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

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

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

Résumé

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

Mots clefs

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

1 Introduction

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

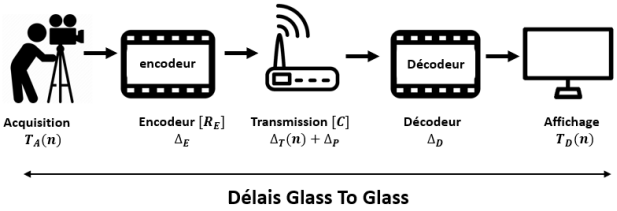


FIGURE 1 – *Latence G2G dans un schéma de transmission vidéo point à point*

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

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

2 Méthodologie

2.1 Modèle d'estimation du délai G2G

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

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

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

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

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

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

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

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

2.2 Méthode de référence : Réduction du débit de codage

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

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

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

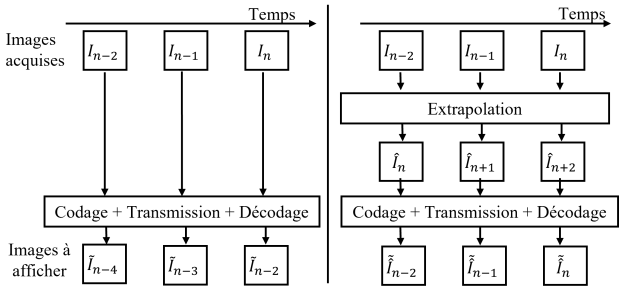


FIGURE 2 – *Illustration de la transmission vidéo sans extrapolation et avec extrapolation. $\hat{I}$ et $\tilde{I}$ indiquent respectivement l'image extrapolée et décodée.*

2.3 Méthode analysée : Extrapolation

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

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

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

3 Configuration de la simulation

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

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

4 Résultats des simulations

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

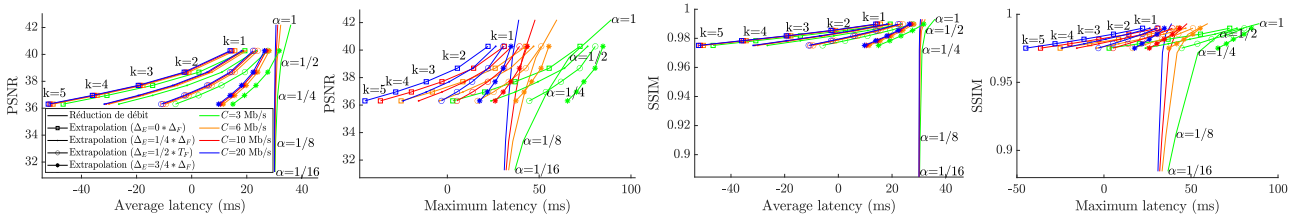
1. encoder_lowdelay_P_vtm.cfg
2. https://drive.google.com/drive/folders/12UJ_Il7yicPklxq1ymioKLj5PgcD06Dz?usp=share_link

5 Discussion

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

6 Conclusion et perspectives

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



(a) PSNR de la séquence *Four people*

(b) SSIM de la séquence *Four people*

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

(a) Image originale

(b) latence moyenne = 35.6 ms

(c) latence moyenne = 37.3 ms

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

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

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

Références

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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