthode commence par extraire les points d'intérêt de l'image avant d'appliquer notre nouvel algorithme de mise en correspondance rapide. Cet algorithme propose de trier les points d'intérêt en fonction de leur orientation au gradient et pour chaque point d'intérêt de ne chercher des correspondants qu'avec les points d'intérêt ayant une orientation proche du point courant. Les nouveaux clusters sont ensuite obtenus avec l'algorithme DBSCAN afin de générer un masque binaire localisant les éventuelles zones falsifiées.

2 Travail proposé

Dans cette section, nous présentons en détail notre méthode de détection rapide de falsification par copié-deplacé.

2.1 Vue d'ensemble

La figure 1 montre une vue d'ensemble du système de détection implémenté. Dans une première étape, celui-ci utilise un détecteur SURF [2] afin d'extraire et décrire les caractéristiques des points d'intérêt de l'image. Un algorithme de mise en correspondance rapide est alors appliqué aux points caractéristiques afin de déterminer, pour chaque point, si d'autres points dans l'image lui sont similaires. Puis, pour chaque paire de points caractéristiques appariés, un segment est implémenté. L'algorithme DBSCAN [12] est appliqué sur ces segments. Enfin, la dernière étape consiste à calculer les enveloppes convexes des points d'intérêt appartenant aux clusters précédemment calculés. Ces enveloppes sont étendues pour obtenir un masque binaire final.

2.2 Extraction des points d'intérêt

La méthode commence donc par extraire les points caractéristiques de l'image à l'aide du détecteur SURF [2]. Le but de cette étape est d'obtenir le plus grand nombre de points d'intérêt possible pour ne manquer aucune falsification. L'algorithme SURF est utilisé car il fournit des

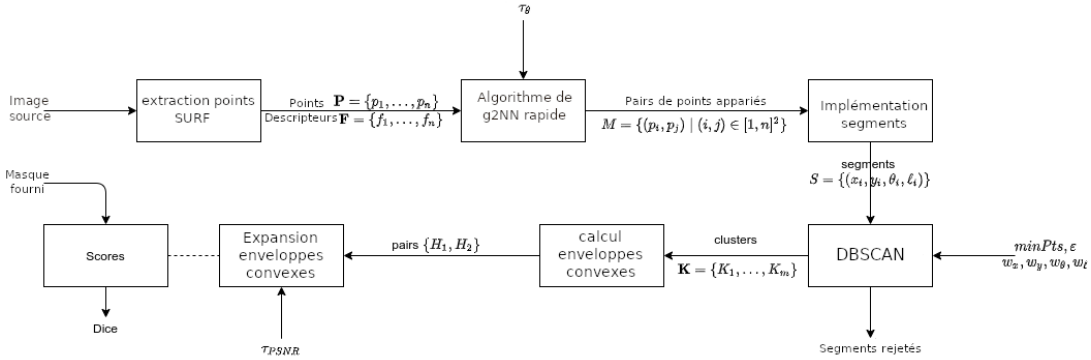


FIGURE 1 – Vue d'ensemble de la méthode.

résultats proches de l'algorithme SIFT [10] tout en étant bien plus rapide. Le calcul des points SURF est basé sur le calcul d'une matrice Hessienne dans le voisinage de chaque pixel considéré. Un pixel est choisi comme point d'intérêt si le déterminant de la matrice Hessienne qui lui est associée est supérieur à un seuil τ_{SURF}. Afin d'extraire un maximum de points d'intérêt, le seuil τ_{SURF} est fixé à 0.

Soit n le nombre de points d'intérêt détectés par l'algorithme. À l'issue de cette étape, l'algorithme extrait les points SURF, notés $\mathbf{P} = \{p_1, \ldots, p_n\}$ et calcule leurs descripteurs et orientations notés respectivement $\mathbf{F} = \{f_1, \ldots, f_n\}$ et $\mathbf{\Theta} = \{\theta_1, \ldots, \theta_n\}$.

2.3 Algorithme de mise en correspondance rapide

Chaque point p_i, $i \in \{1, \ldots, n\}$, est caractérisé par son orientation θ_i et son descripteur f_i. Nous voulons trouver pour chaque point p_i un ensemble de points d'intérêt qui lui sont suffisamment similaires.

La méthode utilise le test g2NN (generalized 2 Nearest-Neighbors) proposé par Amerini *et al.* [1]. Celui-ci calcule pour chaque point p_i la distance euclidienne du descripteur f_i à tous les autres descripteurs f_j, $j \in \{1, \ldots, n\} \backslash \{i\}$, notée $d_j = ||f_i - f_j||_2$. Soit $\mathbf{D}_i = \{d_1, \ldots, d_{n-1}\}$ le vecteur des distances euclidiennes entre le point p_i et les autres points trié par ordre croissant. Le test suivant est itéré tant que le critère (1) est satisfait :

$$\frac{d_j}{d_{j+1}} < \tau_{g2NN}. \tag{1}$$

Ainsi, si le rapport de distance étudié est inférieur à τ_{g2NN}, le processus continue et j est incrémenté. À la fin du test, nous obtenons un vecteur de similarité $\mathbf{D}'_i = \{d_1, \ldots, d_k\}$ avec k le nombre de points d'intérêt appariés au point p_i. Les points d'intérêt mis en correspondance avec le point p_i sont les points associés aux distances présentes dans le vecteur D'_i.

Dans les méthodes standards, chaque point p_i est comparé aux $n-1$ autres points. La complexité de la mise en correspondance est en $\mathcal{O}(n^2)$. Notre méthode réduit la complexité en comparant chaque point p_i à m points d'intérêts, avec $m < n$ la taille de la fenêtre appliqué, réduisant ainsi la complexité à $\mathcal{O}(mn)$. Pour cela, la méthode ne compare p_i qu'avec les points p_j qui ont une orientation θ_j proche de θ_i. La première étape consiste donc à trier les points d'intérêt par ordre lexicographique selon leur orientation dominante pour obtenir le vecteur $\mathbf{P}^\theta$:

$$\mathbf{P}^\theta = \{p_1^\theta, \ldots, p_n^\theta\}, \tag{2}$$

où p_i^θ sont les points d'intérêt ordonnés, avec $i \in \{1, \ldots, n\}$, . Le test de g2NN est effectué sur l'ensemble $\mathbf{P}^\theta_{i,\tau_\theta}$:

$$\mathbf{P}^\theta_{i,\tau_\theta} = \left\{ p_j^\theta \in \mathbf{P}^\theta \;\middle|\; |\theta_i - \theta_j| \leq \tau_\theta \right\}, \tag{3}$$

avec τ_θ le seuil de la fenêtre d'angle.

2.4 DBSCAN

Pour chaque points appariés p_i^θ et p_j^θ tel que $p_i^\theta \leq p_j^\theta$ pour l'ordre lexicographique, un segment v est implémenté, où p_i est le point de départ $(p_s(x_s, y_s))$ et p_j le point d'arrivée $(p_e(x_e, y_e))$.

L'objectif de cette étape est de trouver un cluster de segments correspondants à une falsification et d'éliminer les appariements parasites. Sur la base de copié-déplacés uniquement par translation, les segments commencent et finissent tous dans la mêmes zone, sont parallèles et ont la même longueur. La méthode utilise alors ces trois propriétés pour effectuer le *clustering*. Ainsi, chaque segment $v(x_s, y_s, \phi, \ell)$, est caractérisé par (x_s, y_s) les coordonnées de son point de départ, ϕ son angle et ℓ sa longueur. L'algorithme utilisé pour filtrer les différents appariements est l'algorithme DBSCAN [12] associé à la distance suivante :

$$\begin{aligned} d(v_i, v_j) \quad = \quad & w_x \frac{(x_{s_i} - x_{s_j})^2}{(H+L)/2} + w_y \frac{(y_{s_i} - y_{s_j})^2}{(H+L)/2} \\ & + w_\phi \frac{(\phi_i - \phi_j)^2}{\phi_i} + w_\ell \frac{(\ell_i - \ell_j)^2}{\ell_i}, \end{aligned} \tag{4}$$

avec $w_x + w_y + w_\theta + w_\ell = 1$, H et L la hauteur et la largeur de l'image et $(w_x, w_y, w_\theta, w_\ell)$ les poids associés aux différentes contraintes.

2.5 Génération du masque binaire

La dernière étape consiste à générer le masque binaire $\mathcal{M}$ final permettant de localiser les régions altérées. Pour cela, la méthode traite itérativement tous les clusters trouvés durant l'étape précédente.

Soit $\mathbf{K} = \{v_1, \ldots, v_m\}$ un cluster, où un segment v_i, avec $i \in \{1, \ldots, m\}$ est caractérisé par son point de départ $p_{s_i} = (x_{s_i}, y_{s_i})$ et son point d'arrivée $p_{e_i} = (x_{e_i}, y_{e_i})$. Nous définissons les ensembles suivants : $\mathbf{S} = \{p_{s_i} \mid i \in \{1, \ldots, m\}\}$ et $\mathbf{E} = \{p_{e_i} \mid i \in \{1, \ldots, m\}\}$. $\mathbf{S}$ contient tous les points de départ des segments appartenant au cluster $\mathbf{K}$ et $\mathbf{E}$ les points d'arrivée. Les enveloppes convexes de ces deux ensembles, notées respectivement $\mathcal{H}_S$ et $\mathcal{H}_E$, sont alors calculées. Afin d'améliorer la précision, nous effectuons une expansion de ces enveloppes convexes jusqu'à ce qu'elles atteignent les bords des zones falsifiées. Pour

cela, le voisinage d'un pixel copié est comparé au pixel déplacé correspondant. Si ceux-ci présentent une forte similarité, alors le bord n'est pas atteint et l'enveloppe convexe doit être encore étendue, sinon, le bord de la falsification a été trouvé localement. Nous considérons un voisinage circulaire de taille k autour d'un pixel du bord ($\mathcal{N}_{p_{s_i}}$), et un voisinage de même forme autour du pixel correspondant de l'autre zone à étendre ($\mathcal{N}_{p_{e_i}}$). Deux voisinages sont considérés similaires si le $PSNR$ entre $\mathcal{N}_{p_{s_i}}$ et $\mathcal{N}_{p_{e_i}}$ est supérieur à un seuil τ_{PSNR}.

3 Résultats

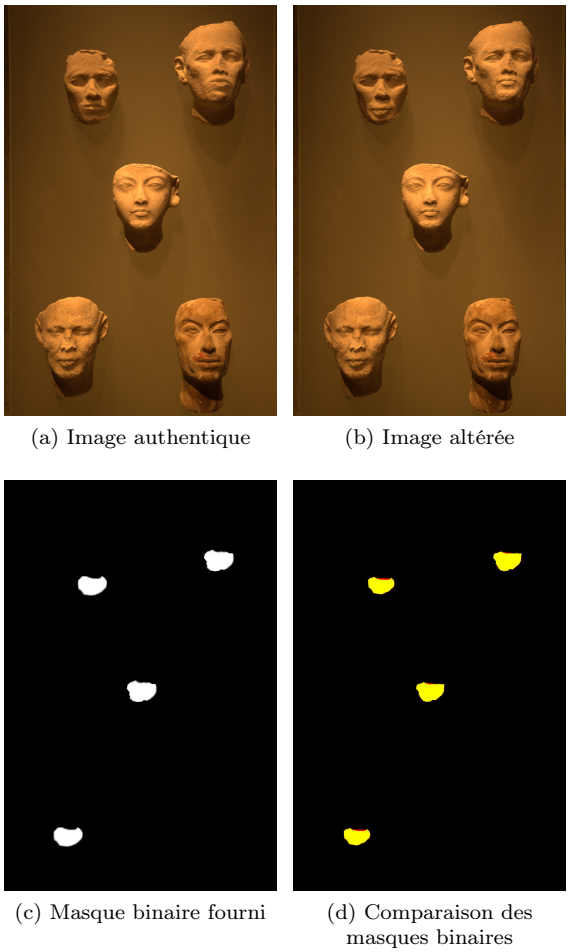
Dans cette section, nous présentons les résultats obtenus par notre méthode, analysons ses performances puis comparons les résultats obtenus avec ceux de l'état de l'art.

Les tests présentés sont effectués au niveau des pixels sur des images issues des bases de données GRIP [5], FAU [4] et de quelques images de grande taille (type 4K) fournies par la Direction Générale de l'Armement (DGA). T_P représente le nombre de vrais positifs, F_P de faux positifs et F_N de faux négatifs. Les trois mesures utilisées sont la précision, le rappel et le F1-score :

$$precision = \frac{T_P}{T_P + F_P}, \tag{5}$$

$$rappel = \frac{T_P}{T_P + F_N}, \tag{6}$$

$$F1 - score = \frac{2 \times precision \times rappel}{precision + rappel}. \tag{7}$$



(a) Image authentique (b) Image altérée

(c) Masque binaire fourni (d) Comparaison des masques binaires

FIGURE 2 – Résultat obtenu par notre méthode sur une image de la base FAU [4].

3.1 Exemple complet

L'algorithme est illustré à l'aide d'une image issue de la base FAU de taille 2014×3039 pixels. La figure 2 est composé de l'image authentique (figure 2a), l'image altérée (figure 2b), le masque binaire localisant les régions falsifiées, soit la vérité de terrain (figure 2c) et la comparaison entre le masque obtenu par notre méthode et la vérité de terrain (figure 2d).

L'image falsifiée, figure 2b possède deux altérations au niveau des lèvres des différents visages. Avec notre méthode, 134 353 points d'intérêt sont détectés, 318 appariements sont réalisés et 301 correspondances sont conservées lors de la phase clustering. Les enveloppes convexes des points d'intérêts liés à la falsification sont calculées puis étendues afin de générer un masque binaire. La figure 2d compare le masque obtenu par notre méthode avec la vérité de terrain. Les pixels en jaune sont les vrais positifs (T_P), en rouge les faux positifs (F_P) et en vert les faux négatifs (F_N). Nous obtenons pour cette image une précision de 0.9441, un rappel de 0.9999 et un F1-score de 0.9712.

3.2 Analyse des performances

Dans cette section, les différents tests ont été effectués en utilisant OpenCV 4.0 et C++17 sur un Ubuntu 16.04.1 64-bit avec un processeur Intel®Core™i7-7820X, 16 coeurs, à une fréquence de 3.60GHz et possédant 110GB de RAM.

Nous comparons dans un premier temps les performances temporelles d'un détecteur utilisant notre algorithme de g2NN rapide (en $\mathcal{O}(mn)$) et celles d'un détecteur utilisant le g2NN standard [1] en $\mathcal{O}(n^2)$. Les tests ont été effectués sur l'ensemble des bases GRIP et FAU ansi que sur 10 grandes images fournies par la DGA. La figure 3 présente le nombre moyen de points d'intérêt SURF extraits ainsi que le temps en secondes mis par un détecteur utilisant l'algorithme de g2NN standard (en bleu) et par un détecteur utilisant l'algorithme de g2NN rapide (en rouge) pour analyser une image selon la taille de cette dernière. Nous pouvons observer que le temps de calcul mis par la méthode standard croit significativement lorsque le nombre de points d'intérêt augmente tandis que notre méthode est capable d'analyser rapidement même une image 4K. Par exemple la méthode standard met en moyenne 141 646 secondes pour analyser une image de taille 6000×4000 pixels, soit 1,6 jours contre 5831 secondes pour notre méthode, soit 1.6 heures.

3.3 Comparaison avec l'état de l'art

Les résultats obtenus au niveau pixels sur les bases d'images FAU et GRIP sont comparés dans les tables 1 et 2 à des méthodes de l'état de l'art. Ces tableaux présentent les F1-scores moyens obtenus sur les 80 images altérées de la base GRIP et sur les 48 images altérées de la base FAU.

La table 1 montre que notre méthode possède un F1-score moyen de 0,9606 sur la base d'image GRIP. Comparé

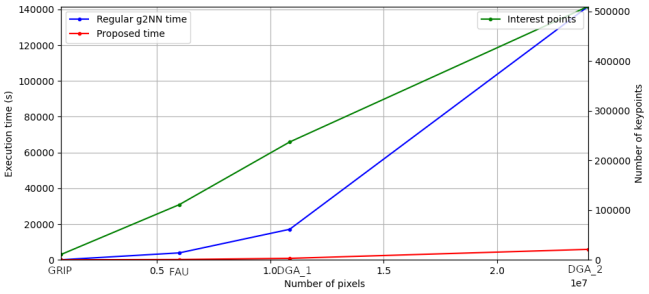


FIGURE 3 – Nombre de points d'intérêt extraits et temps d'exécution des deux approches en fonction de la taille de l'image.

Méhodes	F1-score (Pixels)
Cozzolino *et al.* [5]	0.9299
Li *et al.* [8]	0.2774
Bravo *et al.* [3]	0.8482
Christlein *et al.* [4]	0.8618
Silva *et al.* [15]	0.6662
Zandi *et al.* [16]	0.6444
Li *et al.* [9]	0.9466
Proposée	**0.9606**

TABLE 1 – F1-scores obtenus sur la base GRIP [5].

Méhodes	F1-score (Pixels)
Huang *et al.* [7]	0.6354
Shivakumar and Baboo [14]	0.6954
Zandi *et al.* [16]	0.8607
Li *et al.* [8]	0.7447
Pun *et al.* [13]	**0.8997**
Li and Zhou [9]	0.8838
Lyu *et al.* [11]	0.8142
Proposée	**0.8963**

TABLE 2 – F1-scores obtenus sur la base d'images FAU [4].

aux méthodes de l'état de l'art ([5], [8], [3], [15], [4], [16] et [9]), nous pouvons voir que l'algorithme proposé est plus efficace que le reste des algorithmes. La table 2 montre que notre méthode possède un F1-score moyen de 0.8963 sur la base d'image FAU. Ces résultats sont bien meilleurs que ceux obtenus par les méthodes pionnières basées sur les points d'intérêt ([7] et [14]) ainsi que ceux obtenus par les méthodes [16], [8], [9], [11]. Pour finir, nos résultats sont comparables à ceux de la méthode de Pun *et al.* [13].

4 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode rapide de détection de falsification de type copié-déplacé dans des images 4K Ultra HD. Celle-ci extrait en premier lieu les points d'intérêt SURF de l'image d'entrée. L'algorithme d'appariement proposé dans [1] a été utilisé et amélioré afin de détecter les mises en correspondance en réduisant drastiquement sa complexité. Le nombre de comparaisons effectué pour chaque point d'intérêt a été réduit en introduisant une fenêtre d'angle. L'algorithme de clustering DBSCAN a été appliqué afin de classer les appariements trouvés en différents clusters. Le masque binaire final a été généré en calculant les enveloppes convexes des points présents dans ces clusters puis en faisant une expansion de ceux-ci. Les résultats expérimentaux montrent que notre approche proposée est très efficace et détecte les zones falsifiées très rapidement dans les images 4K, contrairement aux méthodes plus conventionnelles.

Dans de futurs travaux, nous nous intéresserons à la différenciation de la zone copiée de la zone déplacée dans la même image. Nous voulons également améliorer cette méthode pour la rendre robuste à diverses attaques géométriques telles que la mise à l'échelle ou la rotation.

Références

[1] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra. A SIFT-Based Forensic Method for Copy–Move Attack Detection and Transformation Recovery. *IEEE TIFS*, 6(3) :1099–1110, September 2011.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf : Speeded up robust features. In *ECCV*, pages 404–417. Springer, 2006.

[3] S. Bravo-Solorio and A. K. Nandi. Passive forensic method for detecting duplicated regions affected by reflection, rotation and scaling. In *2009 17th EUSIPCO*, pages 824–828, 2009.

[4] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou. An evaluation of popular copy-move forgery detection approaches. *IEEE TIFS*, 7(6) :1841–1854, 2012.

[5] D. Cozzolino, G. Poggi, and L. Verdoliva. Efficient dense-field copy–move forgery detection. *IEEE TIFS*, 10(11) :2284–2297, 2015.

[6] J. Fridrich, D. Soukal, and J. Lukáš. Detection of copy-move forgery in digital images. In *Digital Forensic Research Workshop*, 2003.

[7] H. Huang, W. Guo, and Y. Zhang. Detection of copy-move forgery in digital images using sift algorithm. In *2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, volume 2, pages 272–276, 2008.

[8] J. Li, X. Li, B. Yang, and X. Sun. Segmentation-based image copy-move forgery detection scheme. *IEEE TIFS*, 10(3) :507–518, 2015.

[9] Y. Li and J. Zhou. Fast and effective image copy-move forgery detection via hierarchical feature point matching. *IEEE TIFS*, 14(5) :1307–1322, 2019.

[10] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2) :91–110, 2004.

[11] Q. Lyu, J. Luo, K. Liu, X. Yin, J. Liu, and W. Lu. Copy move forgery detection based on double matching. *JVCIR*, 76 :103057, 2021.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, pages 226–231, August 1996.

[13] C.-M. Pun, X.-C. Yuan, and X.-L. Bi. Image forgery detection using adaptive oversegmentation and feature point matching. *IEEE TIFS*, 10(8) :1705–1716, 2015.

[14] BL Shivakumar and S Santhosh Baboo. Detection of region duplication forgery in digital images using surf. *IJCSI*, 8(4) :199, 2011.

[15] E. Silva, T. Carvalho, A. Ferreira, and A. Rocha. Going deeper into copy-move forgery detection : Exploring image telltales via multi-scale analysis and voting processes. *JVCIR*, 29 :16–32, 2015.

[16] M. Zandi, A. Mahmoudi-Aznaveh, and A. Talebpour. Iterative copy-move forgery detection based on a new interest point detector. *IEEE TIFS*, 11(11) :2499–2512, 2016.

Une approche géométrique pour analyser l'intention sociale à partir du mouvement de marqueurs 3D

Paul Audain Desrosiers [1] , Mohamed Daoudi [2,3], Maria-Francesca Gigliotti[1], Yann Coello [1]
Université de Lille, Sciences Cognitives et Sciences Affectives (SCALab) - UMR 9193

[2] IMT Lille Douai, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France;
[3] Univ. Lille, CNRS, Centrale Lille, Institut Mines-Télécom, UMR 9189 CRIStAL, F-59000 Lille, France.

Résumé : *Dans ce papier, nous proposons un cadre géométrique capable de prédire en temps réel une intention sociale vs une intention personnelle. Le participant doit réaliser un ensemble de gestes comportant une intention sociale ou personnelle, en portant un gant contenant 4 marqueurs passifs. L'utilisation d'un système de capture de mouvement permet d'obtenir la trajectoire de la main du participant contenant les différents marqueurs 3D. Les données 3D obtenues sont définies dans un espace de forme de courbes ouvertes, puis analysées dans une variété Riemannienne. Nous avons obtenu un taux de reconnaissance moyen pour les deux gestes (intention sociale, personnelle) de 68%, ce qui est comparable au score moyen produit par l'évaluation humaine. Les résultats expérimentaux montrent également que le taux de classification pourrait être utilisé pour améliorer la communication sociale entre les agents humains et virtuels. A notre connaissance, il s'agit de la première étude en temps réel qui utilise des techniques de vision par ordinateur pour analyser l'effet de l'intention sociale sur l'action motrice afin d'améliorer la communication sociale entre un humain et un agent virtuel.*

Mots-clés : Intention sociale et personnelle, analyse des trajectoires, géométrie Riemannienne, motion capture (Mocap).

1 Introduction

La reconnaissance des actions et des comportements d'individus est l'un des domaines les plus actifs en vision par ordinateur. Cependant, la motricité volontaire est organisée à partir des intentions motrices et sociales [4]. Des recherches récentes en psychologie cognitive ont montré que lorsque nous réalisons une action avec une intention sociale au lieu d'une intention personnelle, nous amplifions les paramètres spatiaux et temporels de l'action motrice [7]. De plus, un observateur est capable de percevoir ces changements cinématiques et anticiper l'intention sociale dans les actions motrices effectuées par d'autres, afin d'agir de manière complémentaire [5]. Toutefois, les liens entre intention sociale et motricité sont encore mal connus, notamment dans le domaine de la vision par ordinateur. C'est dans ce contexte que Zunino et al. [9] proposent une approche de prédiction d'intention à partir de mouvements représentés par des matrices de covariance. L'utilisation des matrices de covariance est étendue au cas des séquences temporelles d'articulations 3D, en proposant une approche de la reconnaissance de l'action humaine à partir de séquences squelettes 3D extraites de données de profondeur. Contrairement au travail de Zunino et al [9], nous représentons les mouvements de la main par un ensemble de trajectoires [1].

2 Méthodologie

2.1 Représentation du mouvement dans $\mathbb{R}^3$

Nous représentons par P_t l'état du point à un instant t, $P_t = [x_1(t), y_1(t), z_1(t) \dots x_k(t), y_k(t), z_k(t)]^T$. Un mouvement est une séquence de poses et peut être vue comme le résultat d'une trajectoire continue dans l'espace des mouvements. La trajectoire est définie par le mouvement au cours du temps des points caractéristiques encodant les coordonnées 3D des articulations de la main. Soit une trajectoire dans l'espace des actions représentée par une fonction paramétrée $\alpha(t) : I = [0, 1] \rightarrow \mathbb{R}^3$.

2.2 Représentation du mouvement dans l'hypersphère $\mathcal{C}$

Dans le but de modéliser et d'étudier nos courbes, nous adoptons la fonction appelée square-root velocity function (SRVF) [8]. Elle a été exploitée avec succès pour la reconnaissance d'actions humaines [2], la reconnaissance de visages en 3D [3] et plus récemment dans la génération d'expressions faciales [6]. Plus précisément, pour une courbes donnée $\alpha(t) : I \rightarrow \mathbb{R}^3$, la fonction SRVF $q(t) : I \rightarrow \mathbb{R}^3$ est définie par :

$$q(t) = \frac{\dot{\alpha}(t)}{\sqrt{\|\dot{\alpha}(t)\|}} , \tag{1}$$

où, $\| \cdot \|$ est la norme L_2 dans $\mathbb{R}^3$.

Afin d'éliminer la variabilité d'échelle des courbes, nous les mettons à l'échelle pour qu'elles soient de longueur 1. Par conséquent, les SRVF correspondant à ces courbes sont des éléments d'une hypersphère unitaire dans le l'espace de Hilbert $\mathbb{L}^2(I, \mathbb{R}^3)$ comme expliqué dans [8]. Nous appellerons cette hypersphère $\mathcal{C} = \{q : I \rightarrow \mathbb{R}^3 | \ \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$. Chaque élément de $\mathcal{C}$ représente une courbe dans $\mathbb{R}^3$ associée à un mouvement humain. $\mathcal{C}$ est une hypersphère, la longueur ou la distance géodésique entre deux éléments q_1 et q_2 est définie comme suit :

$$d_{\mathcal{C}}(q_1, q_2) = \cos^{-1}(\langle q_1, q_2 \rangle) . \tag{2}$$

Nous définissons les opérations $\log_\mu(.)$ et $\exp_\mu(.)$, les logarithme et exponentielle sur la sphère, utilisées pour projecter les données SRVF dans les deux sens vers l'espace tangent $T_\mu(C)$ à un point de référence μ. Elles sont données par :

$$\begin{aligned} \log_\mu(q) &= \frac{d_{\mathcal{C}}(q, \mu)}{sin(d_{\mathcal{C}}(q, \mu))}(q - cos(d_{\mathcal{C}}(q, \mu))\mu) , \\ \exp_\mu(s) &= cos(\|s\|)\mu + sin(\|s\|)\frac{s}{\|s\|}, \end{aligned} \tag{3}$$

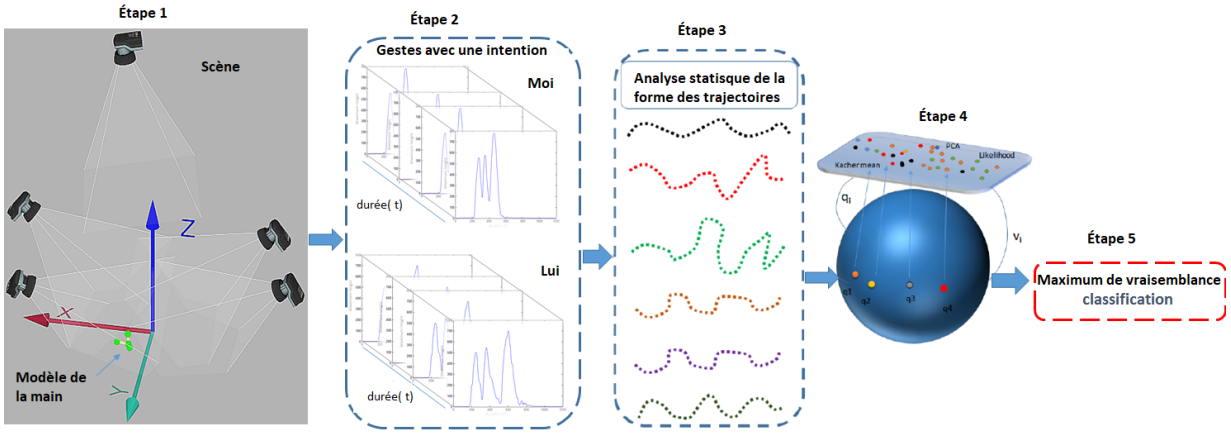


FIGURE 1 – Vue générale de la méthode proposée.

2.3 Analyse statistique des trajectoires

L'objectif principal de notre étude est de classer l'intention de l'utilisateur parmi deux classes c_k qu'on dénote {personnelle, sociale}. Pour cela, nous proposons d'apprendre des distributions représentatives des trajectoires pour chaque classe. La variété $\mathcal{C}$ n'est pas un espace vectoriel. Les structures euclidiennes telles que la norme et le produit scalaire, les algorithmes d'apprentissage automatique, y compris l'analyse en composantes principales (ACP) et l'algorithme de classification par maximum de vraisemblance, ne peuvent pas être appliqués dans leur forme originale sur le variété $\mathcal{C}$. Une approche commune utilisée pour faire face à cette non-linéarité est d'exploiter les propriétés Euclidiennes de l'espace tangent en un point particulier de la variété, par exemple, la moyenne de Karcher des données, μ. Un tel espace tangent est un espace vectoriel linéaire qui est plus pratique pour calculer des statistiques. Par conséquent, afin d'apprendre la distribution des vecteurs dans l'espace tangent, nous pouvons d'abord effectuer une ACP pour apprendre un sous-espace principal appelé $\mathcal{B}$. La matrice de covariance dans cette base est définie par $\Sigma = \sum_{i=1}^{N} v_i v_i^T$, où v_i sont les vecteurs tangents de la projection dans la base $\mathcal{B}$.

Enfin, la distribution normale multivariée de la trajectoire c_k, $p(v|c_k; |\Sigma|)$ est apprise en utilisant la matrice de covariance Σ calculée à partir de l'ensemble des v_i où $|\Sigma|$ est le déterminant de la matrice de covariance Σ, voir l'équation 4.

$$p\left(v \mid c_k; \Sigma\right) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2} v^T \Sigma^{-1} v} \qquad (4)$$

3 Résultats Expérimentaux

3.1 Protocole expérimental

L' expérimentation comprend 2 parties : a) l'acquisition de données et une étape d'apprentissage ; b) la classification et l'analyse du mouvement de la main des sujets pour interagir avec un agent virtuel. Pour une évaluation efficace de notre méthode, nous avons collecté un ensemble de données sur 15 sujets dont l'âge varie de 20 à 50 ans. Toute la scène est couverte par 5 caméras (Infra-rouge) de capture de mouvement (Mocap) qui fonctionnent à une fréquence de 200 Hz chacune. Les sujets ont été invités dans la scène, et à s'asseoir devant une table. Le sujet porte un gant noir qui contient 4 marqueurs passifs. Ces marqueurs sont placés dans une position spécifique sur le gant : l'index (pointe), le pouce (pointe), la main et le poignet. Une tasse est située à une position particulière sur la table. Lorsque le mot («Moi»,«Lui») est diffusé, le sujet doit déplacer la tasse du point A au point B, ces 2 positions étant visibles sur la table, voir figure 2. La distance totale entre le point initial (PI) et le point B est de 48 cm, et la distance entre le point A et le point B est de 24 cm. Les distances ont été choisies en fonction d'une vraie table barman.

FIGURE 2 – Dans cette figure, il est possible d'observer les différentes positions (A, B, position initiale) et mouvements de la main du participant sur la table. PI : position initiale (hand laying), A : déplacement de la main ; B : saisie de la tasse.

1) Avant la réalisation d'un geste, on vérifie bien que la main du participant se trouve dans la position initiale. Lorsque le mot « Moi » est diffusé, le sujet doit déplacer la tasse sur la table du point A vers le point B avec une intention personnelle (déplacer la tasse sans vouloir inclure une autre personne dans l'action).

2) Lorsque le mot « Lui » est diffusé, le sujet doit déplacer la tasse du point A vers le point B avec une intention sociale (avec la volonté d'impliquer une autre personne dans l'action). Les sujets réalisent 50 gestes "Lui" et 50 gestes "Moi" et peuvent commencer aléatoirement par la condition "Lui" ou par la condition "Moi". Pour le traitement des données nous avons éliminé les gestes de retour à la position initiale de la main, ainsi que tous les gestes qui sont incorrects c'est-à-dire : tremblements, hésitations de la main, signaux inexploitables ou données manquantes. Un filtre médian 3D a été utilisé pour éliminer le bruit. Il a été vérifié que la courbe de vitesse de chaque geste contient 4 minimum et 3 maximum ce qui permet de garantir que le mouvement a été réalisé de façon correcte sur le

plan cinématique. La position initiale (PI) de la main du participant détermine le premier minimum, le deuxième minimum correspond au moment où le participant prend la tasse sur la table (position A). Le troisième minimum correspond au moment où le participant dépose la tasse sur la table (position B). Le quatrième minimum correspond au retour à la position initiale (RPI), voir figure 3. Les 3 maximums correspondent aux pics de vitesses des 3 mouvements réalisés (saisie de la tasse, déplacement de la tasse, retour en position initiale). Pour rappel, la fréquence d'échantillonnage du Mocap est de 200 Hz, et les participants sont invités à réaliser le geste dans un intervalle de 0 à 4s. L'axe des abscisses représente la durée (t) du mouvement, et l'axe des ordonnées la vitesse, voir figure 3. Pour analyser l'effet de l'intention sociale sur la cinématique du mouvement, nous représentons le mouvement de la main par une série temporelle des points en 3D. Ensuite nous sélectionnons les moments pertinents du geste (saisie de la tasse, déplacement de la tasse). La question qui se pose est comment classer le mouvement de la main en deux classes (intention personnelle et sociale). Notre approche consiste à représenter cette série temporelle de points en 3D par une trajectoire ou une courbe en 3D. Dans la figure 4, les courbes en rouge correspondent à des trajectoires des gestes «Lui», et les courbes en bleu les trajectoires des gestes «Moi». Nous effectuons par la suite une analyse statistique de la forme de ces trajectoires.

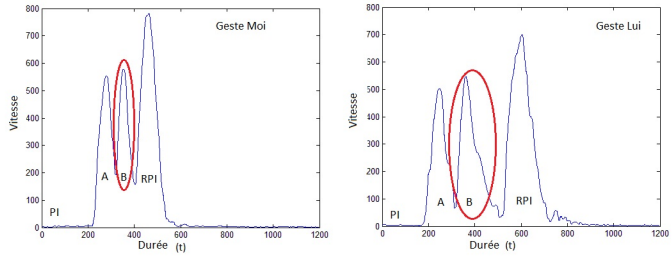


FIGURE 3 – PI : Position Initiale, A : déplacement de la main pour saisie la tasse ; B : saisie de la tasse ; RPI : retour position initiale de la main.

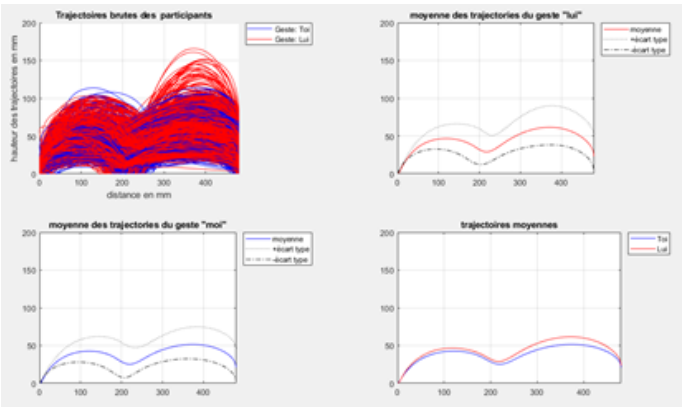


FIGURE 4 – Dans cette figure, il est possible d'observer l'ensemble des gestes (trajectoires) réalisés par les participants avec une intention sociale (les courbes rouges) ou personnelle (les courbes en bleu).

3.2 Résultats

Dans la phase de test, 15 nouveaux participants totalement naïfs ont été invités dans la scène expérimentale qui est identique à l'étape d'apprentissage. Un agent virtuel animé a été projeté sur un téléviseur de 165 cm. L'agent virtuel consiste en un barman dans son propre environnement de travail. Dans la figure 5, on peut observer 3 positions sur la table : 1) La position initiale (PI, Hand laying) ; 2) la position A ; 3) la position B. En effet, le déplacement de la main du participant en partant de la position initiale pour aller vers la position A pour prendre la tasse, et pour la déposer dans la position B, permet de définir la trajectoire parcourue par la main du participant pour réaliser le geste. Ainsi, le participant s'assied devant la table avec la télé. Lorsque le mot (« Lui », « Moi ») est diffusé comme dans l'étape d'apprentissage, le participant déplace la tasse de la position A à la position B avec l'intention sociale ou personnelle. Le participant doit réaliser la bonne intention sociale pour déclencher l'action appropriée du barman virtuel. Ainsi, nous obtenons un score de reconnaissance de 73%. Dans la figure de gauche, le participant réalise une intention personnelle («Moi»), tandis que dans la figure de droite le participant réalise un geste avec une intention sociale («Lui»). L'argent virtuel réagit selon l'intention du geste qui est détectée. L'avantage de la méthode proposée c'est qu'elle est invariante par rapport à la position et la rotation, autrement dit le participant pourrait déposer la tasse n'importe où sur la table sans aucun autre apprentissage supplémentaire. Pour bien mener notre étude, nous avons préféré que les participants fassent tous le même geste avec le même point de départ et même point d'arrivée. Pour rappel, dans cette étude nous avons utilisé le profile de vitesse des deux gestes comme un bon indicateur qui permet d'analyser l'effet de l'intention sur la cinématique, et la trajectoire nous permet d'observer comment le participant module les deux gestes. Dans la figure 5 on observe que les courbes en rouge qui correspondent à une intention sociale possèdent une amplitude plus grande que les courbes en bleu qui représentent une intention personnelle. A partir de ces résultats, il est possible de dire que dans une interaction, parfois on amplifie nos gestes lorsqu'on souhaite inclure une autre personne dans notre action.

FIGURE 5 – Dans la figure de gauche, le participant réalise un geste avec une intention personnelle, donc le barman virtuel le regard. Tandis que dans le figure de droite, le participant réalise un geste avec une intention sociale, et il a été servi par le barman virtuel.

4 Conclusion et perspectives

Dans ce papier, nous avons proposé une approche basée sur l'analyse cinématique et de la géométrie Riemannienne pour analyser le mouvement du bras humain, lorsque les individus réalisent des gestes avec une intention sociale ou personnelle. Les résultats obtenus sur l'ensemble des données nous permettent d'avoir un taux de reconnaissance pour les deux gestes de 73%. Les résultats montrent que la méthode proposée est comparable aux scores produits par [7].

Références

[1] Mohamed Daoudi, Yann Coello, Paul Audain Desrosiers, and Laurent Ott. A new computational approach to identify human social intention in action. In *13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 512–516, 2018.

[2] Maxime Devanne, Hazem Wannous, Stefano Berretti, Pietro Pala, Mohamed Daoudi, and Alberto Del Bimbo. 3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold. *IEEE TC*, 45(7) :1340–1352, 2014.

[3] Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi, and Rim Slama. 3D face recognition under expressions, occlusions, and pose variations. *PAMI*, 35(9) :2270–2283, 2013.

[4] Maria Francesca Gigliotti, Adriana Sampaio, Angela Bartolo, and Coello Yann. The combined effect of motor and social goals on the kinematics of object-directed motor action. volume 10, page 6369, 2020.

[5] Daniel Lewkowicz, Quesque Francois, Coello Yann, and N. Delevoye-Turrell Yvonne. Individual differences in reading social intentions from motor deviants. volume 6, 2015.

[6] Naima Otberdout, Mohamed Daoudi, Anis Kacem, Lahoucine Ballihi, and Stefano Berretti. Dynamic facial expression generation on hilbert hypersphere with conditional Wasserstein generative adversarial nets. *PAMI*, pages 1–1, 2020.

[7] Francois Quesque, Daniel Lewkowicz, Yvonne N. Delevoye-Turrell, and Yann Coello. Effects of social intention on movement kinematics in cooperative actions. volume 7, 2013.

[8] Anuj Srivastava, Eric Klassen, Shantanu H. Joshi, and Ian H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *PAMI*, 33(7) :1415–1428, 2011.

[9] Andrea Zunino, Jacopo Cavazza, Atesh Koul, Andrea Cavallo, Cristina Becchio, and Vittorio Murino. Intention from motion. *CoRR*, abs/1605.09526, 2016.

Recalage de nuages de points 3D pour la comparaison de molécules pharmacologiques

Dominique Douguet
UCA - IPMC - Inserm

Frédéric Payan
UCA - I3S - CNRS

Résumé : *Le calcul de la similarité entre molécules est un sujet très étudié dans le domaine de la pharmacologie et de la modélisation moléculaire. Jusqu'à présent, bien que présentant de nombreux avantages, les représentations 3D des surfaces moléculaires ont été assez peu employées pour trouver des similarités entre molécules. Dans cette étude, nous montrons qu'il est possible d'appliquer des techniques de recalage 3D usuellement utilisées dans d'autres domaines (reconstruction 3D, vision par ordinateur, etc.) pour identifier des similarités entre molécules. Pour cela nous avons représenté ces dernières à l'aide de nuage de points 3D colorés, la couleur étant liée aux propriétés physico-chimiques des atomes composant les molécules. Ensuite, nous les avons alignés en appliquant successivement un recalage global, et un recalage local spécifique aux nuages colorés. En se basant sur des mesures de fitness et de distance géométrique, nous montrons que notre approche est particulièrement efficace pour superposer des molécules de taille semblable mais aussi lorsqu'elles sont de tailles différentes.*

Mots-clés : Nuage de points 3D coloré, recalage, surface moléculaire, descripteurs FPFH.

1 Introduction

Lorsqu'elle est prédite, la ressemblance entre structures moléculaires peut être calculée à partir d'une représentation 1D, 2D ou 3D de celles-ci. La représentation 3D des molécules a un avantage certain. En effet, une similarité géométrique 3D entre deux molécules peut, par exemple, induire des effets biologiques similaires, même si les molécules concernées ont une représentation 2D très différente comme dans le cas de deux molécules de deux familles structurales différentes. Il existe de nombreuses méthodes de calcul de la similarité exploitant la représentation 1D ou 2D, mais très peu utilisant des informations sur la forme géométrique 3D [1, 5].

Le but de notre projet actuel est de développer une méthode originale d'alignement de structures moléculaires pour la recherche de similarité, en s'inspirant des techniques de recalage de points 3D issues de la vision par ordinateur. Le recalage est en effet utilisé depuis plusieurs décennies pour fusionner dans un même repère global des ensembles de points 3D décrivant partiellement la surface d'un même objet ou d'une même scène, mais représentés dans des repères locaux (car capturés selon plusieurs points de vue ou à différents instants). Nous sommes convaincus que ce genre de méthodes peut être utilisé dans le cadre de la recherche de la similarité moléculaire.

2 Travail proposé

A cette fin, nous développons SENSAAS (SENsitive Surface As A Shape) [2], un outil original qui combine des méthodes récentes dédiées au recalage 3D, initialement développées pour la fusion de nuages de points 3D issus de scanners, de type LiDAR par exemple. Considérons deux molécules que l'on souhaite superposer pour rechercher des similarités. L'idée générale est d'aligner des représentations surfaciques de ces deux molécules, pour ensuite calculer un score de ressemblance. La figure ci-dessous décrit le principe général de SENSAAS.

— **Génération d'un nuage de points 3D coloré de la surface moléculaire** de chaque molécule (Figure 1b). La couleur est liée aux propriétés physico-chimiques de l'atome sous-jacent à la surface.

— **Alignement global des 2 surfaces** (Figure 1c) en exploitant uniquement la géométrie des points. Pour cela, nous utilisons les descripteurs locaux FPFH [8] pour caractériser des points d'intérêt sur les surfaces. Les descripteurs de chaque surface sont ensuite appairés avec RANSAC [3], de telle sorte que les surfaces une fois recalées minimisent une distance géométrique.

— **Raffinement de l'alignement** précédent en prenant en compte la géométrie des points 3D mais aussi la couleur associée, liée aux propriétés physico-chimiques (Figure 1d). Pour cela, la technique proposée dans [7] est utilisée. Celle-ci prend en compte la couleur des points lors du recalage de nuages pour améliorer la superposition finale. Ainsi nous obtenons une superposition des surfaces (Figure 1e), que l'on peut reporter sur les graphes 3D des molécules associées (Figure 1f).

3 Résultats

SENSAAS reproduit 89% des superpositions de molécules telles qu'elles sont observées expérimentalement dans le jeu de données AstraZeneca fourni par le CCDC. Ce jeu de données contient 1465 molécules différentes réparties en 121 groupes de molécules alignées car « capturées » dans le site actif d'une même protéine.

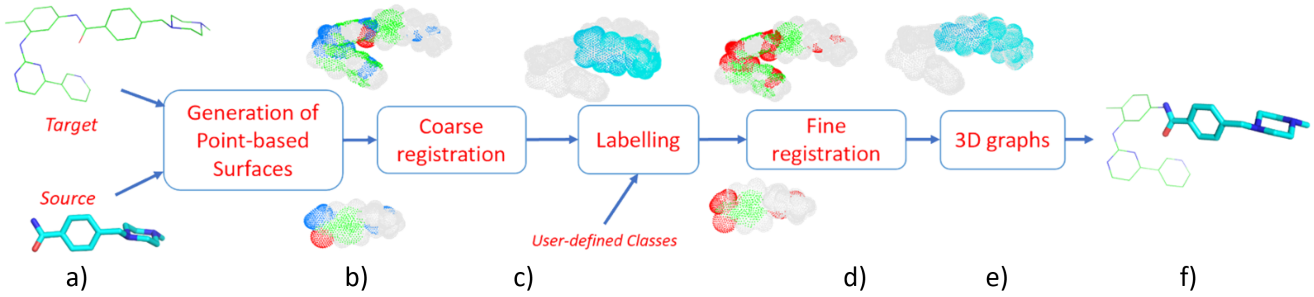


FIGURE 1 – principe général de notre algorithme SENSAAS.

Une superposition est considérée comme reproduite si le RMSD, c'est-à-dire la différence moyenne entre les coordonnées prédites et expérimentales des atomes du graphe 3D, ne dépasse pas 2.0Å. Ces résultats sont comparables à ceux des deux autres méthodes d'alignement testées : 89% pour le programme ShaEP [9] et 87% pour le programme SHAFTS [6]. Cependant, nous avons montré que les alignements proposés par SENSAAS sont plus précis puisqu'un plus grand nombre de molécules reproduites sont dans le premier intervalle de précision avec un $RMSD \leq 0.5$Å.

SENSAAS est capable de recaler parfaitement des molécules dont la structure moléculaire est identique ou très similaire comme une sous-structure (Figure 1) mais aussi lorsqu'elle est différente (2, *Bioisosteric matching*). L'alignement d'une sous-structure ou d'un fragment moléculaire sur une molécule de taille plus grande est ce que l'on appelle un sub-matching. C'est une propriété particulièrement intéressante que SENSAAS permet, contrairement à d'autres méthodes de référence d'alignement par forme moléculaire (logiciels ROCS [4], SHAFTS [6], ou ShaEP [9], par exemple). En effet, une des problématiques à traiter en chimie médicinale est de pouvoir substituer une sous-structure par une autre sous-structure de famille chimique différente afin d'améliorer une propriété pharmacologique donnée et/ou d'acquérir la propriété intellectuelle.

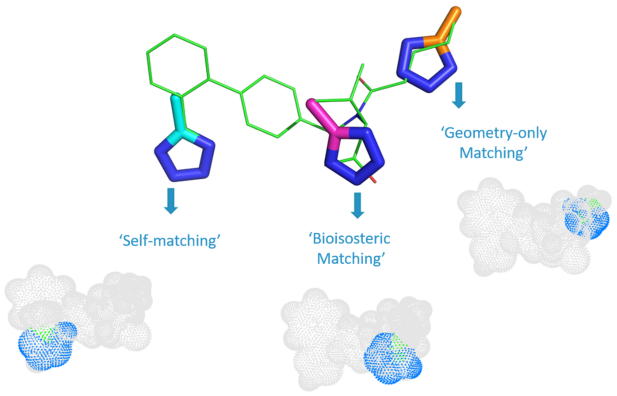


FIGURE 2 – Trois alignements possibles d'un fragment moléculaire (en cyan, magenta et orange) sur une structure moléculaire plus grande, le Valsartan représenté par sa structure en graphe 3D et coloré en vert.

La Figure 2 présente un exemple de bioisostérie bien connue en chimie médicinale qui consiste à remplacer un groupe acide carboxylique par un tétrazole. Dans cet exemple, SENSAAS est capable de proposer des alignements locaux parmi lesquels le meilleur est une superposition du fragment tétrazole sur lui-même puisqu'il est présent dans la molécule target Valsartan (*self-matching*) ; le second est une superposition du fragment tétrazole sur l'acide carboxylique tel qu'on l'attendait (*bioisosteric matching*) ; le troisième est une superposition du fragment tétrazole sur une chaine aliphatique mais qui est moins intéressante car seule la géométrie est similaire (*Geometry-only Matching*, pas de matching de points colorés).

Nos résultats montrent ainsi que SENSAAS apporte une contribution pertinente aux méthodes d'alignement basées sur la forme, en particulier dans le domaine de l'optimisation de molécules où la propriété d'alignement local s'avère importante pour identifier des molécules prometteuses.

4 Conclusion et perspectives

Nous avons présenté SENSAAS, un nouvel outil qui permet d'aligner des structures moléculaires à partir d'une représentation 3D des surfaces associées. Nous montrons que ces alignements, obtenus à partir de techniques de recalage de nuages de points très populaires dans d'autres domaines sont comparables à ceux des techniques de l'état de l'art, voire meilleurs si l'on évalue leur capacité à réaliser des sub-matching. Les perspectives sont nombreuses, tant au niveau de la méthodologie que des applications potentielles en drug discovery.

Le logiciel SENSAAS est le fruit d'une volonté de créer des outils chémoinformatiques de haut niveau accessibles à d'autres chercheurs et ré-utilisables dans le développement d'autres outils, à l'instar de la suite logicielle RDKiT (`http://rdkit.org`). L'accès et l'utilisation libre des bases de données et des outils sont courants dans le domaine de la bioinformatique structurale mais bien moins fréquent dans le domaine de la chémoinformatique où une propriété intellectuelle peut être associée aux petites molécules. **Afin de promouvoir de nouvelles méthodes utilisant des représentations surfaciques des molécules, nous avons créé un dépôt du code sur GitHub** `https://github.com/SENSAAS/sensaas` **accompagné d'une documenta-**

tion détaillée et de tutoriels sur YouTube. Une version en démo est aussi accessible sur le serveur de l'IPMC : `https://chemoinfo.ipmc.cnrs.fr/SENSAAS`.

Références

[1] Daniel Baum and Hans-Christian Hege. A point-matching based algorithm for 3d surface alignment of drug-sized molecules. In Michael R. Berthold, Robert C. Glen, and Ingrid Fischer, editors, *Computational Life Sciences II*, pages 183–193, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[2] Dominique Douguet and Frédéric Payan. sensaas : Shape-based alignment by registration of colored point-based surfaces. *Molecular Informatics*, 39(8) :2000081, 2020.

[3] Martin A. Fischler and Robert C. Bolles. Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6) :381–395, June 1981.

[4] J. A. GRANT, M. A. GALLARDO, and B. T. PICKUP. A fast method of molecular shape comparison : A simple application of a gaussian description of molecular shape. *Journal of Computational Chemistry*, 17(14) :1653–1666, 1996.

[5] Ashutosh Kumar and Kam Y. J. Zhang. Advances in the development of shape similarity methods and their application in drug discovery. *Frontiers in Chemistry*, 6 :315, 2018.

[6] Xiaofeng Liu, Hualiang Jiang, and Honglin Li. Shafts : A hybrid approach for 3d molecular similarity calculation. 1. method and assessment of virtual screening. *Journal of Chemical Information and Modeling*, 51(9) :2372–2385, 2011. PMID : 21819157.

[7] J. Park, Q. Zhou, and V. Koltun. Colored point cloud registration revisited. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 143–152, 2017.

[8] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009.

[9] Mikko J. Vainio, J. Santeri Puranen, and Mark S. Johnson. Shaep : Molecular overlay based on shape and electrostatic potential. *Journal of Chemical Information and Modeling*, 49(2) :492–502, 2009. PMID : 19196024.

Méthode multi-résolution robuste et non linéaire de recalage de nuages de points 3D par ICP

Ketty Favre[1], Muriel Pressigout[2], Eric Marchand[3], Luce Morin[2]
[1] Univ Rennes, CNRS, IETR - UMR 6164, Rennes, France.
[2] Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France.
[3] Univ Rennes, Inria, CNRS, IRISA, Rennes, France.

Résumé : *Dans cet article un algorithme de résolution robuste, non linéaire et multi-résolution du problème de recalage de nuages de points 3D acquis par LiDAR, nommé GNMR-ICP, est proposé. L'algorithme proposé minimise la distance point-à-plan entre deux nuages de points. Afin de résoudre ce problème d'optimisation de moindres carrés non linéaire, l'algorithme de Gauss-Newton est utilisé. Pour rendre l'algorithme moins sensible aux données aberrantes, des fonctions robustes de type M-estimateurs sont intégrées dans la phase de minimisation. En outre, un partitionnement des données basé sur des octrees est utilisé dans le but d'accélérer le processus.*
L'approche proposée est testée de deux manières distinctes : d'abord, la précision de l'algorithme est évaluée sur le jeu de données ASL (Autonomous Sytem Labs) et comparée avec un autre algorithme de l'état de l'art. GNMR-ICP donne des résultats plus précis que l'autre algorithme multi-résolution évalué (81% de succès contre 43%). Ensuite, l'influence du nombre de niveaux à choisir dans un processus multirésolution est évaluée en terme de temps de calcul. Une tendance à la réduction de ce temps est visible, plus particulièrement dans les environnements structurés.

Mots-clés : navigation, localisation, véhicule autonome, ICP, nuages de points 3D

1 Introduction

En robotique, le recalage de formes 3D est un élément clé pour les applications liées à localisation. L'engouement pour les véhicules autonomes fait du recalage 3D un domaine de recherche très étudié. De plus, des modèles 3D de villes ou de bâtiments peuvent être exploités afin d'aider à la localisation ces véhicules. L'une des approches les plus populaires en robotique pour recaler des formes 3D (par exemple des nuages de points issus de LiDAR) est l'ICP (Iterative Closest Point) [1]. Cet algorithme permet de calculer la transformation rigide recalant deux nuages de points. Pour ce faire, on associe chaque point du nuage à recaler avec le point le plus proche dans le nuage cible. Enfin on minimise la distance entre ces points appariés. On itère ce processus jusqu'à ce que l'erreur de minimisation obtenue passe en dessous d'un certain seuil. Dans l'algorithme original, c'est la distance point-à-point qui est minimisée, mais il a été démontré que la distance point-à-plan est plus robuste et converge plus vite [3]. Dans [5], une variante multi-résolution basée sur des octrees est proposée, la méthode de minimisation choisie utilise l'approximation des petits angles. Cette méthode, utilisée comme référence dans les expériences, sera plus loin notée SA-ICP ("Small Angle ICP").

2 Travail proposé

L'objectif est de proposer une variante plus robuste que les méthodes traditionnelles de l'ICP, tout en gardant une structure simple et modulaire qui rend cet algorithme si populaire. Le critère à minimiser choisi est la distance point-à-plan $d_i^{\perp}$, plus robuste et convergeant plus vite que la distance point-à-point [3] :

$$d_i^{\perp} = \| {}^t\mathbf{n}_i^{\top} \cdot ({}^t\mathbf{R}_s \, {}^s\overline{\mathbf{p}}_i + {}^t\mathbf{t}_s - {}^t\overline{\mathbf{p}}_i) \|_2 \quad (1)$$

avec ${}^s\overline{\mathbf{p}}_i$ et ${}^t\overline{\mathbf{p}}_i$ respectivement les coordonnées correspondant aux points source et cible, ${}^t\mathbf{n}_i$ le vecteur normal estimé grâce au voisinage du point cible (grâce à une ACP (Analyse des Composantes Principales)), ${}^t\mathbf{R}_s$ la matrice 3×3 de rotation et ${}^t\mathbf{t}_s$ le vecteur 3×1 de translation liant les nuages de points source et cible. Afin d'améliorer la robustesse de l'ICP, la méthode d'optimisation retenue pour minimiser la distance est l'algorithme de Gauss-Newton. Cette approche nous permet d'éviter les approximations ainsi que d'introduire des M-estimateurs (Cauchy, Tukey et Huber) dans la phase d'optimisation.

Une structuration des nuages de points en octrees permet de résoudre le problème de recalage par un processus multi-résolution : les nuages de points à une résolution moindre sont d'abord recalés grâce à l'ICP. En résulte une première estimation de la transformation rapide à calculer (grâce au nombre de points réduit à ladite résolution) et ainsi de suite jusqu'à la résolution la plus élevée. Cet algorithme est appelé GNMR-ICP (Gauss-Newton Multi-Résolution ICP). La figure 1 présente le processus complet. Un exemple de recalage réalisé avec GNMR-ICP est présenté en figure 2.

3 Résultats

Dans cette section, tout d'abord, GNMR-ICP est comparé avec la version petits angles de l'ICP multi-résolution, SA-ICP, en termes de précision. Les performances de plusieurs M-estimateurs sont également comparées. Ensuite, l'influence du nombre de niveaux choisi dans le processus multi-résolution est évaluée.

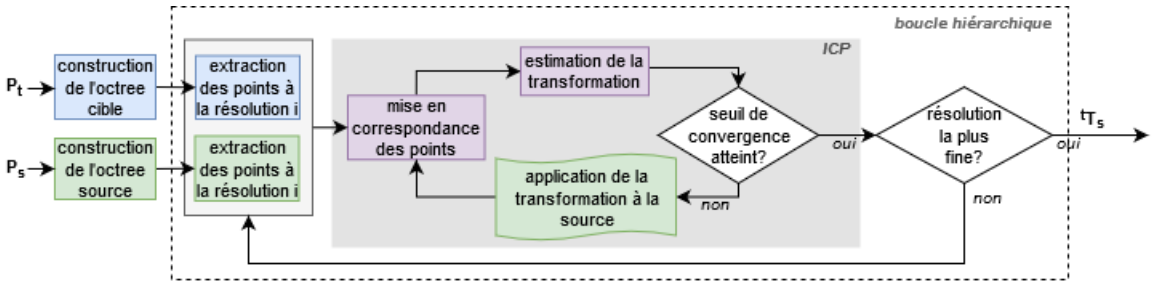
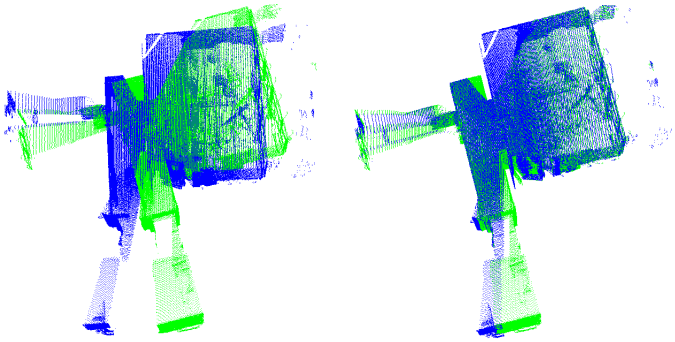


FIGURE 1 – Processus de l'algorithme GNMR-ICP



(a) Avant recalage. (b) Après recalage.

FIGURE 2 – Exemple de recalage avec GNMR-ICP sur 2 scans du jeu de données Autonomous Systems Lab (ASL) *En vert* : Le nuage de points cible. - *En bleu* : le nuage de points source.

3.1 Comparaison de précision sur le jeu de données Autonomous System Labs

Dans cette expérience, les précisions de GNMR-ICP et de SA-ICP sont comparées sur le jeu de données ASL. Ce jeu de données contient la vérité terrain des poses du capteur, cela permet de calculer la distance de l'estimation à la véritable transformation grâce à Δ_t la distance euclidienne et Δ_r la distance géodésique, telles que :

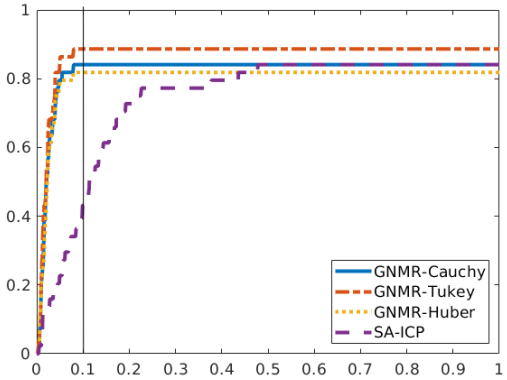
$$\Delta_t = \|{}^t\hat{\mathbf{t}}_s - {}^t\mathbf{t}_s^*\| \tag{2}$$

$$\Delta_r = arccos\left(\frac{trace({}^t\mathbf{R}_s^{*-1}{}^t\hat{\mathbf{R}}_s) - 1}{2}\right) \tag{3}$$

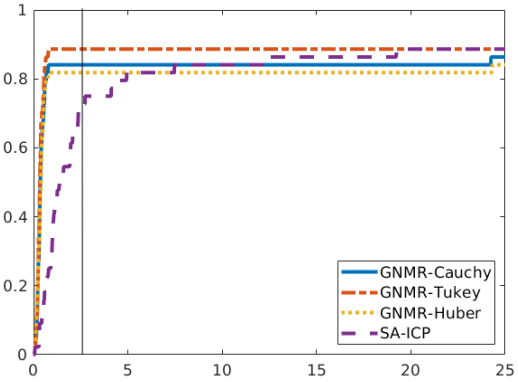
avec ${}^t\hat{\mathbf{t}}_s$ et ${}^t\hat{\mathbf{R}}_s$ les translations et rotations estimées, ${}^t\mathbf{t}_s^*$ et ${}^t\mathbf{R}_s^*$ les translations et rotations de la vérité terrain respectivement. Le recalage est considéré réussi si les erreurs sont inférieures à des seuils respectifs de $t_{tr} = 0,1m$ pour la translation et $t_{rot} = 2.5°$ pour la rotation [2].

Pour ce faire, dans chaque séquence, les nuages de points acquis à deux instants successifs sont recalés grâce aux algorithmes évalués. Le nombre de niveaux de résolution est fixé à 5 pour les deux algorithmes et la résolution des octrees à $3cm$. On ajoute à SA-ICP un filtre de distance maximum lui permettant de supprimer les correspondances de points qui sont éloignés de plus de $1m$.

Les erreurs en translation et en rotation pour GNMR-ICP (avec les trois types de M-estimateurs) et SA-ICP sur la séquences *Apartment* sont données en figure 3. *Notons que les erreurs sont présentées séparément mais que le résultat n'est valide que si la translation et la rotation sont inférieures à respectivement t_{tr} et t_{rot}.* Les courbes



(a) Erreur de translation (m).



(b) Erreur de rotation (°).

FIGURE 3 – Probabilités cumulées des erreurs de translation et de rotation pour la séquence *Apartment*. Les erreurs sur l'axe horizontal et la probabilité sur l'axe vertical. Les barres verticales représentent respectivement les seuils pour un recalage réussi (i.e. t_{tr} et t_{rot}.)

représentent les probabilités cumulées des erreurs. Plus la courbe est en haut à gauche du graphique, meilleur est l'algorithme. Le comportement espéré est d'atteindre 1 avant que le seuil ne soit atteint. Cela signifie que 100% des scans ont été recalés avec une erreur inférieur au seuil, et donc avec succès (selon les seuils t_{tr} et t_{rot}).

Sur la figure 3, on voit que l'algorithme proposé GNMR-ICP est plus précis en translation et en rotation que SA-ICP quel que soit le M-estimateurs choisi. Sur cette séquence, seulement 43% des translations sont estimées précisément avec SA-ICP, alors que pour GNMR-ICP, le moins bon résultat (Huber) atteint 81% de succès en translation. A propos des algorithmes basés Gauss-Newton, leur résultat en terme de précision sont globalement très similaires. Sur cette séquence précise, Tukey donne les

meilleurs résultats et Huber les moins bons. Seules les courbes pour la séquence *Apartment* sont données dans un souci d'espace, mais la tendance est la même pour toutes les séquences.

3.2 Influence de la multi-résolution sur le temps de calcul

Cette expérience a pour but de mettre en évidence l'influence du choix du nombre de niveaux sur le temps de calcul dans un processus multi-résolution.

Ici, GNMR-ICP est évalué avec une résolution d'octree de $3cm$. Le nombre de niveaux de résolution varie de 1 (en d'autres termes, un ICP point-à-plan sans multi-résolution) à 6 niveaux. L'évaluation est faite sur un ordinateur équipé d'un processeur Intel Xeon W-2133, 3.6GHz et 32Go de RAM.

En figure 4 les temps moyen de calcul sur chaque séquence du jeu de données ASL est donné (seul le temps de l'estimation est pris en compte). Le temps d'estimation diminue lorsque le nombre de niveaux augmentent (*Apartment*, *ETH*, *Gazebo*, *Stairs* et *Mountain*).

Le GNMR-ICP avec 4 niveaux est le plus rapide. Cela montre l'intérêt principal de la mutli-résolution : le niveau à plus gros grains est composé de peu de points, rendant l'estimation rapide pour ce niveau. En conséquence, l'estimation pour les niveaux plus fins est initialisée à proximité de la solution attendue donc moins d'iterations sont nécessaires, par rapport à un ICP classique où tous les points sont considérés. On remarque qu'avec 5 et 6 niveaux le temps de calcul augmente légèrement. Sur la séquence *Apartment*, le temps de calcul est divisé par 2 en utilisant 4 niveaux de résolution au lieu de 1. Sur la séquence *Stairs* ce facteur est d'environ 1,8. Il est intéressant de noter que les meilleurs facteurs de réduction sont obtenus sur les séquences les plus structurées. A l'inverse, la séquence *Wood*, qui est la moins structurée, est la seule séquence qui ne suit pas la règle de la diminution du temps de calcul.

Le processus de multi-résolution montre son intérêt dans les environnements structurés car il permet de réduire la présence d'information redondante (comme les points présents sur les mêmes plans) sans trop dégrader les données originales.

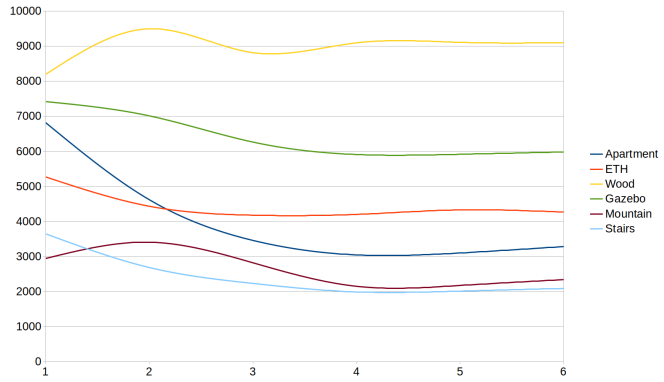


FIGURE 4 – Temps de calcul en fonction du nombre de niveaux utilisés dans GNMR-ICP

4 Conclusion et perspectives

Dans cet article, un algorithme de résolution robuste, non linéaire et multi-résolution du problème de recalage de nuages de points 3D, dénommé GNMR-ICP, est proposé.

La première expérience compare la précision de trois GNMR-ICP utilisant des M-estimateurs différents et la variante de l'ICP utilisant l'approximation des petits angles dans sa version multi-résolution sur le jeu de données ASL [4]. Les trois variantes basées sur un Gauss-Newton génèrent plus d'estimations valides que la variante des petits angles. Concernant le choix des M-estimateurs, en considérant l'ensemble du jeu de données, la différence en termes de précision n'est pas assez significative pour dire si l'un est meilleur que les autres.

La seconde expérience a pour but de montrer l'influence du nombre de niveaux utilisés dans la multi-résolution sur le temps de calcul avec GNMR-ICP. Dans les environnements structurés, l'utilisation d'un processus multirésolution a tendance à diminuer les temps de calcul. Cette tendance se remarque plus particulièrement dans les scénarios comportant des fortes structures planaires (ce qui sous entend de l'information redondante).

L'approche proposée peut être améliorée en intégrant la distance plan-à-plan dans l'étape de minimisation. En effet, dans un environnement urbain, les structures faites par l'homme sont souvent composées de grandes surfaces planaires. Exploiter ces plans amènerait à réduire la dimension du problème.

Références

[1] P. J. Besl and N. D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2) :239–256, February 1992.

[2] Martin Magnusson, Narunas Vaskevicius, Todor Stoyanov, Kaustubh Pathak, and Andreas Birk. Beyond points : Evaluating recent 3D scan-matching algorithms. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3631–3637, May 2015.

[3] F. Pomerleau, F. Colas, and R. Siegwart. A Review of Point Cloud Registration Algorithms for Mobile Robotics. *Foundations and Trends® in Robotics*, 4(1) :1–104, May 2015.

[4] François Pomerleau, Ming Liu, Francis Colas, and Roland Siegwart. Challenging data sets for point cloud registration algorithms. *The International Journal of Robotics Research*, 31(14) :1705–1711, December 2012.

[5] M. Vlaminck, H. Luong, and W. Philips. Multi-resolution ICP for the efficient registration of point clouds based on octrees. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 334–337, May 2017.

A Region of Interest (ROI) based cross-layer system for low latency video streaming over Vehicular Ad-hoc NETworks (VANETs)

Mohamed-Aymen Labiod[1], Mohamed Gharbi[2], François-Xavier Coudoux[2] et Patrick Corlay [2]
[1] LISSI-TincNET Research Team University Paris-Est Creteil, France

[2] Univ. Polytechnique Hauts-de-France, CNRS, Univ. Lille, YNCREA, Centrale Lille,
UMR 8520 – IEMN, DOAE, F-59313 Valenciennes, France

Résumé : *On assiste à une explosion des applications temps réel utilisant les informations vidéo dans les communications véhiculaires notamment avec l'arrivée des véhicules autonomes. Dans ce travail, nous proposons un algorithme inter-couches adaptatif d'acheminement des trames de données vidéo au niveau de la couche MAC de la norme véhiculaire IEEE 802.11p qui permet d'améliorer de telles transmissions. Le système proposé exploite les caractéristiques spécifiques du contenu vidéo à travers la priorisation de la région d'intérêt (ROI) du flux vidéo lors de la transmission. Les résultats de simulations réalistes démontrent que pour les communications vidéo nécessitant une faible latence, le système proposé offre des améliorations significative de la qualité vidéo de bout à bout pouvant atteindre des gains en PSNR de 7 dB pour la ROI.*

Mots-clés : HEVC, Région d'intérêt (ROI), Inter-couches, Réseaux sans-fil Véhiculaire (VANET), faible latence.

1 Introduction

Plusieurs applications émergentes dans les communications véhiculaires nécessitent une transmission vidéo efficace avec des garanties de QoS. Nous pouvons citer : l'assistance au conducteur, le contrôle de véhicule à distance et la vidéo surveillance pour des applications d'urgence. L'inconvénient est que les réseaux véhiculaires (VANET) sont des environnements de transmission sévères avec un pourcentage significatif de perte de paquets. Cela nécessite une stratégie de transmission vidéo de bout en bout adéquate et des normes appropriées [1]. En outre, il est connu que l'œil humain décrit la scène par une succession de fixation sur certaines zones importantes appelés régions d'intérêt (ROI) sur la base des quelles l'observateur va généralement baser son jugement de la qualité vidéo. Ainsi, ces dernières années, un grand nombre de travaux de recherche ont été menés afin de proposer des encodeurs ROI spécifiques. Le principal défi de ces algorithmes est d'assurer un partitionnement du débit sur des régions respectant à la fois la ROI et les contraintes de débit. Ainsi, de nombreuses contributions basées sur les normes d'encodage vidéo H.264 / AVC ou H.265 / HEVC ont introduit des algorithmes de contrôle de débit qui prennent en compte les ROIs pour l'allocation des bits [2]. La norme véhiculaire IEEE 802.11p a une QoS variable qui prend en charge des classes de service différenciées au niveau de la couche MAC. Notamment avec le mode Enhanced Distributed Channel Access (EDCA) qui définit quatre catégories d'accès (AC) aux canaux ou priorités. Ainsi, l'EDCA dispose de quatre files d'attente représentant différents niveaux de priorité où chaque AC a un type de trafic, à savoir : le trafic de fond (AC0), le meilleur effort (AC1), la vidéo (AC2) et la voix (AC3). Dans [3] les auteurs ont proposé un système inter-couches permettant d'acheminer chaque trame de la vidéo transmise vers la file d'attente AC la plus appropriée en prenant en considération : la structure de prédiction temporelle du processus de codage vidéo HEVC, l'importance de l'image dans le flux vidéo et l'état instantané de la charge de trafic du réseau.

2 Travail proposé

Dans ce travail, nous proposons un système inter-couche adaptatif à faible complexité qui exploite les caractéristiques spécifiques du contenu vidéo, avec la ROI, dans l'acheminement des trames de données vidéo au niveau de la couche MAC. En effet, nous intervenons sur deux parties de la chaîne de transmission sous la forme d'un schéma adaptatif inter-couche. Au niveau de la couche d'application, la région d'intérêt est déterminée et est séparée du reste de la scène. Ensuite, les deux régions brutes sont encodées séparément par un encodeur INTRA H.265 / HEVC. Ce choix d'encodage INTRA répond également à la contrainte de latence. Les deux régions sont codées différemment avec une meilleure qualité vidéo d'encodage attribué au flux ROI pour un débit total fixé. À la sortie de l'encodeur, les flux générés sont envoyés au niveau inférieur de la pile de protocoles. Au niveau de l'acheminement des trames nous avons développé un algorithme basé sur celui dans [3] qui permet d'appliquer une discrimination des trames de données vidéo. Afin de protéger le flux ROI au détriment du flux NON-ROI, nous utilisons les autres AC en plus de celui utilisé pour la vidéo. L'algorithme d'acheminement adaptatif proposé attribue dynamiquement pour chaque trame de données vidéo l'AC le plus approprié au niveau de la couche MAC. Il prend en compte l'état de la charge de trafic réseau et la zone de région de chaque paquet de trame, c'est-à-dire si ROI ou NON-ROI. Au niveau du récepteur, si tous les paquets sont correctement reçus, les deux régions du flux peuvent être correctement décodés et peuvent être fusionnées pour reconstruire la vidéo. Dans le cas de perte de paquets, le système proposé applique un mécanisme de dissimulation des erreurs avec une copie d'image permettant ainsi une

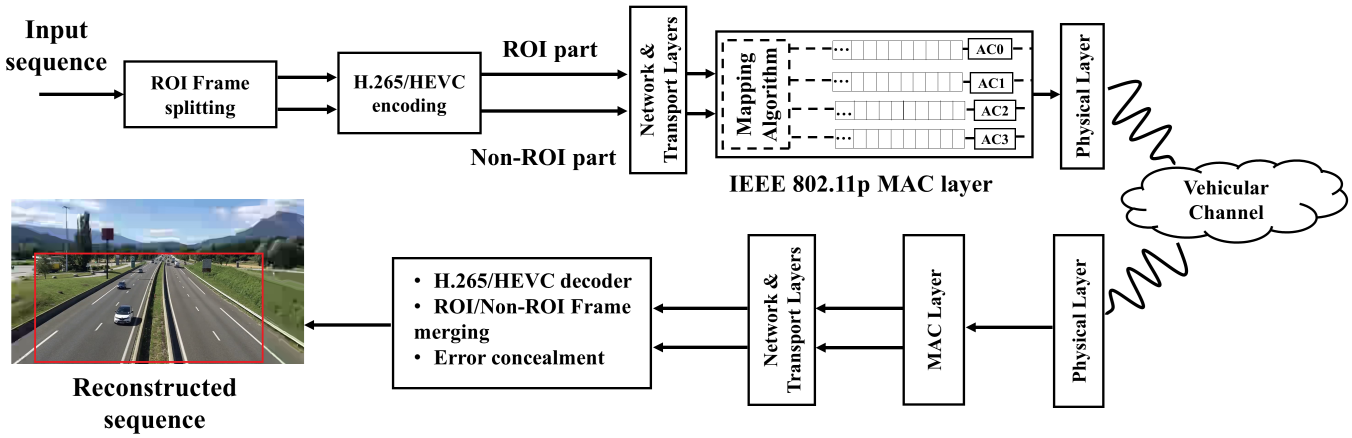


FIGURE 1 – Schéma block du système de transmission basé sur la région d'intérêt

TABLE 1 – Nombre de paquets perdus et PSNR moyen par région pour chaque algorithme d'acheminement.

	Nombre de paquets perdus	PSNR de la ROI (dB)	PSNR de la NON-ROI (dB)	PSNR sur toute l'image (dB)
EDCA IEEE 802.11p	1367	22.76	28.64	27.02
Acheminement adaptatif basé ROI	**269**	**29.52**	**29.13**	**29.16**

séquence reconstruite moins dégradée.

3 Résultats

Nous avons évalué l'efficacité de ce système, en termes de nombre de paquets perdus et également avec la métrique Peak Signal-to-Noise Ratio (PSNR) comme indicateur objectif de la qualité vidéo reçue. Par ailleurs, l'évaluation s'est faite dans le cadre d'un Framework véhiculaire réaliste dédier basé sur le simulateur NS2 [3]. L'efficacité de l'algorithme proposé est étudiée en comparaison à un schéma de transmission conventionnelle qui est l'algorithme EDCA sur lequel se base le IEEE 802.11p.

Dans l'ensemble, nous montrons que la qualité vidéo de la séquence est améliorée. Pour la partie ROI, les gains PSNR moyens peuvent atteindre 6,75 dB par rapport à l'EDCA du l'IEEE 802.11 comme l'illustre le tableau 1. Nous montrons également un gain en ce qui concerne le nombre de paquets reçus et également un gain en PSNR sur toute l'image de l'ordre de 2 dB.

4 Conclusion et perspectives

A travers cette solution, nous proposons un système inter-couche basé sur la ROI de la vidéo, permettant une amélioration de la transmission vidéo dans les applications à faible latence sur les réseaux de véhicules. Le système inter-couches proposé permet une classification des trames basée sur le protocole IEEE 802.11p. En effet, la stratégie s'appuie sur l'acheminent des trames de données vidéo dans les files d'attentes les plus appropriés afin d'offrir une meilleure QoS. Les informations ROI de la vidéo et le remplissage des files d'attentes du buffer au niveau de la couche MAC permettent à l'algorithme proposé de choisir la meilleure option pour l'acheminement des trames de données vidéo. Les résultats établis dans un environnement de véhicule réaliste illustrent une amélioration de la qualité de service et une amélioration de la qualité vidéo de bout en bout.

[1] A. Vinel, E. Belyaev, K. Egiazarian and Y. Koucheryavy, "An Overtaking Assistance System Based on Joint Beaconing and Real-Time Video Transmission," in IEEE Transactions on Vehicular Technology, vol. 61, no. 5, pp. 2319-2329, Jun 2012.

[2] M. Meddeb, M. Cagnazzo, and B. Pesquet-Popescu, "ROI-based rate control using tiles for an HEVC encoded video stream over a lossy network," in 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 1389–1393.

[3] M. A. Labiod, M. Gharbi, F.-X. Coudoux, P. Corlay, and N. Doghmane, "Enhanced adaptive cross-layer scheme for low latency HEVC streaming over vehicular ad-hoc networks (VANETs)," Vehicular Communications, vol. 15, pp. 28–39, 2019.

Conversion thermique-visible en imagerie faciale

Khawla Mallat, Jean-Luc Dugelay
EURECOM
Sophia-Antipolis, France
{mallat, dugelay}@eurecom.fr

Résumé : *La reconnaissance faciale est largement employée dans différents contextes. Cependant, les systèmes de reconnaissance faciale déployés sont principalement conçus pour traiter des données issues du spectre visible. Ces systèmes sont parfois peu performants dans des environnements non contrôlés (e.g. sombres). L'imagerie thermique peut apporter une solution. Cependant les performances en imagerie thermique sont actuellement limitées à cause de la faible résolution des images et du manque de couleur et de texture, et aussi de données pour l'apprentissage. Nous proposons dans cet article de convertir les images de visages acquises en thermique dans spectre visible afin de profiter des récents progrès réalisés en reconnaissance faciale dans le domaine visible et de garantir une intégration simple de la technologie thermique dans les systèmes existants. Nous présentons, également, l'application de la conversion spectrale du visage dans d'autres applications.*

Mots-clés : Imagerie thermique, Reconnaissance faciale, Conversion inter-spectrale.

1 Introduction

Étant donné que les traitements en imagerie thermique restent assez rares, très peu de bases de données publiques sont disponibles. Par conséquent, les technologies d'apprentissage en profondeur ne permettent pas de développer des systèmes de reconnaissance faciale fiables fonctionnant dans le spectre thermique. Dès lors, il est intéressant de proposer une approche par conversion afin de pouvoir bénéficier des nombreux progrès réalisés dans le domaine du traitement d'images dans le spectre visible. Avec la montée fulgurante de l'apprentissage profond, plusieurs travaux se sont basés sur les réseaux adverses génératifs (GAN) [3] pour traduire des images d'un domaine à un autre. Particulièrement, Zhang et al. [8] ont utilisé le réseau pix2pix, proposé par Isola et al. [3], couplée à une fonction de coût de reconnaissance faciale pour préserver les informations liées à l'identité des personnes.

Dans cet article, nous proposons une nouvelle approche pour convertir les images du spectre visible au spectre thermique et vice versa. La conversion-interspectrale a été utilisé dans plusieurs tâches d'analyse d'image de visage, en particulier l'authetification par reconnaissance faciale, la détection de points caractéristiques du visage et l'attaque d'usurpation d'identité.

2 Travail proposé

Pour convertir des images faciales d'un spectre à un autre, nous nous sommes basés sur le réseau de raffinement en cascade (CRN) [2]. Nous avons choisi le CRN comme bloc de base pour notre génération d'images car il prend en compte des informations multi-échelles et est basé sur l'apprentissage d'un nombre limité de paramètres. Cela permet une génération d'images de plus haute résolution avec un nombre limité de données en comparaison avec d'autres solutions basées sur le GAN. Le CRN est un réseau neuronal convolutif qui se compose de modules de raffinement interconnectés. Le premier module considère l'espace de plus faible résolution (4x4 dans notre cas). Cette résolution est augmentée dans les modules suivants jusqu'au dernier module (128x128 dans notre cas), correspondant à la résolution de l'image cible. Pour l'entrainement et l'évaluation de notre modèle de conversion, on a utilisé la base de données VIS-TH [6] qui contient des images de visages dans le spectre visible et thermique acquises simultanément. Les images visibles générées à partir d'images thermiques, présentées dans la colonne (c) de la figure 1, sont de qualité visuelle satisfaisante, en synthétisant des traits visuels informatifs (nez, yeux, bouche...), mais certains attributs sont incorrects comme la couleur de la peau. La conversion du spectre visible au spectre thermique, illustrée dans la colonne (d) de la figure 1, donne des thermogrammes faciaux proches de la vérité terrain. Cela s'explique par le fait que l'on passe d'un domaine riche en information texturale (spectre visible) à un domaine moins informatif (spectre thermique).

3 Résultats et discussion

Dans cette section, on présente l'utilisation de la conversion inter-spectrale dans différentes tâches liées à l'analyse de visage et à la reconnaissance faciale : authentification par reconnaissance faciale, détection des points caractéristiques du visage et usurpation d'identité.

3.1 Evaluation quantitative

Deux métriques de qualité, le rapport signal/bruit de crête (PSNR) et la mesure de l'indice de similarité structurelle (SSIM), sont sélectionnés pour évaluer la qualité visuelle des images visibles synthétisées.

Le tableau 1 rapporte les valeurs PSNR et SSIM obtenues lors de la comparaison entre les images visibles de visages synthétisées, générées à l'aide de différents modèles de synthétisation d'images, et les images visibles qui constituent la vérité-terrain. Les résultats obtenus, ∼17dB

FIGURE 1 – Exemple d'images de visage : (a) thermiques natives (b) visibles natives (c) thermique synthétisées (d) visibles synthéthisée.

pour PSNR et ∼0,65 pour SSIM, ne reflètent pas une grande fidélité des images visibles synthétisées par rapport à la vérité-terrain. Les visages visibles synthétisés sont générés à partir de signatures thermiques faciales, qui représentent des informations différentes. Les modèles de synthèse d'image thermique-visible visent à reproduire une estimation des propriétés du spectre visible, difficile à les reproduire avec précision, comme la texture, la couleur et des informations géométriques plus détaillées.

En comparant les résultats obtenus pour les méthodes de l'état de l'art, le terme de perte d'identité introduit par Zhang et al. [8] dans le modèle proposé par Isola et al. [3] a mené une légère augmentation de qualité. Cependant, une amélioration plus importante existe avec pour notre modèle proposé.

3.2 Authentification par reconnaissance faciale

La synthétisation des visages du spectre thermique au spectre visible est fondamentale pour effectuer la reconnaissance faciale inter-spectrale car elle simplifie l'intégration de la technologie thermique dans les systèmes de reconnaissance faciale déjà déployés et permet la vérification manuelle des visages. On évalue la reconnaissance faciale inter-spectrale : dans notre cas, la personne sera identifiée (on utilise ici LightCNN) en comparant son thermogramme facial acquis au moment de l'authentification à son image d'enrôlement réalisé dans le spectre visible [4]. On montre que l'on peut ainsi obtenir des performances intéressantes par rapport aux méthodes de conversion disponible dans l'état de l'art, illustrées dans le tableau, 1 avec une amélioration relative de 71.43% par rapport au modèle de Zhang et al. [8].

	PSNR	SSIM
Isola et al. [3]	17.247 (±2.855)	0.6485 (±0.123)
Zhang et al. [8]	17.257 (±2.897)	0.6509 (±0.125)
Notre approche	17.8144 (±3.635)	0.6725 (±0.131)

TABLE 1 – PSNR et SSIM calculés sur les images visibles synthétisées obtenues à l'aide de notre approche proposée en comparaison avec les méthodes de l'état de l'art en conversion inter-spectrale.

3.3 Détection de points caractéristiques du visage

La détection de points caractéristiques du visage est une étape essentielle du traitement des images de visage. Avec le succès des approches basées sur l'apprentissage profond, la performances de la détection des points caractéristiques du visage a été considérablement améliorées. Cependant, cette amélioration est principalement dans le spectre visible. Compte tenu du manque de bases de données thermiques annotées en termes de points caractéristiques du visage, il n'y a pas ou peu de méthodes dans la littérature. On propose donc de synthétiser une base de données thermique en convertissant des bases de donnéess conçus pour la détection des points caractéristiques en visible (HELEN et LFPW). En entrainant deux modèles de détection des points caractéristiques sur les bases de donnéess synthétisées, on a amélioré la performance, en terme d'écart quadratique moyen normalisé, d'un facteur de 2 par rapport à celles associées au seul modèle public entrainé sur une base de données thermiques de haute résolution [5].

3.4 Attaques d'usurpation d'identité

La technologie thermique peut également de facto être utilisée comme une contre-mesure efficace face aux attaques d'usurpation d'identité (i.e. leurrage) [1], mais cela n'est vrai que dans le cas d'attaques physiques (masque, photo...). On propose une nouvelle attaque logique en insérant des thermogrammes synthétisés par conversion du spectre visible au spectre thermique afin d'évaluer la robustesse des méthodes de détection d'usurpation d'identité. En classant les images en imposteur et authentique, nous montrons que le taux d'erreurs égal (EER) obtenu augmente sensiblement de 0.2% à 11.6% ; indiquant ainsi que la conversion est suffisamment efficace pour leurrer certains systèmes de reconnaissance faciale [7].

4 Conclusion et perspectives

La conversion inter-spectrale visible–thermique est donc intéressante pour que des systèmes basés sur le spectre thermique puissent suivre l'évolution technologique de ce qui se fait dans le domaine visible. Dans ce résumé, on

	Isola et al. [3]	Zhang et al. [8]	Notre approche
Neutral	48	54	82
Expression	37.33	38.33	67.66
Head pose	14.5	15.5	30
Occlusion	16.4	25	44.8
Illumination	29.6	35.2	63.6
Average	29.166	33.606	57.612

TABLE 2 – Précision de la reconnaissance faciale inter-spectrale à travers de multiples variations en utilisant le modèle LightCNN.

a présenté l'application de la conversion multi-spectrale pour différentes tâches liées à la reconnaissance faciale et l'avantage qu'elle peut apporter pour améliorer les performances pour chacune de ces tâches. Cependant, la qualité des images synthétisés, en particulier les images visibles, reste encore à améliorer afin de réduire les artéfacts de synthétisation. Par conséquent, nous poursuiverons actuellement nos travaux afin de générer une image visible à partir d'un thermogramme facial plus réaliste.

Références

[1] Sushil Bhattacharjee, Amir Mohammadi, and Sébastien Marcel. Spoofing deep face recognition with custom silicone masks. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2018.

[2] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[4] Khawla Mallat, Naser Damer, Fadi Boutros, Arjan Kuijper, and Jean-Luc Dugelay. Cross-spectrum thermal to visible face recognition based on cascaded image synthesis. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[5] Khawla Mallat and Jean-Luc Dugelay. Facial landmark detection on thermal data via fully annotated visible-to-thermal data synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE.

[6] Khawla Mallat and Jean-Luc Dugelay. A benchmark database of visible and thermal paired face images across multiple variations. In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2018.

[7] Khawla Mallat and Jean-Luc Dugelay. Indirect synthetic attack on thermal face biometric systems via visible-to-thermal spectrum conversion. In IEEE, editor, *CVPR 2021, 34th IEEE Conference on Computer Vision and Pattern Recognition Workshops, 19-25 June 2021, (Virtual Conference)*, 2021.

[8] Teng Zhang, Arnold Wiliem, Siqi Yang, and Brian Lovell. Tv-gan : Generative adversarial network based thermal to visible face recognition. In *2018 international conference on biometrics (ICB)*, pages 174–181. IEEE, 2018.

Recalage de deux nuages de points 3D au rapport d'échelle non uniforme

Flora Quilichini, Thomas Fisichella, Frédéric Payan, Marc Antonini
Laboratoire I3S

Résumé : *Ce travail rentre dans le contexte d'un projet de recherche en collaboration avec le CHU de Nice dont l'objectif est le recalage rigide de deux nuages de points 3D représentant des organes du corps humain pour une application en chirurgie augmentée. Plus précisément, l'objectif est d'estimer la transformée $T(s, \Omega, t)$ (i.e. échelle, rotation, translation) entre un objet complet (source) et un objet partiel (cible) dont l'échantillonnage et le facteur d'échelle diffèrent. La première étape du recalage consiste au calcul de descripteurs locaux sur chacun des deux objets qui seront nécessaires pour la mise en correspondance des points. Pour cela, nous avons fait le choix des Growing Least Squares [1] car ce sont des descripteurs robustes au facteur d'échelle. En deuxième étape, nous utilisons un algorithme d'optimisation sur les paires de correspondance afin d'estimer la transformée $T(s, \Omega, t)$. Notre contribution est une extension de l'algorithme Fast Global Registration [3] au recalage de deux objets présentant un rapport d'échelle non uniforme selon les axes X, Y et Z. Nous avons testé notre méthode sur plusieurs modèles synthétiques possédant les propriétés de surface des organes réels (caractère lisse, présence de symétries, et redondance de la géométrie) et nous obtenons des premiers résultats concluants.*

Mots-clés : Recalage, nuage de points 3D, estimation d'échelle

1 Introduction

Le recalage de nuages de points 3D est un sujet largement étudié en robotique et en vision par ordinateur. Les nombreuses contraintes posées par les données (présence de bruit, occlusion, désordre, chevauchement partiel, modalités d'acquisition différentes, déformations) en font un problème vaste et difficile, notamment lorsqu'il s'agit de données médicales acquises à partir de modalités différentes. Les données réelles que l'on doit recaler proviennent d'un scanner à rayons X et d'un système multivues de caméras RGBD et présentent les contraintes suivantes : chevauchement partiel, différence d'échantillonnage, facteur d'échelle différent et distorsion.

Les problèmes de recalage sont classiquement résolus en trois étapes. La première étape consiste en une segmentation où les objets d'intérêt sont extraits de la scène. La deuxième étape consiste à mettre en correspondance un sous-ensemble de points localisés sur les objets d'intérêts par appariement. Cet appariement se fait généralement par la recherche de similarité entre des descripteurs locaux décrivant la surface au voisinage des points d'intérêt. La troisième et dernière étape consiste à calculer, à l'aide d'un algorithme d'optimisation exécuté sur les paires de correspondances, la matrice de transformation permettant de recaler les objets d'intérêt. Nous avons limité notre étude au recalage de deux modèles 3D déjà pré-segmentés (source et cible) et nous nous sommes focalisés uniquement sur les deux dernières étapes. Nous avons repris l'algorithme d'optimisation *Fast Global Registration* (FGR) [3] car il présentait deux avantages majeurs :

- *Robustesse aux paires aberrantes.* Sa fonction objectif possède un terme de pénalisation des mauvaises paires inversement proportionnel au terme favorisant les bonnes paires.
- *Rapidité & Précision.* L'algorithme atteint une précision de recalage comparable à celle d'un algorithme d'ajustement local correctement initialisé, sans nécessiter une telle initialisation (très coûteuse en temps de calcul).

Nous utilisons comme descripteur les *Growing Least Squares* (GLS) [1] car ce dernier est robuste à la quasi-totalité des contraintes mentionnées plus haut. Les auteurs de [1] se servent également des GLS pour estimer le rapport d'échelle, dans leur cas identique selon les axes X, Y et Z, entre la source et la cible. Cela leur permet de recaler des modèles acquis à des échelles différentes.

2 Travail proposé

Ici, nous allons plus loin en proposant une méthode qui recale deux nuages de points 3D dont le rapport d'échelle est différent suivant les axes. Pour cela, nous avons modifié FGR de manière à lui faire estimer la matrice échelle S conjointement à la traditionnelle matrice rotation-translation Ωt. Nous utilisons les GLS mais tout autre descripteur robuste à l'échantillonnage et aux variations d'échelle peut lui être substitué. Pour construire notre extension, nous avons repris la fonction objectif à minimiser définie dans FGR [3] et nous y avons intégré le terme d'échelle (en gras) :

$$E(\Omega t, \mathbf{S}, \mathbb{L}) = \sum_{(p,q) \in pairs} l_{p,q} * ||p - \mathbf{S} * \Omega t * q||^2 + \Psi(l_{p,q}) \quad (1)$$

Dans [3], l'objectif E est minimisé par méthode itérative alternée. Les mises à jour du processus de ligne $\mathbb{L}$ sont obtenues en dérivant E par rapport à $l_{p,q}$ et celles de Ωt en utilisant un algorithme de type Gauss-Newton. Intégrer les paramètres d'échelle dans une transformée globale T (échelle * rotation + translation) est délicat.
En effet, la matrice d'échelle $\mathbf{S}$ va introduire de la non-linéarité dans la méthode de Gauss-Newton (les coefficients de la matrice Jacobienne ne sont plus des constantes mais dépendent des paramètres d'échelle, de rotation et

de translation). A la place, nous avons décidé de scinder T en deux parties (Ωt et **S**). Ωt est mise à jour pareillement à [3], à **S** fixée. Puis nous appliquons la méthode de Gauss-Newton une deuxième fois pour mettre à jour **S**, à Ωt fixée. Le pseudo-code ci-dessous reprend l'algorithme initial présenté dans [3] avec notre contribution indiquée en bleu. Pour plus de détails concernant l'optimisation des *line process* ($l_{p,q}$) ou la méthode Gauss Newton, le lecteur est prié de se référer à l'article de référence [3]. Le descripteur utilisé dans [3] pour déterminer l'ensemble des points d'intérêt et leurs caractéristiques associées est le *Fast Point Feature Histograms* (FPFH) [2] ; or, ce dernier n'étant pas robuste au changement d'échelle, nous lui avons substitué les *Growing Least Squares* [1]. Les paires de points correspondants sont ensuite obtenues en suivant une méthode basée sur la variation d'échelle [1].

Algorithm 1 Proposed FGR-based algorithm

input : A pair of surfaces (**P, Q**)
output : Transformation **T** that aligns **Q** to **P**
Compute GLS features **F(P)** and **F(Q)**
Build K by computing nearest neighbors between **F(P)** and **F(Q)** in log-scale space
$\mathbf{T} \leftarrow \mathbf{I}$, $\mu \leftarrow D^2$
while $iter < N_{max}$ *or* $\mu > \delta^2$ **do**
  $\Omega t \leftarrow \mathbf{0}$, $\delta\Omega t \leftarrow \mathbf{0}$, $\mathbf{S} \leftarrow \mathbf{0}$, $\delta S \leftarrow \mathbf{0}$
  Get which variable to update (Ωt, S or both)
  $(upd\Omega t, updS) \leftarrow updateVariable(iter)$
  for $(p, q) \in K$ **do**
    Compute $l_{p,q}$
    if $is\Omega t$ **then**
      $\delta\Omega t \leftarrow applyGaussNewton(\mathbf{p} - \mathbf{T} * \mathbf{q} * l_{p,q})$
      $\Omega t \leftarrow \Omega t + \delta\Omega t$
    end if
    if isS **then**
      $\delta S \leftarrow applyGaussNewton(\mathbf{p} - \mathbf{T} * \mathbf{q} * l_{p,q})$
      $\mathbf{S} \leftarrow \mathbf{S} + \delta\mathbf{S}$
    end if
  end for
  if $is\Omega t$ **then**
    $\mathbf{T} \leftarrow updateT(\Omega t, \mathbf{T})$
  end if
  if isS **then**
    $\mathbf{T} \leftarrow updateT(\mathbf{S}, \mathbf{T})$
  end if
  Every four iterations, $\mu \leftarrow \mu/2$
  $iter \leftarrow iter + 1$
end while

3 Résultats

Les données cibles ont été générées à partir de données synthétiques sources auxquelles nous avons appliqué - dans l'ordre - les opérations suivantes :

- *Une transformation* Cette première étape nous permet de modéliser un changement de repère lors de l'acquisition de la donnée cible. Nous avons appliqué les mêmes transformations à nos données sources, à savoir : un changement d'échelle (uniforme ou non), suivi d'une rotation, puis d'une translation.
- *Un sous-échantillonnage par deux du nombre de points suivi d'un ré-échantillonnage des points restants.* Ici, l'idée est de modéliser deux acquisitions d'un même objet obtenues au moyen de deux dispositifs de résolutions différentes. Dans notre jeu de données, un maillage cible fixé contient deux fois moins de points que sa source associée, et ses points sont déplacés sur la surface du de manière à ce que leurs positions ne coincident pas avec celles des points de la source.
- *Une coupe (cropping) - facultatif* Cette dernière étape nous permet de modéliser une donnée cible "incomplète" par raport à la source. Nous avons effectué des coupes de la source suivants différents axes. Afin de pouvoir mesurer la robustesse de notre algorithme à la surface de recouvrement (overlapping) entre la source et la cible, chaque donnée cible a été générée, avec une proportion différente du maillage source (fémur : 50%, hanche : 60% et veines : 40%)

Le tableau ci-dessous synthétise les résultats obtenus sur nos modèles. Nous obtenons des résultats de recalage très satisfaisants pour la plupart des tests. Seul le cas d'un fragment de veine - qui possède le pourcentage de recouvrement le plus faible, et une géométrie plutôt redondante dans la donnée complète - échoue. On remarque également que l'erreur de recalage diminue avec l'augmentation de la surface partagée entre la source et la cible.

4 Conclusion et perspectives

Nous avons développé une extension de FGR permettant de recaler deux nuages de points dont le rapport d'échelle est différent selon les axes X, Y et Z. Nous obtenons de bons résultats sur des données synthétiques. Le principal facteur limitant de notre solution est le descripteur GLS, qui n'est théoriquement cohérent que lorsque la matrice de mise à l'échelle entre la source et la cible est uniforme. Afin de présenter une méthode complète, nous avons calculé les GLS en présence de disparités suffisamment faibles des coefficients d'échelle pour que le descripteur reste pertinent. Notons que notre algorithme parvient à estimer de plus grandes disparités, à condition qu'on lui donne un nombre suffisant de bonnes paires de correspondances. Le développement d'un descripteur robuste aux déformations d'échelle est la prochaine étape vers une solution plus générale au problème du recalage avec transformation non uniforme, voire déformations.

Références

[1] Nicolas Mellado, Matteo Dellepiane, and Roberto Scopigno. Relative scale estimation and 3d registration of multi-modal geometry using growing least squares. *IEEE Transactions on Visualization and Computer Graphics*, 22 :1–1, 12 2015.

[2] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009.

[3] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *ECCV*, 2016.

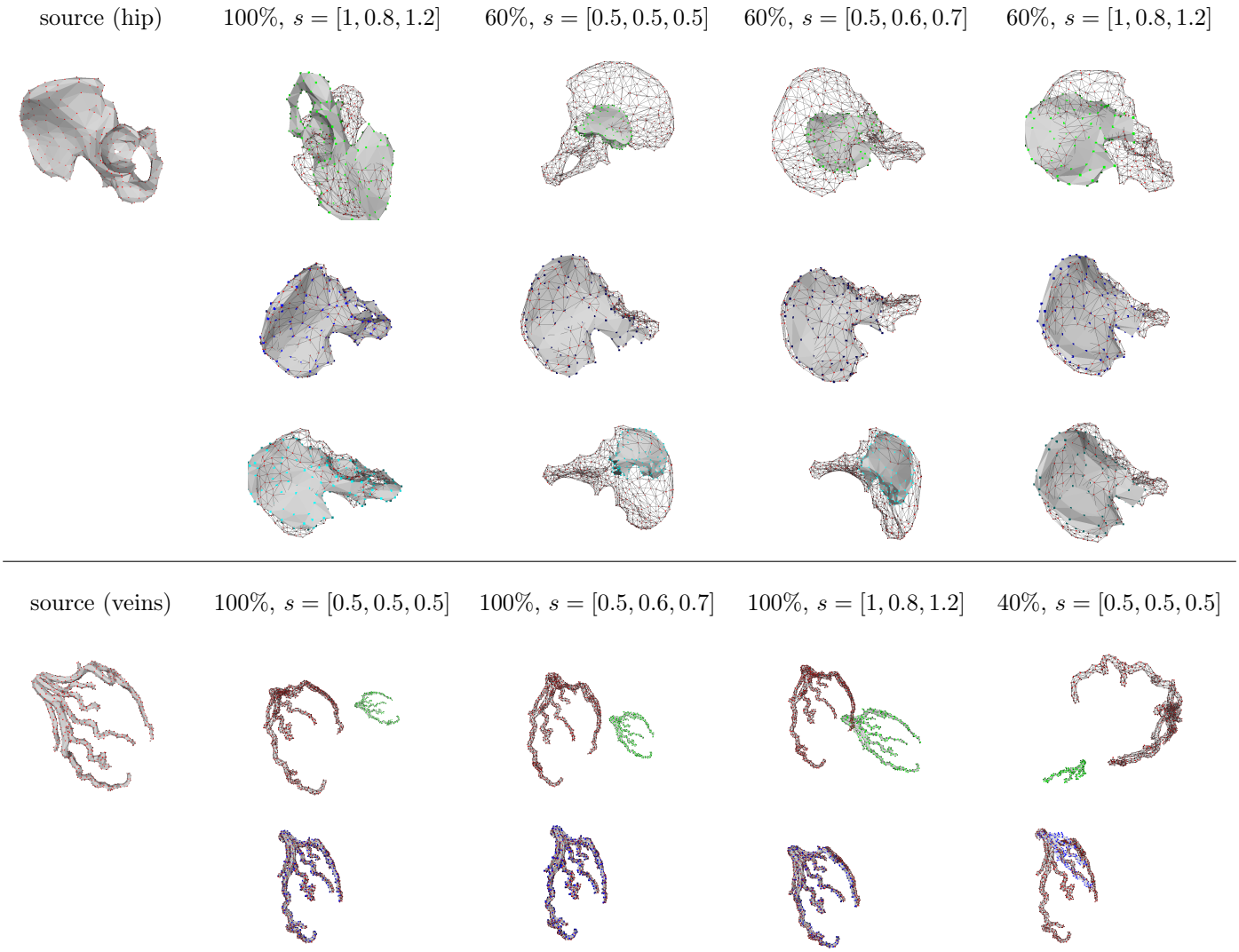


TABLE 1 – Nos données et nos résultats de recalage. Ligne 1 : donnée source (en rouge, à gauche) et données cibles (en vert) avec leur paramètres associés (portion de la surface source conservée et coefficients d'échelle). Ligne 2 : Résultats de recalage (en bleu) avec notre algorithme. Ligne 3 : Résultats de recalage (en cyan) avec FGR.

Source	Cible	MSE FGR+	MSE FGR	Subj eval FGR+	Subj eval FGR
FEMUR (500 pts)	full surface, scale = [0.5 0.5 0.5]	$5.35 * 10^{-4}$	$4.42 * 10^{-2}$	très bonne	mauvaise
	full surface, scale = [0.5 0.6 0.7]	$9.54 * 10^{-4}$	$2.84 * 10^{-2}$	bonne	mauvaise
	full surface, scale = [1 0.8 1.2]	$5.92 * 10^{-4}$	$5.92 * 10^{-4}$	bonne	bonne
	fragment (50%), scale = [0.5 0.5 0.5]	$1.31 * 10^{-3}$	$6.97 * 10^{-3}$	très bonne	moyenne
	fragment (50%), scale = [0.5 0.6 0.7]	$5.21 * 10^{-3}$	$1.84 * 10^{-3}$	bonne	moyenne
	fragment (50%), scale = [1 0.8 1.2]	$3.79 * 10^{-3}$	$4.12 * 10^{-3}$	bonne	bonne
HIP (500 pts)	full surface, scale = [0.5 0.5 0.5]	$6.03 * 10^{-5}$	$5.44 * 10^{-2}$	très bonne	mauvaise
	full surface, scale = [0.5 0.6 0.7]	$1.00 * 10^{-4}$	$2.20 * 10^{-2}$	très bonne	mauvaise
	full surface, scale = [1 0.8 1.2]	$9.85 * 10^{-5}$	$2.54 * 10^{-3}$	très bonne	bonne
	fragment (60%), scale = [0.5 0.5 0.5]	$2.35 * 10^{-3}$	$6.87 * 10^{-2}$	très bonne	mauvaise
	fragment (60%), scale = [0.5 0.6 0.7]	$1.31 * 10^{-3}$	$3.05 * 10^{-2}$	très bonne	mauvaise
	fragment (60%), scale = [1 0.8 1.2]	$1.71 * 10^{-3}$	$3.84 * 10^{-3}$	très bonne	bonne
VEINS (1000 pts)	full surface, scale = [0.5 0.5 0.5]	$1.52 * 10^{-4}$	$4.80 * 10^{-2}$	très bonne	mauvaise
	full surface, scale = [0.5 0.6 0.7]	$1.70 * 10^{-4}$	$2.09 * 10^{-2}$	très bonne	mauvaise
	full surface, scale = [1 0.8 1.2]	$2.36 * 10^{-4}$	$5.63 * 10^{-3}$	très bonne	bonne
	fragment (40%), scale = [0.5 0.5 0.5]	$5.24 * 10^{-2}$	$5.62 * 10^{-1}$	mauvaise	mauvaise
	fragment (40%), scale = [0.5 0.6 0.7]	$1.03 * 10^{-1}$	$5.19 * 10^{-2}$	moyenne	mauvaise
	fragment (40%), scale = [1 0.8 1.2]	$1.77 * 10^{-3}$	$7.69 * 10^{-1}$	mauvaise	mauvaise

TABLE 2 – Comparaison de la qualité du recalage entre notre algorithme (FGR+) et FGR. Nous avons évalué le recalage de manière visuelle et subjective, puis nous avons calculé une métrique géométrique, la MSE point à point symétrique (moyenne de la MSE calculée de la source vers la cible et de la cible vers la source).

Une approche multi-modale à la prédiction des mouvements de tête en réalité virtuelle

Miguel Fabian Romero Rondon, Lucile Sassatelli, Ramon Aparicio-Pardo, Frédéric Precioso
Université Côte d'Azur, CNRS, I3S, 06900 Sophia Antipolis, France

Résumé : *Dans les environnements immersifs comme les vidéos en 360°, il est important de prédire les futures positions de la tête de l'utilisatrice, notamment pour cibler la zone de la sphère à envoyer en haute qualité, et économiser du débit sans envoyer la totalité de la sphère en haute qualité. Nous contribuons à ce problème en proposant une approche multi-modale en concevant un réseau de neurones profond permettant une prédiction meilleure que toutes les méthodes de l'état de l'art, sur différents jeux de données et de large horizons temporels (de 0 à 5s).*

Mots-clés : Vidéos en 360°, prédiction de mouvement de tête, réseaux de neurones récurrents profonds

1 Introduction

Les vidéos à 360° sont une partie importante de l'écosystème de la réalité virtuelle (RV), offrant aux utilisatrices la possibilité d'explorer librement une scène omnidirectionnelle et une sensation d'immersion lorsqu'elle est visionnée dans un casque de RV. Étant donné la proximité de l'écran avec l'oeil et la largeur du contenu, le débit de données requis est de deux ordres de grandeur par rapport à une vidéo ordinaire [1]. Pour diminuer la quantité de données à transmettre, une solution consiste à envoyer en haute résolution uniquement la partie de la sphère à laquelle l'utilisatrice a accès à chaque instant, appelée champ de vision (FoV). Ces approches nécessitent cependant de connaître à l'avance la position de la tête, donc au moment de l'envoi du contenu depuis le serveur. En raison de ce besoin aigu de prédiction des mouvements de la tête dans le streaming vidéo à 360°, un certain nombre d'approches récentes ont proposé des réseaux de neurones profonds destinés à exploiter la connaissance des positions passées et du contenu pour prévoir périodiquement les prochaines positions sur un horizon temporel donné (par exemple, [3, 4, 5, 6]).

2 Travail proposé

Nous formulons d'abord le problème de prédiction considéré : il consiste, à chaque temps de lecture vidéo t, à prédire les positions futures de la tête entre t et $t + H$, comme représenté dans la Fig. 2, en ne connaissant que les positions passées de cette utilisatrice et le contenu complet de la vidéo. Où $\mathbf{P}_t = [\theta_t, \varphi_t]$ désigne les coordonnées vectorielles du FoV à l'instant t et $\mathbf{V}_t$ désigne l'information visuelle considérée à l'instant t : selon dans les hypothèses des modèles, il peut s'agir soit de la carte de saillance issue d'un extracteur de saillance pré-calculé sur le contenu (Content-Based Saliency) soit d'une carte de saillance issue des statistiques des utilisateurs (Ground-Truth Saliency).

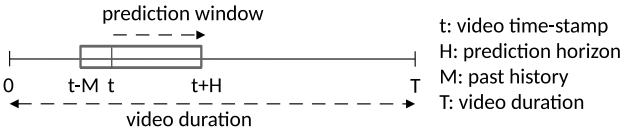


FIGURE 1 – Le problème de prédiction dynamique sur un horizon de taille H, à partir d'un passé de taille M.

3 Résultats

Nous concevons une architecture, représentée en Fig. 2, basée sur un schéma auto-régressif de type encodeur-décodeur, dans laquelle nous dédions un module récurrent au traitement de l'entrée des coordonnées de positions passées et un autre module récurrent à la carte de saillance (obtenue de PanoSalNet). Les deux embeddings ainsi obtenus sont finalement fusionnés pour prédire la position, grâce à un autre module récurrent.

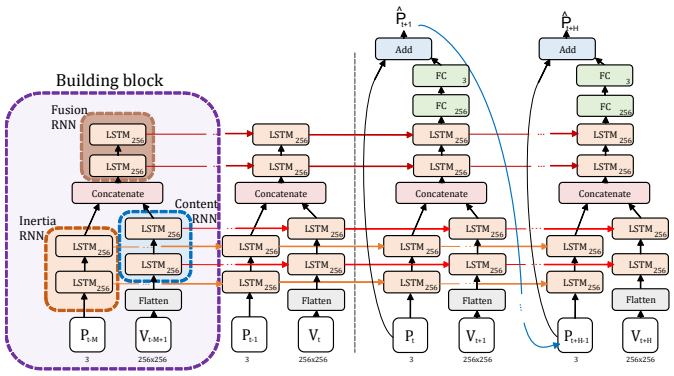


FIGURE 2 – Le réseau profond récurrent proposé. P désigne la position (réelle passée ou future estimée), et V la carte de saillance à chaque instant.

Le réseau est entraîné pour minimiser la distance entre la position réelle et estimée calculée sur les coordonnées euclidiennes. La Fig. 3 montre que notre méthode obtient des meilleures performances dans la configuration originale de [4] : même jeu de données, même horizon de prédiction, même métrique de test. La métrique montrée dans Fig. 3 est *Intersection over Union*, plus ce score est élevé, plus le chevauchement entre le FoV prédit et le FoV réel. La métrique de test montrée en Fig. 4 est la distance orthodromique (distance sur la sphère), plus la distance est courte, plus l'erreur de prédiction est faible. La Fig. 4 montre également des meilleures performances que [6], sur

un jeu de données avec plus de mouvement que l'original (données de [2]), et un horizon de prédiction plus long que l'original : 5s ici. La Fig. 4 montre également que les performances des réseaux se dégradent lorsque nous utilisons *Content-Based Saliency* par raport à *Ground-Truth Saliency*.

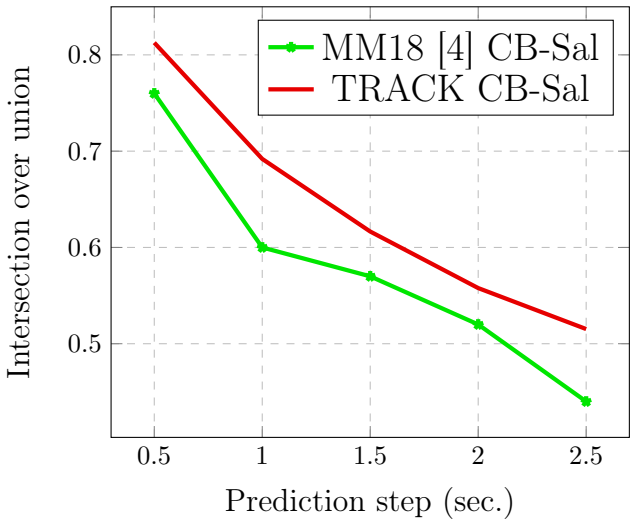


FIGURE 3 – Comparaison avec [4] en utilisant des cartes des saillance basés sur le contenu (Content-Based).

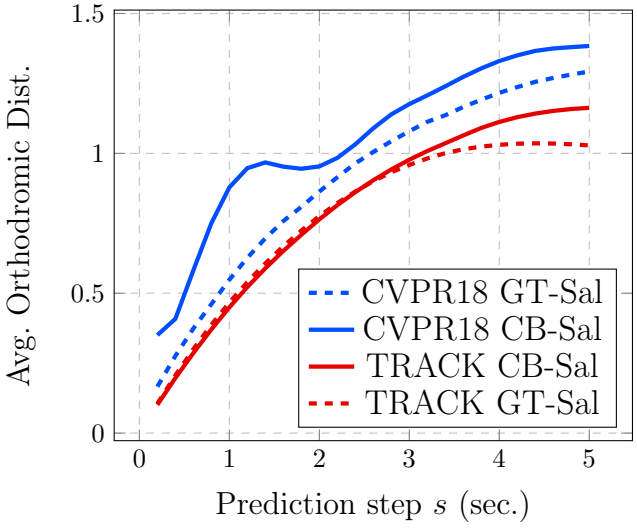


FIGURE 4 – Comparaison avec [6] en utilisant des cartes de saillance basés sur le contenu (Content-Based) ou en utilisant des cartes de saillance basés sur les statistiques des utilisateurs (Ground-Truth)

Pour confirmer l'analyse qui nous a conduit à introduire cette nouvelle architecture (TRACK) pour la prédiction dynamique des mouvements de la tête, nous avons réalisé une étude d'ablation des éléments composant notre architecture : soit nous remplaçons le RNN traitant la saillance CB par deux couches fully-connected (ligne nommée AblatSal dans la Fig. 5), ou remplacer le RNN de fusion par deux couches fully-connected (ligne nommée AblatFuse). Cette étude montre l'intérêt de chacun des modules ajoutés pour les différentes catégories de vidéos. Les vidéos *Exploratory* font référence à des vidéos où la distribution spatiale des positions de la tête des utilisateurs a tendance à être plus répandue, tandis que les vidéos *Static Focus* sont constituées d'un seul objet saillant (par exemple, une personne immobile). La tâche de prédire où l'utilisateur va regarder est plus facile et il y a plus de gains de l'étude d'ablation de notre modèle dans les vidéos *focus* que dans les vidéos *exploratoires*.

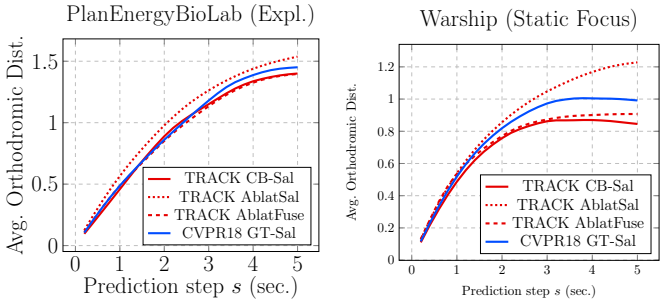


FIGURE 5 – Résultats de l'étude TRACK et ablation pour les vidéos *Exploratory* et *Static Focus*.

4 Conclusion et perspectives

Ce travail permet de ré-envisager les approches multimodales par réseaux de neurones profonds, en montrant l'importance d'une fusion tardive et non pas prématurée, empêchant les modules récurrents de pleinement profiter de chaque modalité. Ces problèmes de prédiction impactent un domaine d'application plus large que les seuls environnements immersifs, et nous pourrons l'illustrer en faisant le lien avec les approches en estimation de pose de squelettes 3D, ou estimation de trajectoire en conduite autonome.

Références

[1] Ejder Bastug, Mehdi Bennis, Muriel Médard, and Mérouane Debbah. Toward interconnected virtual reality : Opportunities, challenges, and enablers. *IEEE Communications Magazine*, 55(6) :110–117, 2017.

[2] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. A dataset of head and eye movements for 360-degree videos. In *ACM MMSys*, pages 432–437, 2018.

[3] Ching-Ling Fan, Jean Lee, Wen-Chih Lo, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. Fixation prediction for 360 video streaming in head-mounted virtual reality. In *ACM NOSSDAV*, pages 67–72, 2017.

[4] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. Your attention is unique : Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *ACM Int. Conf. on Multimedia*, pages 1190–1198, 2018.

[5] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang. Predicting head movement in panoramic video : A deep reinforcement learning approach. *IEEE Trans. on PAMI*, 2018.

[6] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360° immersive videos. In *IEEE CVPR*, pages 5333–5342, 2018.

La biométrie multimodale pour la vérification d'identité et la détection de fraude aux examens à distance

M. A. Haytom[1,2], C. Rosenberger[1], C. Charrier[1]
C. de Jacquelot[2]
[1]Normandie Univ., UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
[2]TestWe, Paris, France

Résumé : *La croissance rapide des nouvelles technologies accessibles par un nombre croissant de personne à travers la planète a permis la multiplication des offres de formation en ligne. Parallèlement, cette évolution rapide a indubitablement démultiplié les tentatives de fraude dans les épreuves à distance. La surveillance en ligne inclut tous les processus automatisés qui aident à sécuriser un événement d'évaluation à distance. En effet, plusieurs organismes reconnaissent enfin les faiblesses de la sécurité de la surveillance traditionnelle. Dans ce travail, nous cherchons à mettre en place un système biométrique multimodal afin de vérifier l'identité des apprenants. Le système a aussi pour objectif de détecter les actions frauduleuses et les comportements inhabituels lors d'un examen à distance.*

Mots-clés : Données personnelles, signature biométrique, comportement de l'apprenant.

1 Introduction

L'objectif visé par la formation à distance est la qualification ou le diplôme. Elle reste un outil primordial pour plusieurs personnes, tels que, les apprenants en situation de handicap, les employés qui souhaitent acquérir de nouvelles qualifications ou encore des étudiants qui préfèrent partir pour se former à l'étranger. Ces derniers sont nombreux à adopter ce mode d'enseignement. Cela leur permet de préparer un diplôme, d'étudier à un rythme adapté à leurs attentes. Ces nouveaux modes pédagogiques permettent aux salariés d'évoluer au sein de leur entreprise par l'acquisition de nouvelles connaissances, sans avoir à s'absenter à de multiples reprises. Le développement continu de l'apprentissage à distance et l'incapacité d'examiner l'environnement de l'apprenant constituent un défi pour les établissements d'enseignement supérieur. Plusieurs études ont été menées pour évaluer les attitudes des apprenants à l'égard de divers problèmes et comportements lorsqu'ils passent un examen en ligne. Les chercheurs confirment que la fraude aux examens à distance a augmenté [5], 73,6% des participants disent qu'il était plus facile de tricher dans une épreuve en ligne que dans un examen traditionnel.
La fraude académique reste un problème majeur. Lorsqu'il s'agit des types de fraude qui ont tourmenté le monde des affaires ces dernières années, les étudiants des écoles de commerce ont souvent été accusés d'être les pires contrevenants. Les chercheurs ont examiné la question de la tricherie et la fraude dans les écoles de commerce en interrogeant 268 étudiants des écoles de commerce et autres écoles professionnelles sur leurs attitudes et leurs expériences [8]. Ils ont trouvé que les étudiants des écoles de commerce trichaient ni plus ni moins que les étudiants des autres écoles. De plus, ils ont énoncé que la plupart des tricheurs dans toutes les écoles étaient souvent les plus jeunes.
Aujourd'hui, l'un des principaux défis de l'éducation en ligne est la difficulté à garantir l'identité des candidats à distance et la confiance liée aux résultats des épreuves réalisées en ligne. Dans la littérature, des chercheurs ont expliqué comment de tels défi peuvent être résolus grâce à une authentification continue à l'aide de technologies biométriques. Une solution multimodale impliquant trois modalités a été utilisée à cette fin [10], le système intègre la dynamique de la souris, la dynamique de la frappe au clavier et la biométrie du visage. Dans une étude similaire [1], des chercheurs ont présenté un système d'analyse multimédia surveillant automatiquement les examens en ligne. Le matériel du système comprend une caméra intégrée, une caméra portable et un microphone pour surveiller l'environnement visuel et acoustique de la salle de l'épreuve. Le système comprend six composants de base qui évaluent en continu les indices de comportement clés : vérification de l'utilisateur, détection de texte, détection de voix, fenêtre active, estimation du regard et détection de téléphone. Plusieurs composants d'estimation ont été combinés pour extraire les caractéristiques afin de prédire si le candidat triche pendant l'examen.
Généralement, la biométrie fait référence à l'exploitation de caractéristiques physiologiques ou comportementales d'un individu à des fins d'authentification ou d'identification d'individus. En premier lieu, la collecte et le traitement des données personnelles doivent respecter le Règlement européen Général sur la Protection des Données (RGPD) [3].

2 Travail proposé

Dans cet article, un système biométrique multimodal a été utilisé afin de détecter la fraude et vérifier l'identité des apprenants.

2.1 Usurpation d'identité

L'usurpation d'identité représente des milliards d'euros chaque année. Alors que tout un chacun peut en être victime, les étudiants sont devenus une cible privilégiée. L'une des principales raisons pour lesquelles les étudiants

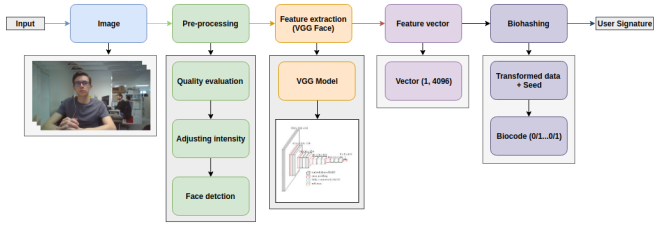


FIGURE 1 – Processus d'extraction de caractéristiques pour la reconnaissance du visage de l'apprenant.

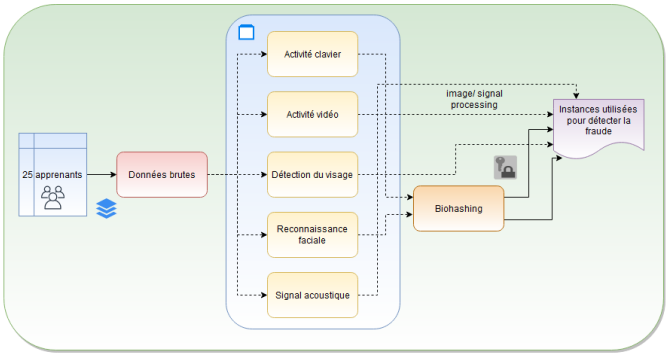


FIGURE 2 – Principe de la méthode proposée

sont plus à risque est liée la quantité excessive de courrier qu'ils reçoivent des sociétés de cartes de crédit et des services multimédia en ligne. Face à ces risques de vol d'identité numérique, une solution alternative peut être l'utilisation de la biométrie. Une signature biométrique (gabarit unique) peut être une réponse appropriée car elle est cohérente pour garantir l'authenticité de l'identité d'une personne. De plus, la signature biométrique semble possible pour l'enseignement à distance ainsi que pour certains actes de la vie sociale, tels que les achats en ligne. Il est possible d'utiliser divers types d'informations biométriques, telles que la reconnaissance faciale, les empreintes digitales, la dynamique de frappe ou la reconnaissance de la voix pour l'identification ou l'authentification.

2.1.1 La signature biométrique

Pour obtenir la signaure de l'apprenant notre choix s'est porté sur l'utilisation des réseaux CNN, à l'aide du modèle VGG pré-entrainé dédié à la reconnaissance des visages. Nous avons choisi cette méthode car elle atteint des performances proches de 99% avec une base de données de 2,6 millions d'images [9]. Pour extraire la signature unique nous utilisons une image d'entrée de taille 224×224×3 après avoir obtenu les coordonnées spatiales de la région d'intérêt (le visage de l'apprenant). Nous avons utilisé la dernière couche du réseau avant le softmax pour obtenir les caractéristiques haut niveau, soit un vecteur de 4096 valeurs. La figure 1 illustre le processus d'extraction de caractéristiques pour la reconnaissance du visage de l'apprenant.

L'intégration de la dynamique de frappe au clavier comme modalité comportementale permettra d'avoir une authentification forte et robuste. Pour extraire les caractéristiques de la DDF, tout d'abord, lorsque l'utilisateur écrit une phrase ou un texte sur le clavier, chaque mot représente une signature. Les mots comportant plus de deux lettres sont traités et analysés. À partir de plusieurs séquences temporelles, on extrait le gabarit de l'apprenant en considérant : 1) l'intervalle de temps entre l'appui et le relâchement sur une touche, 2) l'intervalle de temps entre deux pressions consécutives, 3) l'intervalle de temps entre un relâchement et l'appui sur la touche suivante et 4) la durée globale d'une suite de caractères. A posteriori, nous mettons en œuvre des mesures de sécurité et des processus de protection des données en utilisant le bioHashing afin d'obtenir une signature biométrique chiffrée sous la forme d'un code sécurisé [2].

2.2 Détection de la fraude dans les examens

Aujourd'hui, avec le développement des classes virtuelles pour l'apprentissage à distance, il existe peu de solutions qui peuvent aider les superviseurs à assurer le bon déroulement d'un examen en ligne.
Dans ce travail, nous avons utilisé des données collectées depuis un ordinateur standard tels que le clavier, la webcam et le microphone. Nous allons analyser le signal sonore ambiant, la séquence vidéo de la webcam et l'activité clavier de l'apprenant. La figure 2 présente le principe de la méthode proposée.

Cette combinaison constitue les entrées de notre système pour identifier une tentative de fraude. A partir de ces informations issues des périphériques utilisés, nous allons extraire différentes informations dans ce travail :

— Détection de visages dans le flux vidéo de la webcam : nous analysons une séquence d'images d'une fenêtre de trois secondes lors d'une session d'examen à distance puis nous transformons les vecteurs de détection afin d'obtenir des observations qui vont nous aider à obtenir des estimations (cas suspect/ cas normal). Nous avons utilisé un détecteur de visage plus précis basé sur l'apprentissage en profondeur, ce détecteur utilise en particulier une analyse en un seul coup (SSD) avec ResNet comme réseau de base.

— Analyse de l'activité vidéo pour décrire le comportement de l'apprenant (est il-statique ?) : nous utilisons une mesure statistique simple de distance entre images (éventuellement de la même longueur). Nous nous référons à une solution qui permet de mesurer les points de ressemblance (Erreur Quadratique Moyenne) entre deux images afin de détecter les événements temporels pour une image codée sur 8-bits, d = 255.

— Analyse de l'activité clavier : Nous traitons une activité de trois secondes et créons des attributs d'événement qui seront utilisés comme indicateurs de confiance afin de surveiller la présence de l'apprenant et même de vérifier son identité.

— Analyse du signal sonore : nous analysons la distribution de deux signaux ; un premier vecteur d'une personne qui triche pendant l'examen, puis une deuxième personne honnête qui passe le test sans tricher. L'amplitude du signal a été utilisée pour déterminer le seuil de décision afin d'identifier les actions frauduleuses.

Nous utilisons dans ce système une methode de protection de données biométriques basées sur une transformation non inversible de la signature de l'apprenant. Le biohashing a été utilisé pour transformer le vecteur de donnée d'origine en un code sécurisé d'une longueur plus petite que la taille de la données d'entrée sous la forme d'un code binaire (BioCode) en utilisant une clé spécifique à chaque individu inscrit pour effectuer une épreuve à distance (voir Algorithme 1). Le biohashing tolère des variations de données acquises, ce qui nous permet d'obtenir des séquence de bits hautement corrélés. Le Biohashing présente des avantages fonctionnels significatifs car il permet à la fois de protéger la vie privée de l'apprenant tout en veillant à ce que cette signature n'ait pas été produite par un attaquant pour éviter une détection de la fraude.

Algorithm 1 BioHashing

1: **Inputs**
2: $T = (T_1, \ldots, T_n)$: biometric template,
3: K_z : secret seed
4: **Output** $B = (B_1, \ldots, B_m)$: BioCode
5: Generation with the seed K_z of m pseudorandom vectors $V_1, \ldots, V_m$ of length n,
6: Orthogonalize vectors with the Gram-Schmidt algorithm,
7: **for** $i = 1, \ldots, m$ **do** compute $x_i =< T, V_i >$.
8: **end for**
9: Compute BioCode :

$$B_i = \begin{cases} 0 & \text{if} \quad x_i < \tau \\ 1 & \text{if} \quad x_i \geq \tau, \end{cases}$$

where τ is a given threshold, generally equal to 0.

3 Résultats

3.1 Reconnaissance faciale

Pour l'évaluation des performances de notre système biométrique, nous avons appliqué le processus de vérification d'identité sur trois bases de données de la littérature : 1) la base LFW (Labeled Faces in the Wild), 2) la base de données MED II (Multiple Encounter Dataset II) et 3) la base AR créée par Aleix Martinez et Robert Benavente. Le tableau 1 représente les résultats obtenus après

Table 1 – Valeurs des taux d'erreur (EER) en pourcentage calculées sur les trois bases de données de l'état de l'art testées en utilisant la distance cosinus.

Base de données	Signature	Taux d'erreur
LFW	4096	1.3%
MED II	4096	10.3%
AR Face	4096	0.01%

avoir appliqué notre solution de vérification d'identité sur des bases de données de la littérature. Nous pouvons voir que les taux d'erreurs sont proches de zéro pour les bases LFW et AR Face. En revanche, la base de données MEDS contient beaucoup de variations et nous obtenons au final un taux d'erreur de 10%. Des conditions supplémentaires, tels qu'un mauvais éclairage, une pose extrême, des occultations, une faible résolution et d'autres facteurs importants ne constituent pas une partie importante de la LFW. C'est pour cela nous avons testé notre solution sur une base de données telle que Med II afin d'élargir notre évaluation.

3.2 Dynamique de frappe

Nous allons appliqué notre solution sur des bases de données de DDF au clavier publiées dans l'état de l'art afin d'évaluer la performance de notre système biométrique. Nous appliquons le processus de vérification d'identité sur la base de données GREYCNISLAB [4], et une seconde base créée par Killourhy et Maxion [6] :

— Giot et al [4] ont développé une base de données GREYCNISLAB dédiée à la dynamique de frappe au clavier. Cette base a été générée à l'aide du logiciel GREYC-Keystroke. 133 personnes ont participé à la campagne de test en tapant entre 5 et 107 fois le mot de passe « GREYC Laboratory ».

— Killourhy et Maxion ont collecté des données de DDF. Les données ont été collectées auprès de 51 sujets tapant 400 mots de passe chacun.

Nous avons atteint une précision de 100% en utilisant la base de données GREYCNISLAB. Après avoir lancé les test de performance sur la base de données de Killourhy et Maxion, nous avons obtenu un taux d'erreur de 29,55% en utilisant la distance Euclidienne, ce qui est un taux d'erreur assez élevé par rapport à l'étude des chercheurs (Euclidienne 21.5%, Manhattan 15.3% [7]). En fait, notre protocole d'évaluation de la dynamique de frappe est plus adapté à un champs de texte qu'à un mot de passe. Il est vraiment difficile de comparer une étude utilisant un nombre spécifique de signatures à une étude utilisant un nombre différent de données. La plupart des études de la littérature ont utilisé des protocoles différents pour l'acquisition de leurs données. En effet, l'existence de différents types de systèmes biométriques de DDF nécessitent évidemment des protocoles d'acquisition divers. Pour conclure, on peut dire que la taille des bases de données et le contrôle du processus d'acquisition ont un impact sur les résultats de performance obtenus.

3.3 La fusion multimodale

Les résultats obtenus lors de la fusion des deux modalités visage et DDF sont présentés dans le tableau 2. Comme préconisé dans les sections précédentes, le BioCode est calculé à partir d'un gabarit moyen obtenu à partir des 30 premiers FaceCodes et le gabarit de la DDF est protégé après calcul d'un gabarit moyen généré à partir des dix premiers mots saisis par l'utilisateur. On observe

Table 2 – Mesure de performance (EER) en pourcentage du système multimodal proposé, calculée après fusion des décisions par vote majoritaire en utilisant des données protégées.

Modalité	Gabarit de référence	Taux d'erreur
Visage	(moyenne) 30 signatures	0.09%
DDF	(moyenne) 10 signatures	

une baisse significative de la valeur EER obtenue à 0.09% par rapport aux valeurs obtenues pour les modalités seules (10.17% pour la DDF et 0.09% pour le visage en utilisant notre base de données) ce qui démontre l'intérêt de l'approche multimodale du système développé.

3.4 Surveillance automatisée

Une expérience dans des conditions réelles a été menée avec des étudiants de l'ENSICAEN pour effectuer des simulations de fraude dans un examen à distance. Plusieurs scénarios de fraude ont été envisagés en utilisant des supports non autorisés (téléphone, tableau, deuxième ordinateur, etc.) ou l'aide d'un tiers (communication directe ou indirecte). Nous utilisons la détection du visage, l'activité de la vidéo et la reconnaissance faciale, la dynamique de frappe, l'activité clavier et le signal sonore en effectuant un calcul sur une fenêtre de trois secondes. Après avoir comparé la précision de la détection de différents algorithmes d'apprentissage automatique en ajustant le paramètre de la méthode d'estimation de fiabilité (validation croisée). Nous cherchons à construire un modèle d'apprentissage avec des données protégées. Nous utilisons l'option stratification avec la validation croisée afin d'éviter les problèmes de classification, pour choisir des (k folds) parties avec les mêmes proportions pour chaque classe. Après avoir lancé

Table 3 – Résultats avec protection (valeurs entre 0 et 1).

Modèle	AUC	CA	F1	Precision	Recall
kNN	0.997	0.985	0.985	0.985	0.985
Tree	1	1	1	1	1
SVM	0.565	0.553	0.485	0.526	0.553
NB	0.888	0.622	0.587	0.771	0.622
LR	0.808	0.746	0.745	0.745	0.746

l'apprentissage sur un ensemble de données transformées avec le biohashing, nous avons obtenu une précision de 99% en utilisant la méthode des k plus proches voisins et 100 % avec l'arbre de décision (voir tableau 3). Ces deux méthodes permettent d'identifier quasiment toutes les tentatives de fraude (2067 tentatives de fraude avec kNN et 2100 avec DT).

4 Conclusion et perspectives

Le premier avantage que la biométrie combinée au biohaching peut offrir est qu'il est un excellent outil de reconnaissance et d'analyse de caractéristiques de l'organisme humain. L'utilisation de la biométrie nécessite qu'une personne soit physiquement présente au moment de l'authentification jusqu'à ce qu'elle soit validée. De plus, les informations d'authentification biométrique ne peuvent pas être transférées ou partagées, ils représentent un outil redoutable contre la répudiation. Un système biométrique multimodal avec protection et l'analyse environnementale peut jouer un rôle prépondérant pour améliorer la vérification d'identité des apprenants, détecter les évènements inhabituels et assurer la confidentialité des données échangées. Pour conclure, la solution proposée est efficace contre l'usurpation d'identité et résout une partie importante de la fraude à un examen à distance avec un niveau de précision de détection des comportements inhabituels très élevé.
Concernant les perspectives, la combinaison de nombreuses méthodes biométriques augmente considérablement la sécurité du système, mais cette solution ne couvre pas entièrement toutes les possibilités de fraude dans les examens à distance. Nous avons l'intention d'évaluer les performances du système après avoir fusionné plusieurs modalités (visage, dynamique des frappes, analyse du regard, reconnaissance gestes et d'autres modules biométriques), puis créer un modèle de détection de fraude à l'aide d'une base de données plus large pour augmenter la sécurité du système et rendre la détection robuste et efficace.

Références

[1] Yousef Atoum, Liping Chen, Alex X Liu, Stephen DH Hsu, and Xiaoming Liu. Automated online exam proctoring. *IEEE Transactions on Multimedia*, 19(7) :1609–1624, 2017.

[2] Rima Belguechi, Estelle Cherrier, and Christophe Rosenberger. Texture based fingerprint biohashing : Attacks and robustness. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 196–201. IEEE, 2012.

[3] Peter Carey. *Data protection : a practical guide to UK and EU law*. Oxford University Press, Inc., 2018.

[4] Romain Giot, Mohamad El-Abed, and Christophe Rosenberger. Greyc keystroke : a benchmark for keystroke dynamics biometric systems. In *2009 IEEE 3rd International Conference on Biometrics : Theory, Applications, and Systems*, pages 1–6. IEEE, 2009.

[5] Anja M Jansen, Ellen Giebels, Thomas JL van Rompay, and Marianne Junger. The influence of the presentation of camera surveillance on cheating and pro-social behavior. *Frontiers in psychology*, 9 :1937, 2018.

[6] Kevin S Killourhy and Roy A Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*, pages 125–134. IEEE, 2009.

[7] Kevin S. Killourhy and Roy A. Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP International Conference on Dependable Systems Networks*, pages 125–134, 2009.

[8] Helen A Klein, Nancy M Levenburg, Marie McKendall, and William Mothersell. Cheating during the college years : How do business school students compare ? *Journal of Business Ethics*, 72(2) :197–206, 2007.

[9] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[10] Issa Traoré, Youssef Nakkabi, Sherif Saad, Bassam Sayed, Julibio D Ardigo, and Paulo Magella de Faria Quinan. Ensuring online exam integrity through continuous biometric authentication. In *Information Security Practices*, pages 73–81. Springer, 2017.

Apprentissage automatique

Etude Comparative de l'Apprentissage par Transfert pour l'Identification des Caméras

Alexandre Berthet, Jean-Luc Dugelay
Eurecom, Département Sécurité Digitale, Sophia Antipolis, France.

Résumé : *De nos jours, l'édition d'images est devenue plus facile et plus précise. Les modifications malveillantes sont donc plus accessibles et des méthodes de détection des falsifications ont été développées. L'identification de la caméra source est un domaine important de la criminalistique des images numériques et peut-être réalisée selon la marque, le modèle ou l'appareil exact. Notamment, la classification de modèle de caméras est l'application la plus abordée et a d'abord été étudiée avec des algorithmes classiques, puis avec des réseaux de neurones convolutifs (RNC). Cependant, malgré leur efficacité, les RNC sont dépendants des données et leurs performances diminuent avec le nombre d'objets à classer. Cet aspect est encore plus important avec les caméras puisque chaque appareil possède ses propres artefacts et que les caméras d'une même marque présentent des similarités au niveau de leurs empreintes numériques. Dès lors, un des objectifs de cet article réside dans l'étude de la robustesse des méthodes pour l'identification de caméras. Un domaine récent de l'apprentissage automatique, appelé apprentissage par transfert, offre une alternative intéressante à ce problème. Afin d'étudier pleinement son impact, nous avons appliqué différentes approches de l'apprentissage par transfert à trois architectures de RNC différentes, présentant chacune des particularités. Pour réaliser notre étude comparative, nous avons également proposé un protocole d'évaluation de la robustesse basé sur les deux principales sujet de recherche de l'état de l'art : les caméras inconnues et l'augmentation de caméras à classer.*

Mots-clés : Identification du modèle de caméra ; Apprentissage par transfert ; Criminalistique des images numériques ; Réseaux de neurones convolutifs.

1 Introduction

Grâce au développement des dispositifs numériques (appareils photo, téléphones portables, etc.), l'accès aux images et aux vidéos a augmenté au point de devenir un important canal de communication. D'autre part, la retouche des images numériques est devenue un véritable problème, notamment pour prouver l'authenticité d'une image. Dans le même temps, l'identification de la caméra source, un domaine de la criminalistique des images numériques [11], s'est révélée être une solution pour la détection de ces falsifications. Cette identification peut être faite selon la marque, le modèle ou le dispositif numérique même de la caméra. La classification de modèle de caméras est l'application la plus étudiée dans la littérature et le principe est d'identifier, parmi un groupe de modèles de caméras, celui qui est associé à l'image étudiée. Plusieurs techniques ont été élaborées, utilisant les artefacts laissés lors de l'acquisition d'une image numérique. Notamment, le bruit numérique lié au capteur de la caméra [10] ou à des caractéristiques physiques [6] a d'abord permis d'établir l'empreinte numérique des caméras. Puis, avec la démocratisation de l'apprentissage automatique, les performances ont été améliorées, notamment grâce aux réseaux de neurones convolutifs (RNC). Ces réseaux analysent d'abord les artefacts et les réduisent via un extracteur de caractéristiques, tandis que l'identification est réalisée grâce aux couches de classification. Cependant, même si ces méthodes se sont avérées efficaces pour identifier les modèles de caméras, il existe toujours une forte dépendance aux données. Ce phénomène est plus accentué pour les modèles de caméras puisque chaque dispositif numérique possède ses propres artefacts et que les caméras d'une même marque ont des empreintes numériques proches. Ces phénomènes d'unicité des artefacts et de similarité d'empreinte numérique posent la question de la robustesse de performance des méthodes. Dans cet article, nous avons utilisé l'apprentissage par transfert pour conduire une étude comparative de la robustesse des méthodes d'identification de modèle de caméras. De plus, cette étude est menée avec trois architectures de RNC réputées ainsi que les deux sujets de recherche principaux de l'identification de la caméra source.

La section 1 présente l'identification de la caméra source ainsi que les contributions de cet article. Dans la section 2, nous détaillons les motivations de notre étude comparative, les architectures utilisées ainsi que notre protocole. Les résultats de l'évaluation conduite à l'aide de notre protocole sont détaillés dans la section 3 ainsi que la base de données d'images de Dresde utilisée. Enfin, la section 4 conclut sur l'impact de l'apprentissage par transfert pour l'identification de modèle de caméras.

2 Travail proposé

2.1 Problème adressé

Le problème de similarité des empreintes numériques de modèles de caméras provenant d'une même marque est un des sujets importants de la littérature de l'identification de la caméra source. Dans [12], le problème est abordé à travers une série de trois expériences. La méthode est basée sur un RNC avec un filtre passe-haut utilisé comme module de prétraitement. Le prétraitement est une étape cruciale, voire obligatoire en criminalistique des images numériques [4]. i) Le réseau a tout d'abord été évalué avec 12 caméras de la base de données d'images de Dresde [7] (détaillée en section 3.1). ii) Ensuite, avec deux caméras supplémentaires provenant de la même marque

pour mettre en exergue le problème de similarité des empreintes numériques. iii) Enfin, avec l'ensemble des caméras (33 modèles) pour généraliser ce phénomène. Dans [5], le problème de similarité des modèles provenant d'une même marque est aussi adressé avec la classification de l'ensemble des modèles (27 caméras) de la base de données d'images de Dresde [7]. La méthode est basée sur un RNC pour l'extraction de caractéristiques et une machine à vecteurs de support pour la classification. En outre, le problème de caméras inconnues (c-à-d, non sélectionnées pour l'entraînement du réseau) est également abordé lors d'une expérience. Le problème des caméras inconnues est un sujet important de la littérature de l'identification des caméras. [2] est une étude focalisée sur le scénario de caméras inconnues avec une méthode basée sur un RNC dont la première couche est utilisée comme module de prétraitement. En fait, l'objectif de cette approche est de classer les images comme venant d'un modèle connu ou inconnu.

Le phénomène d'unicité des artefacts de dispositif numérique pose la question de la robustesse de performance des méthodes. Par exemple, une méthode ayant des performances élevées lors d'une évaluation sur une base de données *B1* (ayant servie pour l'entraînement) pourrait subir une baisse de performance sur une nouvelle base de données *B2*. En effet, si les caméras de la base *B2* sont inconnues du réseau, ce dernier pourrait ne pas classer correctement les modèles de caméras. Ces dernières années, l'apprentissage par transfert, un domaine de l'apprentissage profond, a été utilisé pour développer de nouveaux réseaux, de manière plus rapide sans perdre en efficacité. Le principe est de créer un nouveau réseau *M2* en transférant l'architecture et les poids d'un modèle *M1* pré-entraîné sur une base de données *B1*. Le réglage final du modèle transféré *M2* est nécessaire pour l'adapter à une nouvelle base de données *B2*. Il existe trois stratégies différentes de réglage fin pour le modèle *M2*, en fonction de la similarité entre les bases de données *B1* et *B2* ainsi que la taille de *B2*. i) Entraînement complet du réseau (*B1*, *B2* différentes et *B2* importante). ii) Entraînement partiel du réseau : peu de couches (*B1*, *B2* similaires et *B2* importante) ou beaucoup de couches (*B1*, *B2* différentes et *B2* réduite). iii) Entraînement de la classification (*B1*, *B2* similaires et *B2* réduite). En outre, le réseau sera en mesure de fournir de meilleurs résultats si le modèle pré-entraîné présente une diversité de classification et a été entraîné sur une grande base de données. L'apprentissage par transfert a déjà été appliqué pour l'identification de modèle de caméras [1].

2.2 Architectures utilisées

Avec l'émergence de l'apprentissage profond au cours de la dernière décennie, plusieurs défis pour le traitement d'images ont eu lieu, conduisant à l'implémentation de nouvelles architectures de RNC. Certaines sont devenues des standards pour les applications de traitement d'images notamment grâce à leurs performances. Nous avons décidé d'utiliser trois architectures présentant des aspects différents : VGG19 [3] qui est un RNC classique ; ResNet50 [8] qui est un RNC utilisant des sauts de couches pour élargir le domaine des caractéristiques étudiées ; et DenseNet201 [9] qui est un RNC connectant chaque couche avec les suivantes par des sauts de couches, permettant d'obtenir des caractéristiques plus complètes et diverses. Pour chaque architecture, nous avons remplacé la partie classification par une couche d'aplatissement, deux couches denses (de 1028 et 512), deux couches d'abandon (de 0.5) et une sortie de taille N (le nombre de caméras).

2.3 Protocole d'évaluation

TABLE 1 – Description du nombre de modèles de caméras j utilisées pour les modèles pré-entraînés, k inconnues et l totales utilisées pour l'évaluation.

Protocole	**Pré-entrainement**	**Evaluation**	
		Inconnues	**Total**
Expérience 1	j = 8	k = 0	l = 8
Expérience 2	j = 8	k = 8	l = 8
Expérience 3	j = 8	k = 19	l = 27

L'aspect important à considérer pour le protocole d'évaluation est le nombre de modèles de caméras utilisé pour l'apprentissage des caractéristiques et pour la classification. Soit j les caméras utilisées pour l'apprentissage, k et l respectivement les caméras inconnues et le total des caméras pour l'évaluation. Nous avons traité deux sujets réputés de la littérature : l'unicité des empreintes numériques et les caméras inconnues. Tout d'abord, nous avons obtenus des réseaux de référence (un par architecture) à partir d'un ré-entraînement de réseaux pré-entraîné sur ImageNet (détection d'objets). Notre protocole est une série de trois expériences (voir Tab. 1) utilisant les approches de réglage fin. i) Dans un premier temps, nous avons réalisé une évaluation simple de performance grâce à l'apprentissage par transfert avec un entraînement complet du réseau. ii) Puis, nous avons abordé le problème de caméras inconnues. iii) Enfin, la dernière évaluation prend en compte les deux aspects étudiés (caméras inconnues et unicité des empreintes numériques). Le protocole est basé sur l'apprentissage par transfert, dont nous avons présenté les trois possibilités pour régler finement le réseau dans la section 1. Le réglage final du réseau est une approche essentielle de l'apprentissage par transfert et nous avons donc décidé d'évaluer ces trois approches pour déterminer leurs impacts sur les performances des méthodes à l'aide de notre protocole. Les résultats de cette étude sont détaillés dans la section 3.

3 Evaluation expérimentale

3.1 Base de données

Afin d'effectuer une comparaison équitable, tous les réseaux de référence ont été entraînés avec la base de données d'images de Dresde [7]. Elle contient un total de 27 modèles de caméra pour plus de 14 000 images, à partir desquelles nous avons extrait des patchs de taille 128×128 pour les adapter aux entrées des réseaux. Ainsi, le jeu de données final est constitué de 2,6 M de patchs, que nous avons divisés en 3 sous-ensembles : 1,56 M de patchs pour l'entraînement (60%), 0,52 M pour la validation (20%) et 0,52 M pour le test (20%). Pour l'apprentissage, nous

TABLE 2 – Description des sous-ensembles de données pour chaque expérience.

	Transfert	**Validation**	**Evaluation**	**Total**
Exp. 1	448K	149K	149K	746K
Exp. 2	278K	93K	93K	464K
Exp. 3	258K	86K	86K	430K

avons utilisé comme rappel l'arrêt précoce (fin de l'apprentissage si aucune amélioration) et comme optimiseur la descente de gradient stochastique (SGD). Les réseaux ont été entraînés avec une taille de lot de 32 patches et deux GPUs GeForce RTX 2080. Les jeux de données utilisées pour les entraînements ou les transferts de réseaux sont répertoriées dans la Tab. 2.

3.2 Résultats

Dans un premier temps, nous avons réalisé une étude préliminaire dans le but de montrer le problème d'identification lié aux modèles inconnus à l'aide de notre protocole. Notamment nous avons comparé les performances de la première expérience, considérée comme classique (classification de huit modèles connus), avec les deux autres, portées sur les caméras inconnues. L'entraînement des architectures a été effectué avec un apprentissage par transfert et un réglage fin de la partie de classification des réseaux. Les résultats obtenus pour les trois architectures (voir Tab. 3) montrent une perte de performance d'environ 10% (métrique de précision) entre l'expérience 1 et 2, dont le nombre de caméras est similaire (huit modèles). Dès lors, nous en avons conclu que cette baisse de performance était due aux modèles inconnus utilisés dans l'expérience 2. Ce phénomène est encore plus accentué pour l'expérience 3 (19 modèles inconnus) avec une diminution de précision d'environ 17% confirmant le problème soulevé.

TABLE 3 – Résultats de précision des trois expériences du protocole d'évaluation pour une étude préliminaire.

Architectures	*Exp. 1*	*Exp. 2*	*Exp. 3*
VGG 19 [3]	98.47 %	88.52 %	82.66 %
ResNet50 [8]	99.46 %	87.78 %	81.68 %
DenseNet201 [9]	99.49 %	90.02 %	81.82 %
Moyenne	99.14%	88.77%	82.05%

L'étude finale a été mené avec le même protocole pour montrer l'impact des différentes approches sur les performances, mais aussi d'observer le comportement des architectures face au problème de caméras inconnues. Pour l'ensemble des trois expériences, nous avons inclus dans les résultats le temps d'entraînement d'une itération ainsi que la précision des réseaux pour obtenir une comparaison plus complète. Les résultats obtenus montrent que chaque approche de réglage fin des réseaux après l'apprentissage par transfert possède des forces et des faiblesses (voir Tab. 4). Notamment, l'entraînement complet permet d'obtenir de meilleurs résultats, mais nécessite forcément un temps

TABLE 4 – Présentation des résultats pour les architectures choisies et les approches d'apprentissage par transfert en temps d'entraînement par itération, précision (en %) et le nombre de paramètres à entraîner.

Transfert	**Complet**		**Partiel**		**réglage fin**	
	Min.	Acc.	Min.	Acc.	Min.	Acc.
Expérience 2 (8 modèles de caméras inconnues)						
VGG19	21.3	98.02 %	17.7	**97.69 %**	16.6	88.52 %
ResNet50	21.2	97.65 %	17.7	93.84 %	14.3	87.78 %
DenseNet201	**20.5**	**98.85 %**	**12**	96.97 %	**9.5**	**90.82 %**
Moyenne	21	98.11%	15.8	96.17%	13.5	89.04%
Expérience 3 (27 modèles dont 19 caméras inconnues)						
VGG 19	104	**93.48 %**	74.3	**91.22 %**	70.7	**82.66 %**
ResNet50	**88.2**	91.02 %	80.8	87.08 %	68.8	81.68 %
DenseNet201	93	92.58 %	**59.3**	90.05 %	**57.8**	81.82 %
Moyenne	95.1	92.36%	71.5	89.45%	65.8	82.05%

d'entraînement plus long qui s'oppose totalement au réglage fin de la partie de classification, plus rapide, mais aussi moins précis. Le réglage fin partiel, réalisé en incluant le dernier bloc de couches d'extraction de caractéristiques (propre à chaque architecture) offre un compromis entre ces deux approches. L'écart précision-durée pour le transfert complet par rapport au réglage fin partiel est plus avantageux pour l'expérience 2 (2% meilleur pour 6 minutes de plus) que pour l'expérience 3 (3% meilleur pour 25 minutes de plus) encourageant à privilégier le réglage fin partiel pour des bases de données plus importantes. En termes de robustesse aux modèles inconnus, l'architecture VGG19 est plus précise, mais plus longue à entraîner s'opposant à l'architecture DenseNet201. Le constat est similaire aux approches de réglage fin : l'écart précision-durée pour VGG19 par rapport à DenseNet201 est plus avantageux pour l'expérience 2 (0.7% meilleur pour 6 minutes de plus) que pour l'expérience 3 (1.2% meilleur pour 15 minutes de plus). Globalement, ces deux architectures sont adaptées à l'apprentissage par transfert et au réglage fin de réseaux pour l'identification de modèles inconnus.

4 Conclusion et perspectives

Cet article présente une étude sur l'apport de l'apprentissage par transfert pour l'identification de modèle de caméras et notamment pour la classification de modèles inconnus, c'est-à-dire non étudiés lors de l'entraînement du réseau. L'analyse a été effectuée avec notre protocole, de trois expériences basées sur les modèles inconnus, pour différentes architectures de RNC (VGG19, ResNet50 et DenseNet201) afin d'observer l'impact du réglage fin de transfert sur leurs performances. Les résultats montrent dans un premier temps, par le biais d'une étude préliminaire la diminution de performance des réseaux pour la classification de modèles inconnus. L'étude finale atteste de la différence de performance (en temps et en précision) en fonction des approches du réglage fin. Notamment, les résultats montrent que le réglage fin partiel (dernier bloc de l'extracteur de caractéristiques et couches de classification) est à privilégier, au détriment d'un entraînement complet, pour des bases de données conséquentes afin de bénéficier d'un compris efficace. Ce même constat est observable pour les architectures RNC en privilégiant DenseNet201 plutôt que VGG19.

Références

[1] M. H. Al Banna, M. Ali Haider, M. J. Al Nahian, M. M. Islam, K. A. Taher, and M. S. Kaiser. Camera model identification using deep cnn and transfer learning approach. In *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*, pages 626–630, 2019.

[2] B. Bayar and M. C. Stamm. Towards open set camera model identification using a deep learning framework. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2007–2011, 2018.

[3] Yoshua Bengio and Yann LeCun, editors. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] Alexandre Berthet and Jean-Luc Dugelay. A review of data preprocessing modules in digital image forensics methods using deep learning. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 281–284, 2020.

[5] L. Bondi, L. Baroffio, D. Guera, P. Bestagini, E. J. Delp, and S. Tubaro. First steps toward camera model identification with convolutional neural networks. *IEEE Signal Processing Letters*, 24(3) :259–263, March 2017.

[6] T. Filler, J. Fridrich, and M. Goljan. Using sensor pattern noise for camera model identification. In *2008 15th IEEE International Conference on Image Processing*, pages 1296–1299, 2008.

[7] Thomas Gloe and Rainer Böhme. The 'dresden image database' for benchmarking digital image forensics. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1584–1590, New York, NY, USA, 2010. ACM.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[9] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[10] C. Li. Source camera identification using enhanced sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 5(2) :280–287, 2010.

[11] Judith A. Redi, Wiem Taktak, and Jean-Luc Dugelay. Digital image forensics : a booklet for beginners. *Multimedia Tools and Applications*, 51(1) :133–162, January 2011.

[12] A. Tuama, F. Comby, and M. Chaumont. Camera model identification with the use of deep convolutional neural networks. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, Dec 2016.

Amélioration de la robustesse de l'U-Net 3D contre la compression JPEG2000 pour la segmentation des organes pelviens masculins

Karim El Khoury, Martin Fockedey, Eliott Brion et Benoît Macq
ICTEAM, UCLouvain, 1348, Louvain-la-Neuve, Belgium

Résumé : *La segmentation d'organes est un processus essentiel en imagerie médicale pour la planification et le contrôle des traitements. En cas de grandes déformations, les algorithmes classiques de segmentation basés sur le traitement d'images et les atlas échouent. Dans de telles situations, l'apprentissage profond permet de fournir de meilleures solutions. Cependant, l'entraînement des réseaux d'apprentissage profond nécessite une grande quantité d'images. La disponibilité de ces données d'entraînement nécessite un transfert et un stockage de données importants pour lesquels la compression des images est obligatoire. Cependant, les déformations des données causées par la compression peut influencer l es p erformances de l'apprentissage profond. Dans ce travail, nous proposons d'étudier l'impact de la compression JPEG2000 sur la segmentation U-Net 2D et 3D des organes pelviens masculins. Nous montrons que l'utilisation d'un U-Net 3D finement ajusté permettrait de compresser deux fois plus les scans du patient pour la même performance de segmentation par rapport à un U-Net 2D.*

Mots-clés : Segmentation U-Net, JPEG2000, apprentissage profond, compression d'images, imagerie médicale

Les Descripteurs de covariance profonds pour la reconnaissance des expressions faciales

Naima Otberdout , Mohamed Daoudi, Lahoucine Ballihi
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
IMT Lille Douai, Univ. Lille, CNRS, UMR 9189 CRIStAL, F-59000 Lille, France
LRIT - CNRST URAC 29, Univ. Mohammed V de Rabat, Faculté des Sciences, Rabat, Maroc.

Résumé : *Dans cet article, nous proposons une nouvelle approche exploitant la puissance des réseaux de neurones profonds pour encoder les caractéristiques non-linéaires des expression faciales d'un visage et la puissance des matrices de covariance pour encoder les relations entres ces caractéristiques. Notre approche est basée sur l'idée d'encoder les cartes de caractéristiques locales et globales de la couche la plus profonde d'un réseau neuronal convolutif (DCNN) extraites des images fixes dans des matrices de covariance. Ces matrices de covariance sont symétriques définies positives (SPD). En effectuant la classification des expressions faciales en utilisant le noyau Gaussien sur la variété des matrices SPD, nous montrons que notre approche donne des résultats de classification meilleure que celle qui utilise les couches entièrement connectées et softmax. En effectuant des expérimentations sur trois bases de données des expressions faciales (Oulu-CASIA, CK+, SFEW), nous montrons que l'approche proposée atteint des performances compétitives par rapport à l'état de l'art.*

Mots-clés : Apprentissage profond, matrices de covariance, matrices symétriques définies positives, reconnaissance des expressions faciales.

1 Introduction

L'analyse automatique des expressions faciales est un sujet de recherche très important en vision par ordinateur depuis longtemps en raison de ses nombreuses applications potentielles qui vont de l'interaction homme-machine aux investigations médicales et psychologiques. Comme plusieurs autres applications de la vision par ordinateur les méthodes d'apprentissage profond ont beaucoup fait avancer ce problème. Suite à cette motivation, plusieurs recherches ont proposé d'utiliser les réseaux de neurones profonds (Deep Convolutional Neural Networks DCNN) pour la reconnaissance des expressions faciales [2, 1, 4, **?**]. Toutes ces méthodes utilisent une stratégie similaire dans l'architecture de réseau : plusieurs couches convolutionnelles pour l'extraction des caractéristiques ; des couches entièrement connectées avec une couche softmax sont utilisées pour la classification. Cependant, le pipeline de classification des DCNN est basé sur des informations globales et ne prend pas en compte les relations spatiales à l'intérieur du visage. Dans une autre direction, d'autres recherches ont utilisé les descripteurs de covariance pour modéliser les corrélations entre les parties du visage. Ces descripteurs locaux ont montré leur efficacité pour la reconnaissance des expressions faciales. La question est donc, comment est-il possible d'encoder des caractéristiques profondes locales dans une représentation compacte et discriminative pour une classification plus efficace que celle obtenue globalement par la couche softmax classique ? Dans de ce papier nous proposons une nouvelle représentation des expressions faciales qui encode les caractéristiques DCNN globales et locales par des matrices de covariance et les classer à l'aide d'un noyau Gaussian sur la variété des matrices SPD.

2 Travail proposé

Dans ce travail, nous proposons de construire des matrices de covariance des caractéristiques obtenues au niveau des couches profondes pour la reconnaissance des expressions faciales à partir des images. L'idée consiste à encoder les cartes de caractéristiques globales et locales dans une représentation compacte et discriminative. L'utilisation des matrices de covariance avec les cartes de caractéristiques des couches profondes permet de construire une représentation qui combine les caractéristiques de premier et deuxième ordre. En outre, cette méthode permet de combiner l'information locale et globale du visage en construisant des descripteurs locaux sur des régions spécifiques sur le visage (e.x, les yeux, les joues,...). Un modèle DCNN entraîné pour la reconnaissance des expressions faciales est capable de déterminer automatiquement les caractéristiques pertinentes spécifiques à chaque expression. Dans l'approche générale, la classification de ces caractéristiques est effectuée en utilisant des couches totalement connectées pour vectoriser ces caractéristiques, puis une couche softmax est ajoutée pour obtenir une probabilité pour chaque expression faciale. En revanche, dans ce travail, nous proposons d'exploiter les corrélations entre les cartes de caractéristiques des expressions faciales des visages extraites de la dernière couche de convolution en les agrégeant dans des matrices de covariance locales et globales. De ce fait, nous obtenons des représentations locales ainsi que globales compactes, robustes et efficaces. Ces matrices de covariances vivent dans la variété des matrices semi-définies positives, une variété non-linéaire. Nous avons classé ces descripteurs en utilisant les machines à vecteurs de support (SVM) avec un noyau Gaussien défini sur la variété SPD.

3 Résultats

Nous avons effectué les expérimentations sur trois bases de données des expressions faciales ((Oulu-CASIA, CK+ et SFEW) en utilisant deux modèles DCNN (VGG-face et ExpNet). Les résultats ont montré l'efficacité de l'approche proposée. En fait, nos résultats ont surpassé la classification avec la couche softmax classique. En plus, nous avons obtenu des résultats compétitifs en comparaison avec l'état de l'art.

Dataset	FC-Softmax	G-FMs	G+R-FMs
Oulu-CASIA	82.29	83.55	84.80
CK+	90.38	97.07	98.40
SFEW	48.26	49.18	49.18

TABLE 1 – Comparaison des méthodes proposées ; Global(G-FMs) et Global-Local (G+R-FMs)) avec la classification standard en utilisant les couches totalement connectées et softmax.

4 Conclusion et perspectives

Dans ce travail, nous avons proposé une nouvelle représentation compacte et discriminative pour la reconnaissance des expressions faciales. Cette représentation consiste à encoder les caractéristiques profondes du visage dans des matrices de covariance, matrices symétriques définies positives. Nous avons proposé une méthode de classification qui permet de les classer en respectant la géométrie de la variété des matrices SPD.

Références

[1] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet : Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017.

[2] Anis Kacem, Mohamed Daoudi, Boulbaba Ben Amor, and Juan Carlos Alvarez-Paiva. A novel space-time representation on the positive semidefinite cone for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3180–3189, 2017.

[3] Ikechukwu Ofodile, Kaustubh Kulkarni, Ciprian Adrian Corneanu, Sergio Escalera, Xavier Baro, Sylwia Hyniewska, Juri Allik, and Gholamreza Anbarjafari. Automatic recognition of deceptive facial expressions of emotion. *arXiv preprint arXiv :1707.04061*, 2017.

[4] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442, 2015.

[5] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442. Springer, 2016.

Méthod	Kacem et al.[2]	Ding et al[1]	Yu et al.[4]	Zhao et al.[5]	Ofodil et al.[3]	G-FMs	G+ R-FMs
Oulu-CASIA	83.13	82.2	-	84.59	89.60	83.55	84.80
CK+	96.87	98.60	-	97.30	98.70	97.25	98.40
SFEW	-	48.29	52.29	-	-	49.18	49.18

TABLE 2 – Comparaison avec l'état de l'art sur les trois bases de données.

Analyse de l'évolutivité d'un réseau d'apprentissage profond pour la stéganalyse d'images

Hugo RUIZ
Équipe ICAR, LIRMM,
Univ Montpellier, CNRS
Email : hugo.ruiz@lirmm.fr

Marc CHAUMONT
Équipe ICAR, LIRMM,
Univ Montpellier, CNRS, Univ Nîmes
Email : marc.chaumont@lirmm.fr

Mehdi YEDROUDJ
Équipe ICAR, LIRMM,
Univ Montpellier, CNRS
Email : mehdi.yedroudj@lirmm.fr

Frédéric COMBY
Équipe ICAR, LIRMM,
Univ Montpellier, CNRS
Email : frederic.comby@lirmm.fr

Gérard SUBSOL
Équipe ICAR, LIRMM,
Univ Montpellier, CNRS
Email : gerard.subsol@lirmm.fr

Résumé : *Depuis l'émergence de l'apprentissage profond et son utilisation dans le domaine de la stéganalyse, la plupart des travaux ont continué à utiliser des CNNs de taille petite à moyenne, et à les faire apprendre sur des bases de données relativement petites.*
De même, les benchmarks et les comparaisons entre les différents algorithmes de stéganalyse basés sur des CNNs, sont effectués sur des bases de données de petite à moyenne taille. Ceci ne permet pas de savoir, 1) *si le classement entre algorithmes, avec un critère comme l'« accuracy », reste le même si la base de données d'apprentissage est plus grande,* 2) *si l'efficacité des CNNs s'effondre quand la taille de la base d'apprentissage augmente d'un ordre de grandeur,* 3) *et enfin, la taille minimale requise pour obtenir un résultat meilleur que celui obtenu par une prédiction aléatoire.*
Dans cet article, après une discussion sur le comportement observé des CNNs en fonction de leur taille et de la taille de la base de données, nous confirmons que la loi de puissance de l'erreur est également valable en stéganalyse, et ce dans le cas limite d'un réseau de taille moyenne, sur une base de données diverse, de grande taille, et dont le développement est « contrôlé ».

Mots-clés : stéganalyse, passage à l'échelle, million d'images, développement « contrôlé »

1 Introduction

La stéganographie est l'art de dissimuler des informations dans un support anodin de sorte que l'existence même du message secret soit cachée à tout observateur non averti. Inversement, la stéganalyse est l'art de détecter la présence de données cachées dans de tels supports [7]. Tout au long de cet article, les images dites « cover » font références à des images originales et les images dites « stego » sont des images ayant été altérées.

Depuis 2015, grâce à l'utilisation du l'apprentissage profond, les performances de la stéganalyse se sont considérablement améliorées [4]. Néanmoins, dans de nombreux cas, ces performances dépendent de la taille de la base de données d'apprentissage. Il est en effet communément admis que, globalement, plus l'ensemble de données est grand, meilleurs sont les résultats [20].

L'objectif de cet article est de mettre en évidence l'amélioration des performances d'un algorithme de stéganalyse basé sur l'apprentissage profond lorsque la taille de l'ensemble d'apprentissage augmente.

Dans la section 2.1, nous discutons de ces questions et des lois ou modèles qui ont été proposés par la communauté scientifique. Ensuite, dans la section 2.2, nous présentons les tests réalisés pour évaluer la loi de puissance de l'erreur. Nous justifions et discutons les différents choix et paramétrages nécessaires à l'exécution des expériences. Dans la section 3, nous présentons le protocole expérimental et décrivons les expériences menées. Nous analysons ensuite l'évolution de l'*accuracy* en fonction de la taille de l'ensemble d'apprentissage. Enfin, nous concluons et donnons quelques perspectives.

2 Travail proposé

2.1 Passage à l'échelle du modèle et des données

De nombreux articles théoriques et pratiques tentent de mieux comprendre le comportement des réseaux de neurones lorsque leur dimension augmente [3, 16, 2, 11, 11] ou lorsque le nombre d'exemples augmente [8, 15, 12]. De nombreuses expériences sont réalisées afin d'observer l'évolution de l'erreur de test en fonction de la *taille du modèle*, ou en fonction de la *taille de l'ensemble d'apprentissage*. Ces recherches sont essentielles car la découverte de lois génériques pourrait confirmer que les utilisateurs de CNNs appliquent les bonnes méthodologies.

Dans les études sur *la mise à l'échelle du modèle*, les chercheurs ont observé trois régions en fonction de la taille du modèle. Il y a la région de *sous-apprentissage* du modèle, la région de *sur-apprentissage* du modèle, et enfin la région de *sur-paramétrage* du modèle. Le point de transition vers la région sur-paramétrée est appelé le *seuil d'interpolation* (Voir figure 1 dans [13]).

En général, la conclusion est que les réseaux sur-paramétrés (possédant des millions de paramètres) peuvent être en pratique utilisés pour n'importe quelle tâche. On peut citer, par exemple, le réseau EfficientNet [17]). Ce réseau a été massivement utilisé par les compétiteurs [22, 5] du concours Alaska#2 [6].

Dans les études sur la mise à l'échelle des données, les chercheurs ont observé qu'il existe trois régions en fonction

de la taille de l'ensemble de données [8]. Il s'agit de la région à *faible quantité de données*, de la région de la *loi de puissance* (*power law*) et enfin de la région d'*erreur irréductible* (*irreducible error*) (Voir figure 2 dans [13]).. Dans la région de la loi de puissance, plus les données sont nombreuses, meilleurs sont les résultats [15, 19].

Récemment, les auteurs de [12] ont proposé une loi générique qui modélise le comportement lors de la mise à l'échelle intégrant à la fois de la taille du modèle et la taille de l'ensemble de données. Le premier terme est fonction de la taille de l'ensemble de données, notée n, et le second est fonction de la taille du modèle, notée m :

$$\begin{aligned} \epsilon : \mathbb{R} \times \mathbb{R} &\rightarrow [0,1] \\ \epsilon(m,n) &\rightarrow \underbrace{a(m)n^{-\alpha(m)}}_{\text{données}} + \underbrace{b(n)m^{-\beta(n)}}_{\text{modèle}} + c_\infty \end{aligned} \tag{1}$$

avec $\alpha(m)$ et $\beta(n)$ contrôlant le taux de décroissance de l'erreur, dépendant respectivement de m et n, et c_∞ l'erreur irréductible, une constante réelle positive, indépendante de m et n.

Ensuite, les auteurs proposent une simplification de l'expression en :

$$\tilde{\epsilon}(m,n) = an^{-\alpha} + bm^{-\beta} + c_\infty \tag{2}$$

avec a, b, α, et β constantes positives réelles.

Avec un réseau efficace, ayant un nombre conséquent de paramètres, et avec suffisament de données d'apprentissage, on atteind la région de la région de la loi de puissance et ainsi l'équation 2 peut être simplifiée, comme dans [8] :

$$\epsilon(n) = a'n^{-\alpha'} + c'_\infty \tag{3}$$

Dans la suite de notre article, nous observons, dans le contexte de la stéganalyse JPEG, le comportement d'un réseau moyen lorsque la taille du jeu de données augmente.

2.2 Conception des tests de référence

Choix du réseau : Notre objectif est d'évaluer l'*accuracy* (ou de manière équivalente la probabilité d'erreur) en fonction de l'augmentation de la taille du jeu de données. Étant limités par les ressources informatiques, nous avons donc besoin d'un réseau de faible complexité et nous avons donc sélectionné le réseau LC-Net [10], ayant seulement 300 000 paramètres, et reconnu comme l'un des meilleurs CNN en stéganalyse JPEG à la date à laquelle nous avons réalisé les expériences (entre septembre 2019 et août 2020).

Choix de la charge utile : L'objectif, ici, est d'obtenir une « accuracy » comprise entre 60 et 70% [1] pour une petite base de données [2], afin d'observer la progression lorsque l'ensemble de données est échelonné. Après de nombreux ajustements expérimentaux, nous avons trouvé que 0,2 bits par coefficient AC non nul (bpnzacs) était une bonne charge utile pour une image JPEG 256×256 pixels en niveau de gris avec un facteur de qualité JPEG de 75.

Choix relatifs à la base de données : Nous avons décidé de travailler sur des images JPEG en niveaux de gris afin de mettre de côté la stéganalyse couleur, qui est encore récente et pas encore assez comprise théoriquement [1].

Nous avons également décidé de travailler uniquement sur des images de taille 256×256 avec un facteur de qualité 75. Les conclusions obtenues dans ce qui suit s'étendent vraisemblablement sur un petit intervalle autour du facteur de qualité 75 comme observé dans [23].

3 Résultats

3.1 Base de données & matériel utilisés

Les expériences ont été menées sur la base de données LSSD [14]. La base LSSD a l'avantage d'être séparée en plusieurs tailles différentes (10k, 50k, 100k, 500k, 1M et 2M) permettant l'étude du passage à l'échelle. Les images de LSSD ont été obtenues en développant les images RAW [3] des différentes bases de données : Alaska#2, BOSS, Dresden, RAISE, Stego App, et Wesaturate. Dans nos expériences, nous n'avons utilisé que les versions 10k jusqu'à 500k de la base de données « cover » en raison du temps d'apprentissage excessivement long pour les versions 1M et 2M d'images.

La base de données « cover » utilisée pour la phase de test est composée de 100 000 images et sera toujours la même, quelles que soient les expériences. Cette base de données de test est obtenue en développant des images RAW qui n'étaient pas présentes dans la base de données prévue pour l'apprentissage et conserve quasiment la même distribution des bases de données initiales. Ainsi, le scénario de stéganalyse est proche d'un scénario clairvoyant, où le jeu de test et le jeu d'apprentissage sont statistiquement très proches.

L'étude a été réalisée sur un serveur IBM grâce à un docker ayant accès à 144 processeurs Altivec POWER9 supportés (MCP) et à deux cartes graphiques GV100GL (Tesla V100 SXM2 16Go).

3.2 Entraînement, validation et test

Le processus d'insertion a été réalisé par une l'implémentation Matlab de l'algorithme J-UNIWARD [9], avec une charge utile de 0,2 bpnzacs. Il a fallu près de trois jours (2 jours et 20 heures) pour l'incorporation sur un Intel Xeon W2145 (8 cœurs, 3.7 GHz Turbo (max 4.5 GHz), 11M de cache).

Avant d'alimenter le réseau de neurones, les images JPEG doivent être décompressées afin d'obtenir des images spatiales non arrondies en « valeurs réelles ». Notons que l'espace de stockage nécessaire devient important [4]. Afin d'éviter de stocker toutes les images décompressées, il faudrait effectuer une décompression « en ligne » de manière asynchrone couplée à une construction de mini-batch « en ligne », pour alimenter le réseau neuronal « à la volée ».

L'ensemble d'apprentissage est divisé en deux ensembles : 90% pour l'ensemble d'apprentissage « réel » et 10% pour la validation. Comme dit précédemment, l'ensemble de test est toujours le même et est composé de 200k images (« cover » et « stegos »).

3.3 Hyper-paramètres

Pour entraîner notre CNN, nous avons utilisé une descente de gradient stochastique en mini-batch sans dro-

1. « Accuracy » assez éloignée de la région de prédiction aléatoire.
2. Une base de données trop petite pourrait biaiser l'analyse puisqu'il existe une région où l'erreur augmente lorsque l'ensemble de données augmente (voir [11]).
3. Données issues des capteurs de l'appareil photo.
4. Pour une image en niveaux de gris 256×256, la taille du fichier est d'environ 500 kB lorsqu'il est stocké au format MAT en *double*.

pout. Nous avons utilisé la majorité des hyper-paramètres de l'article [10]. Le taux d'apprentissage, pour tous les paramètres, a été fixé à 0,002 et est diminué aux époques 130 et 230, avec un facteur égal à 0,1. L'optimiseur est Adam, et la décroissance des poids est de 5.10^{-4}. La taille du batch est fixée à 100, ce qui correspond à 50 paires cover/stego. Afin d'améliorer la généralisation du CNN, nous avons mélangé l'ensemble de la base d'entraînement au début de chaque époque. La première couche a été initialisée avec les 30 filtres passe-haut de base de SRM, sans normalisation, et le seuil de la couche TLU est égal à 31 comme dans [18, 21]. Nous avons effectué un arrêt précoce après 250 époques comme dans [10]. Une partie du matériel est dispobible ici : http://www.lirmm.fr/ chaumont/LSSD.html.

3.4 Résultats et discussions

Les différents ensembles d'apprentissage, de 20k à 1M d'images (« cover » et « stegos »), ont été utilisés pour tester le LC-Net. Le tableau 1 donne les performances du réseau lorsqu'il a été évalué sur la base de test (200k images) suite à l'apprentissage. Notez que plusieurs tests ont été effectués pour chaque taille de l'ensemble d'apprentissage et que les « accuracys » affichées représentent une moyenne calculée sur les 5 meilleurs modèles sélectionnés grâce à l'ensemble de validation.

TABLE 1 – Accuracy moyenne évaluée sur l'ensemble de test de 200 000 images de cover/stego, en fonction de la taille de la base de données d'apprentissage.

Images	Tests	Accuracy	Std. dev.	Durée
20,000	5	62.33%	0.84%	2h 21
100,000	5	64.78%	0.54%	11h 45
200,000	5	65.99%	0.09%	23h 53
1,000,000	1	68.31%	/	10j

Il faut noter que les temps d'apprentissage deviennent importants (10 jours) dès que le nombre d'images dépasse 1 million. Il s'agit d'un problème important qui ne nous a pas permis de réaliser une évaluation sur les bases de données 2M (1M de « cover » + 1M de « stegos ») et 4 millions.

Les résultats du tableau 1, obtenus pour la charge utile 0.2 bpnzacs, confirment que plus l'ensemble d'apprentissage est grand (100k, 200k, 1M), meilleure est l'« accuracy ». Pour la base de données 20k, l'« accuracy » est de 62% et croit de presque 2% à chaque fois que la taille de l'ensemble d'apprentissage augmente. De plus, l'écart-type devient de plus en plus petit, ce qui souligne que le processus d'apprentissage est de plus en plus stable à mesure que la base de données augmente.

Ces premiers résultats signifient que la plupart des expériences de stéganalyse menées par la communauté, en utilisant un réseau d'apprentissage profond de taille moyenne (mais aussi de grande taille), ne sont pas réalisées avec suffisamment d'exemples pour atteindre la performance optimale, puisque la plupart du temps la base de données est comprise entre 10 000 (ensemble d'apprentissage BOSS) et 150 000 images (ensemble d'apprentissage Alaska#2 avec un seul algorithme d'insertion). Ainsi, dans nos expériences, l' « accuracy » est déjà améliorée de 6% lorsque la base de données passe de 20 000 à 1 million d'images et l' « accuracy » peut probablement être améliorée en augmentant la taille de l'ensemble de données puisque la région d'erreur irréductible n'est probablement pas atteinte.

Ces résultats confirment également qu'un réseau de taille moyenne tel que LC-Net ne voit pas ses performances s'effondrer lorsque la taille de la base de données augmente.

Plus intéressant encore, à partir de ces premiers résultats, nous pouvons estimer la loi de puissance suivante $\epsilon(n) = 0.492415n^{-0.086236} + 0.168059$. Ainsi, si nous choisissons n = 20M d'images, cette loi de puissance prédit une probabilité d'erreur de 28,3%. Si l'on considère une probabilité d'erreur de 28,3% pour 20M d'images, le gain obtenu par rapport à la probabilité d'erreur de 37,7% avec 20k images, correspond à une augmentation de 9% ce qui est une amélioration considérable dans le domaine de la stéganalyse.

En conclusion, la loi de puissance de l'erreur est confirmée pour la stéganalyse avec le Deep Learning, et ce même lorsque les réseaux ne sont pas très grands (300 000 paramètres), même en commençant avec une base de données de taille moyenne (ici, seulement 20 000 images), et même si la base de données est diversifiée. De plus grandes bases de données sont nécessaires pour un apprentissage optimal, et l'utilisation de plus d'un million d'images est probablement nécessaire avant d'atteindre la région d'erreur irréductible [8].

4 Conclusion et perspectives

Dans cet article, nous avons d'abord rappelé les résultats récents obtenus par la communauté travaillant sur l'apprentissage profond, et relatifs au comportement des réseaux d'apprentissage profond lorsque la taille du modèle ou de la base de données augmente. Nous avons ensuite proposé un dispositif expérimental afin d'évaluer le comportement d'un stéganalyseur CNN de taille moyenne (LC-Net) lorsque la taille de la base de données est augmentée.

Les résultats obtenus montrent qu'un réseau de taille moyenne ne s'effondre pas lorsque la taille de la base de données augmente (jusqu'à 1M), malgré une certaine diversité. De plus, ses performances sont accrues avec l'augmentation de la taille de la base de données. Enfin, nous avons observé que la loi de puissance de l'erreur est également valable pour le domaine de la stéganalyse.

Les travaux futurs devront être réalisés sur une base de données encore plus diverse (facteurs de qualité, taille de la charge utile, algorithme d'insertion, couleur, base de données moins contrôlée...), et également avec d'autres réseaux. Plus pratiquement, un effort devra être fait afin de réduire le temps d'apprentissage, et surtout la gestion de la mémoire. Enfin, il reste des questions ouvertes à résoudre telles que : trouver une valeur d'erreur irréductible plus précise, trouver la pente de la loi de puissance en fonction du point de départ du CNN (utilisation du transfert, utilisation du curriculum, utilisation de l'augmentation des données comme les pixels-off [20]).

Remerciements

Les auteurs tiennent à remercier la Direction Générale de l'Armement (DGA) pour son soutien dans le cadre du projet ANR Alaska (ANR-18-ASTR-0009). Nous remercions également IBM Montpellier et l'Institut de Dévelop-

pement et de Ressources en Calcul Scientifique Intensif (IDRISS/CNRS) pour nous avoir donné accès à des ressources de Calcul Haute Performance.

Références

[1] Hasan Abdulrahman, Marc Chaumont, Philippe Montesinos, and Baptiste Magnier. Color Images Steganalysis Using RGB Channel Geometric Transformation Measures. *Security and Communication Networks*, 9(15) :2945–2956, 2016.

[2] Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-Dimensional Dynamics of Generalization Error in Neural Networks. *Neural Networks*, 132 :428–446, 2020.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling Modern Machine-Learning Practice and the Classical Bias–variance Trade-off. *Proceedings of the National Academy of Sciences*, 116(32) :15849–15854, 2019.

[4] Marc Chaumont. Deep Learning in steganography and steganalysis. In M. Hassaballah, editor, *Digital Media Steganography : Principles, Algorithms, Advances*, chapter 14, pages 321–349. Elsevier, July 2020.

[5] Kaizaburo Chubachi. An Ensemble Model using CNNs on Different Domains for ALASKA2 Image Steganalysis. In *Proceedings of the IEEE International Workshop on Information Forensics and Security, WIFS'2020*, Virtual Conference due to Covid (Formerly New-York, NY, USA), December 2020.

[6] Rémi Cogranne, Quentin Giboulot, and Patrick Bas. Challenge Academic Research on Steganalysis with Realistic Images. In *Proceedings of the IEEE International Workshop on Information Forensics and Security, WIFS'2020*, Virtual Conference due to Covid (Formerly New-York, NY, USA), December 2020.

[7] Jessica Fridrich. *Steganography in Digital Media*. Cambridge University Press, 2009. Cambridge Books Online.

[8] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep Learning Scaling is Predictable, Empirically. In *Unpublished - ArXiv*, volume abs/1712.00409, 2017.

[9] Vojtech Holub, Jessica Fridrich, and Tomas Denemark. Universal Distortion Function for Steganography in an Arbitrary Domain. *EURASIP Journal on Information Security, JIS*, 2014(1), 2014.

[10] Junwen Huang, Jiangqun Ni, Linhong Wan, and Jingwen Yan. A Customized Convolutional Neural Network with Low Model Complexity for JPEG Steganalysis. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec'2019, pages 198–203, Paris, France, July 2019.

[11] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep Double Descent : Where Bigger Models and More Data Hurt. In *Proceedings of the Eighth International Conference on Learning Representations, ICLR'2020*, Virtual Conference due to Covid (Formerly Addis Ababa, Ethiopia), April 2020.

[12] Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A Constructive Prediction of the Generalization Error Across Scales. In *Proceedings of the Eighth International Conference on Learning Representations, ICLR'2020*, Virtual Conference due to Covid (Formerly Addis Ababa, Ethiopia), April 2020.

[13] Hugo Ruiz, Marc Chaumont, Mehdi Yedroudj, Ahmed Oulad Amara, Frédéric Comby, and Gérard Subsol. Analysis of the Scalability of a Deep-Learning Network for Steganography "Into the Wild". *Lecture Notes in Computer Science, Springer LNCS*, 12666 :439 –452, January 2021.

[14] Hugo Ruiz, Mehdi Yedroudj, Marc Chaumont, Frédéric Comby, and Gérard Subsol. LSSD : a Controlled Large JPEG Image Database for Deep-Learning-based Steganalysis "Into the Wild". *Lecture Notes in Computer Science, Springer LNCS*, 12666 :470 – 483, January 2021.

[15] Vittorio Sala. Power Law Scaling of Test Error Versus Number of Training Images for Deep Convolutional Neural Networks. In *Proceedings of the Multimodal Sensing : Technologies and Applications*, volume 11059, pages 296 – 300, Munich, 2019. International Society for Optics and Photonics, SPIE.

[16] S Spigler, M Geiger, S d'Ascoli, L Sagun, G Biroli, and M Wyart. A Jamming Transition from Under- to Over-parametrization Affects Generalization in Deep Learning. *Journal of Physics A : Mathematical and Theoretical*, 52(47) :474001, oct 2019.

[17] Mingxing Tan and Quoc Le. EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning, PMLR'2019*, volume 97, pages 6105–6114, Long Beach, California, USA, June 2019.

[18] Jian Ye, Jiangqun Ni, and Y. Yi. Deep Learning Hierarchical Representations for Image Steganalysis. *IEEE Transactions on Information Forensics and Security, TIFS*, 12(11) :2545–2557, November 2017.

[19] Mehdi Yedroudj, Marc Chaumont, and Frédéric Comby. How to Augment a Small Learning Set for Improving the Performances of a CNN-Based Steganalyzer ? In *Proceedings of Media Watermarking, Security, and Forensics, MWSF'2018, Part of IS&T International Symposium on Electronic Imaging, EI'2018*, page 7, Burlingame, California, USA, 28 January - 2 February 2018.

[20] Mehdi Yedroudj, Marc Chaumont, Frederic Comby, Ahmed Oulad Amara, and Patrick Bas. Pixels-off : Data-Augmentation Complementary Solution for Deep-Learning Steganalysis. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, IHMSec '20, page 39–48, Virtual Conference due to Covid (Formerly Denver, CO, USA), June 2020.

[21] Mehdi Yedroudj, Frédéric Comby, and Marc Chaumont. Yedrouj-Net : An Efficient CNN for Spatial Steganalysis. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'2018*, pages 2092–2096, Calgary, Alberta, Canada, April 2018.

[22] Yassine Yousfi, Jan Butora, Eugene Khvedchenya, and Jessica Fridrich. ImageNet Pre-trained CNNs for JPEG Steganalysis. In *Proceedings of the IEEE International Workshop on Information Forensics and Security, WIFS'2020*, Virtual Conference due to Covid (Formerly New-York, NY, USA), December 2020.

[23] Yassine Yousfi and Jessica Fridrich. JPEG Steganalysis Detectors Scalable With Respect to Compression Quality. In *Proceedings of Media Watermarking, Security, and Forensics, MWSF'2020, Part of IS&T International Symposium on Electronic Imaging, EI'2020*, page 10, Burlingame, California, USA, January 2020.

Poster café

Estimation du Regard par un Réseau de Capsules

Vivien Bernard
IMT Lille Douai, Univ. Lille, CNRS
UMR 9189 - CRIStAL
F-59000 Lille, France
vivien.bernard@imt-lille-douai.fr

Hazem Wannous
IMT Lille Douai, Univ. Lille, CNRS
UMR 9189 - CRIStAL
F-59000 Lille, France
hazem.wannous@imt-lille-douai.fr

Jean-Philippe Vandeborre
IMT Lille Douai, Univ. Lille, CNRS
UMR 9189 - CRIStAL
F-59000 Lille, France
jean-philippe.vandeborre@imt-lille-douai.fr

Résumé : *L'information du regard est utilisée dans de nombreuses plateformes, comme les casques de réalité virtuelle par exemple. Afin d'estimer le regard humain, plusieurs solutions ont été proposées, utilisant différents matériels et techniques. Cependant, parvenir à une estimation correcte sur des appareils relativement peu couteux utilisant des caméras RVB permettrait d'ouvrir la porte à de nouvelles interractions sur les appareils mobiles et donc généraliser ces interractions. Nous proposont dans ce papier une nouvelle méthode pour l'estimation du regard fondée sur une nouvelle architecture de réseau neuronnaux : les réseaux de capsules. Ces capsules ont montré de très bons résultats sur des problèmes de classification, mais très peu sur des problèmes de regression. En tirant parti des réseaux de capsules et sa capacité à reconstruire des images, nous pouvons recréer des images simplifiées de l'oeil et d'estimer le regard à partir de celles-ci. Les tests sont effectués sur deux bases de données créées spécifiquement pour la reconnaissance du regard. Des résultats encourageants ont été obtenus sur l'estimation et la reconstruction.*

Mots-clés : regard, capsule, réseaux, neuronnes

1 Introduction

L'estimation du regard est un sujet de recherche depuis de nombreuses années déjà. Et même si certains travaux se concentrent sur l'estimation à l'aide de capteurs particuliers qui sont difficilement accessibles par tout le monde [3, 8], le besoin d'estimer le regard avec une simple caméra RVG augment à mesure que le nombre d'appareils les utilisant augmente. L'utilisation du regard est très variée. Des études de psychologie et de comportement [9, 4] aux interractions assistant l'être humain [2] et ouvrant de nouvelles possibilités avec la réalité augmentée (RA) et virtuelle (RV) [1].

Park *et al.* [5] ont introduit une architecture d'apprentissage profond spécifiquement faite pour l'estimation du regard. Ils génèrent à l'aide de cet apprentissage deux images appelées *gazemaps* à partie des images d'oeil. Il s'agit de deux images binaires, où l'une représente l'iris et l'autre le globe oculaire. Ensuite un autre réseau régresse à partir de ces images les angles composant le regard. En utilisant une représentation intermédiaire de l'oeil, et donc en simplifiant le travail de régression, ils estiment améliorer la reconnaissance du regard. Dans ce papier, nous utilisons la même approche en reconstruisant ces deux *gazemaps* mais utilisons les capsules pour l'estimation.

Les réseaux de capsules sont des architectures d'apprentissage profond introduite Sabour *et al.* [6] supposées corriger les problèmes inhérents aux réseaux convolutionnels (CNN). En effet, même si les CNN affichent d'excellents résultats dans de nombreux cas, ils ont aussi d'énormes limitations. D'abord, s'ils peuvent détecter certains motifs sur une image, ils ne peuvent donner plus d'information sur le motif, comme sa position ou sa rotation entre autre. Aussi, les CNN sont accompagnés d'une opération de *pooling*, qui réduit la taille de l'image pour que la couche de convolution suivante puisse travailler sur une plus grande surface. Cette opération a pour effet de perdre beaucoup d'informations de l'image. Recemment, les réseaux de Capsule ont été employé pour la reconnaissance d'action dans architecture réussissant l'information temporelle aux capsules 2D avec des informations spatiales plus complexes qu'un CNN classique [10].

2 Travail proposé

2.1 Architecture proposée

Notre approche est similaire à celle proposée par Park *et al.* [5], avec en plus l'utilisation des réseaux de capsules. On peut distinguer deux approches résumée dans la figure 1.

Nous avons mis en place deux réseaux différents pour estimer le regard. Dans les deux architectures, la première partie (en bleu dans les schémas) est identique. La première partie commune sont les capsules. Elles sont constituées d'une couche de convolution, une couche de capsule primaire, et de la couche de capsule. La deuxième partie commune est la reconstruction qui génère les *gazemaps*.

Bien que les deux parties mentionnées plus tôt sont partagées par les deux architectures, les parties suivantes sont distinctes. Dans le cas de la première architecture,

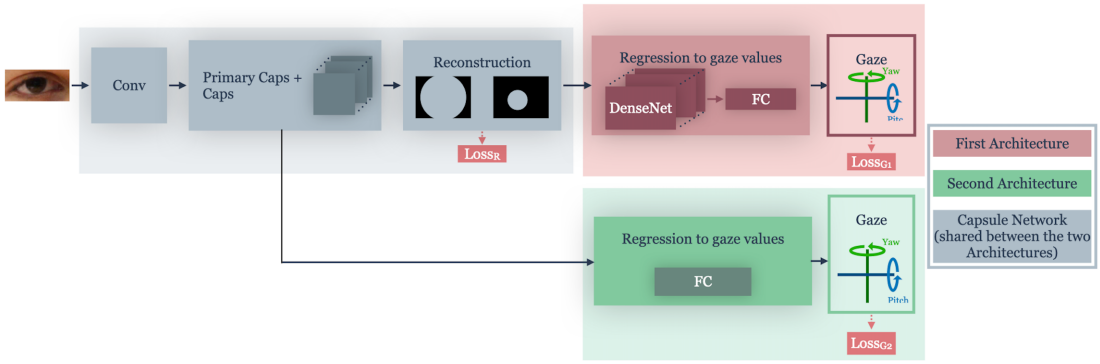


FIGURE 1 – Graphique de l'approche montrant les différences entre les deux architectures et leurs parties communes. Le réseau de capsule est identique pour les deux architectures. La première fonction de perte est la même dans les deux cas, tandis que chaque architecture utilise sa propre fonction de perte pour la deuxième.

un réseau inspiré de DenseNet et un réseau entièrement connecté régresse l'information du regard depuis les images binaires (les *gazemaps*). Dans la seconde, l'information du regard est directement régressée depuis la sortie des capsules grâce à un réseau entièrement connecté. Nous allons maintenant étudier l'architecture en détail.

2.1.1 Réseau de capsules

Le réseau global prend en entrée l'image de l'oeil originale et en ressort les angles du regard. Le réseau de capsules est fait de trois parties.

— **Convolution** : il s'agit de la première partie qui traite sur l'image passée en entrée du réseau directement. Elle permet de détecter les caractéristiques basiques de l'oeil et donc aide les couches suivantes à travailler sur des données plus claires.

— **Capsules primaires** : cette couche travaille sur les caractéristiques détectées plus tôt, et les combine. Il existe plusieurs capsules primaires indépendantes, et elles prennent toutes les mêmes données en entrée.

— **Capsules** : la dernière couche du réseau de capsules. Elle est constituées des capsules elles-mêmes.

2.1.2 Régression

Après le réseau de capsules, l'information du regard, constituée des deux angles définissant la direction, est régressée. Comme mentionné plus tôt, chacune des deux architectures régresse le regard à partir de points différents du réseau. Des détails sur l'implémentation sont données en figure 2.

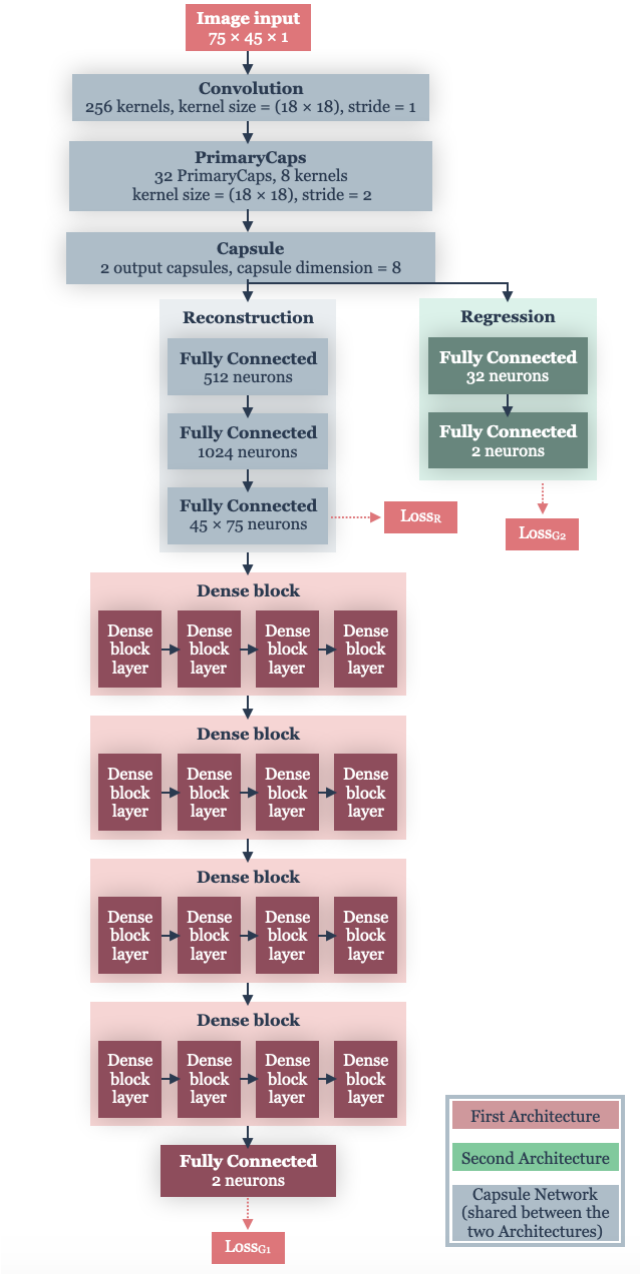


FIGURE 2 – Details de l'implémentation des deux architectures. Certaines parties sont communes alors que d'autres sont spécifique à l'une ou l'autre.

3 Résultats

Nous avons évalué notre approche à l'aide de deux bases de données : MPIIGaze [11] et Columbia [7]. Nous avons ensuite comparé nos résultats avec ceux de l'état de l'art.

3.1 Bases de données et expérimentations

La base de données MPIIGaze est composée de 213 659 images obtenues à partir de 15 participants. Afin de mesurer l'efficacité de notre approche sur cette base, nous

Modèle	Nombre param	Entrées	Err. (degrés) MPIIG.	Colum.
kNN [12]	0	OV	7.2	-
RF [12]	-	OV	6.7	-
[11]	1.8M	OV	6.3	-
AlexNet	86M	O	5.7	4.2
VGG-16	158M	O	5.4	3.9
GazeNet [12]	90M	OV	5.5	-
DeepPict. [5]	0.7M	O	4.5	3.8
1ère arch.	32M	O	5.7	6.1

TABLE 1 – Résultats de la première architecture. *Entrées* correspond au type de données données au réseau : **O** correspond à une image d'oeil seulement, et **OV** à une image d'oeil et à la pose du visage. L'erreur est exprimée en degrés.

Modèle	Nombre param	Entrées	Err. (degrés) MPIIG.	Colum.
kNN [12]	0	OV	7.2	-
RF [12]	-	OV	6.7	-
[11]	1.8M	OV	6.3	-
AlexNet	86M	O	5.7	4.2
VGG-16	158M	O	5.4	3.9
GazeNet [12]	90M	OV	5.5	-
DeepPict. [5]	0.7M	O	4.5	3.8
2ème arch.	32M	O	6.0	7.2

TABLE 2 – Résultats de la deuxième architecture. *Entrées* correspond au type de données données au réseau : **O** correspond à une image d'oeil seulement, et **OV** à une image d'oeil et à la pose du visage. L'erreur est exprimée en degrés.

utilisons le protocol de *leave one out* en validation croisée. Ainsi, le premier participant est utilisé pour le test, alors que les 14 autres sont utilisés pour l'entrainement. Les jeux de test et d'entrainement ne s'intersectent donc pas. En répétant ce processus pour chaque personne, nous obtenons 15 résultats différents, et il nous suffit de faire la moyenne de ceux-ci pour obtenir le résultat final.

La base Columbiaest quant à elle constituée de 56 participants pour un total de 5 880 images de très bonne qualité. Nous ne pouvons pas tout à fait utiliser la même méthode sachant que chaque participant possède bien moins d'image. ainsi, nous regrouppons les participants en 5 groupes distincts, et appliquons le protocol sur les groupes, plutôt que sur les personnes. Ainsi, alors qu'un groupe sert de jeu de test, les autres servent à l'entrainement. En prenant comme pour MPIIGaze la moyenne des résultats, on peut obtenir le résultat final.

Notre implémentation a été réalisée en Python avec Tensorflow et Keras, et a été entrainée sur une NVIDIA GeForce RTX 2080 Ti avec 11 Go de mémoire GPU.

Le tableau 1 donne les résultats obtenus sur la première architecture, et le tableau 2 ceux obtenus sur la deuxième architecture.

4 Conclusion et perspectives

Nous avons proposé une architecture de réseau de capsules capable de produire des résultats très encourageant. Même pour un problème de régression, il est possible d'en tirer parti. Cette architecture, qui se sert directement des capsules pour l'estimation du regard, montre que les capsules sont capables d'apprendre les informations nécessaires à cette régression. En revanche, nous n'arrivons tout de même pas à atteindre l'état de l'art. Afin de s'en rapprocher, diverses solutions sont envisagées. Tout d'abord, il faut essayer plusieurs configurations de capsules (changement de dimensions, changement du nombre de capsules et du nombre de couches). Aussi, nous envisageons l'utilisation du *transfer learning* pour les premières convolutions.

Références

[1] Istvan Barakonyi, Helmut Prendinger, Dieter Schmalstieg, and Mitsuru Ishizuka. Cascading hand and eye movement for augmented reality videoconferencing. *2007 IEEE Symposium on 3D User Interfaces*, pages 71–78, 01 2007.

[2] Craig A Chin, Armando Barreto, J Gualberto Cremades, and Malek Adjouadi. Integrated electromyogram and eye-gaze tracking cursor control system for computer users with motor disabilities. 2008.

[3] Kenneth A Funes-Mora and Jean-Marc Odobez. Gaze estimation in the 3d space using rgb-d sensors. *International Journal of Computer Vision*, 118(2) :194–216, 2016.

[4] Sabrina Hoppe, Tobias Loetscher, Stephanie A Morey, and Andreas Bulling. Eye movements during everyday behavior predict personality traits. *Frontiers in human neuroscience*, 12 :105, 2018.

[5] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 721–738, 2018.

[6] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017.

[7] Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. Gaze locking : passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280, 2013.

[8] Li Sun, Zicheng Liu, and Ming-Ting Sun. Real time gaze estimation with a consumer depth camera. *Information Sciences*, 320 :346–360, 2015.

[9] Jos Nicolaas Van Der Geest, Chantal Kemner, Marinus N Verbaten, and Herman Van Engeland. Gaze behavior of children with pervasive developmental disorder toward human faces : a fixation time study. *Journal of Child Psychology and Psychiatry*, 43(5) :669–678, 2002.

[10] Théo Voillemin, Hazem Wannous, and Jean-Philippe Vandeborre. 2d deep video capsule network with tem-

poral shift for action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3513–3519, 2021.

[11] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.

[12] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze : Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1) :162–175, 2017.

Utilisation conjointe de données textuelles et modèles 3D pour l'aide à la rédaction de plans de traitement orthodontiques

Maxime Chapuis[1, 2], Noura Faraj[1], Mathieu Lafourcade[1], William Puech[1], Stéphane Fourcade[2], and Gérard Guillerm[2]

[1]LIRMM, Université de Montpellier, CNRS
[2]Groupe Orqual

Résumé : *Ce document décrit le pipeline de traitements envisagé pour le développement d'un système d'aide à la rédaction de plans de traitement orthodontiques. Les analyses et raisonnements de ce système reposent sur l'utilisation d'un diagnostic, d'un modèle 3D de la dentition du patient, d'un plan de traitement et d'une base de connaissances (connaissances spécifiques à l'orthodontie et connaissances de sens commun). La particularité du projet est de s'appuyer sur un modèle 3D pour consolider les informations textuelles.*

Mots-clés : Simulation 3D ; Traitement automatique du langage naturel ; Base de connaissances

1 Introduction

La planification de traitements orthodontiques assistée par ordinateur est de plus en plus répandue dans les cabinets d'orthodontie. Parmi les solutions commerciales existantes, nous pouvons citer 3Shape [1] Orthodontic Planer et SureSmile [2] Ortho pour les traitements à base d'appareils multi-attaches, ou encore Invisalign Clincheck [3] et SureSmile Aligners pour la simulation de traitements par gouttières. Ces outils offrent beaucoup de contrôles à l'utilisateur, ainsi qu'une grande diversité de traitements mais requièrent un nombre élevé d'actions de sa part. Le but principal de nos travaux, est de mettre au point un outil d'aide à la rédaction de plans de traitement utilisant des données textuelles (diagnostic, plan de traitement proposé, base de connaissances) et 3D (modèle 3D de la dentition du patient acquis à l'aide d'un scanner intra-oral) afin de réduire le nombre d'actions de l'utilisateur et rendre le processus le plus automatique possible. Le travail proposé dans les sections suivantes se distingue donc part (a) le caractère automatique des traitements et (b) l'utilisation conjointe de données textuelles et de modèles 3D.

2 Travail proposé

Afin de déterminer le pipeline de traitements nécessaires au développement d'un tel système, nous considérons trois cas d'utilisation de sophistication croissante. Dans le premier cas, le système dispose d'un diagnostic, d'un plan de traitement (indiquant les corrections à effectuer : multiattaches maxillaire et mandibulaire, cales occlusales, tractions intermaxillaires élastiques, *etc.*), du modèle 3D de la dentition du patient (modèle surfacique segmenté comme illustré en figure 1), et doit être capable de (1) déterminer si le plan de traitement est cohérent avec le diagnostic, (2) simuler le plan de traitement, et (3) produire une animation montrant l'évolution de la dentition (*i.e.* le résultat de la simulation). Dans le deuxième cas, le système ne dispose que d'un diagnostic et d'un modèle 3D de la dentition, et doit être capable de proposer de manière automatique un plan de traitement valide permettant de corriger la situation décrite dans le diagnostic. Enfin, dans le troisième cas, le système dispose du modèle 3D ainsi que des données de céphalométrie, et doit pouvoir dresser un diagnostic lexicalisé.

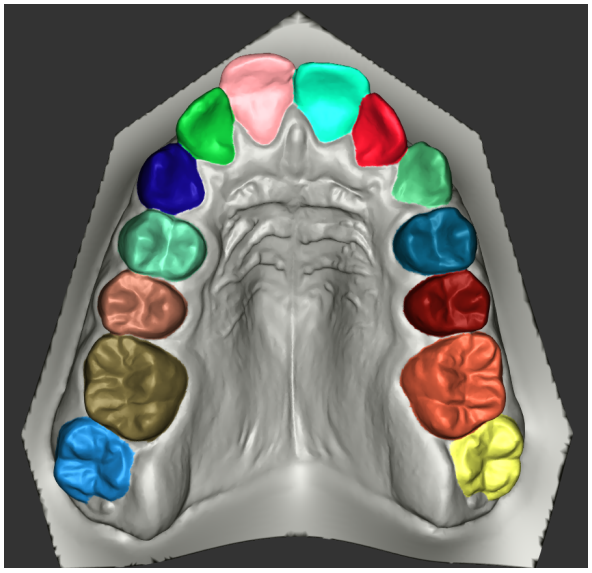


FIGURE 1 – Arcade dentaire maxillaire segmentée et étiquetée

Nous identifions trois verrous scientifiques à la réalisation d'un tel système. Pour pouvoir vérifier la cohérence des documents, simuler les traitements et proposer des modifications, le système doit avoir accès à des informations spécifiques à l'orthodontie (informations sur les pathologies, les traitements, les effets mécaniques des traitements sur les dents, *etc.*). Or, à notre connaissance, il n'existe pas de base de connaissances française contenant les informations recherchées. Il est donc nécessaire de la construire. Le réseau lexico-sémantique RezoJDM (projet JeuxDeMots

1. https ://www.3shape.com/fr/software-overview
2. https ://www.suresmile.com/en/
3. https ://www.invisalign.ca/

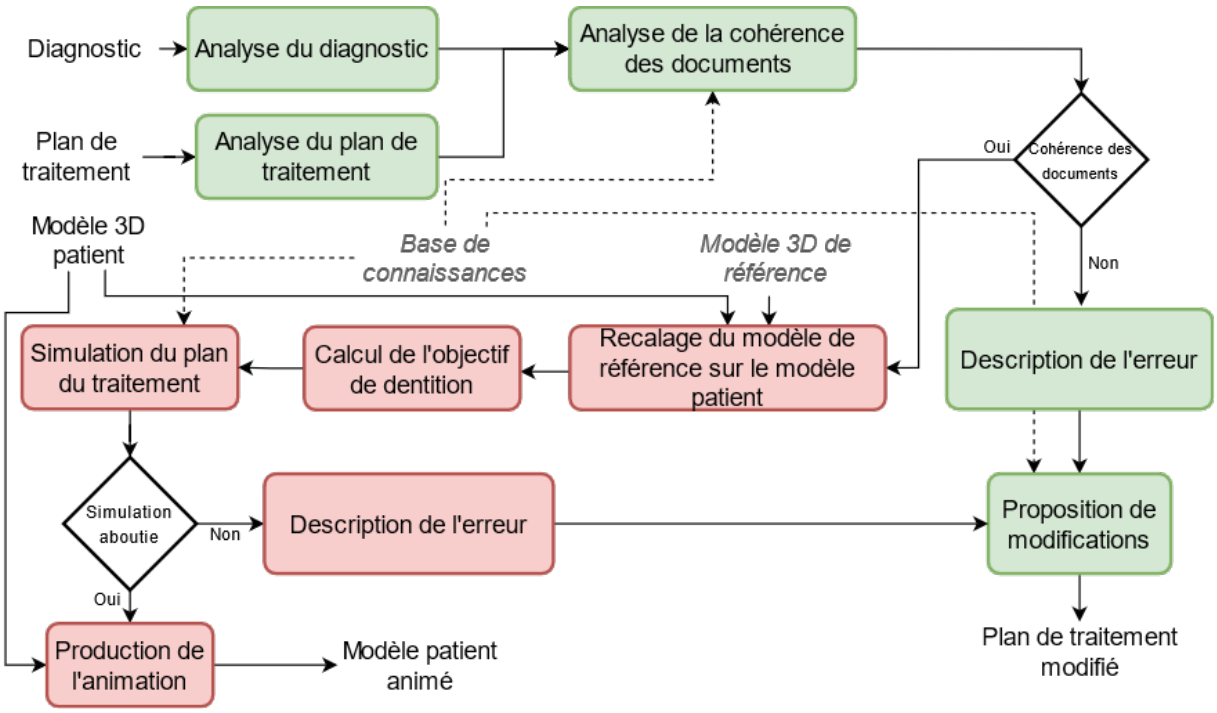


FIGURE 2 – Pipeline - En vert, les étapes relatives à l'analyse de textes. En rouge, celles relatives à l'analyse du modèle 3D.

[4]) est notre de point de départ. Un deuxième verrou scientifique concerne la nature des diagnostics et des plans de traitement. En effet, ces documents se rapprochent plus souvent de notes personnelles que de comptes rendus correctement rédigés. Il faut donc faire face à une syntaxe approximative, des abréviations personnelles, des imprécisions ainsi que des informations manquantes ou implicites. Pour ces raisons, le système que nous proposons s'appuie sur le modèle 3D afin de compléter les informations textuelles. Enfin, il est nécessaire pour le système d'avoir une notion de configuration dentaire idéale. Nous proposons de la calculer en adaptant un modèle de dentition de référence au modèle 3D patient.

3 Résultats

Le pipeline de traitement proposé pour le premier cas d'utilisation est illustré en figure 2.

D'abord une analyse du texte du diagnostic et du plan de traitement est effectuée. Il s'agit ici de repérer les éléments connus de diagnostic et de plan traitement (*i.e.* les éléments présents dans la base de connaissances). Cette analyse s'effectue en plusieurs passes, de manière à garantir une certaine robustesse aux termes *a priori* inconnus (corrections orthographiques, expansion d'abréviations, *etc.*).

Ensuite, une deuxième étape évalue la cohérence du plan de traitement par rapport au diagnostic, en mettant en relation les éléments du diagnostic précédemment identifiés avec ceux du plan de traitement. Cette mise en correspondance est réalisée grâce à la base de connaissances, qui contient des relations sémantiques permettant de déterminer si le plan de traitement est en mesure de traiter les pathologies du diagnostic. Deux issues sont possibles. La première, le plan de traitement n'est pas cohérent avec le diagnostic. Dans ce cas là, le système décrit le problème,

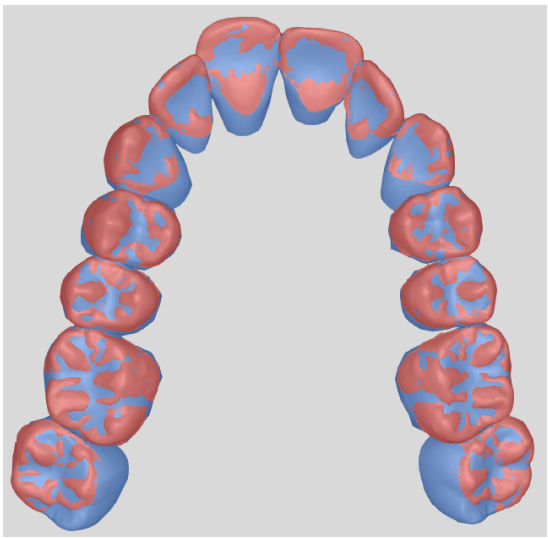


FIGURE 3 – Résultat du recalage de l'arcade maxillaire du modèle de référence (bleu) sur celle du modèle patient (rouge).

et propose une ou des modifications à apporter au plan de traitement. Sinon, si le plan de traitement est cohérent avec le diagnostic, le système passe à l'analyse du modèle 3D.

Ce processus débute par le recalage d'un modèle 3D de référence sur le modèle patient. Le modèle de référence représente une configuration dentaire générique, dans laquelle les dents sont correctement orientées et positionnées. Le recalage s'effectue en deux étapes. Une première, *rigide*, permet d'aligner et d'orienter chaque dent de référence sur la dent patient correspondante l'aide de l'algorithme ICP [1, 2]. Une seconde, *non-rigide*, vise à déformer les dents de références pour approcher la forme des dents du patient. Cette déformation est réalisée avec la méthode de projection APSS [3]. À l'issue du recalage, les dents du modèle de référence sont dans la même configuration que les dents du modèle patient (figure 3). Cela permet entre autre de calculer le défaut d'orientation et de position de chaque dent, par rapport à une configuration *standard*.

Le modèle de référence étant une dentition générique, il doit être adapté au patient avant de pouvoir servir d'objectif de dentition. Il est donc déformé en tenant compte de la forme de l'arche dentaire du patient, et de contraintes d'alignement et de positionnement dentaires.

Une fois l'objectif calculé, le système peut procéder à la simulation du plan de traitement sur le modèle de référence recalé et déformé. Pour cela il est nécessaire que les effets mécaniques des traitements sur les dents soient correctement renseignés dans la base de connaissances. Si la simulation aboutie, le plan de traitement est considéré comme valide, et le système produit une animation du traitement en appliquant les transformations effectuées sur le modèle de référence au modèle patient. Dans le cas contraire, le système présente le problème (violation d'une contrainte physique, *etc.*), le plan de traitement est considéré comme invalide, et une ou des modifications sont proposées à l'utilisateur.

4 Conclusion et perspectives

Dans cet article nous avons présenté un pipeline de traitements nous permettant de réaliser une première version d'un système d'aide à la rédaction de plan de traitement. Étant donné un diagnostic, un plan de traitement, un modèle 3D de la dentition du patient et une base de connaissances de traitements orthodontiques, celui-ci peut : (a) simuler le plan de traitement, et (b) proposer des modifications du plan de traitement en cas d'impossibilité.

Références

[1] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2) :239–256, 1992.

[2] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3) :145 – 155, 1992. Range Image Understanding.

[3] G. Guennebaud and M. Gross. Algebraic point set surfaces. *ACM Trans. Graph.*, 26(3) :23–es, July 2007.

[4] M. Lafourcade. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand, December 2007.

Une nouvelle métrique de qualité vidéo sans-référence basée bitstream pour les vidéos de vidéosurveillance

Hugo Merly, Alexandre Ninassi et Christophe Charrier
Normandie Univ., UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

Résumé : La problématique de l'évaluation de la qualité continue sans vidéo de référence est abordée dans cet article. L'approche retenue consiste à extraire des caractéristiques directement du bitstream de la vidéo au format H.264/AVC, puis à utiliser un algorithme de machine learning afin de prédire la qualité de la vidéo de manière continue. Les résultats obtenus montrent que l'approche retenue permet d'obtenir une très forte corrélation entre les scores ainsi prédits et la vérité terrain.

Mots-clés : Évaluation de la qualité vidéo sans référence, vidéoprotection, bitstream, Perceptron multicouche.

1 Introduction

Lors d'enquêtes judiciaires, les forces de l'ordre collectent habituellement un maximum de vidéos provenant de caméras de *vidéo-surveillance* autour des lieux d'intérêts. Des téra-octets de vidéo, correspondant à des milliers d'heures d'enregistrement, sont analysés à la recherche d'indices clés. La qualité d'encodage, de transmission et les conditions météorologiques sont autant de facteurs qui peuvent rendre ces vidéos inexploitables. Dans ce contexte, il devient alors crucial de disposer d'un outil permettant aux enquêteurs de pouvoir passer les séquences de vidéo qui sont trop dégradées pour être exploitables. Il convient alors de disposer d'un système d'évaluation de la qualité de la vidéo en continue.

Les méthodes d'évaluation de la qualité des vidéos peuvent être classées en trois catégories : 1) les méthodes avec référence (FR-VQA, *Full-Reference Video Quality Assessment*), 2) les méthodes avec référence réduite (RR-VQA, *Reduced-Reference* VQA)et 3) les méthodes sans référence (NR-VQA, *No-Reference* VQA). Dans les deux premiers cas, on suppose que l'on a accès à une vidéo, ou une représentation parcellaire de la vidéo, dite de référence qui est réputée être exempte de toutes dégradation et être d'excellente qualité.

Dans notre contexte, étant donné que l'on traite la vidéo qui arrive directement d'un système d'enregistrement, nous ne disposons pas de la vidéo de référence. Qui plus est, parmi les méthodes d'évaluation de la qualité des vidéos sans référence, deux approches co-existent. Une première approche, qui est basée pixel, consiste à décoder la vidéo avant d'appliquer un schéma algorithmique de notation de la qualité. La seconde stratégie consiste à appliquer un processus algorithmique pour calculer la note de qualité sans décoder le bitstream, *i.e.*, dans le domaine compressé. Le processus d'évaluation est ainsi plus rapide, et permet une meilleure adaptation à une certaine gamme de distorsions vidéos [5]. Ces approches deviennent nécessairement dépendantes de la technologie utilisée (H.264/H.265 par exemple).

Dans la littérature, plusieurs modèles d'évaluation de la qualité des vidéos sans référence se bansant sur le bitstream ont été développés. Ces travaux se basent principalement sur la norme H.264/AVC [6, 7], et également sur la norme H.265/HEVC [9]. Les modèles présentés sont entraînés à partir de données subjectives, *i.e.*, le *Mean Opinion Score* (MOS), et/ou à partir de données objectives, *i.e.*, obtenues à l'aide de méthodes FR VQA. La prédiction de la qualité globale d'une vidéo esst alors calculée en fonction des caractéristiques extraites du bitstream. Ces caractéristiques sont généralement liées aux paramètres d'encodage [6, 9] mais aussi aux erreurs de transmission [7].

L'objectif étant d'évaluer de manière continue la qualité de vidéos provenant de caméras CCTV, une méthode d'évaluation de la qualité basée sur une approche sans référence dans le domaine compressée est appliquée. De plus, la norme H.264/AVC étant la plus utilisée, les travaux présentés se focalisent sur l'exploitation d'un bitstream H.264/AVC afin de déterminer la qualité de la vidéo courante.

2 La mesure de qualité continue basée bitstream

La figure 1 présente le fonctionnement du modèle proposé. Pour chaque GOP (*Group Of Pictures*) de la vidéo, un processus d'extraction de caractéristiques du bitstream est appliqué. Ensuite, une régression basée sur l'utilisation d'un perceptron multi-couche est réalisée afin d'obtenir la note finale du GOP. La compilation des notes de chaque GOP permet d'aboutir une notation continue de la qualité de la vidéo.

2.1 Extraction des caractéristiques vidéos

Dans la norme H.264/AVC, une *image* (frame) est composée d'une ou plusieurs *tranches d'image* (slices), qui sont elles-mêmes composées de plusieurs *macroblocs* (MB). La norme fournit ainsi plusieurs niveaux de granularité. Pour bénéficier de la précision d'analyse conférée par cette décomposition, les caractéristiques du bitstream sont extraites directement aux niveaux des slices, et non au niveau des frames. Pour caractériser l'influence des paramètres d'encodage, les caractéristiques utilisées dans l'approche proposée sont les suivantes : **(a)** le bitrate moyen, le *paramètre de quantification* (QP) et la variation du QP au sein des slices, reflétant le niveau global et local de compression, **(b)** la longueur moyenne et maximale des

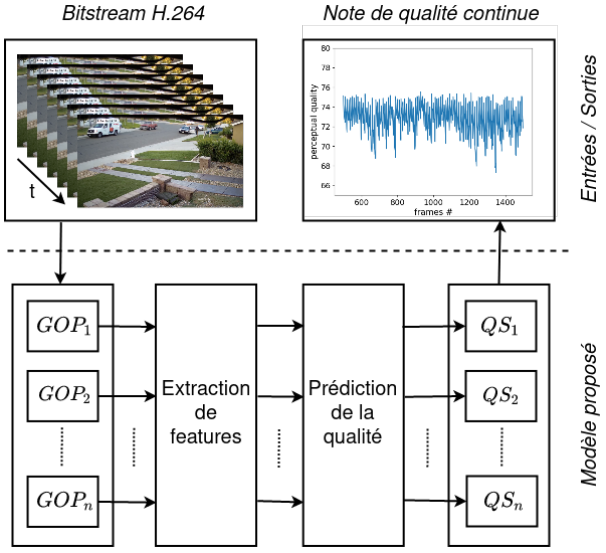


FIGURE 1 – Illustration du modèle proposé

vecteurs de mouvement (MV) et la longueur de l'erreur de prédiction sur ces vecteurs, quantifiant ainsi la présence de mouvement, **(c)** le type de partitionnement (en pourcentage) des MB et de leurs subdivisions, suivant les tables 7.11 à 7.18 de la recommandation ITU-T H.264 [4], représentant la nature de la prédiction des MB, et enfin **(d)** le pourcentage de macroblocs codés en mode intra, inter ou non codés (mode skip).

Au total, 27 caractéristiques f_i sont extraites sur chaque image I, P ou B de chaque GOP. Un vecteur de caractéristiques pour chaque GOP k, V^k_{GOP}, est finalement obtenu en moyennant ces résultats sur la taille N du GOP courant :

$$V^k_{GOP} = \left(\frac{1}{N} \sum_{l=1}^{N} f_i^l \right)^k \quad (1)$$

où f_i^l représente la caractéristique i de l'image l du GOP k.

Le décodeur H.264 de référence [1] a été modifié afin de collecter les caractéristiques précédemment citées.

2.2 Prédiction de la qualité

Une fois les différents vecteurs de caractéristiques extraits du bitstream, il reste à calculer une note finale de qualité pour chaque GOP.

Un réseau de neurones artificiel, à savoir *le peceptron multicouche* (MLP), est développé afin de prédire la qualité vidéo. Les paramètres du MLP sont établis empiriquement, cherchant le meilleur compromis entre complexité et performance. Ainsi, le MPL utilisé est composé de trois couches cachées, composées de 27 neurones chacune utilisant la fonction *unité linéaire rectifiée* (ReLU) en tant que fonction d'activation. La rétropropagation du gradient permet d'actualiser le poids des neurones afin de réduire la fonction de coût, ici *l'erreur quadratique moyenne* (MSE), entre les valeurs attendues et prédites. Pour éviter le phénomène de sur-apprentissage, une méthode d'arrêt automatique (early stopping) est utilisée lors de la phase d'apprentissage. La figure 2 présente l'architecture du MLP décrit ci-dessus.

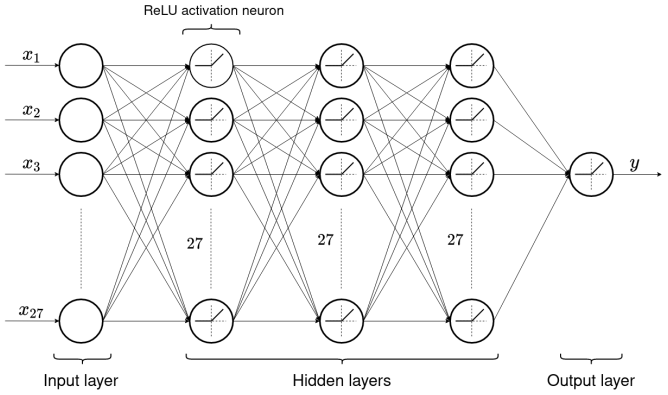


FIGURE 2 – Architecture du MLP développé

3 Construction d'une base de données et d'une vérité-terrain

L'évaluation des performances du modèle décrit dans la partie 2.2 requiert l'existence d'une base de données vidéo provenant de caméras CCTV avec, pour chacun des GOP de vidéo, une vérité-terrain associée, *i.e.*, une note de qualité. Cette dernière est souvent fournie par le MOS ou le *Differential MOS* (DMOS) obtenus grâce à l'évaluation d'observateurs humains.

A notre connaissance, aucune base de données vidéo provenant de caméras CCTV avec un MOS continue sur chaque vidéo n'existe à ce jour. Ainsi, 20 vidéos provenant de caméras CCTV, obtenues sur Internet, ont été sélectionnées. La localisation de la scène, la mobilité des caméras (fixe, rotative, possibilité de zoomer dans l'image) et la luminosité selon l'heure de la journée ont été trois critères de sélection afin de créer une base de données au contenu varié et tangible. Ces vidéos, de 30 secondes à 5 minutes, possèdent toute une résolution de 1280x720p. Pour constituer la base de données finale, chaque vidéo est encodée suivant 11 niveaux de compression par le logiciel x264 [3]. Ce niveau varie de $QP = 20$, équivalent à une faible niveau de compression, à $QP = 51$, équivalent à un fort niveau de compression. Une référence, avec un niveau de compression nul, *i.e.*, $QP = 0$, est aussi créée pour chacune des vidéos. La nouvelle base de données, nommée G-CCTV, est ainsi composée de 320 vidéos dégradées par l'algorithme de compression H.264 et de 20 références associées.

Au lieu de calculer la qualité subjective de chaque GOP de la base de données, nous avons choisi de quantifier le biais introduit par l'usage de métriques FR-VQA pour créer notre vérité-terrain. L'idée est de déterminer si les valeurs prédites, objectives et subjectives, sont statistiquement différenciables. Présentant de fort taux de corrélation avec le DMOS, la métrique *Video Multimethod Assessment Fusion* (VMAF) [2] est sélectionnée pour construire notre vérité-terrain. Développée et utilisée par Netflix, elle présente ainsi une corrélation (*Spearman Rank Order Correlation Coefficient* - SROCC) de 0.872 sur la base de données LIVE Mobile et de 0.953 sur la base NTFX-TEST [8]. Ces performances sont obtenues grâce à la fusion d'au moins deux méthodes avec référence *d'évaluation de la qualité d'image* (FR-IQA) et d'une fonction mesurant les différences temporelle de luminance (MCPD) avec une *machine à vecteurs de support*

t-test	VMAF	VIF	SSIM	MS-SSIM
DMOS	0	1	1	1

TABLE 1 – Résultats du test t entre le DMOS et les quatre métriques FR-VQA.

(SVM). L'idée sous-jacente est d'exploiter les forces de chacune des méthodes FR-IQA afin de fournir une métrique précise et performante. In fine, la qualité vidéo est donnée entre 0 et 100, de manière similaire au DMOS.

VMAF est donc utilisé pour calculer les *scores de qualité* (QS) sur chaque GOP de la base G-CCTV, constituant ainsi notre vérité-terrain. À titre de comparaison, trois autres métriques FR-IQA sont aussi utilisées : SSIM [14], MS-SSIM [15] et VIF [10].

4 Résultats

Avant de présenter les performances du modèle proposé, une étude est conduite entre les QS subjectifs (*i.e.*, le DMOS), et les QS objectives (*i.e.*, les scores de qualité calculés par les quatre métriques FR-I/VQA présentées en 3). Cette étude est basée sur le test t, ou test de Student, qui compare les moyennes de deux groupes d'échantillons. Il s'agit ainsi de savoir si ces moyennes sont significativement différentes d'un point de vue statistique. Ce test a été réalisé avec les données recueillies sur la base LIVE Wild Compressed Video Quality Database [16]. Cette base de données est composée de 55 vidéos tirées de la base LIVE-VQC [11, 13, 12]. Les vidéos ont ensuite été dégradées par l'algorithme de compression H.264 de quatre manières différentes pour former la base de donnée finale. Le DMOS est fourni pour chaque vidéo de la base.

Les résultats du test sont présentés Tab.1. Une valeur de 1 indique que les deux groupes sont statistiquement différentiables, une valeur de 0 indique qu'ils sont statistiquement indifférentiables.

Seul VMAF ne présente pas statistiquement de différence avec le DMOS. Ainsi, l'utilisation des QS générés par VMAF n'engendre pas de biais statistiquement significatif par rapport à l'usage du DMOS, contrairement aux autres métriques présentées. Le modèle proposé est donc entraîné à prédire des QS basés sur l'emploi de VMAF.

4.1 Performances du modèle

Afin d'évaluer correctement les performances du modèle, la base G-CCTV est divisée en deux sous-ensembles distincts : un réservé à la phase d'apprentissage et un à la phase de test. Le premier représente 70% de la base, *i.e.*, il est composé de vidéos issues de 14 vidéos de référence sélectionnées aléatoirement, et le deuxième est constitué des 30% restants, *i.e.*, il est composé de vidéos issues des 7 vidéos de référence restantes. La prédiction des QS, obtenue par le modèle, est répétée 1000 fois, avec pour chaque itération un nouveau découpage aléatoire de la base d'apprentissage et de test.

La figure 3 représente les performances du modèle en terme de corrélation entre les valeurs de QS attendues et les QS prédits par le modèle, avec leur intervalle de confiance de 99%. Cette corrélation est exprimée par le

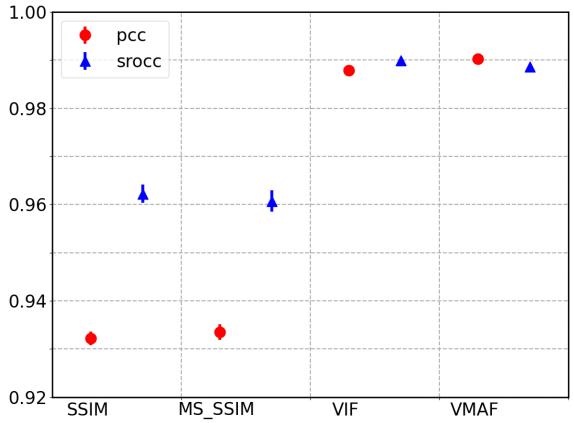


FIGURE 3 – Performances du modèle utilisant chacune des quatre méthodes FR-VQA en tant que vérité-terrain. Le PCC et le SROCC sont ici représentés avec leurs intervalles de confiance de 99%.

coefficient de corrélation linéaire, de *Pearson* (PCC) et le *coefficient de corrélation de rang*, de *Spearman* (SROCC) pour chacune des quatre vérités-terrains utilisées, à avoir VMAF, SSIM, MS-SSIM et VIF.

Le plus important taux de corrélation est obtenu lorsque VMAF est utilisé par le modèle afin de prédire les QS, avec un score de 0.99. Avec la métrique VIF, les taux de corrélation sont d'environ 0.98. VIF faisant partie intégrante de VMAF, il n'est donc pas surprenant d'obtenir des taux de corrélation comparables. La même remarque vaut aussi pour SSIM et MS-SSIM, où les taux de corrélation, même s'ils sont plus faibles, sont très similaires, allant jusqu'à 0.96.

Ces valeurs confirment ainsi nos précédents résultats, à savoir la pertinence de l'utilisation de VMAF à la place du DMOS, en tant que vérité-terrain.

4.2 Généralisation du modèle

Pour évaluer la capacité du modèle proposé en généralisation, une analyse sur son indépendance face à la base de données utilisée est conduite. Pour ce faire, la base LIVE VQC [11, 13, 12] est utilisée. Au vue des performances du modèle illustrées par la figure 3, VMAF est utilisé pour générer la vérité-terrain sur la nouvelle base de données LIVE VQC.

Les performances du modèle sont calculées **(1)** à partir du modèle entraîné sur toute la base G-CCTV puis testé sur toute la base LIVE VQC et, inversement, **(2)** à partir du modèle entraîné sur toute la base LIVE VQC puis testé sur toute la base G-CCTV. De la même manière que dans la partie 4.1, les prédictions sont réalisées 1000 fois par le modèle et les performances moyennes sont représentées par le PCC et le SROCC. Les performances du modèles dans ces deux cas sont ainsi présentées Tab.2. A ces résultats s'ajoutent ceux obtenus dans la partie 4.1 avec uniquement la base G-CCTV mais aussi, de la même manière, avec la base LIVE VQC.

Les taux de corrélation obtenus, que ce soit dans le cas **(1)** ou le cas **(2)**, restent consistants malgré l'utilisation de bases de données d'entraînement et de test très diffé-

	PCC		SROCC	
Test / Train	G-CCTV	LIVE-VQC	G-CCTV	LIVE-VQC
G-CCTV	0.991	**0.957**	0.988	**0.948**
LIVE-VQC	**0.948**	0.979	**0.938**	0.969

TABLE 2 – Performances du modèle utilisant différentes bases de données, avec une vérité-terrain générée par VMAF.

rentes. Les taux oscillent autour de 0.95 et sont très similaires dans les deux cas, avec des résultats légèrement supérieurs lorsque le modèle est entraîné sur la base LIVE VQC. Cette base étant plus riche en type de contenu, il n'est pas étonnant d'observer ce phénomène. Ces résultats démontrent que la base de données utilisée n'influe que très peu sur les performances du système. Ainsi, le modèle développé peut être utilisé avec n'importe quel jeu de données vidéo tout en conservant des performances très élevées.

5 Conclusion et perspectives

Un modèle d'évaluation continue de la qualité vidéo, sans-référence, a été présenté. En utilisant un nombre réduit de caractéristiques, directement extraites du bitstream vidéo, le modèle développé est capable de prédire la qualité d'une vidéo de manière précise et rapide.

Les caractéristiques du bitstream, relatives aux distortions induites par le processus de compression, sont combinées par un algorithme de machine learning dans le but d'estimer la qualité d'une vidéo. La vérité-terrain utilisée a été généré par VMAF, une méthode d'évaluation de la qualité vidéo avec référence, qui a montré de très fortes similarités statistiques avec l'appréciation humaine.

Partant d'un contexte judiciaire, où les vidéos d'entrées sont des vidéos provenants de caméras de vidéo-surveillance, la généralisation du modèle vers la prédiction de la qualité pour tout type de contenu vidéo a été clairement démontrée. De plus, le modèle propose des temps d'exécution au moins trois fois plus rapides que le temps réel pour des performances compétitives avec les méthodes reposant sur l'approche basée pixel.

Références

[1] H.264/AVC reference software decoder. `http://iphome.hhi.de/suehring/tml/`. Version : JM 19.0.

[2] VMAF - Video Multi-Method Assessment Fusion. `https://github.com/Netflix/vmaf/blob/master/resource/doc/references.md`. Version : VMAF model v0.6.1.

[3] x264 software - Encoding video streams into the H.264/AVC compression format. `https://www.videolan.org/developers/x264.html`. Version : x264 core 155 r2917.

[4] Advanced video coding for generic audiovisual services. Recommandation ITU-T H.264, International Telecommunication Union, May 2003.

[5] Christian Keimel. *Design of Video Quality Metrics with Multi-Way Data Analysis : A data driven approach*. Springer Singapore.

[6] Christian Keimel, Manuel Klimpke, Julian Habigt, and Klaus Diepold. No-reference video quality metric for hdtv based on h.264/avc bitstream features. In *2011 18th IEEE International Conference on Image Processing*, pages 3325–3328, 2011.

[7] Katerina Pandremmenou, Muhammad Shahid, Lisimachos Kondi, and Benny Lövström. A no-reference bitstream-based perceptual model for video quality estimation of videos affected by coding artifacts and packet losses. volume 9394, 02 2015.

[8] Reza Rassool. Vmaf reproducibility : Validating a perceptual practical video quality metric. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–2, 2017.

[9] Muhammad Shahid, Joanna Panasiuk, Glenn Van Wallendael, Marcus Barkowsky, and Benny Lövström. Predicting full-reference video quality measures using hevc bitstream-based no-reference features. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–2, 2015.

[10] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2) :430–444, 2006.

[11] Zeina Sinno and Alan C. Bovik. Large scale subjective video quality study. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 276–280, 2018.

[12] Zeina Sinno and Alan C. Bovik. LIVE Video Quality Challenge Database. `http://live.ece.utexas.edu/research/LIVEVQC/index.html`, 2018.

[13] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2) :612–627, 2019.

[14] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4) :600–612, 2004.

[15] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[16] X. Yu, N. Birkbeck, Y. Wang, C. G. Bampis, B. Adsumilli, and A. C. Bovik. Predicting the quality of compressed videos with pre-existing distortions. In *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, volume 30, page 7511–7526, 2021.

Compression sans perte de données GNSS au format RINEX dans un contexte applicatif de véhicules autonomes

Arnaud SOULIER, Pauline PUTEAUX, Frédéric COMBY, William PUECH
LIRMM, Univ. Montpellier, CNRS – Montpellier, France

Résumé : *L'essor des technologies de positionnement et d'automatisation a permis le développement de véhicules entièrement autonomes (train, bus, voiture, etc.). Il est indispensable pour ces véhicules de communiquer avec les systèmes de positionnement le plus rapidement possible ce qui engendre des échanges de données parfois volumineux. Dans cet article, nous présentons une nouvelle méthode de compression sans perte de données GNSS au format RINEX. Pour cela nous exploitons au maximum le caractère prédictif des données en utilisant les erreurs de prédiction par un modèle polynomial. Nos résultats expérimentaux montrent un taux de compression median de 7,86, contre 7,31 pour les méthodes de l'état de l'art.*

Mots-clés : Données satellites, RINEX, Compression sans perte, Modèle polynomial, Véhicules autonomes.

1 Introduction

Les véhicules autonomes représentent aujourd'hui l'avenir technologique du transport. Pour assurer leur bon fonctionnement, ils ont besoin de se repérer dans l'espace tout au long d'un trajet. Ils utilisent pour cela un dispositif de géolocalisation et navigation par un système de satellites (GNSS), comme le GPS. Les données transmises par ce type de système peuvent s'avérer très volumineuses. Afin d'améliorer l'efficacité de leur transmission et de permettre un fonctionnement en temps réel, il est intéressant de compresser au préalable les données.

Un des formats utilisés dans le stockage des données GNSS pour l'estimation de la position est le RINEX (*Receiver Independent Exchange Format* [4]). Un satellite peut fournir la phase du signal transmis, ainsi que sa distance avec le récepteur. En utilisant ces informations avec au moins trois satellites, il est alors possible de déterminer, par triangulation, la position du récepteur à un instant t. Chaque satellite supplémentaire permet d'augmenter la précision de la position. Pour le GPS, quatre satellites sont constamment visibles simultanément. Le fichier RINEX stocke ces informations pour N instants consécutifs.

Hatanaka a proposé en 2008 une méthode de compression des fichiers au format RINEX se basant sur une opération de différences entre les valeurs successives [2] (*cf.* figure 1). Cette opération permet de réduire les valeurs pour obtenir des nombres plus petits codés sur moins d'espace mémoire. Il applique également une élimination de la redondance des données, préparant ainsi le fichier pour une compression par codage entropique comme Gzip. Cette différence s'applique sur plusieurs rangs (*order* dans la figure 1). Pour chaque rang, la différence entre la valeur du rang inférieur Y_i^{n-1} et la valeur précédente associée Y_{i-1}^{n-1} (flèches rouges dans la figure 1) est calculée. Il faut donc au préalable avoir traité les quatre premières valeurs pour calculer le troisième rang des différences. Dans son

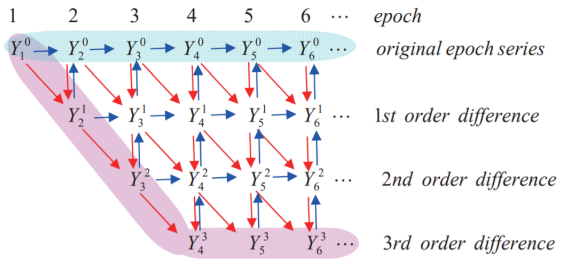


FIGURE 1 – Méthode des différences d'Hatanaka [2].

article, Hatanaka précise que plus de trois rangs ne permettent pas de réduire davantage la taille des valeurs. Le décodage des valeurs originales se fait en additionnant la valeur du rang précédent Y_i^{n-1} et la valeur précédente sur le même rang Y_{i-1}^n (flèches bleues dans la figure 1). Les équations 1 et 2 présentent respectivement les opérations nécessaires pour la compression et la décompression :

$$\begin{cases} Y_i^1 = Y_i^0 - Y_{i-1}^0, \\ \quad \dots \\ Y_i^n = Y_i^{n-1} - Y_{i-1}^{n-1}, \end{cases} \tag{1}$$

$$\begin{cases} Y_i^{n-1} = Y_{i-1}^{n-1} + Y_i^n, \\ \quad \dots \\ Y_i^0 = Y_{i-1}^0 + Y_i^1. \end{cases} \tag{2}$$

Il existe à ce jour peu de méthodes pour la compression des fichiers au format RINEX. De plus, la compression proposée par Hatanaka [2] n'exploite pas au maximum le caractère prédictif des données.

2 Travail proposé

Nous avons observé que l'évolution au cours du temps des données stockées dans le format RINEX, notamment celle des distances aux satellites, peut être modélisée par des paraboles (*cf.* figure 2).

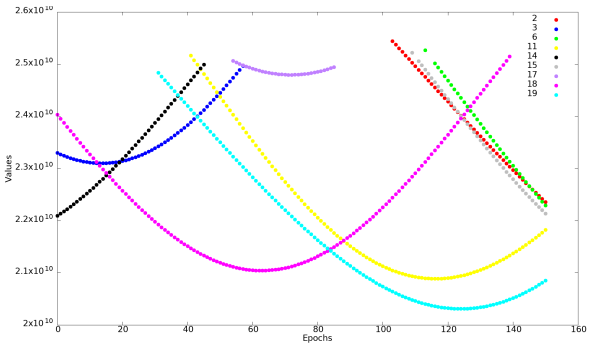


FIGURE 2 – Exemple de signaux (distances) dans un fichier RINEX (chaque satellite est associé à une couleur).

Nous proposons une nouvelle méthode de compression des fichiers au format RINEX s'appuyant sur une interpolation polynomiale de degré 2 des courbes des données

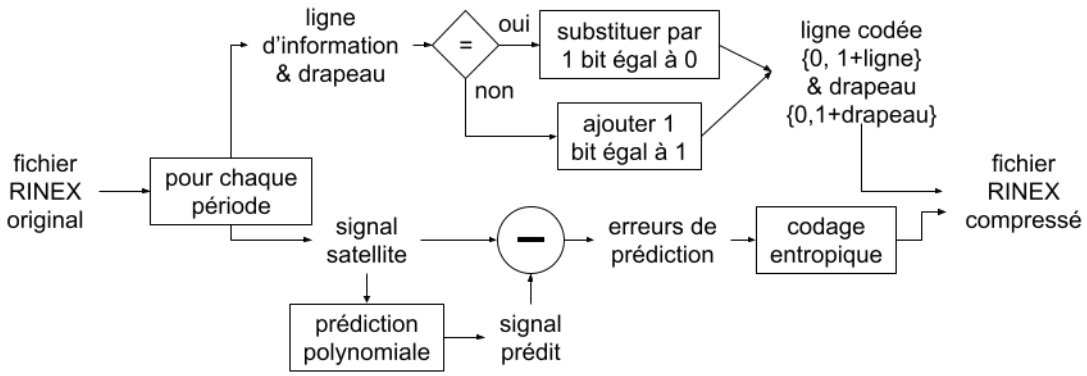


FIGURE 3 – Vue d'ensemble de la méthode proposée pour la compression de fichier RINEX basée sur une prédiction du signal par interpolation polynomiale de degré 2.

des satellites. Nous commençons par interpoler chaque signal (phase ou distance) de chaque satellite pour pouvoir prédire les données des signaux. Nous calculons alors les erreurs de prédiction. Par ailleurs, un codage de Huffman avec la même table que celle utilisée dans la méthode de compression JPEG [3] est utilisé pour compresser les erreurs de prédiction. Nous appliquons également une élimination de la redondance dans les données d'information de chaque acquisition. Une vue d'ensemble de la méthode proposée est présentée dans la figure 3.

2.1 Interpolation polynomiale de degré 2

Dans une première approche, nous utilisons trois valeurs consécutives afin de construire un modèle par interpolation polynomiale et prédire la valeur suivante. Nous considérons l'équation polynomiale :

$$p_{2\_t}(t) = a_t \cdot t^2 + b_t \cdot t + c_t, \tag{3}$$

avec $p_{2\_t}(t)$ la prédiction de la valeur au temps t et a_t, b_t et c_t les coefficients polynomiaux.

Les résultats de cette interpolation sont illustrés par la figure 4. Chaque courbe représente le signal d'un satellite.

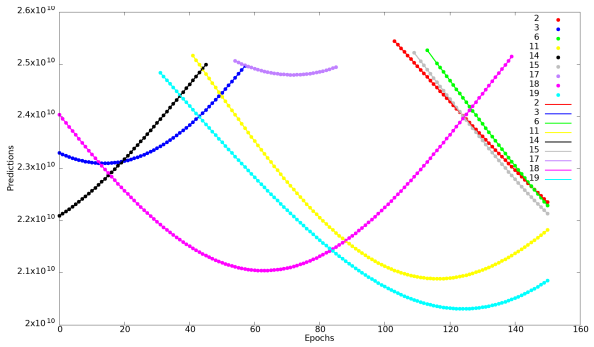


FIGURE 4 – Interpolation polynomiale de degré 2 des signaux d'un fichier RINEX en utilisant trois valeurs.

Les points représentent les valeurs réelles et les courbes continues correspondent aux interpolations obtenues. Le R^2 score moyen de ce modèle est de 0,9999 ce qui confirme la validité du modèle. Une fois notre modèle interpolé, nous calculons les erreurs de prédiction ϵ_t entre les valeurs originales des signaux satellites et les valeurs interpolées :

$$\epsilon_t = p_{2\_t}(t) - y_t, \tag{4}$$

avec ϵ_t l'erreur de prédiction, $p_{2\_t}(t)$ la prédiction du modèle et y_t la valeur observée du signal satellite au temps t.

Ces erreurs de prédiction représentent un échantillonnage de valeurs réduites (*cf.* figure 5). Les données originales

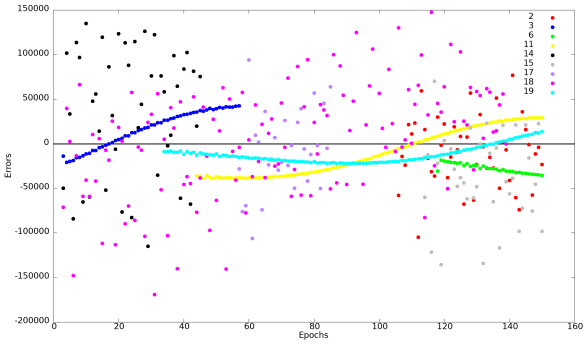


FIGURE 5 – Erreurs de prédiction du modèle polynomial de degré 2 interpolé en utilisant trois valeurs.

des signaux satellites sont de l'ordre de 10^{10}. Les erreurs de prédiction de notre modèle interpolé sont de l'ordre de 10^5. Grâce à ces erreurs de prédiction nous pouvons désormais obtenir de meilleurs résultats avec un codage entropique. Nous avons également expérimenté une interpolation polynomiale de degré 2 en utilisant plus de trois valeurs précédentes. Les résultats de l'interpolation utilisant toutes les valeurs précédentes sont illustrés dans la figure 6 et les erreurs de prédiction sont données sur la figure 7. D'après nos observations, plus le nombre de va-

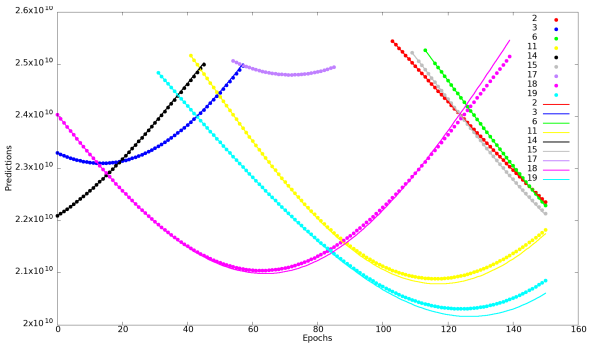


FIGURE 6 – Interpolation polynomiale de degré 2 des signaux d'un fichier RINEX en utilisant toutes les valeurs.

leurs utilisées pour interpoler le signal est élevé, plus la précision de la prédiction diminue.

2.2 Codage entropique

Nous représentons alors chaque erreur de prédiction calculée par un couple en-tête/valeur $(head_t, value_t)$, corres-

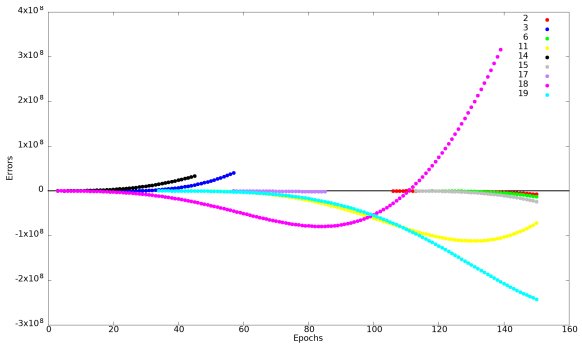


FIGURE 7 – Erreurs de prédiction du modèle polynomial de degré 2 interpolé en utilisant toutes les valeurs.

pondant à sa version compressée. La partie en-tête $head_t$ indique le nombre minimum de bits requis pour coder la valeur de l'erreur de prédiction (*cf.* table 1).

$head_t$	Plage de valeurs
1	[-1] & [1]
2	[-3 ; -2] & [2 ; 3]
...	...
8	[-255 ; -128] & [128 ; 255]
...	...

TABLE 1 – Table de Huffman utilisée pour le codage de l'erreur de prédiction [3].

La partie valeur $value_t$ correspond à l'index de l'erreur de prédiction dans la plage de valeurs correspondante à son en-tête $head_t$. Elle est donnée par la formule :

$$value_t = \begin{cases} 2^{head_t} - 1 - \mid \epsilon_t \mid & \text{si } \epsilon_t < 0, \\ \epsilon_t & \text{si } \epsilon_t > 0. \end{cases} \quad (5)$$

Une fois le couple en-tête/valeur construit, il est écrit à la suite du fichier compressé. Le nombre de bits requis pour écrire l'en-tête $head_t$ dépend de la valeur maximale des $head_t$. Cette valeur est stockée sur 1 bit après les métadonnées du fichier RINEX. La valeur $value_t$ est écrite sur un nombre de bits égal à la valeur de $head_t$.

2.3 Suppression de la redondance

Les premières lignes de chaque acquisition stockée dans un fichier RINEX contiennent des données redondantes comme la date et l'heure de l'acquisition de l'epoch. Ces données peuvent être retrouvées à partir de la première date stockée et de l'intervalle de temps entre deux acquisitions présent dans l'en-tête du fichier. Nous avons donc conservé cette première information en la recopiant sous forme binaire pour réduire sa taille et supprimé toutes les suivantes.

La première ligne d'une acquisition contient également des informations sur les satellites présents pour cette acquisition (nombre de satellites et identifiant de chaque satellite). Ces informations sont susceptibles de ne pas changer d'une acquisition à l'autre. Nous avons donc utilisé un bit pour indiquer si elles sont identiques à l'acquisition précédente. Si ces informations sont exactement les mêmes que précédemment, alors le bit est mis à 0 et rien d'autre n'est écrit. Dans le cas contraire, si elles ont été modifiées, le bit est mis à 1 et les nouvelles informations sont écrites.

3 Résultats

D'après nos expérimentations, la méthode d'interpolation polynomiale utilisant les trois valeurs précédentes est effectivement la plus efficace et permet d'obtenir un modèle plus proche des données (*cf.* figure 4 et figure 5). Même si l'écart entre les valeurs observées et leurs versions interpolées en utilisant toutes les valeurs précédentes est globalement faible (figure 6), certaines erreurs de prédiction nécessitent d'être codées sur un grand nombre de bits, comme illustré en figure 7.

Dans le tableau 2, nous comparons les taux de compression obtenus avec la méthode proposée en utilisant différents paramètres pour l'interpolation avec ceux des approches de l'état de l'art. Nos expérimentations ont été effectuées sur une base de 151 fichiers RINEX de taille comprise entre 58,3 Ko à 1,9 Mo, avec une taille moyenne de 1,5 Mo et une taille médiane de 1,7 Mo.

Méthode		Min	Max	Moyenne	Médiane
Hatanaka [2]		2,53	3,36	3,01	3,07
Hatanaka [2] + Gzip [1]		5,44	8,56	7,31	7,46
Degré 2	**3 valeurs**	**5, 82**	**27, 33**	**8, 29**	**7, 86**
	4 valeurs	5,68	27,07	8,09	7,70
	5 valeurs	5,55	26,68	7,94	7,57
	Toutes les valeurs	3,30	17,82	4,85	4,57

TABLE 2 – Comparaison des taux de compression obtenus avec notre méthode (en utilisant différents paramètres pour l'interpolation) avec ceux de l'état de l'art.

Le taux de compression médian obtenu avec notre méthode basée sur une interpolation polynomiale quadratique en utilisant les 3 valeurs précédentes est de 7,86 contre 7,46 avec la méthode proposée par Hatanaka [2] combinée avec Gzip [1]. Notons également que la méthode proposée permet d'obtenir un taux de compression bien plus élevé pour certains fichiers.

4 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode efficace de compression sans perte de fichiers au format RINEX. Nous avons montré que nos résultats étaient meilleurs que ceux de l'état de l'art. En utilisant les trois valeurs précédentes pour interpoler le modèle, nos courbes des différences peuvent être approchées par des équations polynomiales. Dans de futurs travaux, nous envisageons d'utiliser un modèle sur les erreurs de prédiction dans la perspective d'améliorer notre taux de compression.

Références

[1] J.-L. Gailly. Gnu gzip – the data compression program for Gzip version 1.9. Technical report, Free Software Foundation, 2018.

[2] Y. Hatanaka. A compression format and tools for GNSS observation data. *Bulletin of the Geographical Survey Institute*, 55, 2008.

[3] G. K. Wallace. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1) :18–34, 1992.

[4] G. Werner and E. Lou. RINEX : The receiver independent exchange format version 2.11. Technical report, Astronomical Institute, University of Bern, 2007.

Deep Video Capsule Network avec Décalage Temporel pour la Reconnaissance d'Action

Théo Voillemin[1], Hazem Wannous[1], Jean-Philippe Vandeborre[2]
[1] Université de Lille - CNRS, IMT Lille Douai - UMR 9189 - CRIStAL - Lille
[2] IMT Lille Douai - Université de Lille, CNRS - UMR 9189 - CRIStAL - Lille

Résumé : *La reconnaissance d'action en temps réel est un domaine en plein essor ces dernières années. L'apprentissage profond et les Réseaux de Neurones Convolutifs (CNNs) permettent l'obtention de bons résultats même si des limitations intrinsèques aux CNNs plafonnent les performances étant donné que les CNNs 2D ne peuvent pas capturer l'information temporelle et que les CNNs 3D [4] demandent trop de ressources. Les Réseaux de Capsules, évolution des CNNs, ont déjà prouvé des améliorations sur de petits jeux de données, tant en taille qu'en informations à traiter comme la base de chiffres manuscrits MNIST. Leur vrai potentiel n'a pas encore été prouvé même si les récents Réseaux de Capsules Profond ont montré des résultats prometteurs. Dans ce travail, nous proposons d'explorer plus en profondeur les réseaux de capsules pour le problème de reconnaissance d'actions et de gestes à partir des flux vidéos en intégrant un module de décalage temporel. Avec cette nouvelle architecture, que nous appelons Deep Video Capsule Network, nous réussissons à intégrer de l'information temporelle aux capsules 2D pour un coût de calcul nul et ainsi conserver la légèreté des capsules ainsi que leur avantage à capturer des informations spatiales plus complexes qu'un CNN classique. Notre méthode surpasse ou se rapproche de l'état de l'art sur des bases de reconnaissance de gestes et d'actions tout en ayant 10 à 40 fois moins de paramètres.*

Mots-clés : Reconnaissance d'action, apprentissage profond, réseau de capsule

1 Introduction

Les méthodes d'apprentissage profond sont devenues l'une des techniques les plus performantes pour la compréhension d'images et de vidéos, particulièrement depuis le développement des réseaux neuronaux convolutifs [5]. Ces méthodes ont aussi prouvé leur avantage à être déployées en temps réel sur des applications, tout du moins pour de relativement petits réseaux, puisque seule la phase d'apprentissage est principalement longue et coûteuse, l'inférence quant à elle consiste simplement en une suite d'opérations et de calculs. Cependant, l'étude d'une vidéo, ainsi que sa dimension temporelle en plus à prendre en compte, complexifie le problème. En effet, si les architectures de convolutions 3D ont le bénéfice de prendre en compte les informations spatiales et temporelles, elles sont cependant généralement trop lourde pour être appliquées à des cas d'utilisation en temps réel surtout avec une puissance de calcul raisonnable. Inversement, les CNNs 2D, grâce notamment au partage des poids propre à leur architecture, demande peu de puissance de calcul, sont parfait pour un entraînement rapide et de l'inférence en temps réel, mais leur implémentation basique ne permet pas l'extraction et l'analyse d'information temporelle puisqu'ils ne prennent qu'une unique image en entrée. L'arrivée des technologies de réalité augmentée et leurs appareils autonomes demande des solutions de plus en plus performantes mais aussi de plus en plus légères pour pouvoir être déployée et utilisée sur ces appareils.

Pour répondre à ces défis, deux travaux nous ont servi de base et de point de départ. Le premier est le réseau neuronal à capsule développé par Sabour *et al.* [7] qui propose une nouvelle architecture aux résultats comparables à ceux d'un CNN 2D mais avec un nombre de paramètres à entraîner bien moindre, la contrepartie étant un temps d'entraînement plus long à cause de l'algorithme de routing mais qui n'a aucune incidence au moment de l'inférence. Etant donné que cette nouvelle architecture, à l'instar du CNN 2D, n'extrait que des informations spatiales, nous nous sommes intéressés aux travaux de Li *et al.* [6] et de leur module de décalage temporel qui permet, au sein d'un CNN 2D, d'apporter du traitement d'informations temporelles en décalant les résultats des opérations d'une couche de convolution appliquées à une image de la vidéo vers la même couche mais au moment du traitement de l'image suivante et/ou précédente, le tout sans incidence sur le nombre de paramètres à gérer et donc sur la puissance de calcul nécessaire.

2 Travail proposé

Nous proposons ainsi une nouvelle architecture, 2D Deep Video Caspule Network (voir Figure 1), pour le traitement de vidéos et plus particulièrement pour la reconnaissance d'actions et de gestes. A l'instar du module de décalage temporel sur CNN 2D [6], nous avons implémenté un décalage temporel sur capsule pour ajouter du traitement d'information temporelle aux capsules.

Ce module permet aux capsules traitant la trame n de la vidéo d'acquérir une partie de l'information des capsules ayant traitées les trames $n-1$ et $n+1$. Nous avons implémenté et testé trois différentes configurations (voir Figure 2), une première où seuls les premiers noyaux des premières capsules d'une couche de capsule sont décalés, dans cette implémentation, une capsule de la couche sur laquelle est appliquée le module est considérée de la même manière qu'un unique filtre de convolution. La seconde implémentation décale les premiers noyaux de toutes les capsules de la couche sont décalés, ici, chaque capsule est considérée comme une couche de convolution à part entière ainsi toutes les capsules subissent donc un décalage

de leur premiers filtres. La dernière implémentation décale l'entièreté des premières capsules de la couche, c'est l'implémentation qui se rapproche le plus naïvement de celle appliqué aux couches de convolution classiques.

Afin de ne pas perdre l'information de la trame étudiée en cours au moment du décalage, ce dernier est implémenté au sein d'une branche résiduelle dans ce que nous avons appelé une ShiftConvCapsule où sont effectués les décalages puis l'opération de convolution au sein des capsules avant d'être additionné avec l'entrée de la capsule contenant l'information originale. 2D Deep Video Capsule Network (voir Figure 1) est une architecture constituée de deux premières couches de convolutions pour extraire les caractéristiques spatiales primaires, suivi d'un réseau de capsules profonds où sont implémentés les décalages temporels avant d'aboutir sur une couche de classification et d'un réseau de reconstruction des images d'entrée pour permettre une régularisation durant l'entraînement des capsules. Les fonctions de pertes utilisées sont les mêmes que pour l'architecture de réseau neuronal à capsule traditionelle [7], à savoir la margin loss pour la partie encodeur et classification du réseau :

$$L_k = T_k max(0, m^+ - ||v_k||)^2 + (1 - T_k) max(0, ||v_k|| - m^-)^2 \quad (1)$$

avec les valeurs habituelles de 0.9 pour m^+ et 0.1 pour m^- et de 0.5 pour α. L'erreur quadratique est utilisée pour la reconstruction.

Etant donné que l'architecture est composée principalement d'opérations de convolution, elle accepte en entrée des images, que ce soit en couleur, noir et blanc ou encore de profondeur. Cependant, puisque le réseau est orienté vers la compréhension de vidéos, ces dernières doivent être échantillonée puis envoyée image par image à l'architecture. Celle-ci analyse les images indépendamments les unes des autres et c'est au moment des applications du module de décalage temporel que les résultats des opérations de convolution appliquées à chaque image sont récupérés pour pouvoir être ensuite décalés.

3 Résultats

La première base testée est First Person Hand Action Dataset (FPHA) [3]. Il s'agit d'une base de données d'actions effectuées à la première personne dans un cadre de manipulation de différents objets avec les mains dans le contexte d'une cuisine. La base contient 1 175 séquences réparties sur 45 classes où le sujet manipule un des 28 objets via des fluxes vidéo couleur et de profondeur ainsi que de séries de positions du squelette de la main effectuant l'action. Dans notre cas nous n'utilisons que la modalité couleur pour entraîner notre architecture. Nous observons ainsi que nous obtenons une meilleure précision de classification que toutes les autres méthodes de l'état de l'art qui utilisent aussi le flux vidéo couleur ou de profondeur, le tout, avec 6 à 45 fois moins de paramètres à gérer et à entraîner pour le réseau (voir Table 1). Nous avons aussi appliqué la méthode TSM sur la base FPHA afin de comparer le bénéfice de notre développement du module de décalage temporel sur capsule plutôt que sur convolution simple, bénéfice prouvé par les 5% supplémentaires de bonnes classifications. Nous pouvons observer que, même si l'amélioration n'est pas aussi notable, de l'ordre de 1%,

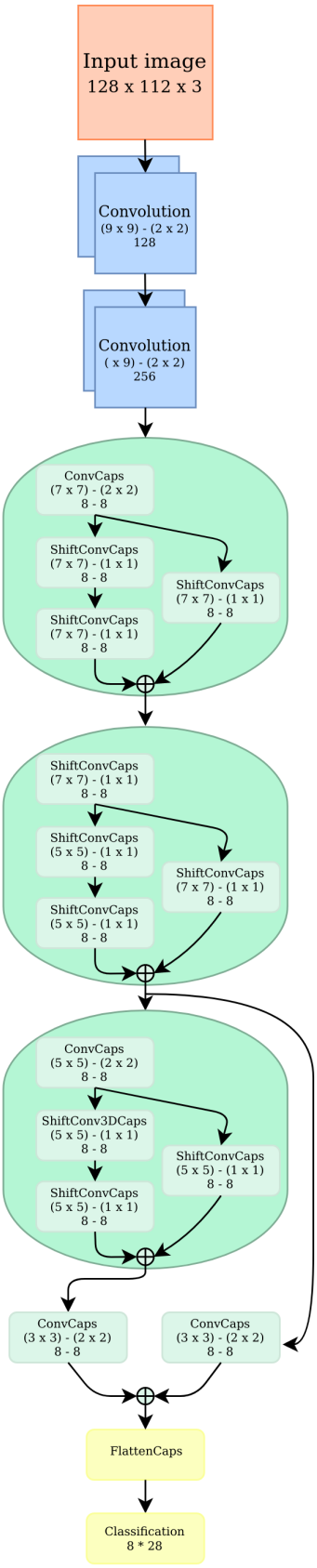
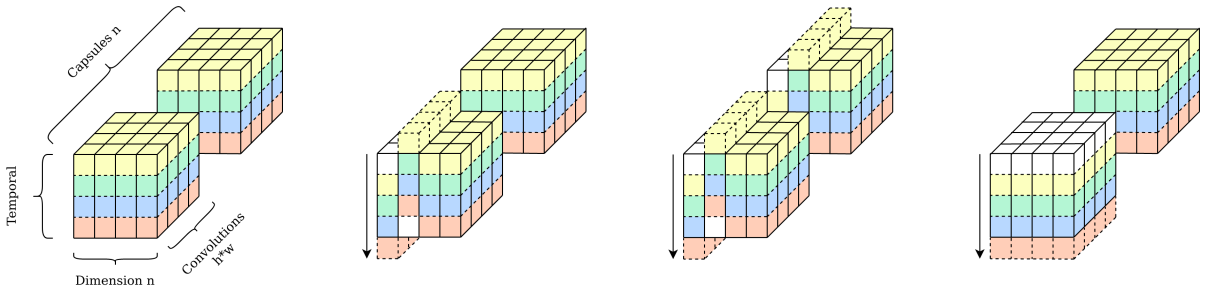


FIGURE 1 – Notre architecture 2D Deep Video Capsule Network



(a) Original convolution capsule layer

(b) Offline shift on first kernels of the first capsule

(c) Offline shift on first kernels of every capsules

(d) Online shift on every kernels of the first capsules

FIGURE 2 – Implémentation de notre module de décalage temporel appliqué sur des couches de capsules. Une tranche horizontal du cube principal représente une capsule, ce même cube représentant la même capsule à des temporalités différentes. Les deux cubes de chaque implémentation représentent la première et dernière ShiftConvCapsule d'une couche.

face à la méthode de Feichtenhofer *et al.* [2] qui propose de combiner le résultat de deux CNN 2D parallèles s'exécutant sur deux modalités différentes, l'une de couleur et l'autre de profondeur, non seulement notre architecture n'utilise qu'une seule de ces deux modalités, la couleur, mais en plus, c'est ici que la différence du nombre de paramètres est la plus notable avec notre architecture demandant 45 fois moins de paramètres à entraîner.

Méthode	Nb paramètres	Précision (%)
Two stream-color [2]	46M	61.56
Two stream-flow [2]	46M	69.91
Two stream-all [2]	181M	75.30
TSM [6]	24M	71.57
2D CVNN	**4M**	**76.72**

TABLE 1 – Comparaison de 2D DVCN sur la base FPHa [3] comparé à l'état de l'art

La deuxième base de données testée est Dynamic Hand Gesture (DHG) [1]. Il s'agit cette fois-ci d'une base de gestes de la main en vue à la troisième personne. Nous avons notamment choisi cette base de données pour démontrer que notre méthode n'est pas seulement efficace dans un unique contexte de reconnaissance d'action à la première personne et que le fait qu'elle accepte tout type d'image en entrée la rend portable et adaptable à bon nombre de problématiques. Cette base contient 2800 séquences réparties sur 28 classes. Parmi ces 28 classes, on retrouve 14 gestes distincts chacun effectué avec un seul doigt et le reste de la main fermée ou effectué avec la main complétement ouverte. Deux modalités sont fournies pour chacune des séquences, soit un flux vidéo de profondeur que nous utiliserons pour notre réseau, soit une séquence de positions du squelette de la main effectuant le geste. De la même manière que pour FPHA, nous avons comparé notre méthode aux autres de l'état de l'art. Ainsi, sur la base DHG, si nous ne battons pas l'état de l'art, à savoir la méthode de Abadi *et al.* [**?**] qui propose cette fois-ci de fusionner le résultat de deux CNN, l'un 2D et l'autre 3D, nous nous approchons tout de même de leur résultat de classification avec une architecture proposant 20 fois moins de paramètres à entraîner. Nous avons aussi, une fois de plus, tester le module TSM sur CNN 2D sur cette base et montrons, de la même manière que pour FPHA, que l'application du module de décalage temporel appliqué sur des capsules plutôt que sur de simples convolutions offre des bénéfices avec ici un apport de l'ordre de 10% de bonnes classifications supplémentaires (voir Table 2).

Méthode	Nb pramètres	Précision (%)
TSM [6]	24M	58.66
2D-3DCNN Fusion [8]	140M	**74.41**
2D CVNN	**7M**	**68.98**

TABLE 2 – Comparaison de 2D DVCN sur la base DHG28 [1] comparé à l'état de l'art

4 Conclusion et perspectives

Nous proposons une nouvelle architecture pour la compréhension de vidéos et pour la reconnaissance de gestes ou d'actions de la main. Le réseau 2D Deep Video Capsule Network avec module de décalage temporel peut, pour la première fois pour une architecture de réseau neuronal à capsule, à la fois d'analyser des vidéos à l'aide uniquement d'opérations de convolution 2D, mais aussi traiter des données de grande dimensionnalité sans explosion du gradient. Nous prouvons ainsi les bénéfices de notre méthode sur deux bases de données en montrant que notre architecture 2D DVCN s'approche, voire surpasse l'état de l'art en proposant un réseau contenant 10 à 40 fois moins de paramètres à gérer.

Références

[1] Quentin De Smedt, Hazem Wannous, and Jean-Philippe Vandeborre. Skeleton-based dynamic hand gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2016.

[2] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[3] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

[4] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1) :221–231, 2012.

[5] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[6] Ji Lin, Chuang Gan, and Song Han. Tsm : Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019.

[7] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017.

[8] Erhu Zhang, Botao Xue, Fangzhou Cao, Jinghong Duan, Guangfeng Lin, and Yifei Lei. Fusion of 2d cnn and 3d densenet for dynamic gesture recognition. *Electronics*, 8(12) :1511, 2019.

Liste des auteurs