Utilisation de la coordination oeil-main pour la prédiction d'un geste de préhension

M. Carrasco [1] X. Clady[2]

[1] Escuela de Ingeniería Informática, Facultad de Ingeniería,
Universidad Diego Portales, Av. Ejército 441, Santiago, Chile,
[2] Université Pierre et Marie Curie-UPMC
Institut des Systèmes Intelligents et de Robotique (ISIR), UMR7222 CNRS

mlacarrasco[at]gmail[dot]com, xavier[dot]clady[at]upmc[dot]fr

Résumé

La coordination oeil-main est primordiale chez l'homme dans l'action de préhension. Il est donc nécessaire d'en tenir compte dans l'analyse de tels gestes. Dans ce papier, nous proposons un capteur de vision permettant l'étude des mouvements de l'oeil et de la main de manière simultanée afin de détecter le mouvement d'atteinte de l'objet et donc de prédire l'intention de préhension chez l'utilisateur. Cette prédiction pourrait être utilisée au sein de systèmes robotiques, tels que des orthèses pour la rééducation fonctionnelle du membre supérieur, ou plus généralement dans des interfaces homme-machine. Notre solution effectue une fusion entre des informations extraites de deux systèmes différents : un eye-tracker nous fournit une vue "utilisateur" de la scène où la direction du regard est ainsi connue et une caméra fixée au poignet de l'utilisateur nous fournit une vue "main" de cette même scène. Des informations visuelles extraites de ces deux vues, nous caractérisons le mouvement de la main en conjonction avec les mouvements du regard au travers de modèles de Markov cachés. Dans nos expérimentations, nous démontrons que combiner ces deux sources d'informations permettent de reconnaître un mouvement d'atteinte, ainsi que l'objet cible.

Mots Clef

Capteur de vision, reconnaissance de geste, coordination oeil-main, mouvement d'atteinte, préhension.

Abstract

The eye-hand coordination is primordial in the action of reach-to-grasp an object by a human hand. So it is important to take it account in the analysis of this type of gesture. This paper proposes a visual sensor allowing the simultaneous analysis of the hand and eye motions in order to recognize the reach-to-grasp movement, i.e. to predict the grasping gesture . Our solution performs a fusion between two viewpoints taken from the user's perspective. First, by using an eye-tracker device from the user's head ; and Second, by using a wearable camera from the user's hand. We use the information from these two viewpoints, and we characterize multiple hand movements in conjunction with eye-gaze movements through a Hidden-Markov Model framework. In some experiments, we show that combining these two sources, allows predicting a reach-to-grasp movement and the object wanted.

Keywords

Visual system, gesture recognition, eye-hand coordination, reach-to-grasp movement.

1 Introduction

Depuis quelques années, les projets d'orthèses actives [1, 2, 3] pour la rééducation fonctionnelle du bras fleurissent en robotique. La plupart des études s'intéressent à leur design ou à leur commande. Certaines montrent comment l'estimation du mouvement humain peut être utilisée dans des schémas de commande ; par exemple, [4] démontre comment l'utilisation de la prédiction du mouvement humain dans une commande par anticipation, peut accroître la transparence du robot pour l'assistance à la manipulation. Ces approches nécessiteraient d'identifier de manière précoce le modèle de mouvement requis, c'est-à-dire le geste voulu par l'utilisateur. Il s'agirait donc de prédire l'intention gestuelle de l'utilisateur.

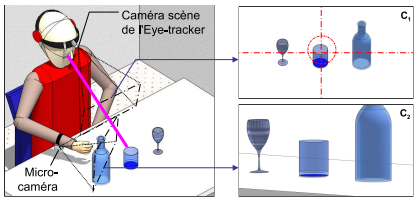


FIGURE 1 – Schéma illustrant le système combinant un *eye-tracker* et une caméra au poignet lors d'un mouvement d'atteinte

Dans ce papier, nous décrivons une nouvelle approche pour reconnaître cette intention à l'aide d'un capteur de vision

embarqué original. Il est composé d'un *eye-tracker* fournissant une vue "utilisateur" de la scène (où la direction du regard est estimée) et une caméra fixée au poignet de l'utilisateur qui nous fournit une vue "main" de cette même scène, tel que présenté dans la figure 1). Les informations visuelles extraites de ces deux points de vue vont nous permettre d'exploiter la coordination oeil-main, via des modèles cachées de Markov, pour prédire les gestes de préhension (à la fois le geste et l'objet cible). Tout d'abord, une méthode utilisant uniquement l'information visuelle de la caméra au poignet est proposée pour reconnaître le type de mouvement réalisé par la main. Ensuite, cette information est combinée avec un système de reconnaissance d'objets et une analyse du mouvement des yeux pour distinguer les instants où l'utilisateur ne fait que fixer les objets qu'on lui présente de ceux où il initie un geste de préhension. L'ensemble du système a été testé dans un contexte expérimental de préhension planaire (objets sur une table) qui correspondrait à un protocole thérapeutique classique.

Dans la section suivante, nous allons situer ce travail vis-à-vis de ceux entrepris sur la coordination oeil-main dans le cadre de la préhension et de ceux en reconnaissance de gestes par vision. La section 3 présentera les méthodologies développées pour reconnaître les mouvements de la main puis pour prédire l'intention d'un geste de préhension. Les expérimentations réalisées pour les évaluer, ainsi que les résultats obtenus, seront exposées et discutées dans la section 4.

2 Contexte du travail

La coordination oeil-main dans le geste de préhension : Les êtres humains ont développé une aptitude à prendre des objets quelques soient les conditions, en tenant compte de toutes les variations possibles de position, de forme ou d'orientation. Cette aptitude naturelle est appelée la coordination Bras-Oeil. Le mouvement de préhension est initié et planifié au sein du cerveau bien avant le geste lui-même, puis est régulé via l'interaction de plusieurs systèmes sensorimoteurs tels que la vue, les systèmes vestibulaires et proprioceptifs qui travaillent en conjonction avec le processus de contrôle de la tête, des yeux, du bras et de la main[5]. Dans notre travail, nous limitons l'intention à l'action consciente d'atteinte et qui est précursive à la saisie de l'objet. Elle peut être définie ainsi : lorsque le sujet initie un geste d'atteinte de l'objet, son regard se fixe sur celui-ci pendant un court temps tandis que sa main effectue une trajectoire stable et rectiligne vers l'objet[6]. Ces deux caractéristiques (regard et mouvement de la main) sont essentielles pour prédire l'intention du sujet.

L'analyse du mouvement humain par vision : Les études dans le domaine de l'analyse par vision du mouvement chez l'homme peuvent se diviser en trois principaux paradigmes : Passif, Actif ou de Pointage, en fonction de la position du sujet par rapport à la caméra. L'approche Passive consiste à capturer le mouvement humain par une ou plusieurs caméras déportées et stationnaires[7]. Il s'agit alors soit d'identifier et suivre les différentes parties du corps [8] afin d'en extraire des caractéristiques (trajectoires, dynamiques,...) soit d'extraire des descripteurs globaux[9], semi-globaux [10] ou locaux [11] du mouvement observé dans l'image. L'approche Active utilise des dispositifs externes embarqués sur le corps humain. Leur objectif est de fournir une représentation de l'environnement de l'utilisateur ; par exemple, pour la navigation de personnes non-voyantes [12, 13]. Ces caméras transportables et souvent actives offrent de nouvelles voies pour l'interaction homme-machine le laissant libre de ses déplacements. L'approche dite de Pointage est fondée sur le concept de *Ce que je regarde est ce je veux*. Pour cela, l'outil le plus utilisé actuellement est l'*eye-tracker*. Il consiste à suivre les mouvements des yeux afin de fournir une estimation de la direction du regard, ou indirectement la matérialisant dans une image courante de la scène observée sous la forme d'un curseur. La plupart des systèmes de capture du mouvement sont actifs ou passifs. Dans ce papier, nous avons développé un système original, illustré dans la figure 1 : il combine les paradigmes Actif, via une micro-caméra installée au poignet du sujet, et de Pointage, via un *eye-tracker* fournissant une image de la scène et l'information sur la localisation du regard dans cette image (cf. Figure). A notre connaissance, nous ne connaissons aucun autre travail qui propose de prédire les intentions de préhension chez un sujet humain qui exploitent la coordination Bras-Oeil dans un tel système mixte.

3 La méthode proposée

Notre approche consiste à détecter un mouvement d'atteinte avant la préhension elle-même. Pour cela, nous avons séparé l'analyse en deux parties. La première s'intéresse uniquement au mouvement de la main ; l'idée est de détecter le mouvement d'approche de l'objet via un modèle de Markov cachés (HMM). Ensuite, la prédiction du geste de préhension utilise un second HMM qui combinera l'information du mouvement de la main avec celui fourni par l'*eye-tracker*.

3.1 Reconnaissance du mouvement

Un mouvement d'atteinte se traduira dans l'image de la caméra-poignet par un *zoom* constant sur l'objet désiré. Autrement dit, tous les autres objets disparaîtront de la scène observée. De plus, si le mouvement est trop stochastique, il ne s'agit probablement pas d'un mouvement d'atteinte. Cette dernière remarque est à l'origine de notre système. Celui-ci se base sur une analyse du suivi de points d'intérêt, réalisé à l'aide du descripteur SURF[14] bien connu pour sa relative faible complexité algorithmique et sa robustesse en échelles et en rotations. Expérimentalement, même si cette méthode est efficace, sur des objets de la vie quotidienne et dans le cadre de notre travail, le nombre de points d'intérêt suivis, donc mis en correspondance, est relativement faible et sont souvent mal distri-

bués dans l'image. De plus, nous travaillons à une distance proche des objets et nous ne souhaitons pas construire une méthode nécessitant une modélisation 3D des objets afin d'en conserver une certaine souplesse au niveau applicatif. Dans ces conditions, les méthodes classiques d'estimation de mouvements de caméra ne sont pas adaptées car elles supposent soit un modèle *a priori* de la scène soit un grand nombre de points. Aussi nous avons choisi une approche que l'on qualifiera de sémantique, en extrayant d'une représentation simplifiée du mouvement global observé sur une Fenêtre Temporelle Glissante (FTG) plusieurs indicateurs des mouvements de la main que l'on veut reconnaître. Ces mouvements sont les mouvements d'atteinte, d'éloignement, de rotation ou de translation. Pour des gestes naturels, ces mouvements sont souvent très faiblement couplés, sauf dans des tâches complexes qui sont hors de propos ici. Lors d'une préhension simple (sans obstacles), la main se positionne en direction de l'objet avant le mouvement d'atteinte proprement dit. La figure 2 présente un aperçu de la méthode développée.

Représentation du mouvement sur une Fenêtre Temporelle Glissante : Dans ces travaux, nous avons choisi de traiter les séquences vidéo selon une approche de Fenêtre Temporelle Glissante (FTG). Sur chaque fenêtre, des mises en correspondances multiples de points d'intérêt sont réalisés afin d'estimer des caractéristiques du mouvement global observé sur chaque fenêtre. En premier lieu, les points d'intérêt sont extraits et caractérisés selon l'algorithme SURF sur chaque δ-*frame* de la FTG. Nous notons alors $\mathbf{p}_1^j = [x_1^j, y_1^j, 1]^\top$ la position (en coordonnées homogènes) du $j^{\text{ème}}$ point d'intérêt au temps $t = 1$, et $\mathbf{p}_n^j$ le point d'intérêt au temps $t = n$ (la dernière *frame* de la FTG) qui peut être mis en correspondance avec celui-ci selon le critère NNDR (*Nearest-Neighbor with Distance Ratio* [15]). En sélectionnant dans les autres images, les points vérifiant aussi une forte similarité avec ce point, nous obtenons un

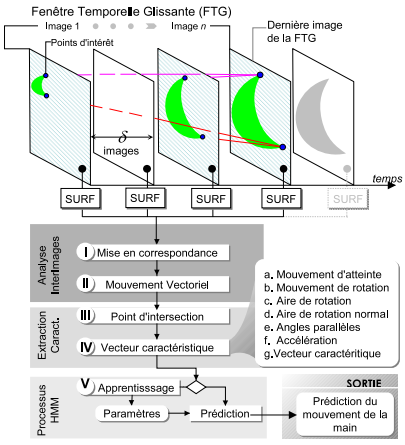


FIGURE 2 – Synopsis de la méthode de reconnaissance du mouvement de la main

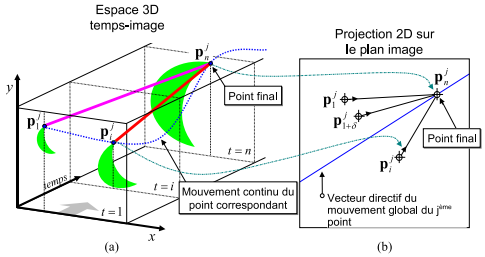


FIGURE 3 – Illustration de la mise en correspondance des points d'intérêts sur une Fenêtre Temporelle Glissante. (a) Points mis en correspondance dans l'espace 3D image-temps ; (b) Projection dans le plan image.

faisceau de vecteurs (notés $\mathbf{q}_{i,n}^j$ avec $i \in \{1, \ldots, n-\delta\}$ [1]) convergeant (cf. Figure 3) vers le point $\mathbf{p}_n^j$ de la dernière image de la FTG. Ce faisceau de vecteur est noté $\mathbf{Q}_{1\mapsto n}^j$, correspondant à la matrice regroupant tous les vecteurs homogènes $\mathbf{q}_{i,n}^j$:

$$\mathbf{Q}_{1\mapsto n}^j = \left[\mathbf{q}_1^j, \ldots, \mathbf{q}_i^j, \ldots, \mathbf{q}_{n-\delta}^j\right]^\top$$

Du fait des distorsions géométriques et photométriques, ainsi que des occlusions, les correspondances ne sont pas forcément réalisées pour chaque image. Aussi nous définissons un paramètre ρ qui correspond au nombre minimum de colonnes de la matrice $\mathbf{Q}_{1\mapsto n}^j$ (autrement dit le nombre de correspondances établies) pour lequel ce faisceau sera conservé.
L'étape suivante consiste à réduire ce faisceau de vecteur en un seul vecteur directeur, représentatif du mouvement global observé pour le $j^{\text{ème}}$ point. Ce vecteur correspondra à la moyenne des vecteurs $\mathbf{q}_{i,n}^j$ pondérés par la pertinence de la mise en correspondance. Pour construire cette pondération, nous considérons l'angle $\mathbf{F}_{1\mapsto n}^j$ entre les descripteurs (issus de l'algorithme SURF) des points $\mathbf{p}_i^j \mapsto \mathbf{p}_n^{j'}$. On aura alors en regroupant ces angles dans un vecteur :

$$\mathbf{F}_{1\mapsto n}^j = \left[\mathbf{F}_{1,n}^j, \ldots, \mathbf{F}_{i,n}^j, \ldots, \mathbf{F}_{n-\delta,n}^j\right] ,$$

Si l'angle est faible, la correspondance entre les points est forte. Pour inverser cette relation, nous définissons $\tilde{\mathbf{F}}_{1\mapsto n}^j$ tel que :

$$\tilde{\mathbf{F}}_{1\mapsto n}^j = 1 - \frac{\alpha \mathbf{F}_{1\mapsto n}^j}{\max\left(\mathbf{F}_{1\mapsto n}^j\right)}. \qquad (1)$$

Expérimentalement, α a été fixée à 0.98, afin de ne pas brutalement exclure les points avec une correspondance faible. Nous établissons alors le vecteur moyen pondéré

1. $\mathbf{q}_{i,n}^j$ est le vecteur homogène qui relie $\mathbf{p}_i^j \mapsto \mathbf{p}_n^{j'}$ et est défini par $\mathbf{q}_{i,n}^{j,j'} = \mathbf{p}_i^j \times \mathbf{p}_n^{j'} = [x_i^j, y_i^j, 1] \times [x_n^{j'}, y_n^{j'}, 1]$; il est noté $\mathbf{q}_{i,n}^j$ par simplification et suppose implicitement une mise en correspondance correcte entre les $j^{\text{ème}}$ et $j'^{\text{ème}}$ points entre les temps $t = i$ et $t = n$

$\mathbf{v}^j_{1\mapsto n} = \mathbf{Q}^{j\top}_{1\mapsto n}\mathbf{N}^{j\top}_{1\mapsto n}$ où $\mathbf{N}^{j\top}_{1\mapsto n}$ est le vecteur normalisé issu de $\tilde{\mathbf{F}}^j_{1\mapsto n}$ tel que :

$$\mathbf{N}^j_{1\mapsto n} = \frac{\tilde{\mathbf{F}}^j_{1\mapsto n}}{\sum_{i=1}^{inlier} \tilde{\mathbf{F}}^j_{1\mapsto n}(i)} \tag{2}$$

avec donc $\sum_{i=1}^{inlier} \mathbf{N}^j_{1\mapsto n}(i) = 1$.
De la même manière, en considérant les vecteurs normaux $\mathbf{Q}_{\perp}{}^{j}_{1\mapsto n}$, nous définissons un vecteur de mouvement normal $\mathbf{v}_{\perp}{}^{j}_{1\mapsto n} = \mathbf{Q}_{\perp}{}^{j\top}_{1\mapsto n}\mathbf{N}^{j\top}_{1\mapsto n}$ qui servira à détecter les mouvements rotationnels.

Points d'intersection : Une fois établi selon la procédure précédente l'ensemble des vecteurs de mouvement $\mathbf{v}^{\Theta}_{1\mapsto n}$ (où $\Theta = \{1, \ldots, j, \ldots, k\}$ est l'ensemble des points d'intérêt détectés entre les temps $t = 1$ et $t = n$, et k le dernier point mis en correspondance), nous estimons le *point d'intersection* de ces vecteurs. Expérimentalement, pour un mouvement d'atteinte, l'ensemble de ces vecteurs tendra à converger vers ce point. Nous définissons la matrice $\mathbf{A}^{\Theta}_{1\mapsto n}$ de dimensions $(k \times 3)$ les regroupant :

$$\mathbf{A}^{\Theta}_{1\mapsto n} = \left[\mathbf{v}^1_{1\mapsto n}, \ldots, \mathbf{v}^j_{1\mapsto n}, \ldots, \mathbf{v}^k_{1\mapsto n}\right]^{\top}.$$

Pour estimer le point d'intersection, nous formulons le système non-homogène d'équations suivant :

$$\underbrace{\begin{bmatrix} & \mathbf{A}^{\Theta}_{1\mapsto n} & \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{H}} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{0_{k\times 1}} \\ 1 \end{bmatrix}}_{\mathbf{b}}. \tag{3}$$

En posant cette équation sous la forme matricielle, (3) peut être exprimée comme $\mathbf{Hm} = \mathbf{b}$, où $\mathbf{H}$ est la matrice surdéterminée des $\mathbf{A}^{\Theta}_{1\mapsto n}$ vecteurs ; Pour résoudre ce système, nous utilisons la méthode $\mathbf{QR}$ du fait de sa stabilité numérique [16].
De la même manière, nous estimons le *point d'intersection normal* $\hat{\mathbf{m}}_{\perp}$ comme le point d'intersection des vecteurs normaux $v_{\perp}{}^{\Theta}_{1\mapsto n}$.

Indicateurs extraits : A partir de la représentation du mouvement et des points d'intersection précédemment déterminés, nous proposons donc d'extraire huit indicateurs qui nous permettrons de prédire les types de mouvements de la main, i.e. d'approche (ou atteinte), d'éloignement, de rotation ou de translation. Nous rappelons qu'il s'agit ici de détecter le mouvement principal de la main, donc d'attribuer une sémantique au geste effectué, et non pas d'estimer le mouvement de la caméra. Aussi les éléments estimés sont issus de l'observation des gestes réalisés. Ils ont été essentiellement déterminés de sorte qu'ils soient caractéristiques de l'un ou l'autre de ces mouvements ou qu'ils permettent de les distinguer entre eux.

-Mouvement d'atteinte **:** Les deux premiers paramètres proposés sont liés au mouvement d'atteinte. Lors d'un mouvement d'atteinte, l'ensemble des points suivis tendent à s'éloigner du point d'intersection $\hat{m}$ précédemment déterminé. Aussi nous allons caractériser cet éloignement. Tout d'abord, nous calculons sur la FTG, une position moyenne pondérée de chaque point suivi, en considérant l'ensemble des points $\mathbf{P}^j_{1\mapsto n}$ sous la forme d'une matrice de dimension $(inliers \times 3)$:

$$\mathbf{P}^j_{1\mapsto n} = \left[\mathbf{p}^j_1, \ldots, \mathbf{p}^j_i, \ldots, \mathbf{p}^j_n\right]^{\top}.$$

En utilisant la pondération selon les angles de l'équation 2, nous obtenons cette position moyenne, notée $\mathbf{p}^j_{1\mapsto n} = \mathbf{P}^{j\top}_{1\mapsto n}\mathbf{N}^{j\top}_{1\mapsto n}$. Etendant cette procédure à tous les points, nous obtenons la matrice suivante :

$$\mathbf{p}^{\Theta}_{1\mapsto n} = \left[\mathbf{p}^1_{1\mapsto n}, \ldots, \mathbf{p}^j_{1\mapsto n}, \ldots, \mathbf{p}^k_{1\mapsto n}\right]^{\top}$$

Ainsi, en tenant compte des positions au temps $t = n$, notées $\mathbf{p}^{\Theta}_n = \left[\mathbf{p}^1_n, \ldots, \mathbf{p}^j_n, \ldots, \mathbf{p}^k_n\right]^{\top}$, nous définissons deux distances Euclidiennes par rapport au point d'intersection $\hat{m} : d_{1,m}(j) = \|\mathbf{p}^{\Theta}_{1\mapsto n}(j) - \hat{\mathbf{m}}\|$ et $d_{n,m}(j) = \|\mathbf{p}^{\Theta}_n(j) - \hat{\mathbf{m}}\|$ pour chaque point. Si la première est inférieure à la seconde, le point j se sera globalement éloigné du point d'intersection. Pour coder cette interprétation, nous définissons la fonction $v(j)$ telle que :

$$v(j) = \begin{cases} 1 & \text{si } d_{n,m}(j) \geq d_{1,m}(j) \\ 0 & \text{sinon.} \end{cases}$$

Ainsi nous pouvons définir nos deux paramètres (f_1, f_2) qui seront simplement la moyenne $f_1 = \mu(v)$ et la variance $f_2 = \sigma^2(v)$ sur l'ensemble des points. En effet, $f_1 \mapsto 1$ dans le cas d'un mouvement d'approche (et inversement, $f_1 \mapsto 0$ dans le cas d'un mouvement d'éloignement) ; une faible valeur de f_2 doit confirmer cette prédiction.

-Mouvement de rotation **:** Le troisième paramètre se base sur une approximation de la vitesse de rotation de chaque point, construite de manière à ce qu'elle soit indépendante du sens de rotation. Soient deux points $\mathbf{p}^j_{\lambda} \mapsto \mathbf{p}^j_n$ mis en correspondance et séparés de λ-images,nous définissons s^j_i et s^j_{λ} tels que :

$$s^j_i = \frac{y^j_i - y^j_n}{x^j_i - x^j_n}, \qquad s^j_{\lambda} = \frac{y^j_{\lambda} - y^j_n}{x^j_{\lambda} - x^j_n}.$$

Ceci nous permet de calculer l'angle entre les vecteurs $\mathbf{q}^j_{i,n}$ et $\mathbf{q}^j_{\lambda,n}$:

$$\theta^j_{i,\lambda} = \arctan\left|\frac{s^j_i - s^j_{\lambda}}{1 + s^j_i s^j_{\lambda}}\right|,$$

Ce qui nous permet d'approximer les vitesses angulaires $\omega^j_{i,\lambda} = \frac{\triangle\theta^j_{i,\lambda}}{\triangle t_{i,\lambda}}$ avec $i = 1, \ldots, inliers$ et $\triangle t_{i,\lambda}$ la différence temporelle entre les images. Nous obtenons alors la troisième caractéristique f_3 :

$$f_3 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{inlier} \sigma^2(\omega^j_{i,\lambda})}{\sum_{j=1}^{k}\sum_{i=1}^{inlier} \sigma^2(\|\mathbf{p}^j_i - \mathbf{p}^j_{\lambda}\|)}, \tag{4}$$

Elle distingue les mouvements en rotation de ceux en translation car elle tends vers 0 pour les seconds et est supérieure à 1 pour les premiers, sachant que les gestes naturels sont composés d'enchaînement d'accélérations et de décélérations.

-Aire de rotation **:** L'aire du triangle composée par le point d'intersection, la position moyenne des points et la position du point final est un bon indicateur si le mouvement est en direction de l'objet ou non. Aussi la quatrième caractéristique est définie formallement comme :

$$f_4 = \frac{1}{2k} \sum_{j=1}^{k} d_{1,m}(j) d_{n,m}(j) \sin(\phi_{1,n}^j) \tag{5}$$

avec $\phi_{1\mapsto n}^j$ est l'angle du triangle au point $\hat{\mathbf{m}}$, et $d_{1,m}(j)$ et $d_{n,m}(j)$ les longueurs des segments adjacents.

-Variation ... **:** Pour gérer le cas d'une rotation pure, nous proposons un indicateur similaire au précédent mais en remplaçant le point d'intersection par le point d'intersection des normales ; on a alors :

$$f_5 = \frac{1}{2k} \sum_{j=1}^{k} d_{1,m_\perp}(j) d_{n,m_\perp}(j) \sin(\rho_{1,n}^j) \tag{6}$$

où $\rho_{1,n}^j$ est l'angle au point $m_\perp$. Cet indicateur prends une valeur élévée lorsque le mouvement n'est pas une rotation puisque le point d'intersection normal n'existe théoriquement pas, et devient stable lorsque celui-ci s'en rapproche. En combinant les angles $\rho_{1,n}^j$ et $\phi_{1,n}^j$ de la manière suivante :

$$f_6 = \frac{\sum_{j=1}^{k} \phi_{1,n}^j}{\sum_{j=1}^{k} \rho_{1,n}^j}, \tag{7}$$

nous obtenons un paramètre dont les variations sont de bons indicateurs du type de mouvement. Pour des mouvements en rotation, f_6 devient constant avec une faible valeur. Dans le cas de mouvement en translation, f_6 prends une forte valeur et lors d'un mouvement d'atteinte ou d'éloignement, il croît ou décroît respectivement.

-Angles parallèles **:** Ce septième indicateur est défini comme tel :

$$f_7 = \frac{\sum_{j=1}^{k} (\|p_{1\mapsto n}^j - p_n^j\|)}{k\sigma^2(\psi)} \tag{8}$$

où ψ est l'angle absolu du vecteur $\overrightarrow{p_{1\mapsto n}^j p_n^j}$. Ce rapport tends vers 0 lors d'un mouvement en rotation et est élevé lorsque le mouvement est purement en translation.

- Accélération **:** Comme nous l'avons dit, les gestes naturels sont composés de phases d'accélération et de décélération non continues de la main. Le huitième indicateur est donc composé pour détecter ces variations :

$$f_8 = \frac{\sigma^2(a_x^i)}{\sigma^2(a_x^i) + \sigma^2(a_y^i)}, \tag{9}$$

où a_x^i et a_y^i sont des moyennes d'estimations des accélérations selon les axes du repère image, moyennes calculées sur l'ensemble de la FTG pour chaque point suivi, i.

- Vecteur des caractéristiques **:** Les huit indicateurs précédents ont été estimés sur une FTG ; ils sont regroupés dans un vecteur $\mathbf{o}_1 \equiv \mathbf{o}_{1\mapsto n} = [f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8]^\top$. Ils seront fournis comme entrées à un HMM. Cependant, ceci requiert une séquence de données, donc une séquence de fenêtres glissantes. Nous noterons $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$ cette séquence d'observations, où $T+1$ est le nombre d'images de la séquence vidéo utilisée alors pour la construire.

La reconnaissance du geste : Actuellement les HMMs sont employés dans un large spectre d'applications, et spécialement lorsqu'il s'agit de tenir compte de phénomènes variant dans le temps et l'espace. Ils sont notamment utilisés dans la reconnaissance de gestes [17] ou d'actions [10]. Un HMM est composé par un nombre de N états $\{S_1, S_2, \ldots, S_N\}$ connectés via des transitions. A chaque transition est associée une probabilité définie dans une matrice A ; les probabilités d'émettre une observation connaissance l'état, définies par la matrice B ; et une distribution de l'état initial $\pi = \{\pi_i\}$. Dans sa notation compacte, un HMM est un triplet $\Lambda = (A, B, \pi)$.

Notre problème est de classifier chaque classe comme un mouvement de l'utilisateur. En premier lieu, nous apprenons un HMM pour chaque catégorie en utilisant l'algorithme de Baum-Welch (aussi connu sous le nom de l'algorithme *Forward-Backward*[18]). Une fois établis les paramètres des HMMs lors de la phase d'apprentissage, notre objectif en prédiction est de reconnaître une sequence d'observations $\mathbf{O}$ comme une classe de mouvement. Il s'agit de calculer $p(\mathbf{O}|\Lambda_i)$ pour chaque Λ_i et nous choisissons la classe avec la probabilité maximale : $class = arg\max_i(p(\mathbf{O}|\Lambda_i))$.

3.2 Reconnaissance de l'intention d'un geste de préhension

Cette section décrit la procédure (cf. Fig.4) mise en oeuvre pour détecter l'intention de préhension, c'est-à-dire l'objet cible et le mouvement d'atteinte, en utilisant les informations issues de l'*Eye-tracker* (position du regard et scène observée) et notre estimation du mouvement de la main. Il s'agira comme précédemment de définir un ensemble d'indicateurs que nous fournissons ensuite à un nouvel HMM.

Fixations du regard :

Lorsque qu'un humain initie une préhension, son regard se focalise sur l'objet à saisir[19]. Cependant, celui-ci reste saccadé afin d'explorer l'objet ou son environnement proche. Aussi nous proposons comme indicateur de fixation h_1 la variance des vitesses estimées sur une FTG. h_1 aura une valeur faible lors d'une fixation sur un objet et haute lorsque les saccades servent à explorer l'environnement. Cependant ceci ne nous indique pas à plus long terme si le regard se fixe sur le même objet d'intérêt.

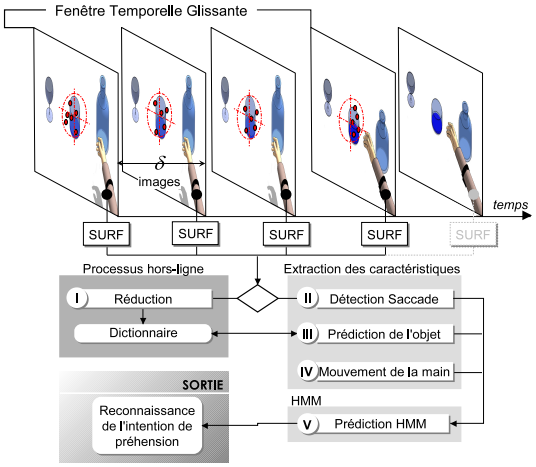


FIGURE 4 – Processus de reconnaissance de l'intention de préhension

Construction du dictionnaire visuel : Dans un premier temps, il s'agit ici de déterminer les descripteurs les plus pertinents pour reconnaître l'objet désiré au cours d'une séquence vidéo. Nous utilisons une méthode similaire à celle proposée par Sivic et Zisserman[20] pour construire un dictionnaire visuel.
Il y a plusieurs manières de créer un dictionnaire[21]. Dans notre approche, nous commençons par extraire aléatoirement des images d'une séquence vidéo acquise par la caméra scène de l'*Eye-tracker* lorsque le regard est focalisé sur un objet en particulier. Nous utilisons la position du regard dans l'image pour extraire uniquement des descripteurs SURF proches de celle-ci, donc liés à cet objet. Ensuite nous utilisons l'algorithme *Vector Quantification* (VQ) pour créer un dictionnaire pour cet objet. Reproduisant cette approche pour chaque objet, nous regroupons les descripteurs obtenus via l'algorithme VQ dans une matrice $\mathbf{D_n}$ qui constituera le dictionnaire pour l'ensemble des objets qui constitueront notre scène expérimentale. A chaque descripteur de ce dictionnaire est bien entendu associé l'objet d'où il est issu.

Reconnaissance de l'objet regardé : Nous extrayons tout d'abord les descripteurs proches de la position du regard sur l'ensemble des images d'une FTG acquise d'une séquence fournie par la caméra scène de l'*Eye-tracker*. Ceci permet d'accroître la probabilité de classer l'objet puisque nous disposons alors de plus de descripteurs. Ensuite, nous utilisons une distance correspondant au cosinus de l'angle entre les vecteurs pour comparer ces descripteurs à ceux du dictionnaire, et nous associons à chaque descripteur extrait, la classe du descripteur le plus proche dans le dictionnaire. Finallement, nous réalisons une simple sommation pour chaque classe. L'indicateur h_2 correspondra la classe ayant obtenu la somme la plus élevée.

Estimations du mouvement de la main : Les quatre derniers indicateurs correspondent aux probabilités en sortie des HMM décris dans la section précédente pour la reconnaissance du mouvement de la main ; nous avons donc pour chacun des mouvements un indicateur : $h_3 = p(\mathbf{O}|\lambda_1)$, $h_4 = p(\mathbf{O}|\lambda_2)$, $h_5 = p(\mathbf{O}|\lambda_3)$ et $h_6 = p(\mathbf{O}|\lambda_4)$.

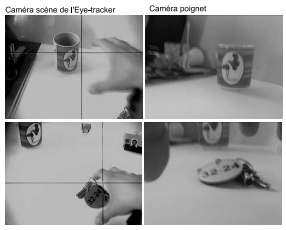


FIGURE 5 – Exemples d'acquisitions obtenues de la scène par le système.

Définition du HMM : Dans le même esprit que dans la section précédente, nous regroupons les indicateurs extraits dans une FTG dans un vecteur d'observations : $\mathbf{o}_1 \equiv \mathbf{o}_{1 \mapsto n} = [h_1, h_2, h_3, h_4, h_5, h_6]^\top$, puis nous définissons une séquence d'observations sur les séquences d'images en faisant glisser la fenêtre. Nous avons considéré deux types d'actions pour lesquels nous avons appris des HMMs : l'action de fixation et l'action de préhension. A l'aide de ces HMMs, nous procédons alors de la même manière que pour la reconnaissance du mouvement.

4 Résultats expérimentaux

Cette section présente les résultats obtenus au cours de deux expériences. La première a consisté à tester notre approche de reconnaissance du mouvement de la main sur des objets de la vie quotidienne. La seconde porte sur la prédiction de l'intention de préhension (action d'atteinte et objet cible), en couplant les informations fournies par l'*Eye-tracker*[2] et la caméra fixée au poignet. La figure 5 montre deux exemples d'images de la scène capturées par ce système lors d'une préhension.

Expérience 1 : Nous avons défini quatre mouvements que nous voulons reconnaître : l'atteinte, l'éloignement, la translation et la rotation. Notre objectif ici est de vérifier si le mouvement a été correctement prédit pour chaque FTG sur des séquences test durant lesquelles nous avons réalisés ces mouvements en présence d'objets divers. Les séquences ont été acquises à 30 images/secondes avec une résolution de 320x200 pixels en niveau de gris. Elles représentent au total 21544 images, ce qui correspond à 7131 FTGs analysées (chaque FTG dure 10 images, et est séparée de $\delta = 3$ images de la précédente). Pour entraîner les HMMs dédiés à chaque mouvement, nous utilisons 1466 FTGs acquises en réalisant ces mouvements en présence d'une tasse à café (cf. image en haut à droite de la fig. 5). Ceci afin de vérifier l'indépendance du système vis-à-vis du type d'objets en présence. De plus, puisque les performances d'un HMM varie en fonction des données fournies en entrée, nous avons appliqué une approche type valida-

2. nous avons utilisé le système ASL Eye-Trac 6.

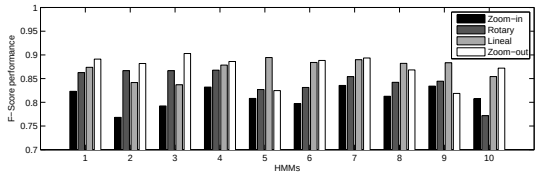


FIGURE 6 – Performances obtenues pour la reconnaissance du mouvement de la main pour les 10 HMMs, en moyenne sur les 5 objets

tion croisée en n'apprenant de manière tournante que sur 90% des données. Nous avons donc testé 10 HMMs différents. L'évaluation a été réalisée avec 5 objets différents (une tasse, une bouteille d'eau, une tasse à café, une boîte de bonbons et une bouteille de dédodorant).
La figure 6 montre les résultats [3] obtenus des 10 différents HMMs en moyenne sur ces objets. Nous constatons que ces performances varient autour de 0.85. Le mouvement d'atteinte obtient les moins bons résultats ; il est souvent confondu avec la rotation. Les figures 7a et 7b montrent les résultats obtenus en différentiant à présent les objets. Nous constatons que la bouteille d'eau obtient les moins bons résultats. En fait, il s'agit d'un objet sur lequel relativement peu de descripteurs SURF sont extraits. A l'inverse, la tasse de café et la boîte de bonbon du fait de la présence de plus de texture obtiennent les meilleurs résultats. Dans les figures 7b et 7c, nous avons différentié la moyenne sur les 10 HMMs, le HMMs ayant obtenu le meilleur F-Score et celui qui a obtenu le meilleur Taux de Vrais Positifs (TPC). Nous pouvons constater dans ces figures que le second obtient une performance supérieure de 2% par rapport à la moyenne et que le troisième à une performance supérieure de 4%. C'est ce dernier HMM qui sera utilisé dans la seconde expérience.

Expérience 2 : Dans cette seconde série d'expériences, nous testons le sytème complet décrit dans la section 3.2.Nous avons placé quatre objets différents [4] sur une table uniforme à une distance les uns des autres de l'ordre de 15cm (sans obstacles). Puis équipé du système, un utilisateur devait simuler la préhension d'un objet (sans les saisir) puis retirer sa main et recommencer avec un autre objet.
Des séquences synchronisées des images (et données sur le regard) fournies par les deux caméras ont été crées. Elles sont composées de 404 FTGs. Sur ces séquences, nous avons évalué en procédant par validation croisée les performances du HMM pour la prédiction du geste de préhension. Comme précédemment, ces performances sont calculées en fonction du nombre de FTGs correctement prédites. Il s'agit certes de résultats préliminaires et demanderaient dans l'avenir à être étendus à une base de données plus importante, mais ils permettent de réaliser une première validation de l'approche.

3. en terme de F-score $= 2(\text{précision.rappel})/(\text{précision} + \text{rappel})$.

4. Une tasse de café, un trousseau de clés, une boîte de bonbons et une carte d'accès ; pour lesquels nous avons appris un dictionnaire visuel à partir d'autres vidéos où les objets ont été placés en différentes positions.

TABLE 1 – Performances obtenues pour la reconnaissance de l'intention de préhension

	Classé comme		Performance	
Classe	Fixation	Préhension	**TVP**	**TFP**
Fixation	270	54	83.3%	1.9%
Préhension	6	74	92.5%	16.7%

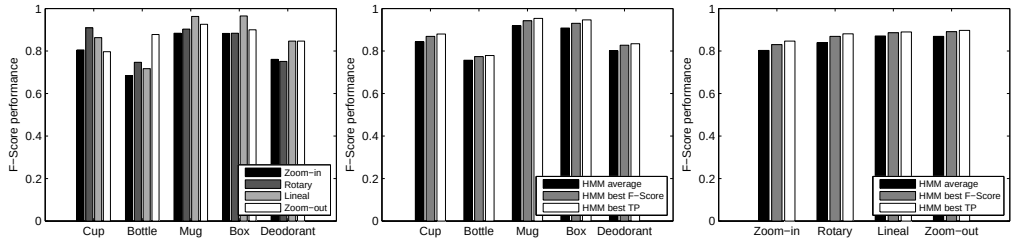
TABLE 2 – Résultats concernant la reconnaissance des objets

	Classé comme				Performances	
Objet	Tasse	Clés	Boîte	Carte	**TVP**	**TFP**
Tasse	119	1	2	1	96.7%	0.7%
Clés	1	128	3	4	94.1%	2.2%
Boîte	0	2	74	1	96.1%	1.5%
Carte	1	3	0	64	94.1%	1.8%
				Moyennes	95.3%	1.6%

Dans le tableau 2, nous pouvons constater que la méthode de reconnaissance d'objets utilisant un dictionnaire construit avec l'algorithme VQ et l'extraction de descripteurs sur une FTG semble être une bonne approche. Il s'agit cependant d'un problème relativement simple : distinguer 4 objets différents entre eux. Dans l'avenir, pour une meilleure généricité applicative, il serait intéressant de considérer plutôt des classes d'objets.
Le tableau 1 présente les résultats obtenus pour la reconnaissance de l'intention sous la forme d'une matrice de confusion et des performances en Taux de Vrais Positifs (TVP) et Taux de Faux Positifs (TFP). Nous pouvons y constater que le geste de préhension est reconnu avec un fort taux ; cependant proportionnellement, il y a un taux de faux positifs assez important. En fait, nous avons constaté qu'il s'agissait d'événements relativement isolés, se produisant souvent lorsque la personne se positionne pour prendre l'objet mais n'amorce pas le geste d'atteinte. Ce comportement de l'utilisateur, non prévu dans le protocole, est sans doute induit par le stress expérimental. Nous pensons donc que ce taux de faux positifs pourrait être nettement diminué en considérant l'immobilité de la main dans les mouvements à reconnaître.

5 Conclusions

Les expérimentations démontrent qu'il est possible de prédire le geste de préhension en utilisant notre capteur de vision combinant un *eye-tracker* et une caméra fixée au poignet. De plus, le paradigme de Fenêtre Temporelle Glissante s'avère efficace dans le cadre de la reconnaissance de mouvement, de geste ou d'objets. Il permet d'accroître le nombre et la diversité des descripteurs visuels et temporelles par rapport à une approche image-par-image. Nous obtenons ainsi des taux de reconnaissance entre 80% et 90%. Nous avons présenté par ailleurs plusieurs pistes pou-



(a) (b) (c)

FIGURE 7 – (a) Performances obtenues pour la reconnaissance du mouvement avec les 5 objets, en moyenne sur les 10 HMMs ; (b) Performances obtenues sur les 5 objets selon le HMM choisi (en moyenne sur le type d'actions) ; (c) Performances obtenues sur les 4 actions différentes selon le HMM choisi (en moyenne sur les 5 objets)

vant servir à améliorer notre système (notamment diminuer le taux de fausses alarmes, encore élevé) et à élargir son cadre expérimental. Aussi, il pourrait être étendu à d'autres types d'applications en robotique [22, 23] ou de manière plus générale en interaction homme-machine. Dans des travaux futurs, nous envisageons notamment d'exploiter la redondance d'informations visuelles entre les deux scènes perçues par l'*eye-tracker* et la caméra fixée au poignet.

Références

[1] M. Mihelj, T. Nef, and R. Riener, "Armin ii 7 dof rehabilitation robot : mechanics and kinematics," in *IEEE ICRA*, 10-14 April 2007.

[2] J. Perry, J. Rosen, and S. Burns, "Upperlimb powered exoskeleton design," *IEEE trans. on Mechatronics*, vol. 12, no. 4, pp. 408–417, Aug. 2007.

[3] N. Jarrassé, J. Robertson, P. Garrec, J. Paik, V. Pasqui, Y. Perrot, A. Roby-Brami, D. Wang, and G. Morel, "Design and acceptability assessment of a new reversible orthosis," in *IROS*, 2008, pp. 1933–1939.

[4] N. Jarrassé, J. Paik, V. Pasqui, and G. Morel, "How can human motion prediction increase transparency ?" in *ICRA*, 2008, pp. 2134–2139.

[5] J. Crawford, W. Medendorp, and J. Marotta, "Spatial transformations for eyehand coordination," *Journal of Neurophysiology*, vol. 92, pp. 10–19, 2004.

[6] R. Johansson, G. Westling, A. Bäckström, and J. Flanagan, "Eye-hand coordination in object manipulation," *The Journal of Neuroscience*, vol. 21, no. 17, pp. 6917–6932, 2001.

[7] J. Aggarwal and Q. Cai, "Human motion analysis : A review," *CVIU*, vol. 73, pp. 90–102, 1997.

[8] K. Kim, K. Kwak, and S. Ch, "Gesture analysis for human-robot interaction," in *ICACT*, vol. 3, 2006.

[9] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE PAMI*, vol. 23, 2001.

[10] C. Achard, X. Qu, A. Mokhber, and M. Milgram, "Action recognition with semi-global characteristics and hidden markov models," in *ACIVS*, 2007.

[11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, June 2008.

[12] S. Treuillet, E. Royer, T. Chateau, M. Dhome, and J.-M. Lavest, "Body mounted vision system for visually impaired outdoor and indoor wayfinding assistance," in *CVHI*, 2007.

[13] S. Muhammad Hanif and L. Prevost, "Texture based text detection in natural scene images - a help to blind and visually impaired persons," in *CVHI*, 2007.

[14] H. Bay, T. Tuytelaars, and L. Gool, "Surf : Speeded up robust features," in *ECCV*, May 2006.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, 2004.

[16] N. Higham, *Accuracy and Stability of Numerical Algorithms*. SIAM, 1996.

[17] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *CVPR*, June 1992.

[18] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, Feb. 1989.

[19] M. Hayhoe, A. Shrivastava, R. Mruczek, and J. Pelz, "Visual memory and motor planning in a natural task," *Journal of Vision*, vol. 3, 2003.

[20] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE PAMI*, vol. 31, no. 4, April 2009.

[21] A. Gersho, A. and R.M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Press, 1992.

[22] Y. Tamura, M. Sugi, J. Ota, and T. Arai, "Estimation of user's intention inherent in the movements of hand and eyes for the deskwork support system," in *IEEE/RSJ IROS*, Nov. 2007.

[23] D. Sasaki, T. Noritsugu, T. Masahiro, and H. Yamanmoto, "Wearable power assist device for hand grasping using pneumatic artificial rubber muscle," in *IEEE ROMAN*, 2004.