Suivi robuste d'objets en temps-réel : une approche hybride 2D-3D

Éric Marchand[1], Patrick Bouthemy[1], François Chaumette[1], Valérie Moreau[2]

[1] IRISA / INRIA Rennes
Campus de Beaulieu, 35042 Rennes Cedex

[2] EDF - Pôle Industrie- Division Recherche et Développement
6, Quai Watier, 78401 Chatou Cedex

Eric.Marchand@irisa.fr

Résumé – Dans cet article, nous présentons une méthode originale de suivi d'objets complexes approximativement modélisés par un polyèdre. L'approche repose sur l'estimation du mouvement de l'objet dans l'image, ainsi que sur un calcul de pose. La méthode proposée permet un suivi fiable et robuste en temps réel comme le montrent les résultats présentés.

Abstract – We present an original method for tracking in an image sequence complex objects which can be modeled approximately by a polyhedral shape. The approach relies on the estimation of the object image motion and the computation of the object pose. The proposed method fulfills real-time constraints as well as reliability and robustness requirements.

1 Introduction

Nous avons développé une méthode permettant un suivi robuste et rapide d'objets complexes pouvant être approximativement modélisés par une forme polyédrique. Elle repose sur l'estimation du mouvement 2D de l'objet et sur le calcul de sa pose 3D. Un modèle de mouvement affine 2D est estimé, à partir des déplacements orthogonaux calculés le long des projections des arêtes du modèle polyédrique dans l'image, grâce un algorithme robuste. Le modèle de mouvement affine ne permettant pas de représenter complètement le mouvement 3D de l'objet, une seconde étape consistant à recaler la projection du modèle de l'objet dans l'image est nécessaire. Cette étape revient en fait à calculer la pose de l'objet par rapport à la caméra. Ceci est réalisé par la minimisation itérative d'une fonction d'énergie non linéaire par rapport aux paramètres de pose.

Les principaux avantages de cette méthode peuvent se résumer ainsi. L'étape d'estimation du modèle de mouvement permet de prendre en compte de grand déplacement et évite ainsi une étape de prédiction de type Kalman [4, 11]. Le résultat de cette étape est exploité pour fournir une initialisation correcte du calcul de pose. La seconde étape reposant sur l'utilisation du modèle CAO ne nécessite qu'une calibration grossière de la caméra et un modèle approximatif de l'objet considéré. Les deux étapes (estimation du mouvement et de la pose) ne requièrent pas de phases explicites de segmentation, notamment d'extraction de contours dans les images successives (on utilise directement les niveaux de gris). Les deux étapes sont robustes aux occultations partielles de l'objet. Finalement, l'algorithme peut être exécuté à une cadence très rapide (10Hz sur un PC à 450Mhz).

Un des objectifs de cet algorithme est d'extraire des informations fiables de l'image dans l'optique d'une application d'asservissement visuel. De tels systèmes automatiques ou semi-automatiques intéressent la division R&D d' EDF en particulier pour des tâches de maintenance et de surveillance en environnement hostile.

2 Suivi 2D : estimation d'un modèle de mouvement

Nous considérons dans un premier temps que la transformation entre deux projections successives de l'objet dans le plan image peut être représenté par un modèle de déplacement affine 2D. L'objectif de cette première étape est d'estimer les paramètres de transformation 2D même en cas de grands déplacements de l'objet. Contrairement aux méthodes reposant sur l'utilisation du filtrage de Kalman pour prédire les positions successives de l'objet [6, 10, 11], cette méthode ne requiert pas l'introduction d'un modèle d'état (par exemple, un modèle à vitesse ou accélération constante), ni l'initialisation, souvent problématique, des matrices de covariance des bruits associés aux modèles d'état et de mesure.

Modèle de transformation affine. Considérons le vecteur $\mathcal{X}^t = [X_1^t, \ldots, X_n^t]^T$ composé par l'ensemble des pixels X_i^t le long de la projection des arêtes du modèle de l'objet à l'instant t. La projection de l'objet dans l'image à l'instant $t+1$ notée $\mathcal{X}^{t+1}$ sera donnée par :

$$\mathcal{X}^{t+1} = \Psi_\Theta(\mathcal{X}^t) \tag{1}$$

où Ψ_Θ est une transformation affine 2D donnée par :

$$\begin{bmatrix} x_i^{t+1} \\ y_i^{t+1} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} x_i^t \\ y_i^t \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \end{bmatrix} = \mathbf{W}(X_i^t)\Theta \quad (2)$$

avec $\Theta = (a_1, a_2, a_3, a_4, T_x, T_y)^T$, $X_i^t = (x_i^t, y_i^t)^T$, $X_i^{t+1} = \Psi_\Theta(X_i^t)$, et

$$\mathbf{W}(X) = \begin{bmatrix} x & y & 0 & 0 & 1 & 0 \\ 0 & 0 & x & y & 0 & 1 \end{bmatrix}$$

Cette transformation est linéaire en Θ, et le déplacement $d_i(X_i) = X_i^{t+1} - X_i^t$ peut être réécrit comme suit :

$$d_i(X_i) = \mathbf{W}(X_i)\Theta' \quad (3)$$

où $\Theta' = \Theta - (1, 0, 0, 1, 0, 0)^T$.

L'algorithme permettant l'estimation de cette transformation affine 2D représenté par le vecteur de paramètres Θ est structuré en deux étapes. La première calcule les déplacements orthogonaux aux arêtes de l'objet selon la méthode décrite dans [3] dite des ECM, tandis que la deuxième étape consiste à utiliser ce champ des déplacements orthogonaux pour estimer Θ par une technique d'estimation robuste adaptée de celle introduite dans [13]. Nous décrivons maintenant ces deux étapes.

Calcul des déplacements orthogonaux. L'un des avantages de la méthode des ECM est qu'elle ne nécessite pas une étape, souvent coûteuse, d'extraction des contours spatiaux dans l'image. De plus, elle peut être implantée en temps réel, les calculs à effectuer se limitant à de simples convolutions par des masques précalculés [2, 3].

Nous voulons un algorithme qui soit à la fois rapide et robuste à des occultations partielles de l'objet ainsi qu'à des erreurs locales de mise en correspondance. Nous considérons donc une liste L^t de pixels le long de la projection des arêtes du modèle CAO de l'objet dans l'image. L'ajustement de ce modèle dans la première image est effectué manuellement. Le procédé consiste à rechercher le "correspondant" P_i^{t+1} dans l'image I^{t+1} de chaque pixel $P_i^t \in L^t$. Nous déterminons un intervalle de recherche 1D $\left\{Q_i^j, j \in [-J, J]\right\}$ dans la direction δ normale à l'arête projetée. Pour chaque pixel P_i^t de la liste L^t, et pour chaque position Q_i^j dans l'image I^{t+1} dans la direction δ, nous calculons le critère de mise en correspondance qui s'exprime comme la maximisation d'un rapport de vraisemblance ζ^j. Ce rapport (en fait sa racine carré) peut se réécrire comme la valeur absolue de la somme des convolutions dans l'image calculées en P_i et Q_i^j avec un masque précalculé M_δ dépendant de l'orientation de l'arête projetée. La nouvelle position P_i^{t+1} est donnée par :

$$Q_i^{j^*} = \arg \max_{j \in [-J,J]} \zeta^j \text{ avec } \quad \zeta^j = \mid I^t_{\nu(P_i)} * M_\delta + I^{t+1}_{\nu(Q_i^j)} * M_\delta \mid$$

sous la contrainte que la valeur ζ^{j^*} soit supérieure à un seuil λ. $\nu(.)$ désigne un voisinage du pixel considéré. Le pixel P_i^{t+1} donné par $Q_i^{j^*}$ est ensuite placé dans la liste L^{t+1}.

Après cette étape, nous disposons d'une liste de k pixels ainsi que des déplacements associés dans la direction orthogonale aux arêtes de l'objet : $(P_i^t, d_i^\perp)_{i=1\ldots k}$. Cette approche locale, qui ne considère jamais une approximation polygonale des contours de la projection de l'objet dans son ensemble, mais en fait des pixels "indépendants", assure une certaine robustesse vis-à-vis des occultations partielles et des absences partielles de mesure.

Estimation de la transformation affine. En utilisant $(P_i^t, d_i^\perp)_{i=1\ldots k}$, nous pouvons estimer la transformation affine 2D. À partir de l'équation (3), nous avons :

$$d_i^\perp = \mathbf{n}_i^T \mathbf{d}(P_i) = \mathbf{n}_i^T \mathbf{W}(P_i)\Theta' \quad (4)$$

où $\mathbf{n}_i$ est un vecteur unitaire orthogonal à l'arête projetée de l'objet au point P_i. Partant de (4), nous pouvons utiliser un estimateur robuste (un M-estimateur ρ) pour obtenir $\widehat{\Theta}'$:

$$\widehat{\Theta}' = \arg\min_{\Theta'} \sum_{i=1}^{k} \rho(d_i^\perp - \mathbf{n}_i^T \mathbf{W}(P_i)\Theta')$$

Cette approche statistique robuste permet de ne pas être affectée par des mesures localement incorrectes (dues aux ombres, à des erreurs de mise en correspondance, à des occultations, etc.).

3 Suivi 3D : calcul de pose

Connaissant la position des arêtes projetées de l'objet $\mathcal{X}^t$ à l'instant t et l'estimation $\widehat{\Theta}$ de la transformation affine entre les instants t et $t+1$, il est possible de calculer la position $\mathcal{X}^{t+1}$ des points de l'objet à l'instant $t+1$: $\mathcal{X}^{t+1} = \Psi_{\widehat{\Theta}}(\mathcal{X}^t)$. Une transformation affine 2D ne prend cependant pas en compte complètement les transformations subies par la projection de l'objet (effet de perspective, rotations importantes, etc.). Après quelques itérations, le procédé de suivi peut alors être mis en échec. Pour résoudre ce problème, dans une première version de cet algorithme [7, 14], le modèle de déplacement affine 2D était complété par des déformations locales 2D. La contrainte de rigidité n'était cependant plus assurée. De plus cette méthode était très coûteuse en temps de calcul. C'est pourquoi nous avons développé une seconde étape reposant sur l'utilisation d'un modèle CAO approximatif de l'objet. Nous avons donc à calculer la pose Φ (c'est à dire translation et rotation 3D) de l'objet par rapport à la caméra à partir des positions $\mathcal{X}^{t+1}$ obtenues à l'issue de la première étape de l'algorithme décrite en Section 2. Nous utilisons la méthode conçue par Dementhon [5] suivi de la méthode de Lowe [12] en choisissant comme mesure les sommets du modèle polyédrique. Nous obtenons alors une première évaluation Φ_{init}^{t+1} des paramètres de pose qui doit être affinée pour correspondre le plus précisément possible au nouvel aspect de l'objet. Cette étape consiste à recaler la projection du modèle CAO sur les gradients d'intensité de l'image. Ceci est réalisé par une minimisation itérative d'une fonction d'énergie non linéaire fonction de Φ avec Φ_{init}^{t+1} comme valeur d'initialisation. Quand les paramètres de pose sont disponibles, il est par ailleurs aisé de déterminer les faces de l'objet visibles et invisibles, ce que nous exploiterons dans le suivi.

Plus précisément, nous estimons les paramètres $\widehat{\Phi}$ tels que $\widehat{\Phi} = \arg\min_\Phi \left\{E(d_\Phi^{t+1})\right\}$ où la fonction d'énergie $E(d_\Phi^{t+1})$

est définie par :

$$E(d_\Phi^{t+1}) = -\int_{\Gamma_\Phi} \|\nabla I_{\pi_\Phi(s)}(t+1)\| ds \qquad (5)$$

où

- Γ_Φ représente la partie visible des arêtes du modèle CAO de l'objet pour la pose Φ.
- $\nabla I_{\pi_\Phi(s)}$ représente le gradient spatial de l'intensité lumineuse au point $\pi_\Phi(s)$ le long de l'arête projetée $\pi_\Phi(\Gamma_\Phi)$ où π_Φ est la fonction de projection perspective.

La fonction d'énergie définie par l'équation (5) est la réponse la plus "directe" à notre problème. Cependant, il est possible d'utiliser une information plus riche que la norme du gradient d'intensité. En effet, lors de la projection des arêtes du modèle pour une pose Φ donnée, il est possible de calculer la direction attendue de l'arête projetée en un site $\varsigma = \pi_\Phi(s)$. Si l'on note $\mathbf{n}$ un vecteur unitaire correspondant à cette direction attendue, le produit scalaire $\nabla I_\varsigma.\mathbf{n}$ doit être nul. On en déduit une autre expression de la fonction d'énergie :

$$E(\Phi) = \int_{\Gamma_\Phi} \frac{|\nabla I_\varsigma.\mathbf{n}|}{\|\nabla I_\varsigma\|^2} ds \qquad (6)$$

Nous ne considérons dans le calcul de cette énergie que les site ς pour lesquels $\|\nabla I_\varsigma\| > \varepsilon$. Cette formulation de la fonction d'énergie donne de meilleurs résultats dans des environnements texturés.

La fonction de projection π_Φ dépend des paramètres intrinsèques $\mathcal{I}$ de la caméra. Le modèle de la caméra peut être utilisé tel quel dans l'optimisation de la fonction d'énergie si la calibration de la caméra est connue. Les paramètres intrinsèques peuvent aussi être estimés en ligne. Dans ce cas, la fonction à optimiser peut être récrite :

$$\left(\widehat{\Phi}, \widehat{\mathcal{I}}\right) = \arg\min_{(\Phi,\mathcal{I})} \left\{ E(d^{t+1}_{(\Phi,\mathcal{I})}) \right\} \qquad (7)$$

Dans le cas général, nous avons 11 paramètres à estimer (si nous intégrons la distorsion radiale). Dans la pratique, nous avons seulement réalisé des expériences considérant l'évaluation en ligne de la distorsion radiale.

La minimisation de E est réalisée par un algorithme de recherche déterministe dérivé de l'algorithme ICM.

4 Résultats

La plupart des expériences présentées considèrent un écrou comme objet d'intérêt. Précisons que le suivi de cet objet dans les séquences d'images est compliqué par la présence de très faibles contrastes, d'ombres portées, de fortes spécularités,... Par ailleurs, l'écrou n'est pas exactement polyédrique (les arêtes sont arrondies). Finalement, la calibration de la caméra n'est pas connue de façon précise.

La figure 1 montre le résultat du suivi de l'écrou sur une séquence de 44 images. La figure 1.a montre le résultat du suivi obtenu seulement à partir de l'estimation des mouvements 2D. Dans ce cas, le suivi est réalisé à la cadence vidéo (25Hz). Cependant, après quelques images, l'algorithme n'est plus capable de suivre efficacement l'objet. Cela est principalement dû au fait qu'un modèle de mouvement affine 2D ne peut rendre complètement compte des mouvements 3D de l'objet. La figure 1.b contient le résultat du suivi en considérant à la fois l'estimation du mouvement 2D et des paramètres de pose 3D. Dans ce cas, le suivi est effectué à la cadence de 10Hz.

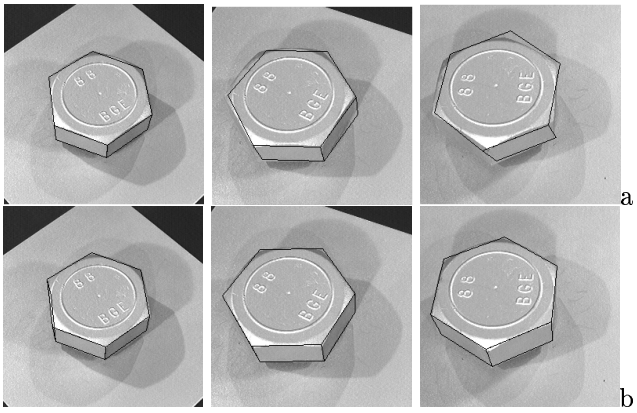

FIG. 1: *Suivi de l'écrou: (a) suivi à partir de la seule estimation du mouvement 2D (b) suivi avec utilisation du mouvement 2D et du calcul de la pose 3D*

Nous avons aussi testé le comportement de notre algorithme en présence d'un certain nombre de difficultés. Dans la séquence de la figure 2.a, le mouvement de la caméra est une rotation autour de l'axe y[1]. Une face apparaît alors qu'une autre face disparaît. Dans la séquence de la figure 2.b, la difficulté principale réside dans la très forte rotation autour de l'axe x. L'estimation de la pose devient alors très complexe à cause de la disparition de la face supérieure. De plus, les conditions d'illumination ne sont pas constantes. Dans la séquence de la figure 2.c, la difficulté provient d'une occultation partielle de l'écrou. Finalement, dans la séquence de la figure 2.d, l'écrou est suivi dans un environnement très texturé lors d'une expérience d'asservissement visuel (la position desirée est superposée à l'image).

L'algorithme présenté a été testé sur plusieurs objets. Sur la figure 3 sont présentés les résultats du suivi d'un banc micrométrique pendant une expérience d'asservissement visuel visant un positionnement particulier de la caméra par rapport à cet objet à partir d'une position initiale relativement éloignée. De plus divers outils sont placés sur le banc pendant le suivi provoquant des occultations partielles de celui-ci. La caméra effectue de grands déplacements par rapport à l'objet tant en rotation qu'en translation. Le suivi est réalisé à la cadence de 5Hz (ce temps de calcul plus élevé est principalement dû au fait que le modèle CAO de l'objet est plus complexe, ce qui induit un nombre de site ς beaucoup plus important).

Nous avons également essayé d'estimer en ligne la distorsion radiale. Pour cela, nous avons considéré un objet simple (une plaque métallique carrée) et une caméra avec une distorsion radiale non négligeable (voir figure 4.b). La valeur de la distorsion radiale est initialisée à zéro. Elle

1. l'axe z suit l'axe optique de la caméra alors que l'axe x est parallèle au lignes de l'image et l'axe y est parallèle au colonnes de l'image.

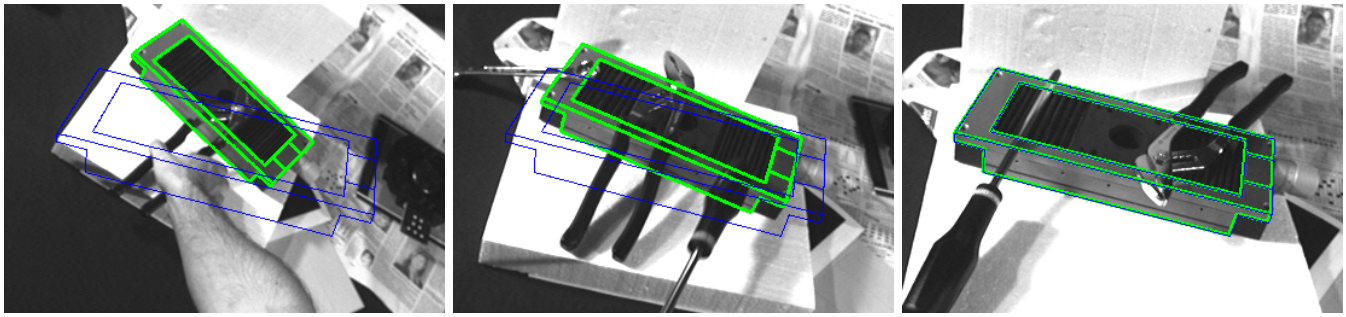

FIG. 3: *Asservissement visuel sur un banc micrométrique (position initiale, intermédiaire et finale), la position désirée apparaît en traits fins*

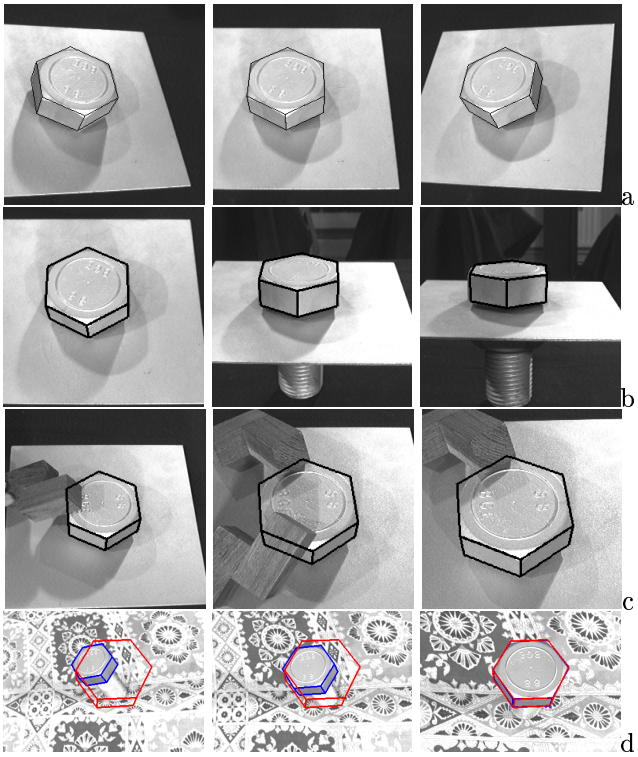

FIG. 2: *Suivi d'un écrou : diverses difficultés (voir le texte pour les détails)*

converge vers sa valeur nominale ($K \simeq 0.2 \pm 0,02$) quand l'objet se déplace vers le bord de l'image (où une meilleure évaluation de ce paramètre est possible). Dans ce cas, le suivi est réalisé à 1Hz.

Remerciement. Cette étude a reçu le soutien de EDF – Pôle Industrie – Division Recherche et Développement (contrat 1.97.C234.00).

Références

[1] B. Bascle, P. Bouthemy, N. Deriche, and F. Meyer. Tracking complex primitives in an image sequence. *ICPR'94*, pp. 426–431, Jerusalem, Oct. 1994.

[2] S. Boukir, P. Bouthemy, F. Chaumette, and D. Juvin. A local method for contour matching and its parallel implementation. *Machine Vision and Application*, 10(5/6):321–330, Avr. 1998.

[3] P. Bouthemy. A maximum likelihood framework for determining moving edges. *IEEE Trans. on PAMI*, 11(5):499–511, Mai 1989.

[4] N. Daucher, M. Dhome, J.T. Lapreste, and G. Rives. Modelled object pose estimation and tracking by monocular vision. In *British Machine Vision Conf., BMVC'93*, pp. 249–258, Guildford, Sept. 1993.

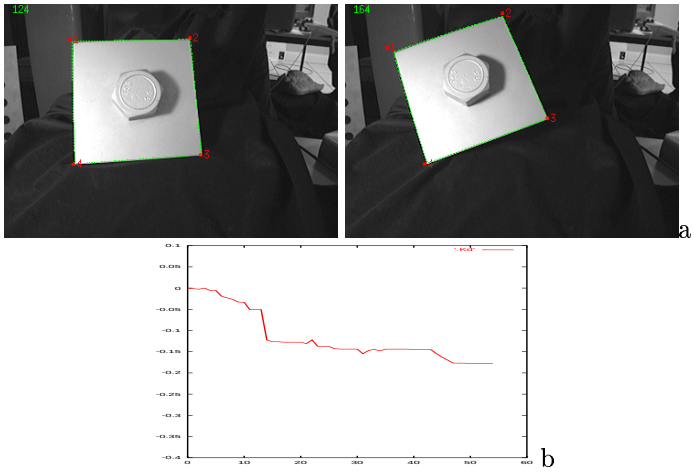

FIG. 4: *Suivi d'une pièce métallique avec une caméra présentant une distorsion radiale non négligeable (a), estimation du paramètre de distorsion (b)*

[5] D. Dementhon and L. Davis. Model-based object pose in 25 lines of codes. *Int. J. of Computer Vision*, 15:123-141, 1995.

[6] D.B. Gennery. Visual tracking of known three-dimensional objects. *Int. J. of Computer Vision*, 7(3):243–270, 1992.

[7] N. Giordana, P. Bouthemy, F. Chaumette, F. Spindler, J.-C. Bordas, and V. Just. 2D model-based tracking of complex shapes for visual servoing tasks. In G. Hager and M. Vincze, editors, *IEEE Workshop on Robust Vision for Vision-Based control of Motion*, Leuven, May 1998.

[8] G. Hager and K. Toyama. The X-Vision system: A general-purpose substrate for portable real-time vision applications. *Computer Vision and Image Understanding*, 69(1):23–37, Jan. 1998.

[9] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV'96,* pp. 343–356, Cambridge, 1996.

[10] C. Kervrann and F. Heitz. A hierarchical Markov modeling approach for the segmentation and tracking of deformable shapes. *Graphical Models and Image Processing*, 60(3):173–195, Mai 1998.

[11] H. Kollnig and H.-H. Nagel. 3D pose estimation by fitting image gradients directly to polyhedral models *ICCV'95*, pp. 569-574, Boston, Mai 1995.

[12] D.G. Lowe. Robust model-based motion tracking through the integration of search and estimation. *Int. J. of Computer Vision*, 8(2):113–122, 1992.

[13] J.-M. Odobez, P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. of Visual Communication and Image Representation*, 6(4):348-365, Dec. 1995.

[14] J.-M. Odobez, P. Bouthemy, E. Fleuet. Suivi 2D de pièces métalliques en vue d'un asservissement visuel. *RFIA'98*, Vol. 2, pp. 173-182, Clermont Ferrand, Jan. 1998.