Suivi et reconnaissance de gestes par vision monoculaire en temps réel : application à la formation des *chargés de manœuvres* pour la conduite des ponts polaires

tracking and gesture recognition by monocular vision : application to the training course for the polar crane driving in nuclear plant

T. Chateau[1] F. Jurie[1] R. Marc[2] M. Dhome[1]

[1] Lasmea,
[2] EDF

Lasmea, UMR6602 du CNRS/Université Blaise Pascal, 63172 Aubiere Cedex, France
thierry.chateau@lasmea.univ-bpclermont.fr

Résumé

Nous présentons une application de suivi et de reconnaissance de gestes humains.
Les objectifs visés sont doubles. D'une part, nous souhaitons obtenir une reconstruction 3D temps réel des membres supérieurs d'une personne, en vision monoculaire. D'autre part nous souhaitons également interpréter ces gestes, en exploitant une base de gestes types.
Ce système est utlisé dans une application de formation de stagiaires chargès de manœuvres (conduite d'un pont de levage) en réalité virtuelle. Le suivi de geste permet l'animation d'un avatar qui sera vu par l'ensemble des acteurs, dans la simulation ; la reconnaissance de gestes permet le déclanchement d'évènements liés à la formation.
Le système proposé, qui doit être facile à transporter et simple à mettre en oeuvre, fonctionne en temps réel à une cadence de 15 images par seconde.

Mots Clef

Suivi de gestes, reconnaissance de gestes, reconstruction 3D, réalité augmentée.

Abstract

The application described here is about gesture tracking and recognition.
Two functions have to be provided by the final system. On one hand, a 3D real time reconstruction of the upper limbs must be produce from a monocular image. On the other hand, we propose a method to recognize gestures from a typical gestures basis.
The system is used into an application of training course for the polar crane driving in nuclear plant.
Real time gesture tracking provides information to an avatar which is seen by all the actors into the virtual nuclear plant reactor.
Gesture recognition allows the control of events linked to the training.
The proposed system, which have to be transported easiely, is working in real time at 15 images by seconde.

Keywords

gesture recognition, gesture tracking, 3D reconstruction, advanced reality.

1 Introduction

Le bâtiment du réacteur des centrales nucléaires est équipé d'un pont polaire (pont de levage) servant à transporter des charges lourdes. La conduite de ce pont est assurée par plusieurs personnes. D'une part, le "*pontier*", assis en hauteur dans une cabine (à 20m de hauteur), est aux commandes du pont. D'autre part, un *chargé de manœuvres* guide ce dernier, à partir du sol, par un langage gestuel.
Dans le cadre de la formation de ces deux personnes, EDF souhaite mettre en place un système de formation en réalité virtuelle, qui doit être à la fois simple à transporter et à mettre en œuvre (utilisable par un non spécialiste). A cette fin, le chef de manœuvres doit pouvoir être vu dans la simulation par le pontier. C'est pourquoi ce système de formation repose sur un dispositif de capture de gestes (bras uniquement) en temps réel, dispositif permettant l'animation de l'avatar du *chargé de manœuvre* dans la simulation. D'autre part, le pontier peut dans certains scénarios être remplacé par l'ordinateur. La reconnaissance des gestes du *chargé de manœuvres* permet alors d'interpréter les gestes de commandement et ainsi de réaliser le scénario de manière cohérente. La figure 1 montre les différents gestes possibles.

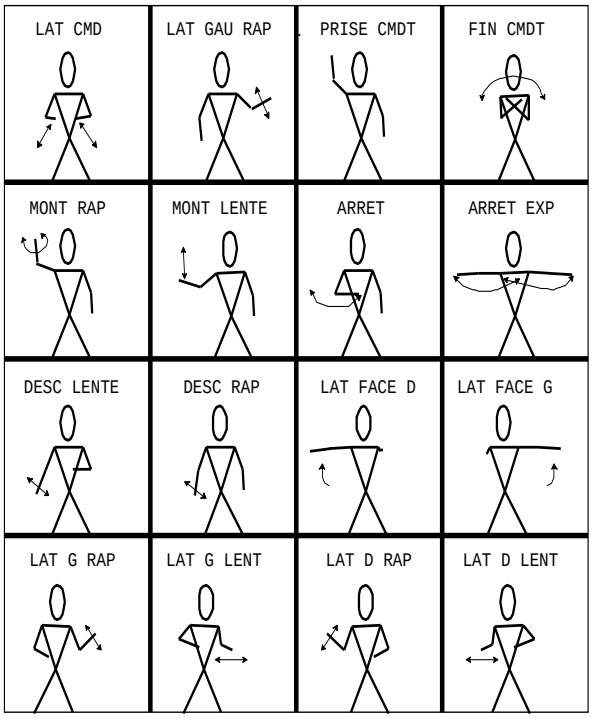


FIG. 1 – Base de gestes utilisés.

L'analyse de mouvements humains en temps réel est un domaine de la vision par ordinateur encore jeune. L'évolution des puissances de calcul rend possible la réalisation de ce type d'applications en temps réel. Dans [5], Gravila cite, entre autres, les applications suivantes :

– les mondes virtuels interactifs,
– les jeux vidéo,
– la surveillance des magasins,
– les interfaces hommes-machines,
– le langage des signes pour les mal-entendants,
– l'analyse de gestes sportifs (tennis, golf).

Parmi ces différents domaines, on peut distinguer deux groupes. Celui où il est nécessaire d'avoir une approche en trois dimensions (c'est par exemple le cas pour l'analyse de gestes sportifs) et celui où une approche en deux dimensions est suffisante (comme par exemple la surveillance des magasins). Les problèmes se rapportant au second groupe sont de manière générale bien mieux traités que ceux se rapportant au premier groupe. En effet, lorsqu'on travaille en deux dimensions, il suffit d'analyser l'image envoyée par la caméra sans se soucier de la notion de profondeur. En revanche, pour un problème comme celui de l'analyse de gestes sportifs, la notion de profondeur est essentielle.

Dans le cas du système présent, l'estimation de la pose 3D se situe dans le premier groupe. Par contre, la reconnaissance du geste effectué par le stagiaire peut être réalisée, comme nous le verrons, directement dans le plan image.

Certains travaux comme [9] utilisent des approches monovision pour extraire des informations 3D à partir de séquences video. Dans [3], les auteurs proposent, à partir

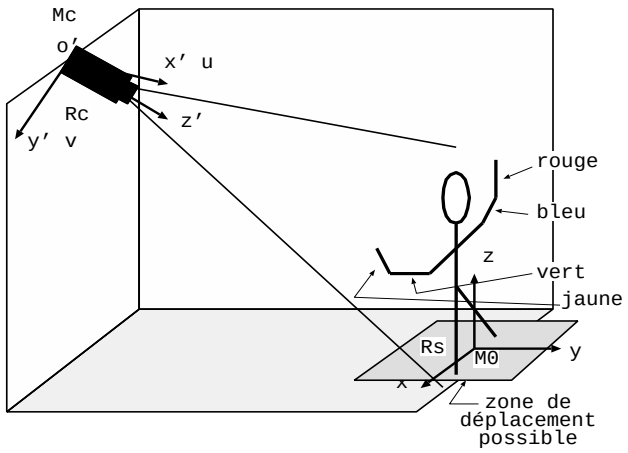


FIG. 2 – Position relative caméra stagiaire, couleur des manchons et disposition du repère scène et du repère caméra.

d'un système stéréo, d'extraire un modèle 3D simplifié d'humain.

Dans l'approche que nous proposons, le calcul de la pose 3D des bras de la personne se fait à partir d'informations issues d'une seule caméra. Pour que cela soit possible, des contraintes géométriques ont été injectées. De plus, le suivi de la pose 3D est assuré par un filtre à particules.

Dans une première partie, nous décrivons l'initialisation du système et le traitement bas niveau des images conduisant à la détection des manchons dans l'image. Puis, nous exposons la méthode utilisée pour estimer la pose 3D du stagiaire à partir d'informations 2D. Une méthode de suivi de la pose, basée sur l'utilisation d'un filtre à particules est également présentée. Dans une quatrième partie, nous décrivons le système de reconnaissance de gestes, qui s'appuie sur une comparaison des trajectoires des poignées dans le plan image avec une base de gestes de référence.

2 Initialisation du système et traitement bas niveau

La figure 2 illustre le système, composé d'une caméra couleur et d'un habit comportant des manchons colorés. La calibration de la caméra est supposée connue. L'initialisation se décompose en trois phases :

1. la modélisation du fond,
2. la localisation de la caméra par rapport à la scène,
3. l'apprentissage des couleurs des manchons et le calcul des mensurations du stagiaire.

Cette section aborde successivement chacune de ces trois phases.

2.1 Apprentissage du fond

Afin de faciliter la détection et le suivi du stagiaire, un prétraitement des images est effectué pour supprimer le fond de l'image.

Soit $\mathbf{i}_t(\mathbf{u})$ ($\mathbf{u} = (u,v)^T$) le vecteur couleur (rouge, vert, bleu) du point image de coordonnées $\mathbf{u}$ dans le repère associé au plan image. Le fond de la scène est modélisé par la moyenne et l'écart type de ses composantes couleur calculés sur un ensemble de $(tb_f - tb_0)$ images sans la présence du stagiaire :

$$\mathbf{b}(\mathbf{u}) = (\mathbf{b_m}(\mathbf{u}), \mathbf{b_s}(\mathbf{u}))^T \tag{1}$$

avec

$$\begin{array}{l} \mathbf{b_m}(\mathbf{u}) = \frac{1}{tb_f - tb_0} \sum_{t=tb_0}^{tb_f} \mathbf{i}_t(\mathbf{u}) \\ \mathbf{b_s}(\mathbf{u}) = \left[\sum_{t=tb_0}^{tb_f} (\mathbf{i}_t(\mathbf{u}) - \mathbf{b_m}(\mathbf{u}))^2 \right]^{1/2} \end{array} \tag{2}$$

En considérant que le bruit est gaussien, nous construisons, pour chaque nouvelle image $\mathtt{I}_t$, la liste des points appartenant au fond :

$$\begin{array}{l} \forall\ \mathbf{u} \in \mathtt{I}_t,\ \mathbf{u} \in B_t \text{ ssi :} \\ \left\{ \begin{array}{l} \mathbf{i}_t(\mathbf{u}) - (\mathbf{b_m}(\mathbf{u}) - 3\mathbf{b_s}(\mathbf{u})) \geq \mathbf{0} \\ \mathbf{i}_t(\mathbf{u}) - (\mathbf{b_m}(\mathbf{u}) + 3\mathbf{b_s}(\mathbf{u})) \leq \mathbf{0} \end{array} \right. \end{array} \tag{3}$$

Rappelons qu'un vecteur est dit positif (resp. négatif) si toutes ses coordonnées sont positives. L'équation précédente est vérifiée si l'inégalité de chaque terme du vecteur est vérifiée. Il faut noter que cette méthode de sélection de fond repose sur une hypothèse d'illumination constante. Si on considère que l'immumination globale de la scène peut varier, il est possible d'utiliser des composantes couleur normalisées.

2.2 Localisation de la caméra par rapport à la scène

Pour reconstruire, sur un avatar 3D, les gestes effectués par le stagiaire, il est nécessaire de localiser la caméra par rapport à un repère lié à la scène. Cela revient à déterminer la matrice $\mathtt{T}$ (trois rotations et les trois translations) qui permet de passer du repère scène au repère caméra. A cette fin, une zone rectangulaire dans laquelle devra évoluer le stagiaire est positionnée sur le sol. Les dimensions de cette zone qui sert de mire de localisation sont supposées être parfaitement connues.

On définit un repère R_s ((M_0xyz)) dans la scène et un repère R_c ($(M_cx'y'z')$) attaché à la caméra. L'axe $(O'z')$ est dirigé selon l'axe optique. Les axes $(O'x')$ est orienté de façon parallèle aux lignes de l'image. De plus, le plan $z = 0$ correspond au plan du sol. La figure 2 présente la position de ces deux repères. Classiquement, un repère associé au plan image, dont l'origine est le coin haut gauche de l'image est également défini.

Pour déterminer les 6 paramètres de localisation, nous utilisons l'algorithme **POSIT**, présenté dans [4] par Dementhon. Nous considérons 4 points 3D situés dans le même plan. Ces points sont les quatre coins de la zone rectangulaire au sol dans laquelle évolue le stagiaire. Les paramètres intrinséques de la caméra sont estimés préalablement.

2.3 L'apprentissage des couleurs des manchons et le calcul des mensurations du stagiaire

L'algorithme de suivi des articulations repose sur la détection de manchons colorés. Lors de l'intialisation, la couleur (teinte et saturation moyenne dans l'espace HSI) de chaque manchon est estimée. Pour cela, le stagiaire tend les bras horizontalement et une localisation supervisée de l'extrémité de chaque manchon est effectuée. Cette dernière permet de définir des zones de recherche initiales. Un premier critère est alors fixé : tout point dont la composante rouge (reps. vert, bleu) est "très" supérieure aux deux autres composantes est considéré comme appartenant au manchon rouge (rep. vert, bleu). Puis, un filtre morphologique basé sur une séquence de L érosions suivies de L dilatations avec un masque horizontal permet d'éliminer les pixels parasites. Le résultat de ce pré-traitement est l'extraction de quatre régions $\mathtt{R}^1$ (manchon coude poignet droit), $\mathtt{R}^2$ (manchon coude poignet gauche), $\mathtt{R}^3$ (manchon épaule coude droit), $\mathtt{R}^4$ (manchon épaule coude gauche). Pour chaque Région, on construit un vecteur de paramètres $\mathbf{S}^i$ composé de sa teinte (*hue*) et de sa saturation moyenne $(\overline{h}^i, \overline{s}^i)$, ainsi que l'écart type associé $(\sigma h^i, \sigma s^i)$. Ainsi, la couleur de chaque manchon est modélisée par un vecteur de dimension 4 :

$$\mathbf{S}^i = (\overline{h}^i, \overline{s}^i, \sigma h^i, \sigma s^i)^T \tag{4}$$

Pour l'image t, les 4 régions $(\mathbf{r}_t^i)$ peuvent maintenant être extraites en fonction d'un critère de distance entre chaque pixel et les paramètres $\mathbf{S}_t^i$:

$$\begin{array}{l} \forall\ \mathbf{u} \in \mathtt{I}_t,\ \mathbf{u} \in \mathtt{r}_t^i \text{ ssi :} \\ \left\{ \begin{array}{l} \mathbf{h}_t(\mathbf{u}) - \begin{pmatrix} 1 & 0 & -3 & 0 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{S}^i \geq \mathbf{0} \\ \mathbf{h}_t(\mathbf{u}) - \begin{pmatrix} 1 & 0 & 3 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{S}^i \leq \mathbf{0} \end{array} \right. \end{array} \tag{5}$$

$\mathbf{h}_t(\mathbf{u})$ est le vecteur des composantes hsv (teinte, saturation, luminance) du pixel de coordonnées $\mathbf{u}$. Pour estimer les extrémités de chaque manchon, l'axe d'inertie des pixels composant chaque ensemble $\mathbf{r}_t^i$ est calculé. La position des extrémités est alors estimée de telle façon que 95% des points soient compris entre ces deux extrémités.

Lors de cette initialisation, le stagiaire se trouve au centre de la zone rectangulaire au sol, donc dans un plan vertical connu. Ainsi, dans l'hypothèse où les bras sont parallèles à l'axe (Ox), il est possible de remonter aux coordonnées 3D des extrémités des manchons. La dimension de chaque manchon peut alors être estimée par cette méthode.

3 Suivi 3D temps réel de gestes humains

Dans notre cas, il faut, à partir d'informations 2D issues d'une caméra, reconstruire une attitude 3D. Pour cela, nous

θ_2
θ_1
α
z
Tx
Ty
y

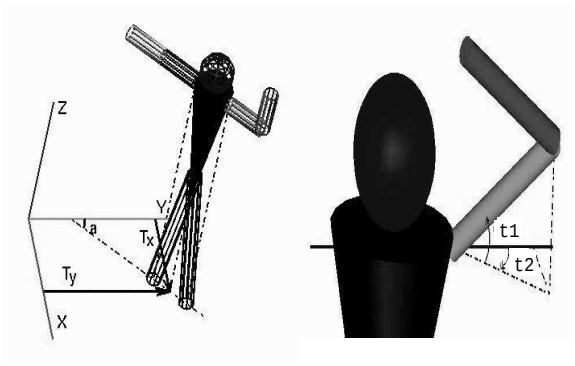


FIG. 3 – Choix des 11 paramètres pour l'estimation de la pose 3D du stagiaire.

FIG. 4 – Le plus souvent, 4 solutions sont possibles pour chaque bras

.

fixons un certain nombre d'hypothèses qui contraignent plus ou moins le système :
– la longueur des membres est supposée constante,
– la hauteur d'épaule est supposée constante.
Si la première hypothèse semble vraie, la seconde est bien plus contraignante. Néanmoins, elle permet de résoudre notre problème. La première phase consiste à choisir un modèle d'avatar articulé. Puis, chacun des paramètres de ce modèle est estimé, soit à partir de l'analyse de l'image courante dans le cas de la première itération, soit à partir d'un filtre à particules dans le cas des images suivantes.

3.1 Choix du modèle d'avatar

Le modèle choisi pour l'animation de l'avatar est de dimension onze. Il peut se déplacer sur le plan horizontal(Oxy) (Tx et Ty) et pivoter autour de l'axe (Oz) (α). Couramment, trois rotations sont choisies pour situer un coude par rapport à une épaule, mais nous pouvons n'en considérer que deux. La première (θ_1) est une rotation autour de l'axe vertical. Elle permet de situer la projection du coude dans le plan des épaules $Z = Z_{Ep}$. La seconde (θ_2) est une rotation permettant de connaître l'inclinaison du bras par rapport au plan Z = Z_{Ep}. Ainsi, les deux degrés de liberté permettant de situer un coude sont ici des mesures d'angles. De la même manière, deux angles sont suffisants pour situer un poignet par rapport à un coude. Les 11 paramètres sont illustrés sur la figure 3.

3.2 Estimation de la pose 3D

La phase suivante consiste à estimer les coordonnées 3D des articulations dans la première image du flot vidéo.
Les épaules évoluent dans un plan connu $Z = Z\_Ep$, dans l'hypothèse où leur hauteur est constante. Pour une épaule, les inconnues sont donc son abscisse et son ordonnée. Sa projection (u, v) dans l'image étant connue, trouver les coordonnées 3D de l'épaule revient à calculer l'intersection entre le faisceau projeté en (u, v) et le plan $Z = Z\_Ep$. Ensuite, il nous faut rechercher les coordonnées 3D des coudes. La projection d'un coude dans l'image est connue, ce qui revient à dire que l'équation du rayon optique issu de la caméra passant par ce coude est connue. De plus, la longueur des bras du stagiaire a été estimée lors de l'initialisation. Le coude se trouve donc à l'intersection du rayon optique et d'une sphère centrée sur l'épaule de rayon la longueur du bras.

De la même manière, un poignet se trouve à l'intersection du rayon optique passant par ce poignet et d'une sphère centrée sur le coude de rayon la longueur de l'avant-bras.

Lors du calcul du point d'intersection entre le rayon optique et la sphère, trois cas de figures peuvent être envisagés :
– Le résultat de l'intersection est un ensemble vide. Ce cas de figure ne devrait en principe jamais se produire. Il signifie que le rayon optique n'intersecte pas la sphère, et donc qu'il n'y a aucune position possible pour le coude ou le poignet considéré. Cette incohérence est dûe principalement aux erreurs de mesure de la position des manchons. Une erreur de l'ordre du pixel dans la direction du manchon peut avoir un impact plus important lors de la reconstruction 3D de l'avatar.
– Le résultat de l'intersection est un point. Cela signifie que le rayon optique est tangent à la sphère. Ce cas est rare
– Le résultat de l'intersection est deux points. Ce cas de figure signifie que deux attitudes sont possibles mathématiquement pour le stagiaire, car se projetant de manière identique dans l'image. La figure 4 montre, que pour chaque bras, 2 solutions sont donc possibles pour le coude et 2 autres solutions pour le poignet. Ainsi, 4 configurations sont mathématiquement plausibles. L'ambiguité est levée en considérant, pour le coude, la solution la plus éloignée de la caméra, et pour le poignet, la solution la plus proche de la caméra. Ce choix empirique peut être discutable mais il conduit souvent à la bonne solution.
A ce stade, les onze paramètres permettant de coder la pose ont été déterminés. Pour les images suivantes de la séquence vidéo, un algorithme de suivi basé sur l'utilisation d'un filtre à particules est appliqué.

3.3 Suivi de la pose 3D à l'aide d'un filtre à particules

La précision de la pose 3D estimée pour la première image repose en grande partie sur la précision de la reconstruction des épaules. Si une erreur est commise à ce niveau, elle se répercute sur les autres articulations. Nous proposons l'utilisation d'une méthode d'optimisation stochastique afin de modifier les reconstructions de manière à satisfaire au mieux à un critère de reprojection globale dans l'image. Cette optimisation exploite un filtre à particules. Ce choix repose sur le fait que le filtrage particulaire est capable de prendre en compte des erreurs de mesures non gaussiennes et qu'il est particulièrement adapté au suivi multicibles.

Le principe de cet algorithme est décrit dans deux articles, l'un [1] présentant de manière assez globale l'ensemble des méthodes dites de Monte Carlo, l'autre [7] permettant une approche plus précise du filtre à particules dans le cadre du suivi de primitives visuelles. En traitement d'images, le filtre à particules, plus connu sous le nom d'algorithme CONDENSATION est utilisé dans de nombreux travaux [8, 11].

Le suivi consiste à déterminer la densité de probabilité à *posteriori* $P(\mathbf{X_t}|\mathbf{Z_t})$, à partir de la densité de probabilité $P(\mathbf{Z_t}|\mathbf{X_t})$; où $\mathbf{X_t}$ est le vecteur d'état composé des paramètres du modèle à suivre et $\mathbf{Z_t}$ représente le vecteur d'observation $\mathbf{Z_t} = \{\mathbf{z_1}, ..., \mathbf{z_t}\}$ regroupant l'historique des mesures.

L'idée de base du filtre à particules est d'approximer la distribution de probabilités liée au vecteur d'état par un ensemble fini d'éléments $S = \{(s^{(n)}, \pi^{(n)})|n = 1...N\}$. Chaque élément est composé d'une hypothèse sur la valeur du vecteur d'état (s) et d'une valeur sur la probabilité (π) liée à cette hypothèse (avec $\sum_{n=1}^{N} \pi^{(n)} = 1$). L'évolution de l'ensemble S est obtenue en propageant chaque élément de l'ensemble en fonction d'un modèle d'évolution. Un poids est alors affecté à chaque élément en fonction des observations. L'ensemble S est alors regénéré par un tirage de N éléments avec *remise* où chaque élément possède une probabilité d'être tiré de $\pi^{(N)} = P(z_t|X_t = s_t^{(N)})$. Le vecteur d'état moyen est alors estimé à chaque itération par :

$$E(S) = \sum_{n=1}^{N} \pi^{(n)} s^{(n)}$$

Une particule est une estimation de la position 3D du stagiaire. Suite à la description d'une position par des degrés de liberté, nous pouvons considérer qu'une particule évolue dans un espace de dimension onze. N particules sont définies.

Pour la première image, chaque particule est initialisée à la position calculée par la méthode décrite dans le paragraphe précédent. Une nouvelle position est ensuite prédite pour chaque particule, à l'aide d'un modèle d'évolution (dans le cas présent, nous considérons un modèle à position fixe). Les particules sont ensuite diffusée, afin de disposer d'un nuage de particules distinctes autour de la position calculée. Cette diffusion se fait en ajoutant un bruit gaussien à chacune des onze coordonnées de chaque particule. On calcule alors pour chaque particule la distance séparant sa projection dans l'image de la position 2D du stagiaire mesurée dans l'image 2. Puis on affecte à chacune un poids d'autant plus fort que cette distance est faible et tel que la somme des poids des particules soit unitaire.

Après itération, nous avons un ensemble de particules pondérées regroupées autour de la position calculée du stagiaire. La pose 3D estimée est représentée par le centre de gravité de l'ensemble de particules.

La seconde itération commence avec une phase de *drift*. Elle consiste en deux étapes. Tout d'abord un tirage au sort avec remise de N particules ($100 < N < 200$ dans le cas présent), parmi la totalité des particules. Cette sélection tient compte du poids des particules. Plus une particule a un poids élevé, plus elle a de *chances* d'être choisie plusieurs fois. Puis, les nouvelles particules sélectionnées sont déviées (par le modèle d'évolution et par le bruit gaussien), la distance de reprojection dans la nouvelle image est calculée, de nouveaux poids sont affectés et le nouveau centre de gravité des particules est calculé. Avec cette méthode, il n'est plus utile d'estimer la pose 3D par la méthode présentée dans le paragraphe 3.2 pour chaque image de la séquence vidéo. En effet, les seules primitives utiles sont les positions des manchons dans l'image afin de calculer les distances de reprojection entre les particules et la position des articulations dans l'image 2D. La figure 5 illustre le fonctionnement du filtre. La figure 6 est un exemple de résultat obtenu. L'image représentée sur la sous-figure (6.a) montre la scène observée par la caméra et le stagiaire en action. La sous-figure (6.b) est une représentation 3D de l'avatar reconstruit. La figure 7 est un autre exemple de reconstruction, où une vue de dessus de l'avatar est proposée.

4 Reconnaissance automatique des gestes

Le but du système présenté dans ce chapitre est de reconnaître, à partir d'un flux vidéo temps réel, l'ensemble des gestes réalisés par le stagiaire chef de manoeuvres.

Il s'agit en fait d'un double problème :

– il faut d'une part segmenter temporellement la séquence en gestes, c'est à dire repérer les moments où la personne passe d'un geste à un autre,
– d'autre part, pour chaque segment de séquence, il faut reconnaître le geste parmi une base de gestes connus.

Nous nous intéressons ici uniquement au second problème. Le premier est réglé «artificiellement» par une segmentation supervisée : le stagiaire déclenche lui-même le passage à un nouveau geste. En pratique, et pour des questions de simplicité, ce déclenchement est produit lorsque le stagiaire reste plus de 2 secondes les bras ballants.

Le problème de la reconnaissance des gestes est un problème difficile :

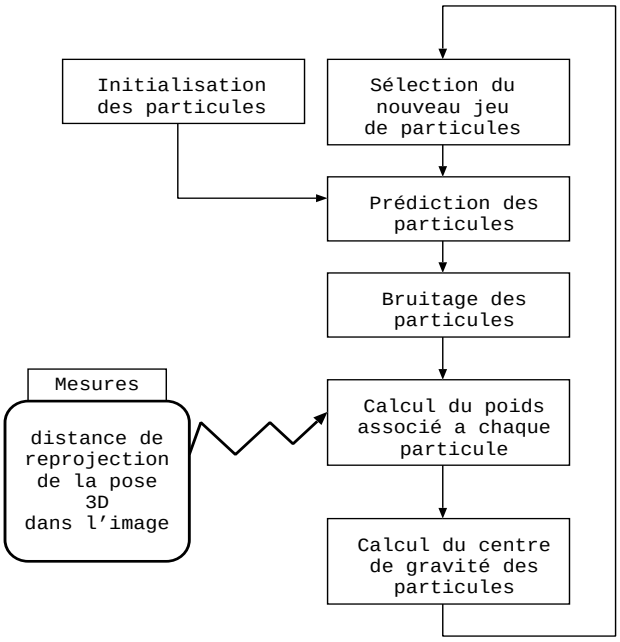
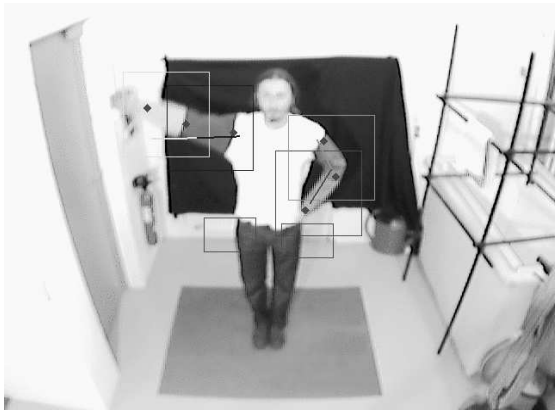
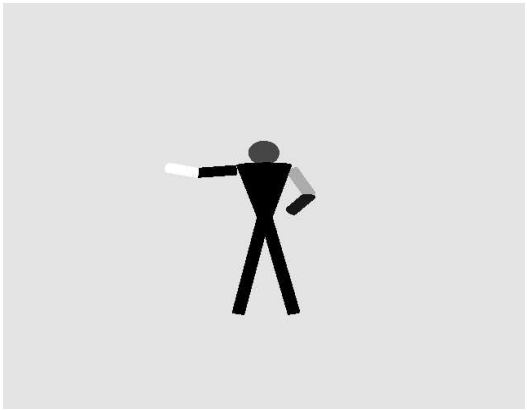


FIG. 5 – Synoptique du filtre à particules.



(a)



(b)

FIG. 6 – Exemple d'une reconstruction 3D. (a) : image originale. (b) : reconstruction de l'avatar correspondant.

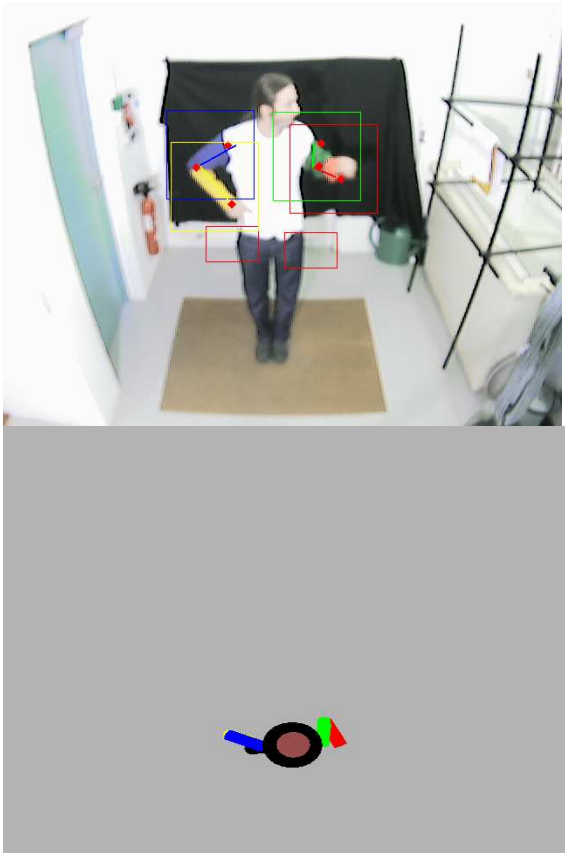


FIG. 7 – Image vidéo et reconstruction du geste vue de dessus

– Certains gestes différents correspondent à des séquences de mouvements proches ;
– Ce qui caractérise le geste ne peut se résumer à un ensemble de postures de la personne ; l'ordre de ces postures et la vitesse de chacun des mouvements jouent un rôle important ;
– Les gestes peuvent être exécutés avec plus ou moins de fidélité par rapport à ceux décrits dans le langage gestuel utilisé.

4.1 Stratégies de reconnaissance

La reconnaissance des gestes est un domaîne de recherche dans lequels de nombreux travaux ont été effectués. Le lecteur peut se reporter à [5, 12] pour obtenir un survol de ces derniers. Classiquement, deux types de stratégie peuvent être utilisés pour aborder ce problème de reconnaissance : une stratégie basée sur une décomposition des gestes comme un enchaînement de postures et une stratégie basée sur une classification des trajectoires réalisées durant le geste.

Reconnaissance basée sur un enchaînement de postures. Si les gestes ne sont pas segmentés au préalable (segmentation temporelle) il est possible de les reconnaître en utilisant des systèmes Markoviens [10]. Dans [6], les auteurs utilisent un tel système dans le cadre de la reconnaissance du language des signes. Il faut noter que cette approches est proche de celles utilisées dans la reconnaissance de la parole.

Cela revient, dans ce cas précis, à supposer que les gestes se décomposent en une succession de postures. A chaque instant la probabilité de chaque posture est calculée, en tenant compte de la probabilité d'avoir cette posture dans l'image et de la probabilité de transition de la posture précédente à la posture actuelle pour un geste donné. A chaque instant, la probabilité de chaque posture de chaque geste est ainsi calculée et la posture la plus probable indique le geste ayant le plus de probabilité.

Cette technique pourrait donner d'excellents résultats, mais elle est assez complexe à mettre en oeuvre. En effet, cela nécessite de calculer des probabilités de passage entre postures, qui sont très difficiles à obtenir. Des bases de gestes très importantes, dont nous ne disposons pas, sont nécessaires.

Reconnaissance basée sur une classification des trajectoires. Une autre manière de modéliser un geste est de le représenter comme une trajectoire dans un espace multi-dimensionnel. Dans [2], Black et Jepson proposent un système de reconnaissance de gestes basé sur un modèle de trajectoire de chaque geste. Dans le cas présent il est possible de considérer un espace de dimension 8 (si nous prenons en compte les 4 rotations possibles pour chaque membre). Chaque posture est un point de cet espace.

Comparer deux trajectoires revient alors à comparer deux courbes de cet espace multi-dimensionnel, donc deux ensembles de points dans le cas discret.

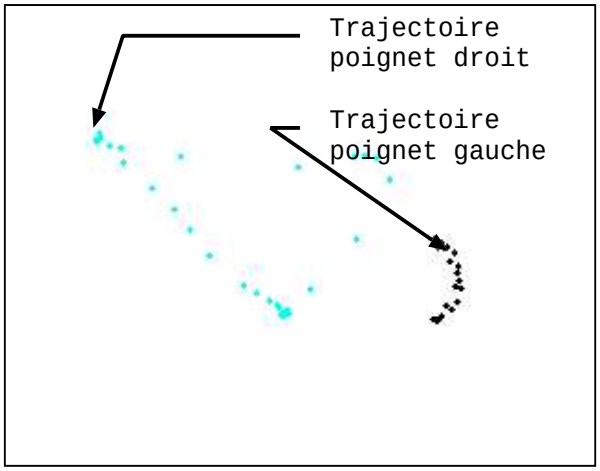


FIG. 8 – trajectoire des poignets dans le plan image pour un geste d'arrêt.

L'approche proposée ici est basée sur ce principe. Cela suppose que le geste a été temporellement segmenté (isolé du reste de la séquence). Par contre, l'espace choisi est de dimension 4. En effet, seule la trajectoire des poignets dans le plan image est utilisée.

4.2 Approche proposée pour la reconnaissance de gestes

L'hypothèse forte émise dans cette méthode réside sur le fait que le geste peut être reconnu uniquement à partir de la trajectoire des deux poignets dans l'image. Cela suppose que la caméra soit placée dans une configuration où les gestes peuvent être interprétés uniquement à partir des images. Par exemple, la figure 8 montre la trajectoire des poignets dans le plan image, pour un geste d'arrêt.

La trajectoire des poignets est exprimée dans le référentiel de l'image. Or la base de gestes types utilisés pour la reconnaissance est une base de gestes 3D.

Afin de pouvoir comparer les gestes, il est nécessaire de représenter les gestes de la base eux aussi dans le repère de l'image. Connaissant la position de la caméra par rapport à la scène et connaissant la calibration de la caméra, il est aisé de projeter la trajectoire des poignets dans le plan image.

Reconnaître un geste va consister à comparer les trajectoires 2D des poignets dans l'image avec la base des gestes types, projetés eux aussi dans le repère de l'image.

Différentes mesures de distance peuvent être envisagées. La mesure de distance utilisée est la distance de Haussdorf. Elle permet de mesurer la distance entre deux nuages de points.

Soit A un ensemble de n point $A = \{a_i, i \in [1...n]\}$, et soit $B = \{b_i, i \in [1...m]\}$ un second ensemble de m points. Nous supposons que chaque point est un vecteur de R^p, et que $D(a_i, b_b)$ représente la distance Euclidienne dans R^p. Avec ces notations, la distance de Haussdorf entre les ensembles A et B est :

$$H(A,B) = \max(h(A,B), h(B,A)) \quad (6)$$

avec :$h(A,B) = \max_i(\min_j D(a_i, b_j))$
et
$h(B,A) = \max_j(\min_j D(a_i, b_j))$
La trajectoire à reconnaître T est comparée avec chacune des trajectoires TR_i de la base des NG gestes. La probabilité du geste i est estimée à :

$$P(TR_i) = \frac{H(T,TR_i)}{\sum_{k=1}^{NG} H(T,TR_k)} \quad (7)$$

La liste des points de la trajectoire T est remise à zéro lorsqu'une nouvelle trajectoire commence. Ensuite, à chaque nouvelle image la trajectoire est augmentée de deux points et les probabilités sont recalculées. Les probabilités des gestes sont donc disponibles en permanence, et par conséquent avant que le geste ne soit terminé.
Nous avons implémenté une distance de Haussdorff modifiée pour h. Plutôt que de la plus grande des plus petites distances, on prend la k-ième plus grande, de manière à être robuste à la présence de k points parasites dans les trajectoires des poignets

4.3 Résultats expérimentaux

Nous avons utilisé, pour constituer la base des trajectoires des poignets correspondant aux gestes types, une base de gestes 3D décrivant une suite de poses de l'avatar pour chaque mouvement. Une nouvelle base a été établie, pour laquelle les gestes 3D de référence ont été projetés dans le plan image, en utilisant les paramètres de la caméra et sa position dans la scène. Nous avons réalisé une évaluation du système de reconnaissance de gestes sur l'ensemble de la base de gestes types.
Le système utilisé est composé d'un PC LINUX à 1.8GHz et d'une caméra numérique IEEE1394 délivrant 15 images par secondes à une résolution de 640×480. Une focale courte permet de disposer la caméra suffisamment près du stagiaire tout en conservant un champ de vision large.
La figure 9 illustre un résultat de reconnaissance. La figure présente la trajectoire à reconnaître (points sombres), la trajectoire la plus proche (points gris), et les probabilités associées à l'ensemble des hypothèses (colonne de gauche - chaque curseur indique la probabilité du geste). Globalement, le système de reconnaissance fonctionne correctement. Néanmoins, pour certains gestes proches (au niveau de la trajectoire des poignets) des erreurs de classification apparaissent. En fait, des valeurs de probabilité très proches sont associées à plusieurs gestes.

5 Conclusion

Nous avons présenté, dans cet article, un système de reconstruction 3D et de reconnaissance de gestes humains, par vision monoculaire. Les résultats obtenus sont satisfaisants. Le système est simple à mettre en place et à utiliser. Néanmoins, cette simplicité a été possible en imposant

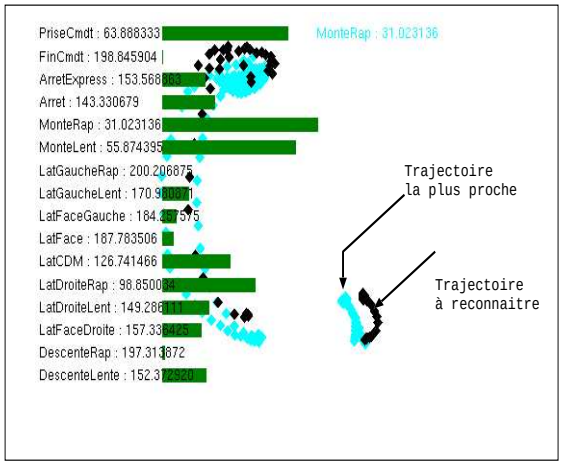


FIG. 9 – Reconnaissance de gestes (voir le texte pour les explications de ce graphique.)

un certain nombre d'hypothèses qui peuvent être contraignantes ou par des choix guidés par la simplicité plutôt que par la performance.
Le choix de travailler en vision monoculaire présente l'intérêt d'être simple à mettre en oeuvre. Par contre, lors du calcul de la pose 3D à partir d'une image monoculaire, deux sources de mauvais fonctionnement apparaissent :
– la hauteur des épaules du stagiaire doit rester constante. Cela n'est pas toujours le cas en réalité ; des imprécisions sur la reconstruction peuvent en découler,
– des ambiguïtés sur la reconstruction 3D ne peuvent pas être facilement levées. Une perspective serait de prendre en compte l'aspect dynamique de la personne pour lever certaines de ces ambiguïtés. Par exemple, certaines successions de postures pourraient être interdites car elles ne sont physiquement pas réalistes.

Pour remonter à une vraie information 3D, une solution serait d'utiliser une base stéréo.
La détection des articulations par une tenue colorée est une bonne solution. Néanmoins, un calcul d'axe d'inertie est utilisé pour localiser chaque manchon. Dans certaines configurations, (bras dirigé vers la caméra) ce calcul s'avère imprécis.
Le processus de reconnaissance de gestes n'a pas encore été testé de manière exhaustive. Cependant, les résultats observés sont satisfaisants. Néanmoins, il est impossible de distinguer deux mouvements réalisés avec les mêmes postures mais avec des vitesses différentes. De plus, lors de la réalisation de certains gestes, des probabilités très proches sur plusieurs gestes ont été observées. Dans ces cas de figures, le modèle (trajectoire des poignets dans le plan image) n'est pas suffisamment complet pour discriminer les différents gestes. Une alternative serait d'utiliser un modèle plus complexe.
Le système proposé peut être étendu à des applications de réalité virtuelle temps réel dans lesquelles un avatar est animé dans un monde virtuel à partir de gestes effectués

par un opérateur. Cet aspect d'intéraction entre les gestes effectués par l'opérateur et la machine fait du système une "interface homme/machine intelligente".

Références

[1] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2) :174–188, 2001.

[2] M. J. Black and A. D. Jepson. Recognizing temporal trajectories using the condensation algorithm. In *Third Int. Conf. on Automatic Face and Gesture Recognition*, pages 16–21, Nara, Japan, April 1998.

[3] D. Demirdjian and T. Darell. 3-D Articulated Pose Tracking for Untethered Deictic Reference. In *ICMI 2002*, Pittsburgh, Pennsylvania, USA, October 2002.

[4] D. Dementhon. *De la vision artificielle à la réalité synthétique : système d'intéraction avec un ordinateur utilisant l'analyse d'images vidéo*. PhD thesis, Université Joseph Fourier (Grenoble I), octobre 1993.

[5] D.M. Gavrila. The Visual Analysis of Human Movement : A Survey. *Computer Vision and Image Understanding*, 73(1) :82–98, January 1999.

[6] H. Hienz and B. Bauer. HMM-based Continous Sign Language Recognition using Stochastic Grammars. In *Human-Computer Interaction International Proceedings of Gesture Workshop, GW99, Lecture Notes in Artificial Intelligence*, volume 1739, pages 185–196, March 1999.

[7] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. In *Int. J. Computer Vision*, volume 29-1, pages 5–28, 1998.

[8] K. Nummiaro, E. Koller-Meier, and L. Van Gool. Object Tracking with an Adaptive Color-Based Particle Filter. In *symposium for Pattern Recognition of the DAGM*, September 2002.

[9] L. Torresani and C. Bregler. Space-Time Tracking. *Computer Vision ECCV 2002*, 1 :801–812, May 2002.

[10] S. Marcel, O. Bernier, J.E. Viallet, and D. Collobert. Hand Gesture recognition using Input/Ouput Hidden Markov Models. In *FG'2000 Conference on Automatic Face and Gesture Recognition*, pages 456–461, March 2000.

[11] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-Based Probabilistic Tracking. In *Computer Vision ECCV 2002*, volume 1, pages 661–675, May 2002.

[12] V. Pavlovic, R. Sharma, and T.S. Huang. Visual interpretation of hand gestures for human-computer interaction : A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) :677–695, 1997.