# DETERMINATION D'UNE CARTE DE PROFONDEUR A PARTIR D'UNE SEQUENCE D'IMAGES

Lionel MARCE*, Patrick BOUTHEMY**

IRISA (*INSA, **INRIA/Centre de Rennes)
Avenue du Général leclerc - Campus de Beaulieu
35042 - Rennes Cedex - France

Nous examinons le problème de la construction d'une carte fiable de profondeur à un instant donné, pour la navigation d'un robot, à partir de plusieurs images ordonnées dans le temps en nombre égal ou supérieur à deux. La caméra est animée d'un mouvement de translation uniforme. Après avoir montré la difficulté du problème en testant la stabilité d'un opérateur d'intérêt lié à un algorithme de corrélation, nous proposons une méthode basée sur une modélisation locale d'éléments de contour spatio-temporels dans la séquence d'images et un critère de maximum de vraisemblance, auquel est intégrée la connaissance du mouvement de la caméra. L'estimation des déplacements de ces éléments permet d'induire la profondeur des points correspondants des objets dans l'espace. Nous donnons les résultats de l'application de l'algorithme sur une séquence d'images de synthèse bruitées.

This paper deals with the issue of building a reliable depth map for robot navigation from a sequence of time-ordered images. More precisely, we investigate the use of successive images, in number equal or more than two, to estimate a depth map at given time. The images are assumed to be acquired by a camera undergoing translation motion. First we show the difficulty of the stated problem considering real images by testing the stability of an interest operator combined with a correlation algorithm. Then we develop a method based on a local modeling of spatio-temporal (or moving) edges. A maximum likelihood scheme, which explicitly incorporates the knowledge of camera motion, allows to determine such features. From the estimated displacement of these edges in the image plane we infer the depth of corresponding object points in space. We give results obtained with a sequence of noisy synthetic images.

## Introduction

- Nous nous plaçons dans le contexte suivant : comment utiliser une caméra unique pour l'évitement d'obstacle dans le cas d'un robot mobile ou d'un robot manipulateur. Dans ce type d'applications, le mouvement de la caméra est approximativement connu, fourni par des capteurs du type odomètre pour les véhicules mobiles ou capteurs optiques incrémentaux pour les robots manipulateurs.

- Il est possible de déduire de cette connaissance du mouvement du capteur, en utilisant le principe de triangulation utilisé en stéréovision, la distance d'obstacles éventuels par rapport à la caméra, les objets dans la scène étant supposés fixes.

- Pour cela, il faut apparier dans deux images successives des éléments d'image, projections de points du même objet dans l'espace.

- Là réside la difficulté principale dans cette mise en correspondance de caractéristiques, conduisant à des erreurs pouvant être très importantes. Une source de mesures de profondeur bruitées est la discrétisation des images, amplifiée par le fait que des images successives correspondent à des positions de la caméra très voisines. Un autre problème dans la phase de mise en correspondance est l'apparition ou la disparition de zones dans l'image, dues à des phénomènes d'occlusion d'un objet par un autre.

- Nous allons tout d'abord mettre en évidence les problèmes associés à ces appariements, à travers la mise en oeuvre sur un exemple, d'un algorithme s'appuyant sur des points d'intérêt, puis décrire la méthode que nous proposons pour réduire cette difficulté.

## I - Suivi de points d'intérêt dans une séquence d'images

### 1) Choix de l'opérateur d'intérêt

Un certain nombre de méthodes d'appariement sont basées sur la recherche de points d'intérêt, telles celle proposée par Moravec qui trouve les points de variance omnidirectionnelle importante ou celles de Kitchen-Rosenfeld ou Nagel qui détectent les coins.

Nous avons testé la qualité de l'appariement par l'algorithme de Moravec bien représentatif de sa classe suivant Thorpe |THO 84|.

Rappelons en brièvement le principe |MOR 81|. Il s'agit de mesurer la valeur d'intérêt d'un point en considérant la variance de la fonction intensité le long de quatre directions (horizontale, verticale et les deux diagonales principales) au sein d'une fenêtre (par exemple cinq par cinq) autour de ce point. Est alors retenue la valeur minimum de ces quatre variances.

Un filtre de maximum local teste chaque point par rapport à ses 24 voisins les plus proches pour éviter des agglomérats de points d'intérêt dans des zones hautement texturées.

Nous avons appliqué l'opérateur de Moravec sur une séquence de 125 images provenant du SRI, acquise avec une caméra Sony de 12,5 mm de distance focale, dans un environnement de bureau en suivant un mouvement uniforme perpendiculaire à l'axe de la caméra |BOL 85|.

A l'issue de cette expérimentation, la stabilité de l'opérateur a été jugée satisfaisante |MAR 81|.

### 2) Couplage de l'opérateur de Moravec à un algorithme de corrélation.

Pour progresser dans le test, il a été ensuite entrepris de retrouver les 5 premiers points d'intérêt déduits de la première image dans toutes les images suivantes à l'aide d'un algorithme de corrélation pseudonormalisée.

La fenêtre de référence est de taille 5x5 pixels dans la première image et la fenêtre de recherche dans la deuxième image de taille 13x13 pixels.

Si les points associés d'image en image représentaient la même partie de l'objet, on devrait obtenir 5 droites parallèles au déplacement sur le graphique des points cumulés, ce qui n'est pas le cas (Fig. 1). Cela signifie que l'on perd des points de temps en temps et que l'on associe des points représentant des parties différentes de l'objet.

On n'a pas constaté de meilleurs résultats en accroissant la dimension de la zone entourant le

point d'intérêt dans la première image |MAR 86|.

On peut néanmoins améliorer le suivi des points d'intérêt en restreignant la zone de recherche de l'algorithme de corrélation à une zone horizontale de 9x1 pixels puisque l'on connaît le mouvement de la caméra.

Mais l'amélioration reste limitée : en effet si l'on examine la déformation de la surface d'intensité autour du point d'intérêt considéré, on s'aperçoit que la valeur d'intérêt associée à cette surface diminue assez rapidement.

## II - Algorithme d'appariement multi-images.

L'introduction de la connaissance du mouvement dans l'utilisation d'un opérateur d'intérêt lié à un algorithme de corrélation diminue le nombre des faux appariements comme le montre l'expérimentation précédente sur des images réelles, ou celle de Bharwani |BHA 86| sur des images de synthèse.

Pour obtenir de meilleurs résultats, nous avons cherché à utiliser simultanément plus de deux images successives pour construire une carte de profondeur à un instant donné. Le propos ici est plutôt d'induire une meilleure robustesse de l'algorithme que d'introduire une estimation récursive comme dans |ESP 87|.

Par ailleurs au lieu de retenir comme primitives les points d'intérêt, nous avons choisi des éléments de contour ce qui permettra d'obtenir une carte de mesure de profondeur de façon beaucoup plus dense. Pour le suivi des primitives à travers la séquence d'images, une technique plus élaborée, comme la relaxation, aurait pu être considérée. Ainsi dans |BAR 80| cette méthode est mise en oeuvre pour apparier des points d'intérêt de Moravec. En fait, nous avons développé une procédure dérivée de celle présentée dans |BOU 86|. Cette dernière est basée sur une modélisation locale d'éléments de contour spatio-temporel, dans une séquence d'images considérée comme un espace à 3 dimensions x, y, t, deux dimensions spatiales et une temporelle. En effet, la caméra se déplaçant et la scène étant fixe, tout contour dans l'image possède un mouvement apparent. La méthode présentée dans ce chapitre a l'avantage de pouvoir prendre en compte sans modification du formalisme défini deux ou plusieurs images successives. De plus la connaissance du mouvement de la caméra peut être explicitement intégrée dans la modélisation considérée.

### 1) Modélisation d'éléments de contour spatio-temporel

Dans l'espace (x, y, t), un élément de contour en mouvement génère une portion de surface engendrée par l'élément de contour spatial représenté par un petit segment de droite dans un plan ($t=t_o$) et par son vecteur vitesse associé.

Examinons de plus près la forme de cette surface. Tout d'abord, soulignons que la recherche des positions successives dans l'image d'un élément de contour 2D donné, est en fait un problème mono-dimensionnel. En effet, la caméra étant animée d'un mouvement de translation connue, on peut mettre à profit l'existence du foyer d'expansion (FOE) qui est l'intersection du plan image de la caméra avec le vecteur translation attaché au centre optique. Les positions successives recherchées sont alors situées sur un rayon d'expansion issu du FOE et passant par la position de référence de l'élément de contour considéré.

Si d est l'amplitude du déplacement du point de l'image entre le temps t et le temps t+1, W le déplacement de la caméra le long de l'axe Z, Z la profondeur du point relative à la caméra au temps t, D la distance du point correspondant de l'image au FOE au temps t+1, on a, dans un repère XYZ approprié, la relation

$$d = D.W/Z \qquad (1)$$

comme l'illustre la Fig. 2.

L'élément de contour ne reste pas parallèle à lui-même. Ceci peut être expliqué par la situation analogue dans le cas uniquement spatial. En effet les projections de droites parallèles dans l'espace réel donnent dans le plan image des droites concourantes au point de fuite (intersection du plan image avec la droite de même direction passant par le centre optique).

La surface générée par l'élément de contour est assez complexe (Fig. 3). On peut la simplifier lorsque le déplacement entre 2 images est faible par rapport à la distance au FOE et que cette distance est grande par rapport à l'élément de contour. Dans ces conditions on peut considérer que celui-ci se déplace parallèlement à lui-même avec une longueur constante. La surface est alors cylindrique avec comme base une hyperbole.

Il s'agit de déterminer cette surface (définie par exemple par l'orientation de l'élément de contour $\Theta$ dans le plan image et les d successifs) à partir de plusieurs projections successives dans le plan image. L'utilisation de la formule **(1)** permet alors d'estimer la distance Z du point considéré à la caméra, W étant connu, d et D déduits de la détermination de la surface, le FOE étant préalablement défini. Ce dernier peut être déduit d'une phase préalable de calibration et de la mesure courante du déplacement de la caméra, ou trouvé à l'aide d'une technique, comme celle décrite dans |JAI 83|.

### 2) Principe de la méthode

La méthode définie permettant la détection directe de telles portions de surface et l'estimation simultanée des paramètres $\Theta$ et d est basée sur un schéma de test d'hypothèses se traduisant par la définition d'un critère de maximum de vraisemblance.

Elle dérive de celle décrite dans |BOU 86|, dans lequel le modèle considéré était simplement une portion de plan, et l'information du mouvement extraite uniquement la composante du vecteur vitesse perpendiculaire au contour, car aucune contrainte supplémentaire n'était prise en compte. Cependant le formalisme développé peut être repris dans le contexte présent.

Etant donné un volume élémentaire $\Pi$ placé au point p dans l'espace x,y,t, deux hypothèses ou configurations locales peuvent intervenir :

**a)** ou bien il n'existe pas d'élément de contour spatio-temporel et la fonction d'intensité est donnée par $c_o+b$ pour tout point de $\Pi$ ($c_o$ est une constante, b un bruit gaussien centré) ;

**b)** ou bien il existe une portion de surfaces divisant $\Pi$ en 2 sous volumes, pour lesquels la fonction d'intensité est définie par $c_1+b$ dans $\Pi 1$ et $c_2+b$ dans $\Pi 2$, (b est un bruit gaussien et $c_1$, $c_2$ des constantes différentes).

A chaque hypothèse est associée une fonction de vraisemblance, respectivement $L_o$ et $L_1$ et l'on considère le rapport logarithmique de $L_1$ sur $L_o$.

Le critère de décision revient en fait à maximiser et à comparer à un seuil l'expression suivante :

$$CRV\ (p, \Theta_j, d_j) = \left| \sum_{m \in M} aj\ (m)\ f\ (p+m) \right|$$

où M représente tous les points de $\Pi$, les aj sont des coefficients dépendant uniquement de la géométrie de la portion de surface S, dont un jeu prédéfini de configurations doit être considéré. Enfin les f(p+m) représentent les intensités observées.

### 3) Mise en oeuvre de la méthode

Les coefficients aj(m) sont calculés hors ligne. L'algorithme s'apparente à une convolution avec un jeu de masques précalculés.

Pour chaque point de l'image courante et pour chaque masque, correspondant à une géométrie donnée ($\Theta$j,dj) on calcule le critère. Est sélectionnée la configuration ($\hat{\Theta}$,$\hat{d}$) qui le maximise à condition que ce maximum soit supérieur à un seuil prédéterminé.

On conclut alors à l'existence d'un élément de contour spatio-temporel au point p de paramètres

$\hat{\theta},\hat{d}$. On peut ainsi déduire la profondeur par rapport à la caméra du point correspondant dans l'espace, auquelle est aussi associé un coefficient de vraisemblance égal à CRV $(\hat{\theta}, \hat{d})$. Lors de la détermination de $\theta$ et d, on balaie en fait l'image à t+1, $I_{t+1}$, le long d'un rayon d'expansion. Les points résultants communs au volume $\Pi$ et à l'image $I_{t+1}$ peuvent ne pas coincider avec la grille d'échantillonnage de l'image. On est donc amené à leur affecter une valeur d'intensité calculée par interpolation des pixels voisins.

Si un tel phénomène n'intervient pas, une implémentation dont la complexité est équivalente à un simple opérateur de gradient spatial peut être proposée.

L'algorithme est étendu sans difficulté à plus de deux images. Ceci peut être aisément mis en évidence de la façon suivante. L'ensemble M entrant dans la sommation de la relation **(2)** peut s'écrire comme U $M_t$, où $M_t$ correspond à l'intersection du volume ${}^t\Pi$ et de l'image $I_t$, avec $t_1 \leqslant t \leqslant t_2$. Le nombre d'images retenues entre $t_1$ et $t_2$ peut être égal à 2 comme supérieur à 2 sans que le formalisme du critère exprimé en **(2)** en soit modifié. Cependant, les déplacements $d_j^{t+s}$ testés entre les images $I_t$ et $I_{t+1+s}$ ne sont pas quelconques ; un déplacement $d_j^t$ étant considéré entre $I_t$ et $I_{t+1}$, ils doivent satisfaire la contrainte suivante :

$$\forall s \; d_j^{t+s} = (1+s)\, d_j^t \, D_{t+1+s} / D_{t+1}$$

## III - Résultats

Pour tester la validité de la méthode, il a été effectué des expérimentations sur des images synthétiques représentant un parallélépipède, le mouvement de la caméra étant parallèle à son axe et perpendiculaire à un plan du parallélépipède. Dans ces conditions, le calcul de l'écart type des profondeurs estimées, est un bon indicateur du comportement de l'algorithme. Une séquence de 6 images a été construite. Les moyennes ont été calculées sur les 200 points correspondant aux plus grandes valeurs du coefficient de vraisemblance.

### 1) Utilisation de 2 images

Le cas de la prise en compte de 2 images successives a tout d'abord été considéré (tableau 1). Le bruit des mesures dû à la discrétisation diminue au fur et à mesure que l'on approche de l'obstacle : ceci est caractérisé par la diminution de l'écart type. Si l'on travaille sur des images prises à t et t+u avec successivement u=2,3,4,5, on observe le même phénomène.

Cette diminution s'explique par le fait que le déplacement d'un élément de contour est d'autant plus grand que l'on est proche de l'obstacle, donc déterminé avec plus de précision car le pas de recherche le long d'un rayon d'expansion est constant.

On remarque également que pour une même distance réelle l'écart type diminue plus les images successives prises en compte sont éloignées. Cela s'explique également par le fait que les distances entre points à apparier sont plus grandes, donc les erreurs de discrétisation ont relativement moins d'importance. Mais il ne sert à rien de prendre des distances plus grandes car les hypothèses postulées pour la mise en oeuvre de l'algorithme ne sont plus vérifiées.

### 2) Utilisation de plus de 2 images

On a successivement appliqué la méthode simultanément sur 3, 4, 5, 6 images. Les résultats obtenus sont réunis dans le tableau 2. L'écart type diminue avec le nombre d'images traitées à peu près de la même manière que lorsque des distances plus grandes entre images sont prises en compte.

### 3) Expérimentation sur des images bruitées

Pour conforter les premiers résultats, ont été menées les mêmes expériences sur des images bruitées par un bruit gaussien d'écart type successivement égal à 10, 20, 30, 40, 50% de l'intensité maximale des pixels dans une image. Les résultats sont donnés dans le tableau 3, dans le cas de 50%.

On a les mêmes tendances que sur les images non bruitées mais avec la différence que l'on gagne un peu moins en utilisant plusieurs images qu'en éloignant les images traitées.

Pratiquement il sera néanmoins beaucoup plus facile de prendre en compte plusieurs images à la fois que de déterminer le meilleur espacement entre images à choisir, en particulier lorsque les obstacles sont nombreux et à des distances différentes. Une difficulté expérimentale sur des images réelles provient de la précision de la détermination du FOE. On peut tenir compte de celle-ci en prenant comme zone de recherche lors de l'appariement à la place d'un rayon d'expansion, un "rayon épaissi" dont l'épaississement dépendrait de la précision de la mesure du FOE.

## Conclusion

Nous avons étudié l'utilisation d'une caméra unique pour l'évitement d'obstacles dans le cas d'un robot mobile ou d'un robot manipulateur dont le mouvement est approximativement connu. L'utilisation de l'appariement de points d'intérêt à travers des images successives permet de déduire la profondeur d'obstacles éventuels. Cependant, le couplage d'un algorithme de corrélation à un opérateur d'intérêt est peu robuste.

Pour remédier à cette situation, un opérateur a été conçu basé sur la modélisation d'un contour spatio-temporel dans lequel est intégrée la connaissance du mouvement de la caméra. Une fois le mouvement de ces éléments de contour déterminé, la profondeur des points correspondants dans l'espace peut être retrouvée. Cet algorithme peut être appliqué à plusieurs images successives, ce qui permet d'améliorer la précision de la méthode.

La méthode a été testée sur une surface parallélipédique synthétique perpendiculaire au mouvement de la caméra colinéaire avec son axe.

## Références


|BAR 80| S.T. BARNARD, W.B. THOMPSON, Disparity analysis of images, IEEE Trans. on PAMI, vol. 2, n° 4, July 1980, pp.323-340.

|BHA 86| S. BHARWANI, E. RISEMAN, A. HANSON, Refinement of environmental depth maps over multiple frames, IEEE workshop on Motion : Representation and Analysis, Charleston, South Carolina, May 1986, pp. 73-80.

|BOL 85| R.C. BOLLES, H.H. BAKER, Epipolar-plane image analysis : a technique for analyzing motion sequences, Proc. of the 3rd IEEE Workshop on Computer Vision : representation and Control, Bellaire, Michigan, Oct. 1985, pp. 168-178.

|BOU 86| P. BOUTHEMY, Determining displacement fields long contours from image sequences, Proc. Conf. Vision Interface'86, Vancouver, May 1986, pp. 350-355.

|ESP 86| B. ESPIAU, P. RIVES, Closed loop recursive estimation of 3d features for a mobile vision system, IEEE Conf. on Robotics and Automation, Raleigh, USA, March 1987.

|JAI 83| R. JAIN, Direct computation of the focus of expansion, IEEE Trans. on PAMI, vol. 5, N° 1, Jan. 1983, pp. 58-64.

|MAR 86| L. MARCE, Sur l'utilisation des séquences multi-images en robotique, Publication interne IRISA n° 293, Avril 1986.

|MOR 81| H.P. MORAVEC, obstacle avoidance and navigation in the real world by a seeing rover robot, Ph.D. Thesis, Stanford University, Sept. 1980.

|THO 84| C.E. THORPE, FIDO : vision and navigation for a robot rover, Ph.D. Thesis, Dpt Computer Science, Carnegie Mellon Univ., CMU-CS-84-168, Dec. 1984.

Tableau 1.

| *Images* | *Distance réelle* | *Distance estimée* | *Erreur relative* |
|---|---|---|---|
| u = 1 | 22.1 | 21.7 | 12% |
| | 20.1 | 22.9 | 13% |
| | 18.1 | 16.3 | 7% |
| | 16.1 | 17.8 | 9% |
| | 14.1 | 15.1 | 5% |
| u = 2 | 20.1 | 20.0 | 7% |
| | 18.1 | 17.3 | 5% |
| | 16.1 | 15.5 | 4% |
| | 14.1 | 14.9 | 4% |
| u = 3 | 18.1 | 16.7 | 6% |
| | 16.1 | 15.7 | 3% |
| | 14.1 | 13.8 | 3% |
| u = 4 | 16.1 | 15.3 | 4% |
| | 14.1 | 13.9 | 3% |
| u = 5 | 14.1 | 13.6 | 3% |

Tableau 2.

| *Images* | *Distance réelle* | *Distance estimée* | *Erreur relative* |
|---|---|---|---|
| 2 images successives | 22.1 | 21.7 | 12% |
| | 20.1 | 22.9 | 13% |
| | 18.1 | 16.3 | 7% |
| | 16.1 | 17.8 | 9% |
| | 14.1 | 15.1 | 5% |
| 3 images successives | 20.1 | 21.8 | 6% |
| | 18.1 | 17.8 | 7% |
| | 16.1 | 16.9 | 5% |
| | 14.1 | 15.1 | 5% |
| 4 images successives | 18.1 | 16.8 | 6% |
| | 16.1 | 16.7 | 4% |
| | 14.1 | 14.5 | 3% |
| 5 images successives | 16.1 | 15.9 | 3% |
| | 14.1 | 14.4 | 3% |
| 6 images successives | 14.1 | 13.9 | 3% |

Tableau 3.

| *Images* | *Distance réelle* | *Distance estimée* | *Erreur relative* |
|---|---|---|---|
| u = 1 | 22.1 | 21.4 | 32% |
| | 20.1 | 20.6 | 30% |
| | 18.1 | 16.0 | 24% |
| | 16.1 | 18.0 | 24% |
| | 14.1 | 14.6 | 16% |
| u = 2 | 20.1 | 19.5 | 17% |
| | 18.1 | 16.9 | 14% |
| | 16.1 | 15.6 | 11% |
| | 14.1 | 14.5 | 10% |
| u = 3 | 18.1 | 16.4 | 11% |
| | 16.1 | 15.5 | 9% |
| | 14.1 | 13.6 | 7% |
| u = 4 | 16.1 | 15.1 | 8% |
| | 14.1 | 13.7 | 6% |
| u = 5 | 14.1 | 13.4 | 6% |

Par erreur relative on entend le rapport de l'écart type par la distance réelle.

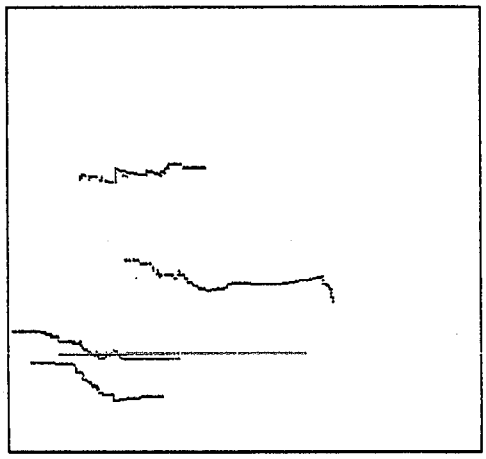

Figure 1. Suivi de 5 points d'intérêt à travers les 125 images.

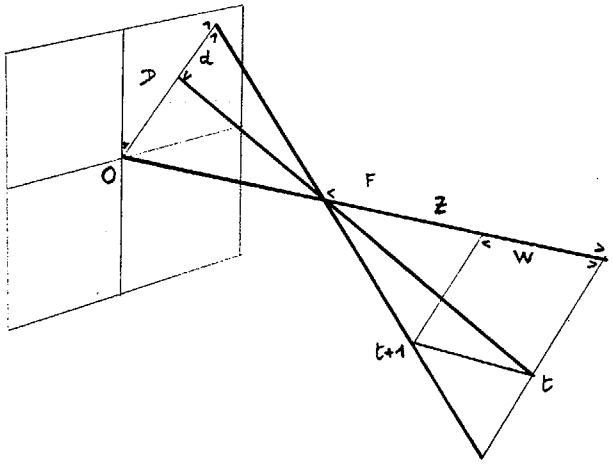


Figure 2. Déplacement d'un point image.

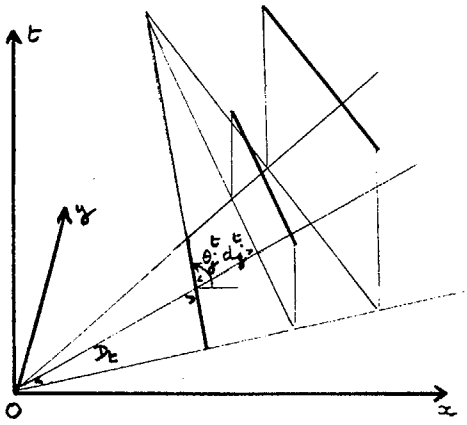


Figure 3. Modélisation d'un élément de contour spatio-temporel.