Des séquences vidéo aux panoramas de mouvement

A. Bartoli N. Dalal R. Horaud

INRIA Rhône-Alpes

655, avenue de l'Europe
38334 Saint Ismier cedex
France
prénom.nom@inria.fr

Résumé

La construction d'un panorama est possible à partir d'une séquence vidéo provenant d'une caméra subissant un mouvement de rotation pure. Un panorama constitue une représentation "minimale" de la séquence vidéo. Les approches existantes permettent de traiter efficacement les séquences vidéo de scènes statiques. Dans ce travail, nous faisons l'hypothèse que la scène observée est constituée d'un fond statique et d'une couche dynamique. Nous proposons un cadre pour la construction de panoramas de mouvement, permettant de représenter de telles séquences de manière efficace.

Mots clefs

Panorama de mouvement, compression, automatisme.

1 Introduction

La construction de mosaïques d'image est intéressante de points de vue tels que la navigation et la compression. Elle consiste a assembler plusieurs images d'une même scène en un panorama [1, 2, 3, 4, 5, 6, 7]. L'information redondante contenue dans les images est alors représentée de manière minimale. Les positions de la caméra correspondant aux différentes images peuvent être estimée conjointement à la construction du panorama. Etant données ces positions, les images originales peuvent être reconstruites à partir du panorama. L'hypothèse de base pour la construction de tels panoramas est que la position de la caméra est statique, en d'autres termes, que le mouvement de la caméra est une rotation. La compression est un intérêt majeur de la construction de panoramas, en particulier lorsque des séquences vidéo complétes sont traitées.

La plupart des algorithmes reposent sur l'hypothèse que la scène observée est statique [1, 2, 5, 6, 7]. Les positions successives de la caméra sont estimées et utilisées pour déformer et coller les images sur une surfaces 2D représentant le panorama.

Dans ce travail, nous nous intéressons à des séquences vidéo de scènes dynamiques [3, 4]. L'observation clef est que beaucoup de séquences vidéo sont constituées d'un fond statique (en mouvement par rapport à la caméra, mais non déformable) sur lequel évolu un premier plan dynamique (en mouvement par rapport à la caméra *et* déformable). Une représentation efficace de telles séquences est un *panorama de mouvement*, constitué par le mouvement de la caméra, un panorama de fond statique et une couche dynamique. Ce concept est illustré sur la figure 1. La séquence vidéo originale est reconstruite par superposition de la couche dynamique sur le fond. Les gains de cette représentation en terme de compression par rapport à une représentation en séquence d'images sont très importants. D'autres applications, telles la visualisation ou la modification de la séquence sont possibles. L'objectif de ce travail est l'élaboration d'un cadre pour la construction de *panoramas de mouvement*, voir [8] pour plus de détails.

Nous présentons les approches existantes en section 2 puis notre approche en section 3. Nous concluons en section 4.

2 Approches existantes

La construction de panoramas passe par le calcul du mouvement de la caméra entre images consécutives de la séquence. Un mouvement de rotation pure induit une déformation homographique representée par une matrice (3×3) notée H. Cette homographie encapsule la rotation de la caméra ainsi que ses paramètres intrinséques, en particulier sa distance focale liée au zoom. Elle permet le transfert de pixels d'une image à une autre. Soient $\mathbf{q}$ et $\mathbf{q}'$ les coordonnées homogènes de deux pixels correspondant entre les deux images :

$$\mathbf{q}' \sim \mathsf{H}\mathbf{q}, \tag{1}$$

où $\sim$ est l'égalité à un facteur multiplicatif près.

La calcul de H est basé sur la minimisation d'une fonction de coût pouvant être de deux types différents. Les méthodes *denses* sont basées sur une erreur exprimée en fonction de la différence d'intensité entre tous les pixels correspondants des deux images. Les méthodes *éparses* sont basées sur une erreur exprimée en fonction de la distance (exprimée en pixels) entre certains des pixels correspondant. Ces deux méthodes sont présentées dans les deux paragraphes suivants.

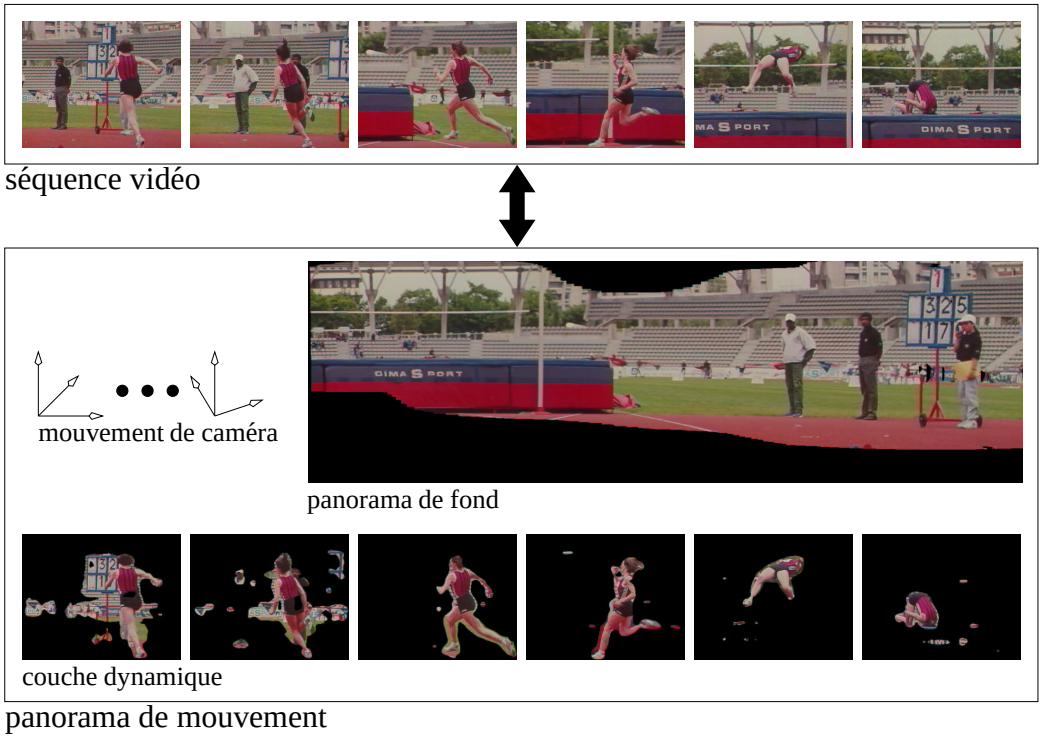


Figure 1 – *La séquence vidéo du saut en hauteur, constituée de 300 images (durée de 12 secondes) et son panorama de mouvement associé.*

2.1 Méthodes denses

Soient $\mathcal{I}$ et $\mathcal{I}'$ les deux images considérées et $\mathcal{I}(\mathbf{q})$ l'intensité du pixel de coordonnées $\mathbf{q}$. D'aprés l'équation (1) :

$$\mathcal{I}(\mathbf{q}) = \mathcal{I}'(\mathbf{q}') + \epsilon = \mathcal{I}'(\mathsf{H}\mathbf{q}) + \epsilon,$$

où ϵ modélise le bruit induit par le système d'acquisition des images. Par sommation sur les pixels communs aux deux images, et en supposant que le bruit ϵ a une distribution Gaussienne, on obtient la fonction de coût suivante :

$$\mathcal{D}(\mathsf{H}) = \sum_{\mathbf{q}} (\mathcal{I}(\mathbf{q}) - \mathcal{I}'(\mathsf{H}\mathbf{q}))^2.$$

Cette fonction de coût doit être minimisée sur les entrées de la matrice H. Elle est non-linéaire. Des algorithmes spéciques ont été proposé pour résoudre le problème [1, 4, 5]. Il faut noter que ces algorithmes sont itératifs et peu robuste, c'est à dire qu'ils ne convergent vers la bonne solution pour H que si quasiment tous les pixels satisfont le modéle de mouvement (hypothèse de scène statique).

2.2 Méthodes éparses

Supposons que l'on dispose d'un ensemble de correspondances de points $\mathbf{q} \leftrightarrow \mathbf{q}'$ entre les deux images. De telles correspondances peuvent être obtenus en utilisant un algorithme de suivi de points d'intérêt, voir par exemple [9]. D'aprés l'équation (1) :

$$\mathbf{q}' \sim \mathsf{H}\mathbf{q} \Rightarrow d(\mathbf{q}', \mathsf{H}\mathbf{q}) = 0,$$

où d est une mesure de distance entre deux points. Par sommation sur les correspondances de points disponibles, on obtient la fonction de coût suivante :

$$\mathcal{E}(\mathsf{H}) = \sum_{\mathbf{q} \leftrightarrow \mathbf{q}'} d^2(\mathbf{q}', \mathsf{H}\mathbf{q}).$$

Comme dans le cas dense, cette fonction de coût doit être minimisée sur les entrées de la matrice H. Lorsque la distance Euclidienne est choisie pour d, $\mathcal{E}$ est non-linéaire. Elle peut être linéarisée et une solution pour H peut être calculée de manière non itérative [6, 7]. Il faut noter que 4 correspondances de points sont suffisantes pour fournir une solution pour H. Cet algorithme peut être rendu hautement robuste avec l'aide de techniques basées sur l'échantillonage aléatoire, tel que les moindres carrés médians.

2.3 Comparaison

Les caractéristiques des méthodes denses et éparses sont les suivantes.

Les méthodes denses sont précises en terme de la finesse de calcul du mouvement de la caméra, c'est à dire que l'alignement des images est bon. La figure 2 illustre cette différence d'alignement avec les méthodes éparses. Cette précision provient du fait que la fonction de coût $\mathcal{D}$ minimisée est exprimée sur les intensités des pixels directement. Les méthodes denses ne tolèrent que très peu de données erronées, c'est à dire de pixels ne satisfaisant par

le modèle de mouvement homographique (1), par exemple ceux relatifs à la couche dynamique.

Les méthodes éparses peuvent être hautement robustes, c'est à dire qu'elles peuvent tolérer un grand nombre de fausses correspondances, jusqu'à 50% par exemple lorsqu'un estimateur basé sur les moindres carrés médians est utilisé. Cependant, elles ne sont pas très précise en terme d'alignement des images, due au fait que la fonction de coût $\mathcal{E}$ considérée n'est pas exprimée directement sur les intensités des pixels.

3 Notre approche

Les méthodes denses et éparses présentées ci-dessus sont complémentaires. La première est précise et la seconde robuste. Ceci suggére l'approche suivante pour la construction de panoramas de mouvement :

1. estimation initiale du mouvement de la caméra par une méthode *éparse*.
2. élimination de la partie dynamique de la scène dans les images, basée sur les images consécutives.
3. construction d'un panorama de fond statique précis avec une méthode *dense*.
4. construction de la couche dynamique, basée sur le panorama de fond.

On observe sur cet algorithme que l'étape 3 utilise une approximation de la couche dynamique fournie par l'étape 2 pour produire le panorama de fond. L'étape 4 utilise ensuite ce panorama de fond pour calculer une meilleure approximation de la couche dynamique. Ceci suggére l'itération des étapes 3 et 4, jusqu'à la stabilisation de la couche dynamique.

L'étape 1 fournie une estimation initiale du mouvement de la caméra. Des correspondances de points sont obtenues par suivi de points d'intérêt. Les distances focales de la caméra avant et aprés le mouvement, ainsi que la rotation, sont estimées de manière robuste.

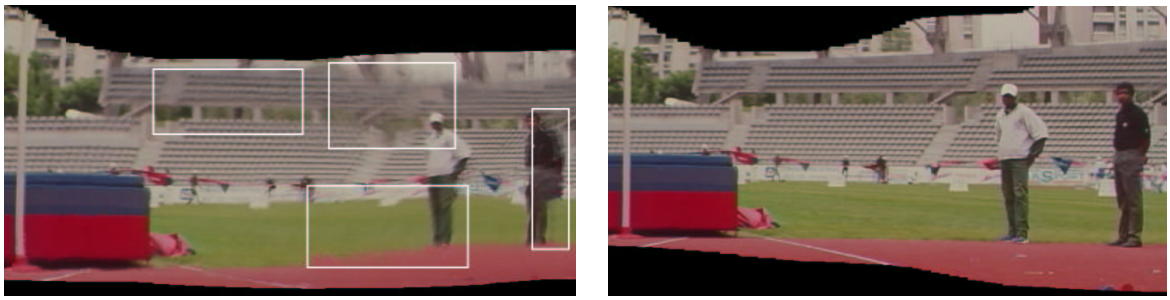
L'étape 2 consiste à éliminer des images la partie dynamique de la scène, c'est à dire délimiter, pour chaque image, quelle est la partie observée en mouvement. A ce stade, une première approximation du mouvement de la caméra est disponible, mais pas encore de panorama de fond. Nous proposons de baser cette délimitation sur les images consécutives. Ceci permet que la fraction des images considérées affectée par la couche dynamique soit réduite. Pour chaque image, nous considérons les images immédiatement précédente et suivante. La fonction de coût $\mathcal{D}$ est calculée pour chaque pixel, puis seuillée par un test du χ^2 pour discriminer les pixels du fond et de la couche dynamique. La figure 3 illustre cette étape. On observe que la délimitation obtenue est très grossière.

Figure 3 – *Couche dynamique extraite en considérant des triplets d'images consécutives.*

L'étape 3 est la contruction du panorama de fond, basé sur une méthode dense, avec une estimation initiale du mouvement de caméra et de la couche dynamique. L'algorithme itératif décrit dans [5] est utilisé. La qualité d'alignement obtenue est meilleure que celle donnée par la méthode éparse, comme l'atteste la figure 2.

L'étape 4 est la construction des images dynamiques, c'est à dire la détermination de la couche dynamique dans chaque image. Elle est basée sur le panorama de fond précédement estimé. De manière similaire à l'étape 2, la fonction de coût $\mathcal{D}$ est utilisée, mais cette fois pour comparer chaque pixel de chaque image au panorama du fond et non plus aux images consécutives. Un test du χ^2 est utilisé pour discriminer les pixels du fond et de la couche dynamique. La figure 4 illustre cette étape. La délimitation obtenue est bien plus précise que celle de l'étape 2 (figure 3).

Figure 4 – *Couche dynamique extraite en considérant le panorama de fond.*



(a) (b)

Figure 2 – *Détail du panorama de fond construit avec une méthode éparse (a) et rafiné par une méthode dense (b).*

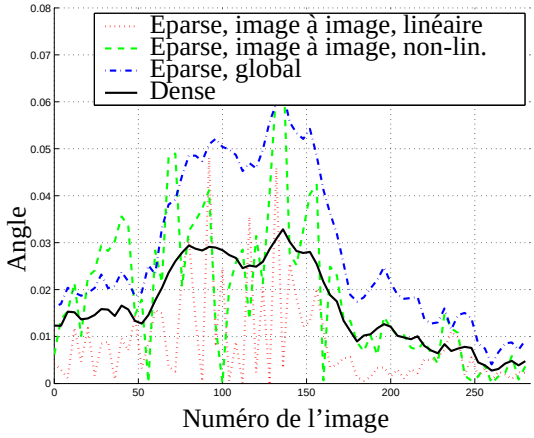


Figure 5 – *Rotation de la caméra par rapport à l'axe vertical. Les 3 premières méthodes sont éparses, et la quatrième est dense.*

La figure 5 montre l'angle de rotation de la caméra autour de l'axe vertical (perpendiculaire au sol), retouvé par les différentes étapes de l'algorithme. La valeur de cet angle reflète la vitesse de l'athlète. En particulier, la vitesse maximale est obtenue juste avant le saut.

4 Conclusion

Nous décrivons une approche pour la construction de panoramas de mouvement, constitués d'un panorama de fond statique et d'une couche dynamique. Cette approche est basée sur un couplage de méthodes éparses et denses, permettant de bénéficier d'une estimation robuste et précise. La robustesse permet la détection de la couche dynamique. Les résultats obtenus sont visuellement bons. Nous pensons que l'utilisation conjointe de méthodes denses et éparses devrait permettre le traitement efficace de séquences vidéo dynamiques.

Références

[1] J.R. Bergen, P. Anandan, K.J. Hanna, et R. Hingorani. Hierarchical model-based motion estimation. Dans *Proceedings of the 2nd European Conference on Computer Vision, Santa Margherita Ligure, Italy*, pages 237–252, 1992.

[2] S.E. Chen. Quicktime VR - an image-based approach to virtual environment navigation. Dans SIGGRAPH *1995, Los Angeles, USA*, pages 29–38, 1995.

[3] M. Irani et P. Anandan. About direct methods. Dans B. Triggs, A. Zisserman, et R. Szeliski, éditeurs, *Vision Algorithms : Theory and Practice*, numéro 1883 dans LNCS, pages 267–277, Corfu, Greece, July 1999. Springer-Verlag.

[4] M. Irani, P. Anandan, J. Bergen, R. Kumar, et S. Hsu. Efficient representations of video sequences and their applications. *Signal Processing : Image Communication, special issue on Image and Video Semantics : Processing, Analysis, and Application*, 8(4), May 1996.

[5] H.-Y. Shum et R. Szeliski. Systems and experiment paper : Construction of panoramic mosaics with global and local alignment. *International Journal of Computer Vision*, 36(2) :101–130, February 2000.

[6] P. H. S. Torr et A. Zisserman. Feature based methods for structure and motion estimation. Dans W. Triggs, A. Zisserman, et R. Szeliski, éditeurs, *Vision Algorithms : Theory and Practice*, numéro 1883 dans LNCS, pages 278–295, Corfu, Greece, July 1999. Springer-Verlag.

[7] I. Zoghlami, O. Faugeras, et R. Deriche. Using geometric corners to build a 2D mosaic from a set of images. Dans *Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA*, pages 420–425, June 1997.

[8] A. Bartoli, N. Dalal, B. Bose, et R. Horaud. From video sequences to motion panoramas. Dans *Proceedings of the* IEEE *workshop on Motion and Video Computing, Orlando, Florida, USA*, 2002.

[9] J. Shi et C. Tomasi. Good features to track. Dans *Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA*, pages 593–600, 1994.