

Reconnaissance de gestes de pointage dans le cadre d'interaction avec un grand écran

S. Carbini

J.E. Viallet

O. Bernier

France Telecom R&D/DTL/TIC

Technopole Anticipa, 2 Avenue Pierre Marzin
22307 Lannion Cedex - France

{sebastien.carbini, jeanemmanuel.viallet, olivier.bernier}@rd.francetelecom.com

Résumé

Parmi l'ensemble des gestes que les hommes accomplissent lorsqu'ils communiquent, les gestes de pointage sont facilement interprétables et peuvent conduire à des interfaces homme-machine plus naturelles et plus puissantes. Nous estimons la direction pointée par l'axe oeil-doigt en détectant et en suivant automatiquement, à une fréquence de 15 Hz, les positions 3D de la tête et des mains obtenues à partir d'une caméra stéréo. A partir de la position de la tête et des contraintes biométriques, on définit une zone de repos et une zone d'action dans laquelle on recherche les mains et détecte l'intention de pointage. La première main spontanément avancée par l'utilisateur est détectée et sert au pointage d'un objet, alors que la seconde est utilisée pour la sélection. Des expériences sur la précision spatiale de notre système montrent que le rayon minimum d'un objet désignable sans gêne pour l'utilisateur est de 5cm.

Mots clefs

Interfaces homme-machine non intrusive, pointage, interaction bi-manuelle, détection, suivi, visage, mains.

1 Introduction

Les techniques de visions par ordinateur appliquées à la reconnaissance de gestes permettent de se passer de câbles et de marqueurs encombrants et permettent donc une plus grande liberté de mouvement aux utilisateurs. Parmi l'ensemble des gestes que les hommes accomplissent lorsqu'ils communiquent, les gestes de pointage sont facilement interprétables et peuvent conduire à des interfaces homme-machine plus naturelles et plus puissantes. Ainsi de nombreuses études ont été menées sur la reconnaissance de gestes de pointage par vision. Nous considérons ici un geste de pointage, non comme la reconnaissance d'une trajectoire temporelle de la main mais ainsi que [1, 2, 3], en tant que lieu pointé à l'issue du geste. Dans [3], la direction pointée est donnée par la direction de l'avant bras de l'utilisateur déterminé par un modèle 3D du buste et du bras. En terme de traitement d'images, l'avant-bras présente

peu d'informations discriminantes et est donc difficile à détecter notamment lorsque l'on pointe vers la caméra. En revanche, la tête a une forme stable et caractéristique que l'on sait bien détecter et suivre. Sans retour visuel, un pointage par un axe ne passant pas par l'oeil (doigt, avant-bras ou bras tendu) est moins précis que celui obtenu en visant naturellement la cible avec l'oeil en se servant du bout du doigt comme 'viseur'.

Nous proposons d'estimer la direction pointée en utilisant la convention 'viseur' et en approximant l'axe oeil-doigt par l'axe tête-main. La première main que tend spontanément l'utilisateur sera utilisée pour pointer et la seconde main sera utilisée pour actionner une commande (équivalent au click de la souris) ou pourrait définir un troisième axe utile notamment pour des interactions avec un monde virtuel 3D. Nous allons détailler la détection et le suivi de la tête et des mains ainsi que les performances obtenues en termes de stabilité et de précision sur le pointage.

2 Méthodologie

Notre méthode ne nécessite pas de connaître la main prédominante de l'utilisateur et fonctionne aussi bien avec les gauchers qu'avec les droitiers. De plus, aucune phase de calibration ou d'initialisation manuelle n'est requise, le visage de l'utilisateur est détecté automatiquement lorsqu'il arrive dans le champ de la caméra (Figure 1). L'utilisateur commence par désigner à l'écran, avec la main de son choix, l'objet de son interaction avant d'interagir avec : la première main avancée et détectée correspond donc à la main de pointage. Dès qu'une partie du corps (visage ou main) est détectée, son suivi est enclenché et la détection n'est réinitialisée qu'en cas de perte. On ne prend en compte la direction tête-main que si la main est suffisamment avancée car l'utilisateur ne souhaite pas forcément pointer en permanence. En position de repos, un bras le long du corps ne traduit pas une intention de pointage (Figure 2-a).

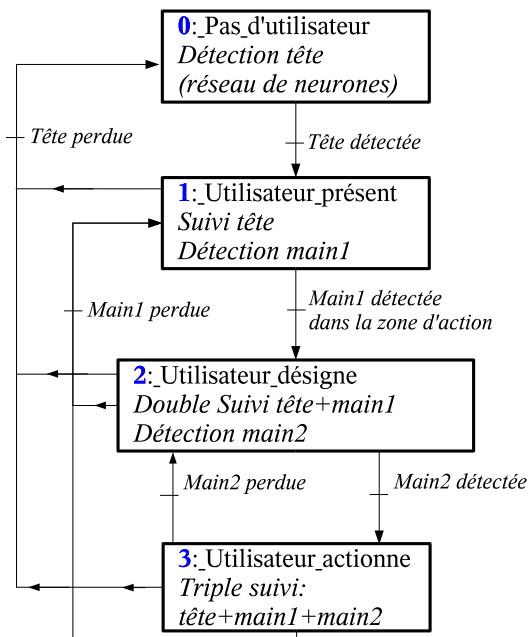


Figure 1 – Les différents états du système de reconnaissance de geste de pointage.

2.1 Détection et suivi de tête

On commence par détecter le visage car sa forme est relativement stable. La détection est effectuée par un réseau de neurones [4]. Le réseau accepte en entrée des imagerie de 15x20 pixels et répond en sortie si l'entrée est un visage ou non. Son architecture privilégie un très faible taux de fausse alarme afin de déclencher uniquement sur des visages. Le suivi de tête est initialisé lors de la détection et utilise les indices teinte-chair et mouvement [5] ainsi que la disparité. Les pixels de teinte-chair (Figure 2-c) sont déterminés grâce à une table de couleur peau représentative et la disparité (Figure 2-d) est obtenue par une caméra stéréo.

2.2 Détection et suivi de main

A la différence du visage, une main a une forme très variable et est donc difficile à détecter dans toutes ses configurations notamment à faible résolution. Ainsi une fois le visage repéré, la disparité permet de déterminer sa position en 3D et celle-ci sert alors de point de repère pour rechercher la main. On considère comme zones candidates de mains uniquement les zones contenant assez de pixels de teinte-chair en mouvement et situés à une certaine profondeur devant la tête. Les contraintes biométriques restreignent la position de la main à un sphéroïde centré sur la tête. Il est raisonnable d'admettre qu'une personne qui interagit avec un grand écran déplacera sa main devant lui et à une distance suffisamment éloignée de la tête (plus de 30 cm). Ce qui restreint l'espace de recherche à l'espace délimité par une sphère et un plan, volume que nous appelons 'zone d'action' (Figure 2-a).

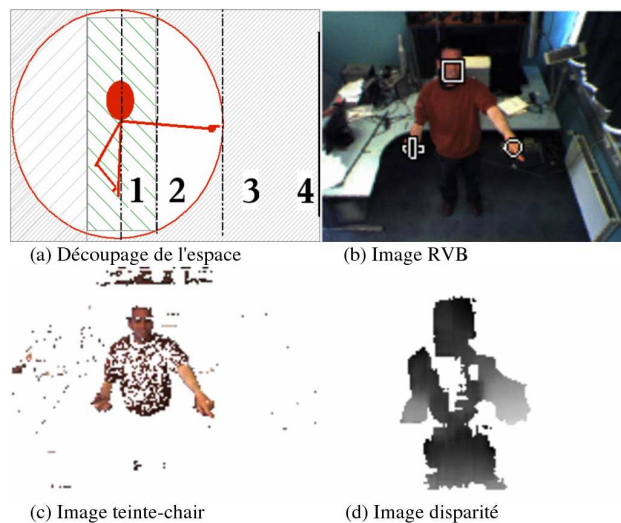


Figure 2 – (a) : 1. Zone de repos, 2. zone d'action, 3. zone de non-détection des mains, 4. écran - (b) : Position du visage (carré), position de la main de pointage (cercle), position de la main de commande (croix) - (d) disparité des pixels en zone de repos ou en zone d'action.

La main est détectée comme la zone de teinte-chair en mouvement la plus proche de l'écran. Cette première main détectée est considérée par convention comme la main de pointage.

Puis la seconde main sera détectée de la même manière, en éliminant de l'espace de recherche la zone englobant la première main ainsi que le bras correspondant. En effet, un bras nu ou recouvert d'un vêtement de teinte-chair peut être la seconde zone de teinte-chair en mouvement la plus proche de l'écran. Pour déterminer le bras, on initialise une zone avec les pixels de teinte-chair de la main, zone que l'on complète, à chaque itération, avec les pixels voisins continus en profondeur et situés devant la tête, tant que la zone croît. Cette seconde main est utilisée pour actionner une commande lorsqu'elle entre dans la zone d'action ou pour contrôler un troisième axe avec sa profondeur.

On pourrait envisager de détecter en permanence la main de pointage puis la main de commande mais lorsque la main de commande est plus avancée que la main de pointage, leurs rôles respectifs seraient inversés. Il est donc nécessaire de suivre au cours du temps les mains présentes, en tant que main de pointage et main de commande et ce, indépendamment de leur distance à l'écran.

Le suivi initialisé lors de la détection, est réalisé de manière analogue au suivi de tête. Lorsque l'utilisateur est bras nu ou porte des vêtements couleur chair, le suivi peut se positionner sur une partie quelconque du bras. Pour retrouver la main, on dirige la solution, en descendant les gradients de profondeur vers l'extrémité du bras la plus proche de l'écran. Ce recadrage est inopérant pour des faibles gradients mais l'avant-bras, alors parallèle à l'écran, est en zone de repos et le pointage non pris en compte.

2.3 Gestion des occultations et des pertes

Les suivis étant basés sur des informations de même nature, en cas d'occultation d'une tête ou d'une main par une main, l'une de ces deux parties du corps est correctement suivie alors que l'autre partie se cale à tort sur la première et ne peut plus être contrôlée par l'utilisateur. Les zones suivies restent fusionnées, même après la fin de l'occultation.

Le cas le plus fréquent d'occultation en situation de pointage est le passage d'une main devant la tête : la main se déplace alors devant la tête et non la tête derrière la main. Pour régler ce problème, on considère que la tête est immobile et on interrompt le suivi de visage en mémorisant sa dernière position connue. Pour le suivi de cette main, on restreint davantage l'espace de recherche en disparité afin de ne pas prendre en compte les pixels de la tête.

Les autres cas d'occultations sont plus compliqués et on se propose de détecter les cas de fusion de zones suivies. Lors d'une fusion tête-main, on constate qu'en général la main estimée se positionne sur le visage et y reste ; alors que lorsque le visage estimé suit la main, le suivi échoue rapidement, cet échec est détecté et force la réinitialisation de la détection du visage. On choisit donc de conserver le visage et de redétecter la main. Lors d'une fusion de deux mains, elles sont toutes deux ré-initialisées.

Les pertes de suivi de tête et de main sont automatiquement détectées en dénombrant les pixels de teinte-chair et de disparité situés en zone de repos ou d'action. Lorsque leur nombre est inférieur à un seuil, l'objet suivi est considéré comme perdu et on déclenche sa redétection (Figure 1).

3 Dispositif expérimental et performances

Nous caractérisons les performances moyennes de notre système de pointage à travers des expériences conduites auprès d'un groupe de 14 personnes, portant sur la stabilité temporelle et la précision spatiale. L'utilisateur est placé à environ 1,5 m d'une image rétro-projetée de 2 x 1,7 m. On emploie une caméra stéréo Bumblebee [6] avec une résolution de 160x120 et placée au dessus de l'écran avec un angle de 45 degrés afin de maximiser les déplacements de la main sur l'image et donc la précision spatiale lorsque l'utilisateur pointe les 4 coins de l'écran. Dans cette configuration, une variation de 1 cm de la position du visage (pour une main fixe) correspond à une variation de 1 cm de l'estimation du point visé sur l'écran. Le système fonctionne à 15 Hz sur un Pentium IV (3 GHz).

Pour caractériser la stabilité temporelle, on demande à chaque utilisateur de pointer une croix au centre de l'écran pendant environ 10 secondes. La distance moyenne au centre est de 2,29 cm et un résultat type est donné Figure 3. Pour évaluer la précision spatiale, chaque utilisateur parcourt des rectangles centrés et de taille décroissante. On évalue la distance moyenne au rectangle. Un parcours type est présenté Figure 3.

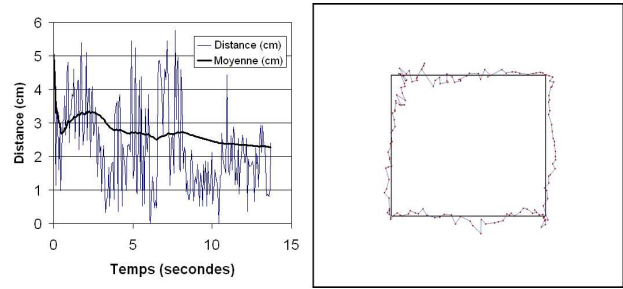


Figure 3 – A gauche : Stabilité temporelle - A droite : Précision spatiale, trajectoire type pour un rectangle de 102x85 cm à l'écran (distance moyenne = 2,75 cm).

LxH (cm)	61x51	102x85	143x120	185x154
D (cm)	2,38	2,70	3,24	3,45

Tableau 1 – Distance moyenne D au rectangle (de largeur L et de hauteur H).

On remarque que la précision lors d'un mouvement est moins bonne que lors d'un pointage fixe et qu'elle est équivalente sur les déplacements horizontaux et verticaux (ce qui justifie le choix d'une caméra placée à 45 degrés). D'autre part la précision décroît continuellement en se rapprochant des bords de l'écran (Tableau 1).

L'expérience suivante a pour but de déterminer la taille minimale des objets que l'on peut désigner avec notre système. On demande à l'utilisateur de pointer successivement des cibles, espacées entre elle de 1 m, dont la taille varie et dont les positions successives sont connues. Une fois une cible désignée, la nouvelle cible apparaît. Une trajectoire type (Figure 4) est constituée de trois parties.

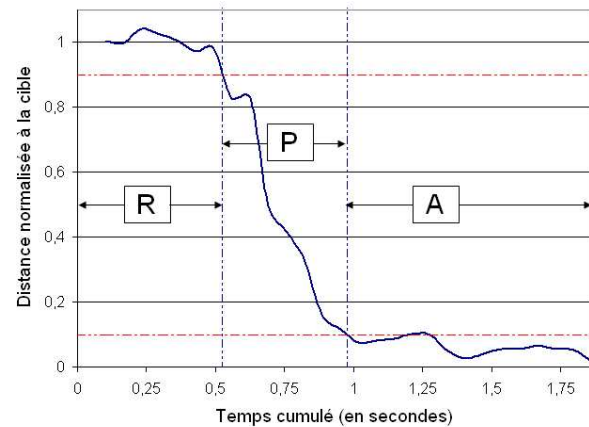


Figure 4 – Trajectoire inter-cible type :
 R : Temps de réaction (distance > 0,9),
 P : Temps de parcours (0,1 < distance < 0,9),
 A : Temps d'ajustement (distance < 0,1).

Une première partie (R) correspond au temps de réaction nécessaire à l'utilisateur pour s'assurer que la cible précédente est atteinte, repérer la cible suivante et débiter son geste. Une seconde partie (P) correspond au temps pendant lequel l'utilisateur parcourt rapidement l'espace inter-cible. Au cours de la dernière partie (A), l'utilisateur ajuste plus finement la position pointée de façon à atteindre la cible. On constate (Figure 5) que si les temps de réaction et de parcours sont indépendants de la taille de la cible, en revanche le temps d'ajustement augmente lorsque la taille de la cible diminue. On considère que lorsque le temps d'ajustement est supérieur au temps de réaction, la précision devient insuffisante et pénalise l'interaction homme-machine.

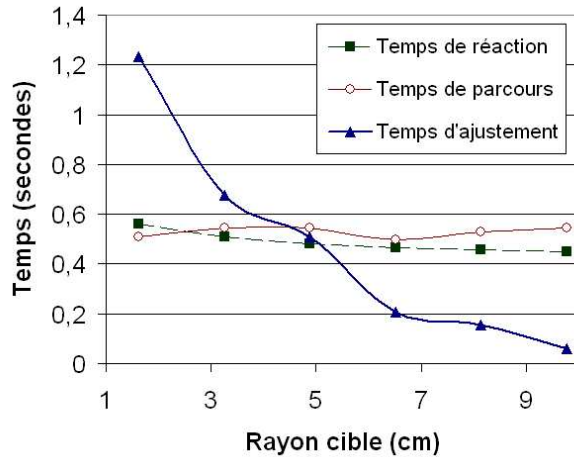


Figure 5 – Evolution des temps de réaction, de parcours et d'ajustement en fonction du rayon de la cible.

4 Conclusion

On déduit de ces différentes expériences, que dans les conditions actuelles de notre système, le rayon minimum d'un objet désignable sans gêne pour l'utilisateur est de 5 cm. Les objets placés en périphérie de l'écran sont plus difficiles à atteindre que ceux au centre.

La stabilité temporelle obtenue permettrait d'envisager la sélection d'un objet en le désignant pendant quelques secondes ainsi que dans [3], mais cette méthode présente le risque de sélectionner accidentellement un objet pointé. Une autre manière de sélectionner serait d'identifier un changement de posture de la main de pointage, mais d'une part la résolution de l'image ne permet pas d'envisager une telle reconnaissance et d'autre part un changement de posture de la main peut en changer son centre et donc le lieu pointé. Le suivi de la deuxième main offre au final une alternative en faisant appel à deux modalités distinctes spatialement pour le pointage et la sélection.

Références

- [1] N. Jovic, B. Brumitt, B. Meyers, et S. Harris. Detecting and estimating pointing gestures in dense disparity maps. Dans *IEEE International Conference on Face and Gesture recognition*, Grenoble, France, 2000.
- [2] K. Nickel et R. Stiefelhagen. Pointing gesture recognition based on 3d-tracking of face, hands and head orientation. Dans *International Conference on Multimodal Interfaces*, pages 140–146, Vancouver, Canada, 2003.
- [3] D. Demirdjian et T. Darrell. 3-d articulated pose tracking for untethered diegetic reference. Dans *Proceedings of International Conference on Multimodal Interfaces*, Pittsburgh, Pennsylvania, October 2002.
- [4] R. Feraud, O. Bernier, J.E. Viallet, et M. Collobert. A fast and accurate face detector based on neural networks. *Pattern Analysis and Machine Intelligence*, 23(1), January 2001.
- [5] O. Bernier, M. Collobert, et D. Collobert. Détection et suivi robuste de visages en temps réel. Dans *CORESA*, Lyon, France, 2003.
- [6] <http://www.ptgrey.com/products/bumblebee/index.html>.