Suivi Tridimensionnel en Stéréovision

Simon CONSEIL[1], Salah BOURENNANE[1], Lionel MARTIN[2]

[1]Groupe Signaux Multidimensionnels, Institut Fresnel, CNRS-UMR 6133
D-U de St Jérôme, F-13397 Marseille Cedex 20

[2]ST Microelectronics
ZI Rousset BP 2, F-13106 Rousset
simon.conseil@fresnel.fr, salah.bourennane@fresnel.fr, lionel.martin@st.com

Résumé – Cet article présente une méthode d'estimation en temps réel de la trajectoire d'une cible en trois dimensions, appliquée au suivi d'un doigt de la main dans un contexte de reconnaissance de gestes. Notre approche est basée sur la vision stéréoscopique, avec deux caméras standards, de type webcams. La main est segmentée par soustraction du fond, et le bout du doigt est détectée avec l'analyse de la courbure du contour. Le suivi du doigt est réalisé par un filtrage de Kalman tridimensionnel, ce qui permet d'améliorer la détection avec une recherche locale basée sur la prédiction de la position 3D et de filtrer la trajectoire pour réduire l'erreur d'estimation.

Abstract – This article presents a real time estimation method of the three dimensional trajectory of a target, applied to fingertip tracking in a gesture recognition context. Our approach is based on stereoscopic vision, with two standards webcams. The hand is segmented by background subtraction, and the fingertip is detected with the analysis of the curvature of finger boundary. The fingertip tracking is carried out by a three dimensional Kalman filter, in order to improve the detection with a local research centered on the prediction of the 3D position and to filter the trajectory to reduce the estimation error.

1 Introduction

Les données en trois dimensions représentent une quantité d'information importante et très intéressante dans de nombreux domaines (reconnaissance de gestes, réalité virtuelle ou augmentée, étude du comportement humain ...). L'acquisition de ces données peut se faire de différentes façons : gants équipés de capteurs, caméra unique associée à des informations géométriques sur la scène [1], ou vision stéréoscopique.

Les gestes de la main sont un vecteur de communication naturel chez l'homme. Ils peuvent être liés à la parole, servir à désigner des objets, et même représenter un langage à part entière (langue des signes). Nous nous intéressons aux gestes déictiques, consistant à désigner un objet avec la main ou le doigt. Le doigt représente en effet un dispositif de pointage naturel et très pratique pour des applications de la vision par ordinateur aux *Interfaces Homme-Machine*.

Parmi les nombreux travaux de ces dernières années, on peut citer le système EnhancedDesk [2] qui permet de suivre plusieurs doigts en deux dimensions, avec un filtre de Kalman pour chaque doigt. Les bouts de doigts sont détectés grâce à une caméra infrarouge et une corrélation normalisée, des gestes symboliques (cercle, carré, triangle) sont ensuite reconnus avec des Modèles de Markov Cachés (HMM). Segen et Kumar [3] utilisent des points de contours caractéristiques pour classifier quatre gestes de la main et déterminer la direction 3D pointée par le doigt, avec deux caméras.

Le système Digital Desk de Crowley *et al.* [4] montre l'intérêt du suivi du doigt pour la réalité augmentée. Le suivi est effectué par corrélation avec un modèle de bout de doigt. Les trajectoires de points ont été utilisées pour la reconnaissance d'écriture 2D avec des HMM [5]. Une application classique dans ce domaine est le remplacement de la souris par la main [6], ce qui permet de valider la bonne localisation du doigt et de tester l'interactivité du système, la position du curseur sur l'écran fournissant un retour d'information à l'utilisateur. On trouvera une étude approfondie des techniques de reconnaissance de gestes dans [7].

Dans cet article, nous nous intéressons au suivi du bout du doigt pour des gestes de pointage. Deux caméras calibrées sont utilisées pour calculer les trajectoires 3D, mais les erreurs de localisation du doigt et l'absence de synchronisation rendent l'estimation de la trajectoire peu précise. L'originalité de notre approche est l'utilisation d'un filtre de Kalman pour réaliser un suivi tridimensionnel, afin d'améliorer la détection avec une recherche locale basée sur la prédiction de la position 3D et de filtrer la trajectoire pour réduire l'erreur d'estimation.

2 Vision Stéréoscopique

Dans le cas général d'un espace projectif, une caméra est caractérisée par sa matrice de projection perspective P, de dimension 3×4. Cette matrice détermine la projection d'un point de l'espace $M(X, Y, Z, 1)$ en un point de l'image $m(u, v, 1)$:

$$\begin{pmatrix} u & v & 1 \end{pmatrix}^T = \mathrm{P} \begin{pmatrix} X & Y & Z & 1 \end{pmatrix}^T \qquad (1)$$

Les paramètres de cette matrice sont calculés par calibration, afin d'obtenir une reconstruction 3D euclidienne. Connaissant cette matrice pour les deux caméras, la reconstruction consiste en une triangulation. Il est donc nécessaire de déterminer le correspondant d'un point de l'image gauche dans l'image droite

et réciproquement. Ainsi pour un point $M(X, Y, Z)$ de l'espace qui se projette en un point $m_1(u, v)$ (resp. $m_2(u', v')$) de l'image gauche (resp. droite), le point 3D reconstruit est celui qui minimise la distance entre les deux droites de vues O_1m_1 et O_2m_2, O_1 et O_2 étant les foyers des deux caméras.

La géométrie épipolaire d'un système stéréoscopique est représentée par la matrice fondamentale F, de dimension 3×3 :

$$m_1^T \mathrm{F} m_2 = 0 \tag{2}$$

La contrainte épipolaire exprime le fait que le correspondant d'un point d'une image se trouve sur une droite dans l'autre image : c'est l'image de la droite de vue du premier point.

3 Détection de la cible

Les interfaces utilisant la reconnaissance de gestes nécessitent une détection robuste et rapide de la main. Une bonne segmentation est essentielle au bon fonctionnement du système. Deux grandes approches sont couramment utilisées : la soustraction du fond [8] et la détection de la couleur de la peau [9]. Ces approches sont sensibles aux variations de luminosité et nécessitent donc un environnement contrôlé. Le seuillage de la disparité obtenue à partir d'une paire stéréoscopique a été étudié [10], mais le calcul d'une carte de disparité est coûteux car il nécessite une reconstruction dense de la scène, ce qui n'est pas utile dans notre cas.

3.1 Détection de la main

Nous utilisons la technique classique de soustraction du fond, qui consiste à soustraire une image de référence, correspondant à la scène sans objets, à l'image courante. L'image de référence est prise à l'initialisation du système et peut être réinitialisée en cas de variation rapide de la luminosité. Afin d'éliminer le bruit du masque binaire obtenu, un filtre médian est appliqué, puis une fermeture morphologique et finalement un étiquetage en composantes connexes. L'intérêt de la soustraction du fond est de ne pas se restreindre au doigt, il est possible d'utiliser d'autres objets pour pointer, comme un stylo. Cette méthode est par contre très sensible à la présence d'ombres.

3.2 Détection du bout du doigt

Lors de l'entrée du doigt dans le champ de vision des caméras, il faut détecter précisément la position du bout du doigt afin d'initialiser le suivi. A partir des images binaires de la main obtenues à l'étape précédente, différentes approches sont possibles pour détecter la position du bout du doigt. La corrélation avec un modèle de bout de doigt a été utilisée [4], mais n'est pas suffisamment robuste aux changements d'échelle et d'orientation.

Nous utilisons une distance euclidienne par rapport au centre de gravité de la région correspondant à la main. Le centre de gravite de la main est calculé à partir des moments géométriques de la région de la main. Le point du contour se situant à la plus grande distance du centre de gravité est choisi comme bout du doigt pour l'initialisation du suivi. Cette méthode est toutefois peu précise, aussi nous affinons la détection en utilisant la courbure du contour, avec la méthode décrite dans [3].

(a) (b) (c)

FIG. 1 – Segmentation de la main : (a) image originale, (b) masque binaire après soustraction du fond et (c) contour et détection du bout du doigt

4 Suivi Tridimensionnel

A partir de la position du bout du doigt dans chacune des deux images, il est possible de calculer sa position 3D. Toutefois les positions 3D ainsi calculées manquent de précision pour plusieurs raisons : détection imprécise du bout du doigt due à une mauvaise segmentation, discrétisation des images (une erreur d'un pixel sur la localisation du bout du doigt peut représenter plusieurs millimètres en 3D), décalage temporel entre l'acquisition des deux images (les deux caméras ne sont pas synchronisées).

Par ailleurs il n'est pas nécessaire de traiter l'image entière alors que nous connaissons la position du doigt. Ainsi la zone de recherche du bout du doigt peut être réduite grâce au suivi du doigt et à la prédiction de sa position à partir de la paire d'images précédente. L'objectif du suivi temporel est donc de faciliter la localisation du doigt et de lisser les trajectoires obtenues.

4.1 Filtrage de Kalman

Notre approche est basée sur un filtre de Kalman [11] tridimensionnel avec un modèle de mouvement uniforme à vitesse constante. Le vecteur d'état $\mathbf{x}_k$ est défini comme :

$$\mathbf{x}_k = (x(k), y(k), z(k), v_x(k), v_y(k), v_z(k))^T$$

avec $(x(k), y(k), z(k))$ la position et $(v_x(k), v_y(k), v_z(k))$ la vitesse du bout du doigt dans l'image k. Le vecteur d'état $\mathbf{x}_k$ et le vecteur d'observation $\mathbf{z}_k$ sont reliés par les équations suivantes :

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathrm{A}\,\mathbf{x}_k + \mathbf{w}_k \\ \mathbf{z}_k &= \mathrm{H}\,\mathbf{x}_k + \mathbf{v}_k \end{aligned} \tag{3}$$

avec $\mathbf{w}_k$ et $\mathbf{v}_k$ les bruits du processus et de mesure, supposés bruits blancs gaussiens, A la matrice de transition de l'état et H la matrice d'observation, avec ΔT la période d'acquisition :

$$\mathrm{A} = \begin{pmatrix} 1 & 0 & 0 & \Delta T & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta T & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta T \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathrm{H} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

En notant $\mathbf{x}_k$ et $\mathbf{x}_k^-$ les estimations *a posteriori* et *a priori*, P_k et P_k^- les covariances des erreurs d'estimation *a posteriori* et *a priori*, Q la covariance du bruit du processus, R la covariance du bruit de mesure et K_k le gain du filtre de Kalman, on obtient les équations suivantes :

Equations de prédiction :

$$\begin{aligned} \mathbf{x}_k^- &= A\mathbf{x}_{k-1} \\ P_k^- &= AP_{k-1}A^T + Q \end{aligned} \quad (4)$$

Equations de mise à jour :

$$\begin{aligned} K_k &= P_k^- H^T (HP_k^- H^T + R)^{-1} \\ \mathbf{x_k} &= \mathbf{x_k}^- + K_k(\mathbf{z_k} - H\mathbf{x_k^-}) \\ P_k &= (I_6 - K_k H)P_k^- \end{aligned} \quad (5)$$

Paramètres. Les trois composantes sont supposées indépendantes, ainsi les matrices de covariance sont diagonales. Comme nous supposons la vitesse constante dans notre modèle, ce qui n'est pas forcément vrai, la covariance du bruit du processus est supposée importante sur la composante de vitesse tandis qu'elle est faible sur la position. Les variances du bruit de mesure sont calculées avec une séquence d'images où le doigt reste fixe. Nous obtenons $Var(X, Y, Z) = (0.31, 2.39, 15.06)$, ce qui montre que l'erreur de mesure est plus importante sur la composante Z que sur X et Y.

4.2 Algorithme développé

La figure 2 résume les différentes étapes du traitement : à partir du calcul de la position 3D avec une paire d'images, le filtre de Kalman permet de prédire la position 3D correspondant à la paire d'images suivante. La prédiction 3D prédite est projetée dans les deux images pour obtenir une prédiction de la position du bout du doigt dans chacune des deux images, ce qui permet de réduire la zone de recherche du bout du doigt. La taille utilisée pour la fenêtre de recherche est de 80×80 pixels. La détection du doigt est alors réalisée avec la méthode décrite dans la section 3. Enfin la contrainte épipolaire est vérifiée pour s'assurer de la bonne détection du bout du doigt dans les deux images.

5 Résultats

Nous utilisons deux caméras de type webcams, au format 352×288. Les images sont transmises par une connexion USB avec une compression MJPEG, ce qui dégrade les images. De plus, les caméras ne sont pas synchronisées, ce qui peut induire une différence de position entre deux images, et peut résulter en une oscillation de la trajectoire du doigt : pendant l'intervalle de temps entre les acquisitions des deux images, le doigt peut avoir bougé plus ou moins en fonction de la vitesse de son mouvement. Par conséquent, la triangulation est biaisée, principalement au niveau de la profondeur (dans la direction correspondant à l'axe optique des caméras).

5.1 Exemple : cas d'un cercle

Afin de pouvoir mesurer les erreurs sur le calcul de la position 3D, il est nécessaire de connaître la vérité terrain, ce qui est souvent difficile en stéréoscopie. Dans notre configuration l'erreur de reconstruction se retrouve principalement sur la composante Z, correspondant à la profondeur (axes optiques), aussi nous nous intéressons à une trajectoire plane : un cercle tracé sur le bureau, correspondant au plan $z = 0$ (Fig. 3).

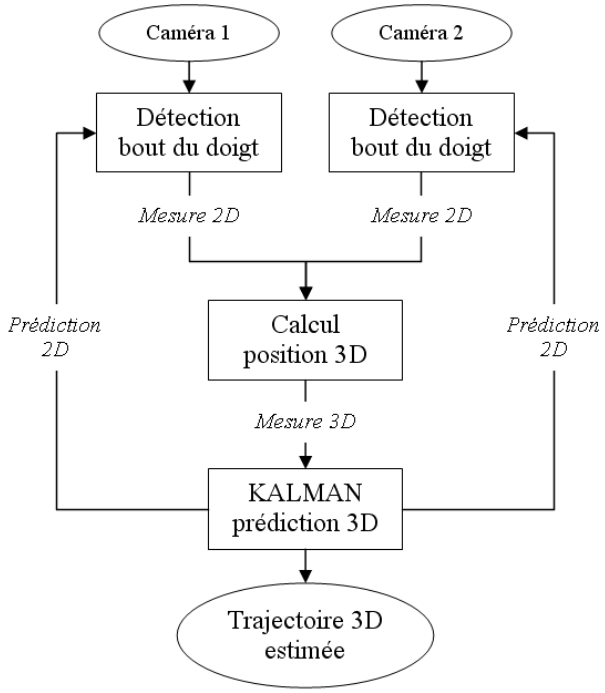

FIG. 2 – Schéma récapitulant les différentes étapes

Nous pouvons voir sur la figure 4 la trajectoire 3D estimée du cercle, ainsi que les points mesurés en pointillés. La figure 5 montre que le filtre de Kalman permet de lisser la trajectoire 3D, et de diminuer l'écart type sur la profondeur qui passe de 9.77 à 5.46.

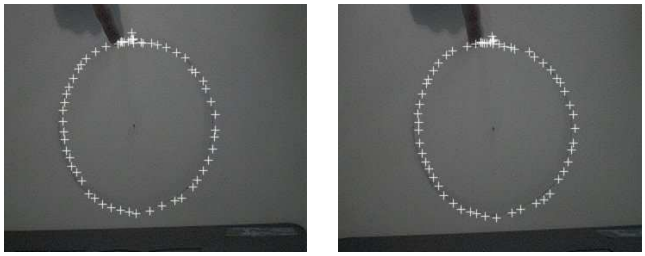

FIG. 3 – Images gauche et droite avec les points mesurés lors du tracé d'un cercle avec le doigt

5.2 Influence de la vitesse

La vitesse du mouvement influe sur l'erreur de reconstruction. En effet, plus le mouvement est rapide, plus le doigt peut avoir bougé entre l'acquisition des deux images, d'où une erreur plus importante.

Le tableau 1 illustre ceci avec l'étude de deux trajectoires planes (cercle et carré), traitée en temps réel (30 Hz). Ces trajectoires sont réalisées avec trois vitesses différentes, ainsi une trajectoire plus rapide est composée d'un nombre de points

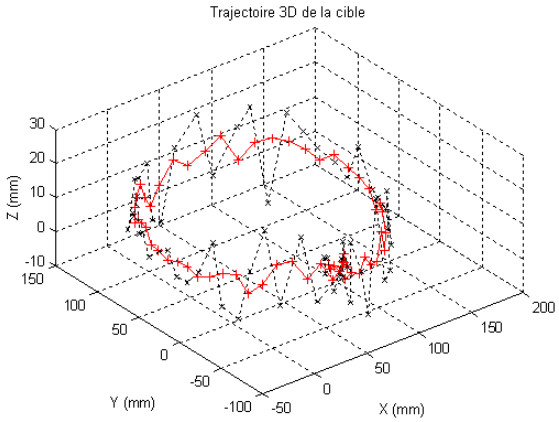

FIG. 4 – Trajectoire 3D (mesures en pointillés, estimations en trait plein)

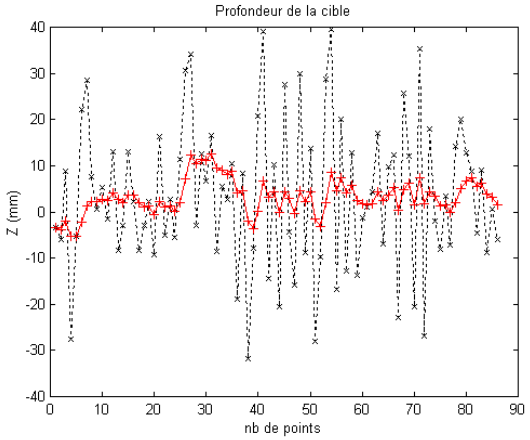

FIG. 5 – Composante Z, correspondant à la profondeur, en mm (mesures en pointillés, estimations en trait plein)

plus faible. L'écart-type sur la composante Z est ensuite calculé pour comparer les erreurs de reconstruction. Dans les deux cas l'écart-type augmente avec la vitesse, et l'écart-type est plus faible pour la trajectoire estimée par le filtre de Kalman que pour les mesures.

Conclusion

Nous avons présenté un système de suivi tridimensionnel d'une cible, qui permet de rendre la détection plus robuste en réduisant la zone de recherche et de réduire l'erreur d'estimation en lissant les trajectoires 3D. Appliqué au suivi d'un doigt de la main, le système fonctionne en temps réel sur des données réelles, avec un PC à 2.6 GHz. Avec une méthode détection adaptée, d'autres applications sont possibles comme le suivi de personnes ou de véhicules.

Pour améliorer le système, une réflexion est menée sur le calcul de la largeur de la fenêtre de recherche afin de l'adapter à la vitesse du mouvement, avant d'étendre le suivi au cas de plusieurs cibles et de traiter le problème des occultations.

TAB. 1 – Évolution de l'écart-type sur la profondeur en fonction de la vitesse de réalisation du mouvement

	Vitesse	Nombre de points	Écart-type Mesures	Écart-type Estimation
Cercle	lent	306	9.7673	5.4587
	moyen	189	11.3158	8.3916
	rapide	108	14.7552	10.8265
Carré	lent	290	10.4771	4.8718
	moyen	185	11.1463	4.4337
	rapide	106	12.2786	6.0401

Références

[1] A. Wu, M. Shah, N. da Vitoria Lobo. *A Virtual 3D Blackboard : 3D Finger Tracking using a Single Camera*. Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2000.

[2] K. Oka, Y. Sato, H. Koike. *Real-time fingertip tracking and gesture recognition*. IEEE Computer Graphics and Applications, Vol. 22, No. 6, pp. 64-71, 2002.

[3] J. Segen, S. Kumar. *Human-Computer Interaction using Gesture Recognition and 3D Hand Tracking*. International Conference on Image Processing, 1998.

[4] J. Crowley, F. Berard, J. Coutaz. *Finger Tracking as an Input Device for Augmented Reality*. International Workshop on Gesture and Face Recognition, Zurich, June 1995.

[5] J. Martin, J.B. Durand. *Automatic Handwriting Gestures Recognition using Hidden Markov Models*. Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2000.

[6] Y. Hung and al. *Free-hand pointer by use of an active stereo vision system*. Proc. of the IEEE Int. Conf. on Pattern Recognition, 1998.

[7] V. Pavlovic and R. Sharma and T. Huang. *Visual Interpretation of Hand Gestures for Human-Computer Interaction : A Review*. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, July 1997

[8] P. Bertolino, G. Foret, D. Pellerin. *Détection de personnes dans les vidéos pour leur immersion dans un espace virtuel*. GRETSI 2001, 18ème Colloque sur le Traitement du Signal et de l'Image, Toulouse (France), 2001.

[9] S.L. Phung and A. Bouzerdoum and D. Chai. *Skin segmentation using color pixel classification : analysis and comparison*. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 27, Jan. 2005

[10] N. Jojic, T. Huang, B. Brumitt, B. Meyers, S. Harris. *Detection and Estimation of Pointing Gestures in Dense Disparity Maps*. Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2000.

[11] G. Welch, G. Bishop. *An introduction to the Kalman filter*. 1995.