TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

Etude cinématique du mouvement humain à partir d'une séquence d'images monoculaires

Juhui WANG, †Guy LORETTE and Patrick BOUTHEMY
IRISA/INRIA
†IRISA/Université de RENNES I
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France

RÉSUMÉ

L'étude présentée dans cet article concerne l'analyse et le suivi d'un mouvement 3D à partir d'une longue séquence d'images monoculaires. La méthode repose sur une mise en évidence des modèles géométriques et cinématiques de la scène. La scène ainsi que son mouvement sont spécifiés par un ensemble de paramètres globaux. L'estimation du mouvement 3D se fait par minimisation d'une fonction de coût qui exprime le changement d'intensité lumineuse dans les images dû au mouvement. Cette méthode travaille directement sur une séquence d'images sans utilisation de primitives intermédiaires extraites de la séquence telles que des contours ou le champ de vecteurs vitesse apparente. Elle permet donc d'éviter les erreurs introduites par le calcul de ces derniers. De plus, nous généralisons la méthode développée dans une procédure de suivi du mouvement 3D dans une longue séquence en prenant en compte la variation éventuelle d'intensité lumineuse due à la présence de bruits et à d'autres causes (rotation des objets, changement d'éclairement etc...). L'expérimentation faite sur une scène de pédalage d'un cycliste montre que cette méthode est robuste et que les résultats sont prometteurs.

ABSTRACT

This study is concerned with the analysis and tracking of 3D motion from a long monocular image sequence. We propose a model-based approach in which particulars about object structures and motions in the scene are formulated as functions known up to the values of a few parameters. The 3D instantaneous motion parameters are estimated by minimizing a goodness-fitting criterion. This criterion relies on the mesurement of the variations of image intensity assumed to be induced by motion of the 3D kinematic model. Our approach works directly on the image intensity without explicitly making use of primitives extracted from sequences such as contours, optical flow etc... thus avoids errors introduced by computing the later. Moreover, taking into account possible changes in brightness due to presence of noise and others in the scene (i.e. object rotations, illumination perturbation etc...), we have embedded the method developed to estimate 3D instantaneous motion parameters in a tracking procedure over a long image sequence. Experiments carried out on a cycling motion sequence show that the approach presented here is robust and the results are promising.

1 Introduction

Dans le domaine de la vision par ordinateur, les séquences d'images numériques offrent un support naturel pour l'analyse du mouvement. L'estimation du mouvement 3D à partir d'une séquence d'images d'intensité lumineuse est un des thèmes les plus étudiés durant ces dernières années [1]. Une des méthodes classiques en ce domaine consiste à exploiter la variation spatio-temporelle de la fonction d'intensité lumineuse dans la séquence. L'algorithme proposé consiste souvent en deux étapes : calcul du champ des vecteurs de vitesse apparents en admettant l'hypothèse de l'invariance temporelle d'intensité lumineuse dans les images et interprétation du mouvement à partir de ce champ en faisant des hypothèses générales sur la scène telle que l'objet et son mouvement sont rigides, la surface de l'objet peut être décomposée en primitives géométriques simples comme par exemple des facettes planes, des cylindres, des sphères etc... Or, le calcul du champ des vecteurs de vitesse apparents ne fournit souvent pas une précision suffisante pour faire une interprétation quantitative du mouvement 3D en présence de bruit. Ceci explique en grande partie l'instabilité dont souffre la solution obtenue selon cette procédure [4].

D'autre part, le problème d'analyse du mouvement humain est difficile et il existe une littérature abondante sur ce sujet, mais aucun d'entre eux ne permet de résoudre de façon satisfaisante les problèmes. Citons l'article de [2] pour une méthode à l'aide des marqueurs spécifiques, et celui de [3] pour une tentative d'utilisation d'un modèle articulé.

Notre analyse repose sur une mise en évidence des modèles géométriques et cinématiques de la scène. La scène et son mouvement sont représentés par un modèle qui est spécifié par un ensemble de paramètres globaux. Une étape d'initialisation interactive permet de trouver les paramètres géométriques du modèle 3D de la scène au tout début de la séquence. L'estimation des paramètres du mouvement 3D se fait ensuite par minimisation d'une fonction de coût qui exprime le changement d'intensité lumineuse induit par le mouvement du modèle 3D de la scène. En outre, nous avons traité le problème du suivi du mouvement 3D dans une longue séquence d'images en introduisant la notion d'image de référence. Contrairement aux cas des courtes séquences, l'hypothèse d'invariance d'intensité lumineuse du mouvement n'est plus valide. Nous avons abordé ce problème en utilisant une procédure de prédiction-compensation qui permet la mise à jour du changement d'intensité lumineuse dû à la présence de bruits et à d'autres causes (rotations des objets, perturbation d'éclairage etc...). L'expérimentation sur une séquence réelle d'un mouvement de pédalage d'un cycliste est présentée.

2 Mouvement 3D et 2D

Dans ce paragraphe, nous expliquons la relation entre les paramètres du mouvement 3D de la scène et le déplacement 2D observé dans les images.

Soit un objet dont la géométrie dans l'espace 3D à l'instant t est définie par un vecteur de paramètres A_t et dont le mouvement 3D entre deux instants t et $t+1$ est représenté par Θ_T. Considérons deux images de la scène I_t et I_{t+1} prises respectivement aux instants t et $t+1$, nous pouvons calculer la relation entre deux points p et p^* respectivement dans I_t et I_{t+1} qui sont les projections d'un même point de la surface d'un objet de la scène dans le plan image aux instants t et $t+1$:
soit

$$p^* = u(p, \Theta_T, A_t, f)) \tag{1}$$

ici f est un vecteur qui définit les paramètres intrinsèques de la caméra, incluant la position de la caméra, le centre, la distance focale de la caméra etc...

3 Estimation du mouvement entre deux instants successifs

Avant aborder le problème de l'estimation et du suivi du mouvement 3D dans une longue séquence, nous traitons le problème de l'estimation du mouvement 3D entre deux instants successifs comme par exemple t_0 et t_1. Nous supposons que les paramètres du modèle géométrique 3D de la scène à l'instant t_0 qui spécifient la géométrie de la scène par rapport à la caméra sont connus. On suppose également que sa projection dans le plan image nous donne une segmentation préalable de l'image I_0 et que la scène est du type lambertienne, ce qui revient à faire l'hypothèse d'invariance d'intensité lumineuse sur le trajectoire du mouvement dans les images.

Soit p et p^* respectivement les projections d'un même points de la surface de l'objet dans le plan image aux instants t_0 et t_1. Etant donné le mouvement 3D de la scène Θ_T, les intensités lumineuses en ces deux points respectives $I_0(p)$ et $I_1(p^*)$ doivent vérifier l'équation suivante :

$$I_0(p) = I_1(p^*) = I_1(u(p, \Theta_T, A_0, f)) \tag{2}$$

soit sous une autre forme:

$$I_0(p) - I_1(u(p, \Theta_T, A_0, f)) = 0 \tag{3}$$

La somme pondérée des termes de la partie gauche de l'équation 3 pour tous les points de la région D, projection du modèle géométrique 3D de l'objet dans le plan image à l'instant t_0, définit une fonction de coût qui exprime la contrainte sur l'intensité lumineuse entre les images due au mouvement de l'objet :
soit

$$\Phi(\Theta) = \sum_{p \in D} w_p \cdot [I_0(p) - I_1(u(p, \Theta, A_0, f))]^2 \tag{4}$$

dans le cas idéal, cette fonction est nulle lorsque le mouvement estimé est le mouvement réel ($\Theta = \Theta_T$). En pratique, cette fonction est toujours positive et atteind son minimum à $\Theta = \Theta_T$. Le problème de l'estimation des paramètres du mouvement 3D revient donc à rechercher les valeurs de Θ qui minimisent la fonction Φ.

Les méthodes pour résoudre ce problème sont classiques, nous utilisons ici une méthode itérative qui consiste à chercher le minimum de la fonction Φ dans la direction opposée à celle du gradient de Φ.
soit

$$\Theta^{n+1} = \Theta^n - \tau \cdot \nabla\Phi(\Theta^n) \cdot \Phi(\Theta^n) \tag{5}$$

ici $\nabla\Phi$ est le gradient de Φ en fonction de Θ ;
τ est le gain de correction.

En ce qui concerne la mise en œuvre de cet algorithme, quelques explications sont nécessaires.

- le choix de la valeur du coefficient de pondération w_p dans (4) s'appuie sur le fait que aux endroits où la fonction d'intensité lumineuse varie brutalement par rapport aux paramètres estimés, la correction des paramètres de mouvement 3D cherchés Θ doit être prudente et donc que la contribution du point à la fonction de coût doit être plus petite. C'est pour cela que nous avons pris comme valeur de w_p l'inverse de la norme du gradient par rapport à Θ . En pratique, pour éviter le passage éventuel à zéro de cette norme, nous prendrons la valeur $\frac{1}{c^2 + \nabla_\Theta I_0^t \cdot \nabla_\Theta I_0}$.

- Coefficient τ : pour s'affranchir partiellement du problème des minima locaux rencontré dans la méthode du gradient et obtenir une vitesse de convergence plus grande, nous utilisons un gain variable. τ est initialisé avec une valeur relativement grande et est multiplié périodiquement par un facteur k ($k < 1$).

- Choix des observations : les points de la région D dont les intensités lumineuses sont trop bruitées pour donner une information fiable sur les gradients ne sont pas pris en compte dans l'estimation des paramètres du mouvement 3D. Pour effectuer ce choix, nous considérons la distribution des gradients spatio-temporels à un point. Nous faisons l'hypothèse que l'information portée par les gradients spatio-temporels est fiable tant que la probabilité, pour que le gradient calculé en ce point n'est pas entièrement perturbé par un bruit, est supérieure à un seuil λ ($\lambda = 0.6$). Dans le cas du bruit gaussien indépendant centré, ce critère se traduit par l'ensemble des tests ci-dessous :

  $$|\nabla_s I_0| > \sigma_s \tag{6}$$

  ici $s = \{x, y, t\}$ et σ_s est la variance de $\nabla_s I_0$.

  En pratique, nous avons également rejeté les points qui sont dans la zone d'occlusion entre l'objet et le fond de la scène. Ce qui se traduit aussi par un test sur la diffirence inter-image déplacée prédite à partir de Θ^n :
  soit

  $$|I_0(p) - I_1(u(p, \Theta^n, A_0, f))| > Seuil \tag{7}$$

4 Suivi du mouvement 3D

Le suivi du mouvement 3D consiste en trois phases : détection des objets à suivre dans la première image, estimation du mouvement 3D de l'objet et mise à jour de l'objet à suivre selon les paramètres du mouvement 3D estimé. Le suivi du mouvement 3D dans une séquence revient à répéter les deux dernières phases.

Dans notre application, nous considérons l'ensemble des objets mobiles dans la scène comme objets à suivre. La localisation de ces objets à l'instant t_0 est obtenue par une mise en correspondance interactive entre le modèle 3D de la scène et l'image de la scène prise à l'instant t_0. A la sortie de ce module, nous avons un vecteur de paramètres géométriques A_0 qui spécifie la position et la géométrie du modèle 3D de la scène que nous supposons en coïncidence avec la scène réelle à l'instant t_0. La projection de ce modèle géométrique 3D dans le plan image donne une segmentation de l'image I_0 que l'on note D et l'association à D de l'intensité lumineuse dans I_0 nous donne une image d'intensité lumineuse du modèle 3D de la scène à l'instant t_0 notée S_0. Dans le cas général, l'image du modèle 3D de la scène à l'instant t, appelée *image de référence*, sera notée par S_t.

Après la phase d'initialisation, nous pouvons présenter notre analyse à l'instant $t+1$ sans perdre la moindre généralité. La phase d'estimation prend comme entrées: les paramètres géométriques du modèle 3D de la scène A_t et l'image S_t. Sous l'hypothèse d'invariance lumineuse du mouvement, l'estimation du mouvement 3D de la scène entre t et $t+1$ se fait en minimisant la fonction de coût à l'aide de la méthode développée au chapitre 3.
soit

$$\Phi(\Theta) = \sum_{p \in D} w_p \cdot [S_t(p) - I_{t+1}(u(p, \Theta, A_t, f)]^2 \qquad (8)$$

La mise à jour de l'objet à suivre consiste à évaluer les variations du modèle géométrique 3D de la scène c'est à dire de A_t et de l'image de référence S_t. La mise à jour de A_t est facile en utilisant les techniques de calcul de la nouvelle position 3D de l'objet à partir de son mouvement 3D, le lecteur se reportera à [6]. Nous nous intéressons ici à la mise à jour de l'image S_t. A cause de la présence de bruits et d'erreurs d'estimation, la mise à jour de S_t à partir du mouvement 3D estimé Θ est difficile. D'une part, nous devons prendre en compte le changement de la distribution d'intensité dans l'image dû au mouvement de la scène et de la variation d'éclairement durant une longue séquence. D'autre part, nous devons aussi prendre en compte le décalage entre la position du modèle géométrique 3D de la scène à l'instant $t+1$ prédite à partir du mouvement 3D estimé et la position 3D réelle de l'objet à l'instant $t+1$.

Nous pouvons envisager deux solutions pour la mise à jour de S_t. La première consiste à calculer la projection du modèle géométrique 3D de la scène dans le plan image à l'instant $t+1$, notée D et prendre D comme la position de S_{t+1} dans le plan image et I_{t+1} comme l'intensité lumineuse de S_{t+1} malgré le décalage entre la position du modèle 3D de la scène prédite et la position réelle de la scène à l'instant $t+1$. Le test que nous avons fait sur cette méthode montre que l'objet à suivre est très vite perdu à cause de l'approximation faite sur le modèle 3D. La seconde méthode repose sur l'utilisation des propriétés photométriques de la scène. Si nous possèdons des connaissances parfaites sur les conditions d'éclairement et les propriétés réfléchissantes de la scène, nous pouvons très bien synthétiser l'image S_{t+1} à partir du modèle géométrique et photométrique de la scène. Or la formation de l'image est une procédure complexe, les hypothèses sur les propriétés photométriques de la scène faites par les techniques existantes ne permettent pas de traiter une scène réelle. Dans cette étude, nous utilisons une procédure de prédiction-compensation pour la mise à jour de l'image S_t. Cette procédure ne demande aucune connaissance sur les propriétés photométriques de la scène. Elle consiste en deux étapes. Dans la première étape, en faisant l'hypothèse d'invariance d'intensité lumineuse sur le trajectoire du mouvement dans les images, S_t se transforme en une image $\hat{S}_{t+1}$ selon les paramètres du mouvement estimé Θ. En fait, la segmentation de $\hat{S}_{t+1}$ est la projection du modèle 3D de l'objet dans le plan image à l'instant $t+1$ et l'intensité lumineuse définie par S_t :
soit

$$\hat{S}_{t+1}(p^*) = S_t(p) \qquad (9)$$

ici p et p^* ont la même signfication que précédemment, $\hat{S}_{t+1}(p^*)$ est l'intensité lumineuse à valeur en point p^* dans l'image $\hat{S}_{t+1}$ et $S_t(p)$ est celle en point p dans l'image S_t

Si l'hypothèse d'invariance lumineuse est vérifiée, nous pouvons prendre $\hat{S}_{t+1}$ comme la mise à jour de l'image S_t à l'instant $t+1$ c'est à dire l'image S_{t+1} et appliquer à nouveau l'algorithme d'estimation du mouvement 3D sur S_{t+1} et I_{t+2}. Malheureusement, dans le cas général on observe toujours un changement "non aléatoire" de la fonction d'intensité lumineuse de la scène à cause de la variation d'éclairement et du mouvement de rotation des objets dans la scène. On utilise donc une deuxième procédure qui permet de compenser ce changement. On calcule d'abord une carte de disparité entre $\hat{S}_{t+1}$ et I_{t+1} selon la méthode développée par Walker et Rao [5] et $\hat{S}_{t+1}$ sera ensuite compensée par I_{t+1} à l'aide de cette carte de disparité pour obtenir S_{t+1} de la façon suivante. Soit δp le vecteur de la disparité obtenu en un point p dans $\hat{S}_{t+1}$, l'intensité de S_{t+1} en p est donnée par $S_{t+1}(p) = I_{t+1}(p + \delta p)$.

La figure 1 montre un schéme synoptique de l'algorithme de l'estimation et du suivi du mouvement 3D. Nous constatons que l'algorithme d'estimation ne s'applique tout en vigueur que dans le cas où l'hypothèse d'invariance de l'intensité lumineuse sur la trajectoire du mouvement dans S_t et I_{t+1} est vérifiée. Or, cette hypothèse n'ést jamais parfaitement valide pour une scène réelle, nous observons toujours une variationd'intensité due au bruits ou à autres causes, la procédure de mise à jour de S_t a alors objectif de compenser cette variation.

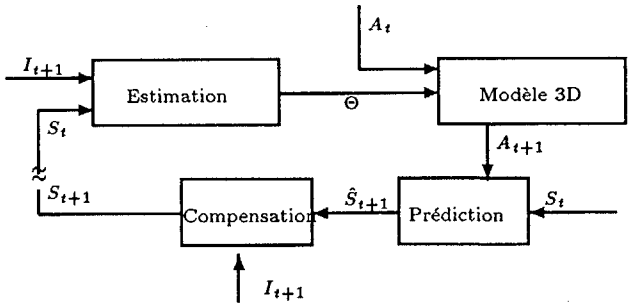

Figure 1: *Scéma synoptique d'analyse*

5 Expérimentation sur une séquence de pédalage

Nous avons testé notre algorithme sur une scène réelle constituée d'un mouvement de pédalage d'un cycliste sur un vélo d'appartement (voir figure 2).

Le choix de ce geste est justifié par ses aspects maîtrisables (sans occlusion ni grand mouvement non-rigide) et représentatifs pour une grande partie des gestes humains. Notre

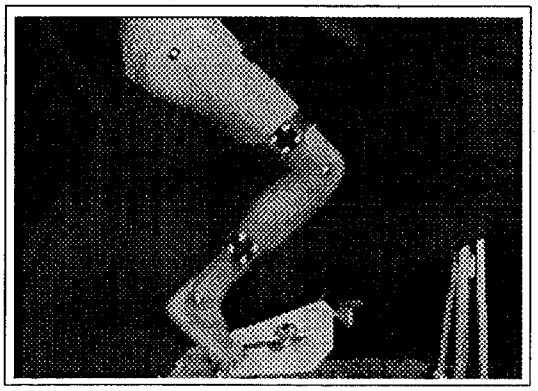

Figure 2: *Une image extraite de la séquence*

objectif est d'estimer le mouvement 3D de la jambe et du mollet. Pour la simplicité et la régularité du système, nous avons modélisé le corps humain par un système d'objets articulés. Chaque partie de la jambe est modélisée par un cylindre et chaque cylindre est défini par un ensemble des paramètres globaux qui sont le rayon, les deux points situés aux extrémités de l'axe de la cylindre. Comme le montre la figure 3, les paramètres du modèle géométriques 3D de cette scène sont $A = (C_1, C_2, C_3, r_h, r_g)$, C_1, C_2, C_3 étant les centres de rotation au niveau de la hanche, du genou et du pied et r_h, r_g les rayons des cylindres représentant la jambe et le mollet.

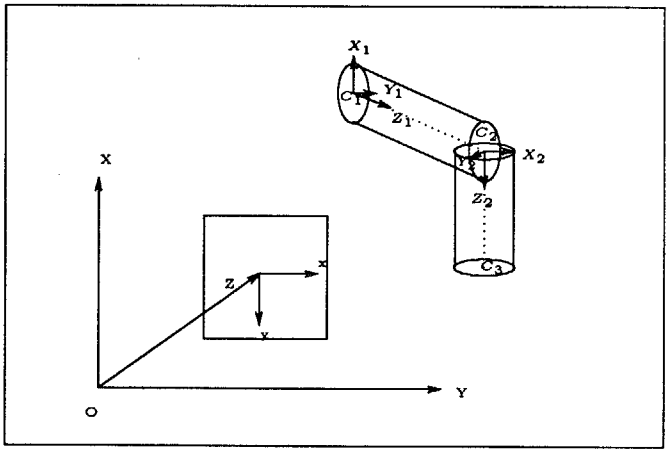

Figure 3: *Modèle du mouvement de pédalage*

Quant à la modélisation cinématique de ce geste, elle consiste à supposer que le centre de rotation C_1 lié à la hanche est fixe et à modéliser le mouvement par trois rotations autour de la hanche et une rotation autour du genou[6]. Les trois rotations au niveau de la hanche sont définies par trois angles $\theta_1, \theta_2, \theta_3$ qui sont respectivement les angles de rotation de la jambe autour des axes x, y, z d'un repère local R_1 lié à la jambe. Le mouvement du mollet est spécifié par une rotation autour de l'axe du genou, ce qui est représenté par un angle de rotation θ_4 autour de l'axe y du repère R_2 qui est un repère local lié au mollet. Ce qui nous donne le vecteur de paramètres du mouvement 3D à estimer $\Theta = (\theta_1, \theta_2, \theta_3, \theta_4)$.

Les résultats obtenus sont donnés dans la figure 4. On remarque que le choix des paramètres dans cet algorithme n'est pas crucial. Les paramètres du modèle géométrique 3D ont été initialisés par une mise en correspondance interactive entre le modèle 3D de la scène et la première image de la séquence, les paramètres du mouvement 3D à l'instant $t + 1$ par ceux obtenus à l'instant t. Pour que la recherche du minimum de la fonction de coût Φ ne soit pas bloquée trop tôt dans un minimum local, nous avons utilisé un gain de correction τ variable : τ a été initialisé avec une valeur relativement grande (0.03) et multiplié par un facteur k (k = 0.86) toutes les 10 itérations. La figure 4 montre les résultats obtenus pour θ_1, θ_2, θ_3, θ_4. Nous voyons bien que le mouvement de pédalage n'est pas un mouvement parfaitement plan. Nous signalons également que le mouvement synthétisé selon les paramètres du mouvement 3D estimé se superpose sur la séquence originale avec une précision tout à fait satisfaisante.

6 Conclusions

L'étude présenté dans cet article concerne l'analyse et le suivi du mouvement du corps humain à partir d'une longue séquence d'images monoculaires. Nous avons montré comment les connaissances a priori sur la scène peuvent être formulées sous forme de représentation mathématique et utilisé dans la reconstruction du mouvement 3D. La méthode développée travaille directement sur les images d'intensité sans utilisation de primitivess intermédiaires extraites de la séquence. Elle permet donc éviter les erreurs introduites par le calcul de ces derniers et donne des résultats plus robustes.

Remerciements: Cette étude est financée par la région Bretagne dans le contexte du projet CBI, contrat No. 189c2710031315061.

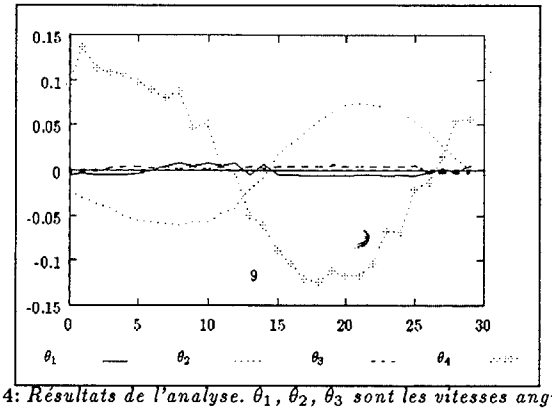

Figure 4: *Résultats de l'analyse. θ_1, θ_2, θ_3 sont les vitesses angulaires de rotation de la jambe autour des axes X_1, Y_1, Z_1 et θ_4 celle du mollet autour de l'axe Y_2*

Bibliographie

[1] J.K. Aggarwal and N. Nandhakumar. On the computation of motion from sequences of images: a review. *Proceedings of the IEEE*, 76(8):917–935, August 1988.

[2] E.K. Antonsson and R.W. Mann. Automatic 6-D.O.F. kinematic trajectory acquisition and analysis. *Journal of Dynamic Systems, Measurement, and Control*, 111:31–39, March 1989.

[3] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, February 1983.

[4] B.K.P Horn and E.J Weldon Jr. Direct methods for recovering motion. *International Journal of Computer Vision*, 2:51–76, 1988.

[5] D.R. Walker and K.R. Rao. Improved pel-recursive motion compensation. *IEEE Transactions on Communications*, 32(10):1128–1134, October 1984.

[6] J. Wang, G. Lorette, and P. Bouthemy. *A Kinematic Study of Cycling Motion of A Cyclist from Long Monocular Image Sequences.* Technical Report, IRISA/INRIA-Rennes, France, 1991. To appear.