Recalage d'un modèle 3D générique sur une séquence d'images 2D

Mikaël Bourges-Sévenier **Patrick Horain**
Françoise Prêteux

Département Signal et Image, Institut National des Télécommunications
9, rue Charles Fourier – 91011 Evry Cedex

Pascal Leray

CNET / DIH, Laboratoires du CCETT
4, rue Clos Courtel – BP 59 – 35512 Cesson-Sévigné Cedex

Résumé :

Nous proposons une méthode permettant de recaler un modèle générique 3D maillé sur un objet présent dans une image et d'en suivre les déplacements au cours de la séquence. L'application envisagée est le suivi d'un visage quel qu'il soit et qu'elles qu'en soient les attitudes dans des séquences d'images de type visiophonie comportant notamment des fonds complexes.

1 Introduction

Cette étude s'inscrit dans le cadre d'une recherche sur la *"Modélisation et représentation hiérarchiques de scènes 3D pour des services multimédia"* associant les laboratoires SYNTIM de l'INRIA, TEMIS de l'IRISA et SIM de l'INT. Les aspects de modélisation géométrique et de mouvement étant abordés respectivement par les deux autres partenaires, le département SIM traite de la modélisation générique d'un objet, de son suivi et de sa déformation dans des séquences d'images sans connaissance a priori sur celles-ci.

Cette communication présente la première partie de l'étude consacrée au recalage d'un modèle générique dans les images et à son suivi à l'ordre zéro au cours de la séquence. Le principe de la méthode proposée consiste à disposer d'un modèle générique 3D maillé de tête, à extraire les contours du visage pour l'image considérée de la séquence, puis à ajuster les limbes projetés du modèle 3D de la tête sur les contours dans l'image.

Le domaine de la reconnaissance du visage et du suivi de ses déformations dans une séquence d'images fait l'objet d'actives recherches et de nombreuses publications. Pour des applications des finalités différentes, allant de la compression de séquences de type visiophonie [GOS97, MM97, WHWV97] à la reconnaissance d'objets 3D [Pen87, SP94, SB90], au suivi de déformations et au recalage 3D [Low91], les principales méthodes proposées relèvent des approches markoviennes [BY95, Bre96, NKH97], de descriptions géométriques par morphologie mathématique [GOS97, MM97, WHWV97], d'apprentissage et reconnaissance par réseaux de neurones [IN97], de formulation énergétique par surfaces actives [TWK87, TM91, Nas94, BN94, YPC92], par déformations de formes libres ou par superquadriques [Bar95].

Pour des raisons de généricité et en l'absence de connaissance a priori, sur les séquences à traiter, nous avons adopté une approche déterministe et géométrique faisant coopérer marquage morphologique et recalage 3D par projection des limbes du modèle.

Dans la première partie de cet article, les contraintes géométriques que le maillage du modèle générique 3D doit satisfaire sont tout d'abord explicitées. La définition d'un limbe est ensuite rappelée, les propriétés des limbes projetés énoncées et la procédure algorithmique d'extraction de ceux-ci en tenant compte des parties cachées décrite. La deuxième partie est consacrée à l'extraction des contraintes spatiales dans l'image. Les contours sont tout d'abord détectés par un opérateur de type Canny-Deriche. Ensuite, un marquage morphologique des yeux et de la bouche est réalisé automatiquement à partir d'un coût de connexion. Enfin, à partir des contours et des marqueurs obtenus, nous définissons une fonctionnelle de coût visant à conditionner l'étape d'ajustement du modèle 3D projeté. Celle-ci, présentée dans la troisième partie s'appuie sur la méthode de résolution de Levenberg-Marquardt [PTVF92, Low91]. Dans la dernière partie, nous présentons et discutons, en terme de robustesse, les résultats obtenus par cette approche sur les séquences `carphone` et `foreman`.

2 Modèle 3D projeté : limbes et contours occultants

Le modèle générique 3D de tête est représenté sous forme d'un maillage triangulé. Les yeux, la bouche et la base du cou correspondent à des zones ouvertes du maillage (figure 1).

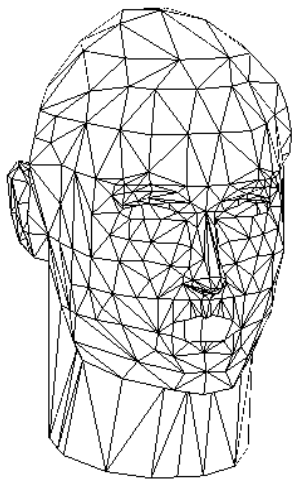

FIG. 1 – *Modèle générique de tête : maillage triangulé (680 facettes).*

Les propriétés suivantes sont imposées au maillage :

- les facettes sont planes (pas de triangles gauches),
- une arête appartient soit à une facette (bord terminal), soit à deux facettes,
- plusieurs facettes ne peuvent pas se raccorder en un point intérieur à une arête (*i.e.* pas de sommet en T) (figure 2),

– la topologie du maillage est quelconque : simplement connexe ou non, localement convexe ou non,

– à chaque facette, on associe une normale qui pointe vers l'extérieur de l'objet.

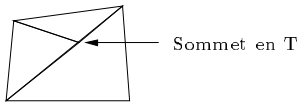

FIG. 2 – *Exemple de configuration interdite (sommet en T).*

Le recalage d'un modèle générique 3D sur l'objet correspondant dans une image 2D s'effectue par projection orthographique (projection parallèle) ou perspective (donnée par la ligne joignant le point focal et le point considéré sur la surface). Pour une direction d'observation donnée, les points de la surface (du modèle) où la normale extérieure forme un produit scalaire positif avec la direction d'observation ne sont pas vus. Si ce produit scalaire est négatif, ces points sont vus pourvu qu'ils ne soient pas cachés par une autre partie de l'objet.

On appelle limbe de l'objet, une ligne de la surface où le produit scalaire précédemment défini change de signe. Sur une surface G^1 continue, les limbes sont des lignes où la direction d'observation est tangente à la surface [PC87]. Sur un maillage triangulé, les limbes sont donc les arêtes communes à deux facettes adjacentes, telles que les produits scalaires des normales aux facettes avec la direction d'observation soient de signes contraires, ce qui s'exprime par :

Soient deux facettes $\mathcal{F}_1$ et $\mathcal{F}_2$ et $\vec{n}_1$ et $\vec{n}_2$ leurs normales associées. Soit $\vec{v}$ la direction d'observation. Si on a :

$$< \vec{v}, \vec{n}_1 > . < \vec{v}, \vec{n}_2 > \leq 0,$$

alors le segment commun aux deux facettes est un limbe de l'objet.

Bien sûr, une formulation équivalente peut être énoncée non plus à partir de facettes mais en considérant deux sommets adjacents [Chi87].

En projection, les limbes forment donc un sur-ensemble de la silhouette : sur un visage observé de 3/4, le nez et les oreilles peuvent générer des limbes qui ne font pas partie de la silhouette. Pour un modèle 3D triangulé, les limbes sont constitués d'arêtes et se projettent sous forme de segments de droite.

Comme dans le cas d'un objet non-convexe, les limbes peuvent être cachés par d'autres parties de la surface, nous proposons de sélectionner les limbes vus en éliminant les limbes occultés par un algorithme de Z-buffer modifié. La première étape consiste à projeter tous les limbes dans le Z-buffer en identifiant l'arête dont ils proviennent. Ensuite, les triangles orientés vers l'observateur ($< \vec{v}, \vec{n} > \leq 0$) sont projetés dans le Z-buffer. Les points de limbes à l'intérieur du triangle projeté sont supprimés lorsqu'ils sont à une profondeur plus grande que celle du point correspondant dans le triangle. Les arêtes restantes dans le Z-buffer forment l'ensemble des limbes vus. Par simplification, les arêtes partiellement vues sont considérées comme entièrement vues.

Les contours terminaux de la surface, qui correspondent ici aux bords des trous formés par la bouche, les narines, les yeux et la base du cou sont rajoutés aux limbes pour former l'ensemble des contours occultants (figure 3).

Les contours occultants proviennent des arêtes du modèle 3D triangulé, et se projettent dans l'image sous forme de segments de droite.

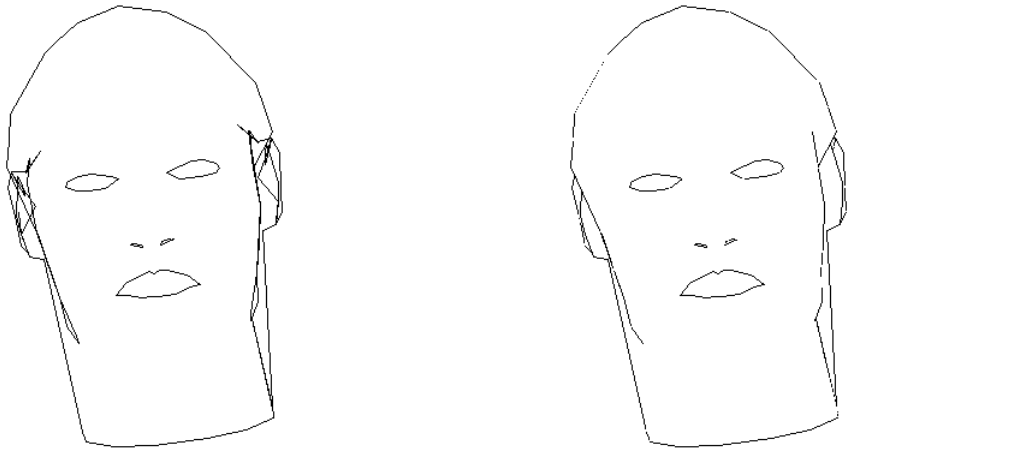

Image des limbes dans le Z-buffer

Image des limbes après élimination des parties cachées

FIG. 3 – *Contours occultants avant et après élimination des parties cachées par Z-buffer modifié.*

3 Contraintes spatiales : contours, marqueurs, cartes de distance

Dans cette étape, il s'agit d'extraire des images 2D de la séquence, les éléments d'information pertinents pour guider la phase finale d'ajustement du modèle 3D sur l'image. Comme les contours occultants du modèle correspondent à des discontinuités dans l'image, nous extrayons dans un premier temps les contours 2D à partir de l'opérateur de Canny-Deriche [Der90], filtré par hystérésis.

Dans un second temps, nous procédons par marquage morphologique à la localisation d'éléments caractéristiques du visage tels que les yeux, le nez et la bouche. Quels que soient le contexte et l'attitude, ces éléments se présentent comme des zones sombres localement contrastées avec l'environnement. Pour leur reconnaissance, nous calculons le coût de connexion par rapport aux bords de l'image 2D filtrée par ouverture [Pre91]. Par différence du coût de connexion avec l'image initiale puis filtrage du résultat, on obtient les marqueurs des yeux, du nez et de la bouche.

Pour exploiter l'ensemble des déformations spatiales ainsi extraites, nous construisons les cartes de distance $\mathcal{D}$ [Bor86] associées aux contours et aux marqueurs.

Les contours occultants et les marqueurs du modèle seront projetés de manière à minimiser la fonctionnelle de coût définie à partir des cartes de distance $\mathcal{D}$ par :

$$F = \alpha \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathcal{D}(x_j) + \beta \sum_{l=1}^{p} \mathcal{D}(m_l)$$

où m est le nombre d'arêtes du modèle formant les contours occultants, n_i le nombre de pixels x_j formant la projection de l'arête i, p le nombre de points caractéristiques du modèle (3 dans le cas des yeux et de la bouche), m_l leur projection dans l'image. Les coefficients α et β servent à pondérer l'influence des contours occultants par rapport aux informations de localisation des marqueurs.

4 Ajustement

Le principe de l'ajustement consiste à minimiser conjointement la distance entre les contours occultants du modèle projetés et les lieux des discontinuités dans l'image ainsi qu'entre les points caractéristiques identifiés sur le modèle et les marqueurs morphologiques localisés dans l'image. Les paramètres utilisés pour la minimisation sont une mise à l'échelle, une translation 3D et une rotation 3D (soit 7 scalaires). Cette minimisation détermine de nouveaux paramètres de transformation.

Le modèle est placé dans son repère propre. Les points dans l'espace sont repérés par leurs coordonnées homogènes. Soit un point $\vec{x}$ du modèle. Les coordonnées de ce point dans le repère absolu sont :

$$\vec{x}_a = \mathbf{T_a}\mathbf{R_a}\mathbf{S_a}\vec{x},$$

et dans le repère de la caméra :

$$\vec{x}_c = \mathbf{T_c}\mathbf{R_c}\vec{x}_a,$$

où $\mathbf{S}$, $\mathbf{T}$ et $\mathbf{R}$ sont respectivement les matrices d'affinité, de translation et de rotation. Cette dernière est représentée par un quaternion. Les coordonnées du point dans l'image sont alors :

$$\vec{x}_I = \mathbf{P}\vec{x}_c$$

où $\mathbf{P}$ est la matrice de projection orthographique ou projective de la caméra. Nous choisissons un modèle orthographique tel que l'axe de visée de la caméra est aligné avec l'axe des z négatifs.

L'expression des coordonnées d'un point (de contour occultant) du modèle projeté dans l'image est alors :

$$\vec{x}_I = \mathbf{P}\mathbf{T_c}\mathbf{R_c}\mathbf{T_a}\mathbf{R_a}\mathbf{S_a}\vec{x}.$$

Notons par q le vecteur des 7 paramètres du modèle, par $\vec{x}_I(q)$ la projection dans l'image d'un point du modèle et par $\mathcal{D}(\vec{x}_I(q))$ la distance fournie par la carte des distances.

L'ajustement est réalisé en appliquant la méthode itérative de Levenberg-Marquardt [PTVF92, Mor77, Low91] dans laquelle le jacobien J de la fonctionnelle F à minimiser par rapport aux paramètres q s'écrit sous la forme suivante :

$$\begin{aligned} J &= \vec{\nabla}_q(F) \\ &= \alpha \sum_i \sum_j \vec{\nabla}_{x_j}\mathcal{D}.\vec{\nabla}_q x_j + \beta \sum_l \vec{\nabla}_{m_l}\mathcal{D}.\vec{\nabla}_q m_l. \end{aligned}$$

Le terme $\vec{\nabla}_{\vec{x}_I}\mathcal{D}$ est le gradient de la transformation distance qui peut être évalué par un gradient numérique sur la carte des distances, avec un filtre de Canny-Deriche. Le terme $\vec{\nabla}_q\vec{x}_I$ est le gradient de la position du point projeté dérivé par rapport aux paramètres du modèle.

Finalement, les 7 paramètres sont combinés à la matrice de transformation du modèle original $\mathbf{S_a}\mathbf{R_a}\mathbf{T_a}$ de telle sorte que l'expression d'un point du modèle dans le repère absolu devient :

$$\vec{x}_a = \mathbf{T_a}\mathbf{R_a}\mathbf{S_a}\mathbf{T_q}\mathbf{R_q}\mathbf{S_q}\vec{x}$$

où $\mathbf{S_q}$, $\mathbf{R_q}$, $\mathbf{T_q}$ sont les matrices d'échelle, rotation et translation (resp.) associées aux 7 paramètres de q.

La nouvelle pose du modèle ainsi obtenue sert d'initialisation pour l'image suivante et le processus recommence. L'initialisation est effectuée manuellement sur la première image.

5 Résulats et discussion

Les paramètres qui interviennent dans la méthode sont :

- les seuils d'hystérésis appliqués sur les discontinuités de l'image,
- le critère de filtrage des marqueurs morphologiques,
- les coefficients de pondération de la fonctionnelle de coût.

Les premiers ont été ajustés de façon à sélectionner les discontinuités les plus marquées et donc les plus significatives (figure 4). Le critère de sélection des marqueurs morphologiques exploite une information de taille au travers d'un filtrage alterné séquentiel d'ordre 1 et de contraste par multiplication avec l'image du coût de connexion. Il est donc auto-adaptatif et automatisé. Les coefficients α et β ont été expérimentalement ajustés à $1/n$ et $1/p$. L'ajustement obtenu sur les séquences "carphone" et "foreman" est présenté figure 5.

"carphone" "foreman"

FIG. 4 – *Image des contours utilisés pour "carphone" et "foreman".*

D'une façon générale, nous avons expérimentalement testé la robustesse de l'approche proposée, en l'appliquant à des séquences d'images vraiment différentes et présentant à chaque fois des environnements complexes. Il en ressort que la principale difficulté réside dans la détection des discontinuités dans l'image. En effet, des contours significatifs peuvent ne pas du tout correspondre aux contours du visage et donc perturber l'ajustement. Dans ce cas, il convient de donner un poids prépondérant au terme de distance dérivé du marquage morphologique. Nous recherchons actuellement un critère de pondération automatique et contextuellement adaptatif.

Références

[Bar95] Bardinet E. – *Modèles déformables contraints: applications à l'imagerie cardiaque.* – Thèse de Doctorat, Univ. Paris IX - Dauphine, décembre 1995.

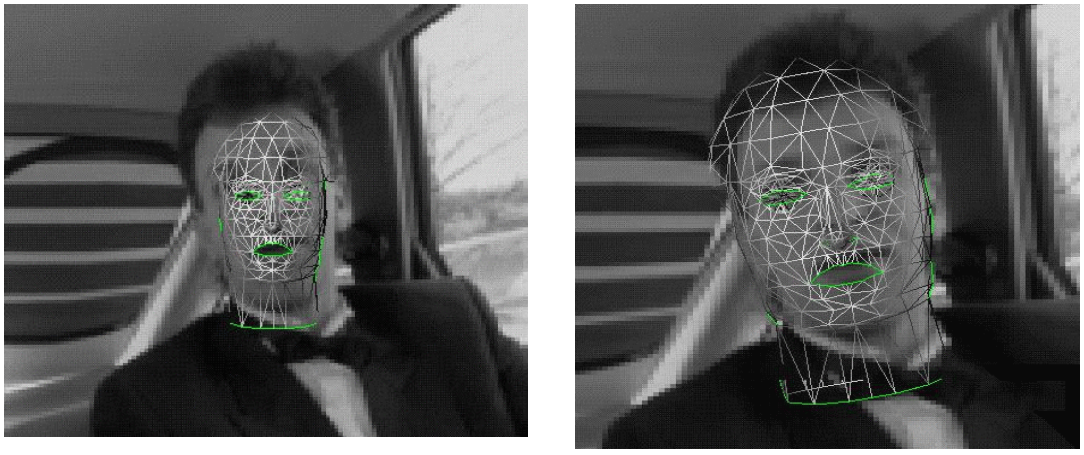

Image de carphone et initialisation du modèle

Ajustement réalisé (zoom)

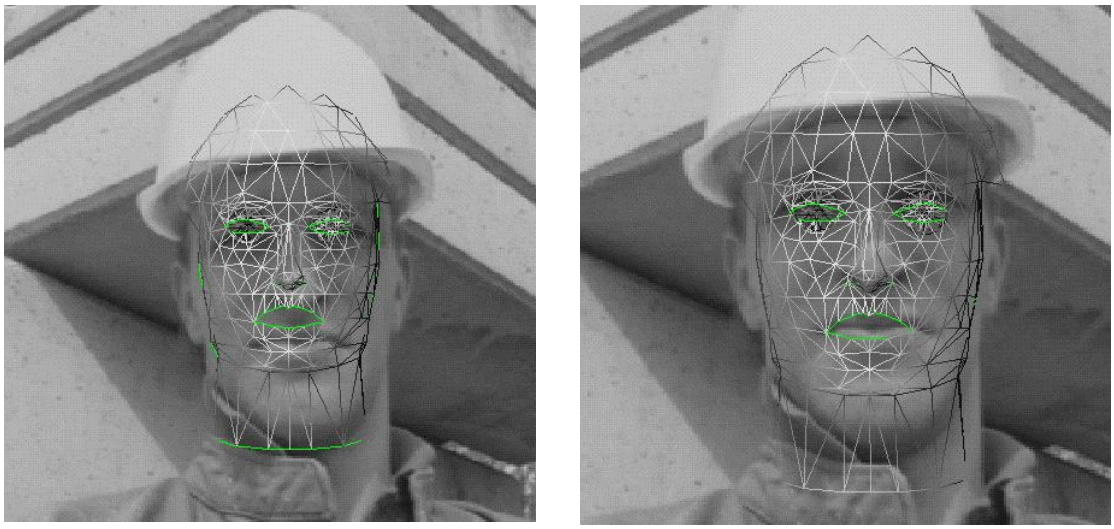

Image de foreman et initialisation du modèle

Ajustement réalisé (zoom)

FIG. 5 – *Exemple d'ajustement sur les séquences "carphone" et "foreman".*

[BN94] Bro-Nielsen M. – *Active nets and cubes.* – Rapport technique, Institute Of mathematical modelling, Tech. Univ. of Denmark, Bldg. 321, DK-2800, Lyngby, Denmark, novembre 1994.

[Bor86] Borgefors G. – Distance transformations in digital images. *Computer Vision, Graphics and Image Processing*, vol. 34, 1986, pp. 344–371.

[Bre96] Bregler C. – *Computer vision in man-machine interfaces*, chap. 7, Probalistic models of verbal and body gestures. – Cambridge University Press, 1996.

[BY95] Black M.J. et Yacoob Y. – *Tracking and recognizing facial expressions in image sequences, using local parametrized models of image motion.* – Rapport technique n° CAR-TR-756, Computer vision laboratory – Center for automation research, janvier 1995.

[Chi87] Chien C.H. – *Reconstruction and recognition of 3-D objects from occluding contours and silhouettes.* – Thèse de Doctorat, Dept. Elec Comput. Eng., Univ. Texas, Austin, 1987.

[Der90] Deriche R. – Fast algorithms for low-level vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, n° 1, janvier 1990, pp. 78–87.

[GOS97] Garrido L., Oliveras A. et Salembier P. – Motion analysis of image sequences using connected operators. *In : SPIE*, pp. 546–557.

[IN97] Iwata H. et Nagahashi H. – Motion tracking of color image sequences using neural networks. *In : SPIE*, pp. 200–210.

[Low91] Lowe D. – Fitting parametrized three-dimensional models to images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, n° 5, mai 1991, pp. 441–450.

[MM97] Marques F. et Molina C. – Object tracking for content-based functionalities. *In : SPIE*, pp. 190–199.

[Mor77] Moré J.J. – Numerical analysis. *In : Lectures notes in mathematics*, éd. par Watson G.A., pp. 105–116. – Springer-Verlag, 1977.

[Nas94] Nastar C. – *Modèles physiques déformables et modes vibratoires pour l'analyse du mouvement non-rigide dans les images multidimensionnelles.* – Thèse de Doctorat, École Nationale des Ponts et Chaussées, juillet 1994.

[NKH97] Nefian A.V., Khosravi M. et Hayes M.H. – Real-time detection of human faces in uncontrolled environments. *In : SPIE*, pp. 211–219.

[PC87] Ponce J. et Chelberg D. – Finding the limbs and cups of generalized cylinders. *International Journal of Computer Vision*, vol. 1, 1987, pp. 195–210.

[Pen87] Pentland A.P. – Recognition by parts. *Proc. IEEE International Conference on Computer Vision*, 1987, pp. 612–620.

[Pre91] Preteux F. – *On a distance function approach for gray-level mathematical morphology*, chap. 10. – Marcel Drekker Inc., 1991.

[PTVF92] Press W.H., Teutolsky S.A., Vetterling W.T. et Flannery B.P. – *Numerical Recipes in C – Second Edition.* – Cambridge University Press, 1992.

[SB90] Solina F. et Bajcsy R. – Recovery of parametric models from range images: the case for superquadrics with global deformations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, n° 2, février 1990, pp. 131–147.

[SP94] Sclaroff S. et Pentland A.P. – *Search by shape examples : modeling nonrigid deformation.* – Rapport technique n° TR94-015, Boston university computer science dept., 1994.

[TM91] Terzopoulos D. et Metaxas D. – Dynamic 3d models with local and global deformations: deformable superquadrics. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, n° 7, juillet 1991, pp. 703–714.

[TWK87] Terzopoulos D., Witkin A. et Kass M. – Symmetry-seeking models and 3d object recognition. *International Journal of Computer Vision*, vol. 1, 1987, pp. 211–221.

[WHWV97] Wang D., Haighton P., Wang L. et Vincent A. – Motion estimation using segmentation and consistency constraint. *In : SPIE*, pp. 568–577.

[YPC92] Yuille A., P.H. et Cohen D. – Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, vol. 8, n° 2, 1992, pp. 99–111.