Nouvelle représentation de données pour les applications interactives de navigation vidéo

Thomas MAUGEY, Pascal FROSSARD

Ecole Polytechnique Fédérale de Lausanne (EPFL), LTS4
Lausanne, Suisse
thomas.maugey@epfl.ch, pascal.frossard@epfl.ch

Résumé – Dans cet article, nous nous intéressons aux schémas de transmission multi-vues dans lesquels l'utilisateur peut, en temps réel, choisir le point de vue depuis lequel il observe la scène 3D. Cette application complexe requière de nouvelles approches dans les méthodes de représentation et codage des données 3D. L'idée fondamentale de ces travaux est de généraliser les représentations basées image, en regroupant un ensemble de vues dans un même signal. Cela permet d'abord de réduire les redondances à l'étape de la représentation, et non au codage comme il est généralement fait dans la littérature. Ensuite, cela permet d'avoir un contrôle plus souple de ce qui est transmis aux groupes d'utilisateurs en fonction de leur chemin de navigation. Enfin, cela ouvre de nouvelles perspectives dans les méthodes de synthèse de vues au décodeur, avec notamment des techniques de remplissage d'images, guidées par de l'information auxiliaire. Les résultats obtenus par notre système sont très encourageants. Le fait de réduire la redondance dans la représentation des données permet aux techniques de codage d'être plus souples et flexibles pour les applications interactives.

Abstract – In this paper, we investigate multi-image transmission schemes that offer the possibility to the users to choose in real-time their viewpoint. This framework requires new approaches in the data representation and coding methods. The main idea of this paper is to generalize image representation by regrouping multiple views together in a unique signal. This permits the system to reduce the redundancies at the representation step. Then, it leads to a more flexible description method well suited for interactive schemes. Finally, this opens new perspectives in view synthesis at the decoder, namely inpainting techniques guided by auxiliary information. Our approach provides a more compact representation which is more adapted to interactivity than the traditional coding schemes.

1 Introduction

Les recherches en techniques 3D se sont récemment orientées vers les schémas à navigation interactive. Contrairement aux schémas dit "classiques", ces derniers permettent l'envoi d'un sous ensemble des images sur demande des utilisateurs. Bien qu'ayant un potentiel évident, ce type de schéma posent de nombreux problèmes le long de la chaîne de transmission, depuis l'acquisition, jusqu'au rendu, en passant par les étapes de compression. Les schémas de codage des séquences vidéo multi-vues traditionnels ont d'abord été développés comme des extensions des techniques mono-vues [1]. Ainsi, l'extraction des corrélations entre les vues s'appuit sur les mêmes techniques de recherche de bloc similaires entre les images. Ces techniques présentent deux inconvénients majeurs. D'abord, les vecteurs de corrélation entre les images créent de fortes dépendances qui rendent difficile l'extraction de seulement une partie des vues. Deuxièmement, ces modèles de corrélation par bloc deviennent peu fiables lorsque la disposition des caméras est complexe (forte rotation, grande distance, zoom, etc.). Dans le cas de schémas à navigation interactive [2], ces deux inconvénients rendent impossible l'utilisation de ces approches. En effet, une navigation de qualité nécessite un grand nombre de vues et donc des transformations complexes entre elles. Ainsi, la transmission de l'ensemble du domaine de navigation est impossible. Bien que des méthodes alternatives existent dans la littérature [3], celles-ci ne considèrent pas des domaines de navigation suffisamment grands et complexes, et restent ancrées à des méthodes de représentation traditionnelles du codage vidéo (*e.g.*, image, image plus profondeur, etc.).

Pourtant, de manière parallèle, de nouvelles méthodes de représentation ont été développées [4]. Certaines restent liées à la représentation image [5], mais d'autres considèrent des signaux regroupant des ensembles de vues [6]. Ceux-ci restent cependant trop distants avec les problématiques de codage, et c'est pourquoi nous proposons ici notre propre méthode de représentation. Celle-ci consiste à diviser le domaine de navigation en segments et à décrire chacun d'eux par un seul signal. Nous choisissons, pour le moment, de travailler dans le cas d'une scène statique (*e.g.*, musée). Un utilisateur, ayant une navigation continue recevra tour à tour les segments de navigation nécessaires. Dans le système développé, les redondances sont éliminées au moment de la représentation des segments et non à l'étape de codage, comme dans les techniques classiques, apportant une plus grande flexibilité au schéma de transmission. Les tailles de stockage et de transmission sont également diminuées comme le montrent les résultats obtenus. Dans cet article, nous présentons d'abord le concept général de

la représentation par segments de navigation, puis une méthode de description de ces segments que nous avons développée. Enfin, nous détaillerons une technique d'optimisation du partitionnement du domaine de navigation en segments. Les résultats obtenus montrent que la description proposée permet d'obtenir des performances encourageantes tout en ayant une flexibilité accrue, nécessaire aux schémas interactifs. Nous finissons par décrire sommairement deux solutions alternatives que nous avons proposées.

2 Segments de navigation

Dans la littérature, l'ensemble des vues atteignables par l'utilisateur est constitué généralement d'une suite discrète de vues alignées ou non éloignées. Afin de garantir une navigation de haute qualité, nous choisissons dans notre approche de travailler avec un ensemble continu et plus vaste de points de vue, appelé *domaine de navigation*. Ce domaine de navigation est paramétré par un ensemble continu correspondant aux positions des caméras. Nous supposons que cette navigation est continue. Deux exemples de domaines de navigation à une et deux dimensions sont donnés en figure 1 pour la séquence *ballet*.

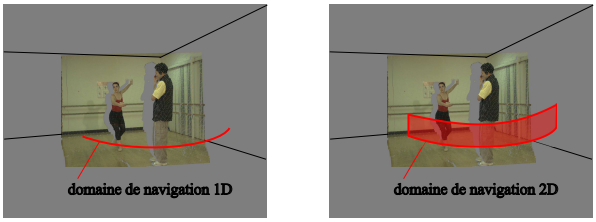

FIGURE 1 – Exemple de domaine de navigation 1D et 2D.

A partir du domaine de navigation, nous définissons des *segments de navigation*. Ceux-ci correspondent à des ensembles connexes de vues qui sont représentés par un signal. Autrement dit, l'information 3D visible par chacun des points de vue d'un segment est regroupée en une seule description, construite comme expliqué dans la section. 3.

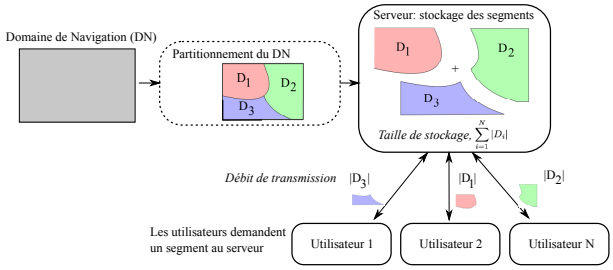

FIGURE 2 – Schéma de transmission proposé.

Une fois les segments construits, nous considérons le schéma de transmission suivant. Chaque segment est stocké sur un serveur accessible directement par les utilisateurs. Les utilisateurs naviguent librement au sein du domaine de navigation et communiquent leur position régulièrement au serveur. Une fois que ceux-ci changent de segment, ils demandent le nouveau au serveur que celui-ci transmet. Ainsi, l'utilisateur peut naviguer librement dans tout le segment sans nécessiter de nouvelles informations. Le schéma de transmission général est illustré en figure 2.

3 Description des segments

Dans la section précédente, nous avons expliqué que chaque segment de navigation était représenté par un seul signal. Autrement dit, nous proposons de regrouper les images en une seule description. Pour cela nous avons développé, entre autres, une approche consistant à fixer un point de vue de référence dans chacun des segments, puis de définir une information auxiliaire pour chacun des segments qui décrirait l'innovation de l'ensemble des points de vues par rapport à cette image de référence (par exemple les occlusions) [7, 8]. Au décodeur, l'utilisateur reçoit le signal décrivant l'ensemble des points de vues d'un segment. D'abord, il extrait la vue de référence, et par des techniques de projection [9], il estime l'image qu'il souhaite atteindre. De cette image, il reconstruit toutes les régions de l'image, visibles depuis le point de vue de référence. Ensuite, il remplit les zones occlusions en utilisant des techniques de remplissage d'image guidée par l'information auxiliaire [7]. Ces techniques originales se reposent sur des méthodes classiques de remplissage, et utilisent l'information auxiliaire transmise pour converger vers une solution cohérente et stable.

4 Partitionnement du domaine de navigation

Les segments de navigation constituent une approche prometteuse pour les schémas interactifs car ils regroupent sans redondance une information qui permet aux utilisateurs une certaine indépendance de navigation dans un voisinage donné. La flexibilité d'une telle description prend tout son sens lorsqu'il s'agit d'établir la taille de ces segments. C'est ce que nous proposons d'étudier dans cette partie. Nous rappelons que chaque segment de navigation $\mathcal{X}(Y_i)$ est constitué d'une image de référence Y_i et d'une information auxiliaire φ_i. L'objectif global est de trouver un partitionnement qui minimise un coût de stockage Γ et un débit R, qui correspond à un coût de transmission moyen. Si $|Y_i|$ et $|\varphi_i|$ correspondent au nombre de bits utilisés pour décrire respectivement Y_i et φ_i, et si $P(\mathcal{X}(Y_i))$ correspond à la probabilité *a priori* que le segment de navigation $\mathcal{X}(Y_i)$ soit choisi, nous pouvons écrire le débit

$$R = \sum_{i=1}^{N_V} P(\mathcal{X}(Y_i))(|Y_i| + |\varphi_i|), \qquad (1)$$

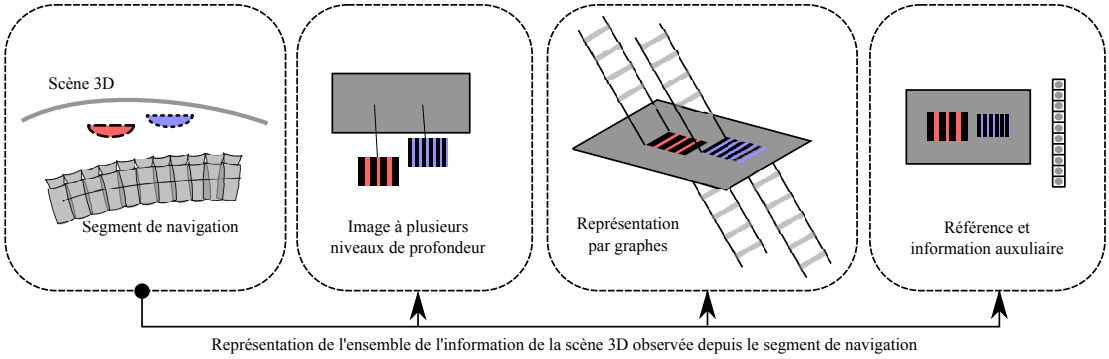

FIGURE 3 – Illustration des trois représentations proposées pour décrire un segment de navigation.

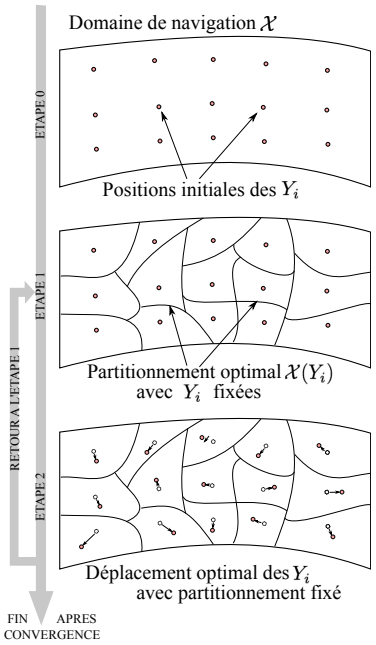

FIGURE 4 – Technique de partitionnement du domaine de navigation analogue à l'algorithme de Lloyd pour la quantification de vecteur [10].

où N_V est le nombre de segments. Avec les même notations, on a l'expression de la taille de stockage :

$$\Gamma(N_V, \{Y_i\}) = \sum_{i=1}^{N_V} (|Y_i| + |\varphi_i|). \qquad (2)$$

Pour trouver une partitionnement qui minimise le coût moyen de transmission sous des contraintes de stockage maximal $\Gamma_{\max}$, nous proposons d'utiliser une méthode analogue à l'algorithme de quantification de vecteurs proposé par Lloyd [10]. L'algorithme proposé est récapitulé dans le figure 4. Dans une étape préliminaire, les images de références Y_i sont positionnées arbitrairement dans le domaine de navigation $\mathcal{X}$. Ensuite, alternativement, l'algorithme modifie la position des frontières entre les segments et la position des images de référence. Cette modification se fait en minimisant le critère de débit moyen donné les équations (1) et (2). L'algorithme se termine lorsque les étapes 1 et 2 ne modifient plus les positions respectives de ces frontières et des images de références. Autrement dit, l'algorithme s'arrête une fois qu'il a convergé vers une solution optimale, du moins d'un point de vue local.

5 Résultats

D'un point de vue performance de codage, la représentation proposée avec image de référence et information auxiliaire permet un gain d'environ 41% par rapport au débit obtenu avec H.264/AVC. Plus de détails pourront être trouvés dans [11].

Nous présentons dans la figure 5 et 6 les performances de notre algorithme de partitionnement en situation d'interaction avec des utilisateurs. Pour la séquence *Ballet*, nous générons 120 vues formant un domaine de navigation unidimensionnel. Nous lançons notre algorithme de partitionnement pour N_V égal à 2 et 3. Nous considérons en plus deux méthodes dites "classiques" à titre de comparaison. La première consiste à coder chaque image capturée en intra et d'envoyer à l'utilisateur les deux images entourant la position demandées. La deuxième consiste à coder les 8 images capturées à l'aide du codec JMVM et de transmettre l'ensemble des vues à l'utilisateur. Nous simulons pour toutes ces méthodes une communication entre un serveur et plusieurs utilisateurs. Dans notre simulateur, nous supposons que 10 utilisateurs arrivent chaque seconde avec une durée de vie égal à T. L'utilisateur navigue dans la scène et nous mesurons la quantité de bits envoyée à tous les utilisateurs par seconde pour chacune des représentations. Les résultats de stockage sur le serveur sont montrés en figure 5 et ceux de débit en figure 6 pour différentes valeurs de T. On voit que pour des résultats de distorsion-stockage similaires, notre représentation utilisant un partitionnement du domaine de navigation obtient des résultats de débit plus efficaces. Les résultats présentés ci-dessus montrent l'efficacité de notre approche. Ils prouvent d'abord l'intérêt de réduire les redondances à l'étape de la représentation, et non du codage. Ensuite, ils valident l'idée d'un schéma de transmission plus général et adapté au codage interactif.

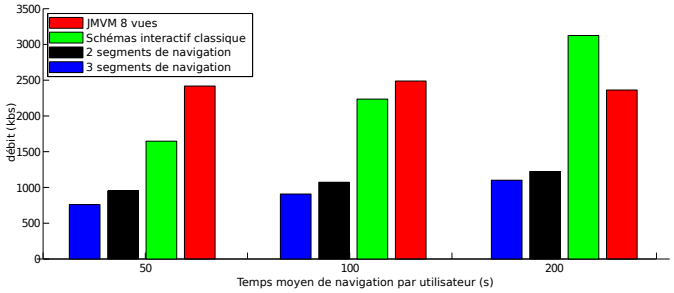

FIGURE 5 – Simulation d'une communication entre un serveur et plusieurs utilisateurs, durant 1000 secondes pour la séquence *Ballet*.

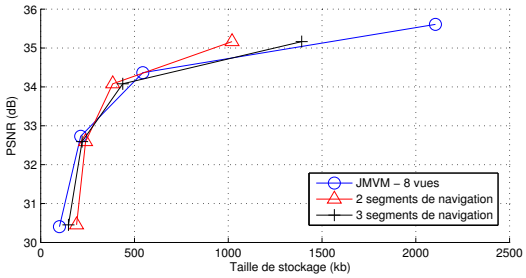

FIGURE 6 – Comparaison de différentes représentations par rapport à leur taille de stockage sur les serveurs.

6 Représentations alternatives

Nous présentons enfin des solutions alternatives que nous avons développées pour décrire un segment de navigation.

Image à plusieurs niveaux de profondeur : Ce type de représentation (IPNP) existe déjà dans certaines références afin d'améliorer les performances de compression des schéma traditionnels [12]. Dans la solution développée ici, nous proposons d'étendre cette notion à la description de segment de navigation. A partir d'un point de vue arbitraire du segment de navigation, nous regroupons l'ensemble des pixels de chaque vue du segment par projection basé-profondeur [9]. Comme illustré dans la figure 3, l'IPNP obtenue regroupe l'ensemble de la scène 3D visible depuis chacun des points de vue du segment de navigation. La forme de chaque niveau est arbitraire. L'IPNP est donc codée avec des outils adaptés type codage à adaptation au contour. La restitution des vues du segment au décodeur se faire en projetant l'IPNP sur le point de vue désiré, en utilisant l'information de profondeur. Nous obtenons que la compression de notre solution améliore les résultats d'une compression classique inter-vues d'environ 30%.

Représentation par graphes Contrairement aux IPNP, la représentation par graphe (RPG) s'affranchit de l'information de profondeur en représentant directement les connections entre pixels. Plus de détails sont fournis dans [13]. L'idée principale est de partir de la description image d'un point de vue de référence arbitraire et de représenter image par image les connections entre l'image précédente et les "nouveaux" pixels de l'image courante. Ces connections et informations de couleurs sont stockées dans un graphe. Comme illustré dans la figure 3, la description résultante a une structure régulière qui permet une bonne compression du graphe. Au décodeur, l'algorithme de reconstruction n'a plus qu'à parcourir le graphe et synthétiser les vues en se reposant sur les connections. De son côté, la représentation par graphe obtient des améliorations allant jusqu'à 33% de débit sauvé par rapport aux représentations classiques multi-vues.

7 Conclusion

Dans ce papier, nous présentons une nouvelle approche dans la représentation de données 3D pour des schémas à navigation interactive au décodeur. La représentation proposée repose sur l'idée générale de regrouper plusieurs vues dans un même segment de navigation. Nous proposons trois approches pour le faire et nous optimisons le partitionnement du domaine de navigation pour une d'entre elles. Les résultats obtenus montrent que notre approche apporte la flexibilité nécessaire aux schémas interactifs.

Références

[1] Y. Chen, YK. Wang, K Ugur, MM Hannuksela, J Lainema, and M Gabbouj, "The emerging MVC standards for 3d video services," *EURASIP J. on Adv. in Sign. Proc.*, vol. 2009, pp. 1–13, 2009.

[2] C. Fehn, R. De La Barré, and S. Pastoor, "Interactive 3-dtv-concepts and key technologies," *Proc. IEEE*, vol. 94, pp. 524–538, 2006.

[3] X. Xiu, G. Cheung, and J. Liang, "Delay-cognizant interactive streaming of multiview video with free viewpoint synthesis," *IEEE Trans. on Multimedia*, vol. 14, pp. 1109–1126, 2012.

[4] A Vetro, A. Tourapis, K Müller, and T. Chen, "3d-tv content storage and transmission," *IEEE Trans. on Broadcasting*, vol. 57, pp. 384–394, 2011.

[5] K. Müller, P. Merkle, and T. Wiegand, "3d video representation using depth maps," *Proc. IEEE*, vol. 99, no. 4, pp. 643–656, Apr. 2011.

[6] M. Tanimoto, "Overview of free viewpoint television," *EURASIP J. on Sign. Proc. : Image Commun.*, vol. 2006, pp. 455–461, 2006.

[7] T. Maugey, P. Frossard, and G. Cheung, "Consistent view synthesis in interactive multiview imaging.," in *Proc. IEEE Int. Conf. on Image Processing*, Orlando, Florida, US, Oct. 2012.

[8] T. Maugey, I. Daribo, G. Cheung, and P. Frossard, "Navigation domain representation for interactive multiview imaging," *accepted in IEEE Trans. on Image Proc.*, 2013.

[9] D. Tian, P. Lai, P. Lopez, and C Gomila, "View synthesis techniques for 3d video," *Proc. of SPIE, the Int. Soc. for Optical Engineering*, vol. 7443, 2009.

[10] A. Gersho and R.M. Gray, *Vector quantization and signal compression*, Kluwer academic publishers, 1992.

[11] I. Daribo, T. Maugey, G. Cheung, and P. Frossard, "R-d optimized auxiliary information for inpainting-based view synthesis," in *IEEE 3DTV-Conference*, Zurich, Switzerland, 2012.

[12] X. Cheng, L. Sun, and S. Yang, "Generation of layered depth images from multi-view video," in *Proc. IEEE Int. Conf. on Image Processing*, San Antonio, Texas, USA, Sep. 2007.

[13] T. Maugey, A. Ortega, and P. Frossard, "Multiview image coding using graph-based approach," in *IEEE Workshop on 3D Image/Video Technologies and Applications (IVMSP)*, Seoul, Korea, Jun. 2013.