Interaction multimodale multiutilisateurs avec un jeu d'échec sur grand écran

S. Carbini O. Bernier J. E. Viallet

France Télécom Recherche & Développement
Technopole Anticipa, 2 avenue Pierre Marzin,
22307 Lannion Cedex, France.

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

Concours Jeune Chercheur : Oui

Résumé

SHIVA (Several-Humans Interface with Vision and Audio) est une interface multiutilisateurs, non intrusive, d'interaction libre par le geste et la parole avec de grands écrans. La tête et les mains de chaque personne sont suivies en temps réel à partir d'une caméra stéréoscopique. A partir de la position 3D de ces parties du corps, le système détermine la direction pointée par chaque utilisateur et les gestes de sélection effectués avec l'autre main sont reconnus. Le geste de pointage est fusionné avec les n-best résultats issus de la reconnaissance de la parole tout en prenant en compte le contexte de l'application. Le système est testé avec un jeu d'échec où deux personnes jouent tour à tour sur un très grand écran mural. Les commandes oro-gestuelles des deux joueurs sont synchronisées et fusionnées en prenant en compte le contexte du jeu. Les commandes sont interprétées et les commandes légales ambiguës, illégales ou impossibles sont représentées de façon à fournir un feedback aux joueurs.

Mots clefs

Suivi multiutilisateurs, détection et suivi de visage et de mains, espace corporel, interface homme-machine non intrusive, multimodale, synchronisation et fusion de modalités, geste de pointage, reconnaissance de parole, jeu d'échec, interprétation de commande ambiguë, feedbacks.

1 Introduction et Travaux antérieurs

Les très grands écrans muraux peuvent être visualisés par plusieurs personnes libres de se déplacer dans une pièce et devraient permettre à plusieurs utilisateurs de travailler ensemble. Mais les utilisateurs sont limités à des interactions via des interfaces de contact et ne peuvent faire appel aux moyens naturels et efficaces de communication, tels que la voix et le geste, utilisés lorsqu'ils collaborent entre eux. SHIVA (Several-Humans Interface with Vision and Audio) est une interface multimodale conçue pour permettre à plusieurs utilisateurs d'interagir librement par le geste et la parole avec de grands écrans et nous présentons ici sa déclinaison pour deux utilisateurs.

Les auteurs de [4] présentent un suivi de personnes multiples, basé sur un filtre à particules, avec une composante visuelle et audio (position, hauteur, et état locuteur de la personne). Dans [13], les auteurs suivent deux personnes dans un environnement extérieur afin d'identifier leur interaction. Le suivi est basé sur l'extraction de blobs et sur une soustraction de fond à partir d'images monoscopiques en niveaux de gris. Dans [2], un système multi-caméra et une fusion bayésienne sont utilisés pour suivre plusieurs personnes dans une pièce. Le principal inconvénient des méthodes précédentes est que chaque personne est considérée comme un seul objet et qu'aucune information sur la position des parties du corps n'est disponible.
Dans [12], les parties du corps de plusieurs personnes sont suivies, en traitant certains cas d'occultations grâce à une technique de suivi multiple de pistes et une fonction de contrainte de cohérence de trajectoire. Mais cette technique intéressante requiert une cadence d'acquisition élevée de façon à vérifier l'hypothèse de mouvements fluides. De plus, le suivi de blobs de couleur chair se fait sans identification (tête ou main ou personne auquel il appartient). Dans [9], après une étape de segmentation basée sur la teinte chair, la tête et les mains sont localisées en s'appuyant sur des heuristiques liées à la morphologie humaine et au contexte applicatif. Le suivi temporel temps réel de plusieurs personnes est réalisé par filtrage de Kalman partiel est robuste aux occultations entre personnes. Mais ainsi que le précise les auteurs, les modèles de la teinte chair sont assez sensibles à l'environnement et la précision obtenue peut diminuer lorsque la couleur des vêtements est proche de la teinte chair ou en présence de bras nus.
Comparé aux travaux antérieurs décrits ci-dessus, SHIVA détecte et suit la tête et les mains de deux utilisateurs, comme dans [12], mais en assignant chacune des parties du corps suivies à l'une ou l'autre personne. Cette interface s'appuie sur les techniques de détection et suivi des parties du corps d'une personne décrits dans [1]. Ainsi que dans [5], une caméra stéréo est utilisée à une résolution 320x240. Le suiveur est robuste à des variations raisonnables de luminosité, aux vêtements de teinte chair et aux fonds complexes. L'ensemble des processus de détection,

de suivi des parties du corps et de détection des pertes est entièrement automatique et fonctionne en temps réel.
Aucune calibration ou adaptation préalable à un utilisateur est nécessaire. Le système conserve le même comportement alors que les utilisateurs se déplacent librement dans la pièce comme dans [15] (tant qu'il n'y a pas d'occultations entre eux). L'axe tête-main est utilisé comme convention de pointage comme dans [7]. La fonction sélection est assurée avec la seconde main ou par la reconnaissance de la parole. Un vocabulaire dédié à l'application permet d'exprimer des commandes plus directes que celles obtenues par une souris gestuelle [5], par exemple pour le contrôle multimodal d'environnement virtuel [11]. Les n-best résultats de la reconnaissance de la parole sont fusionnés avec le geste et le contexte de l'application de façon à autoriser des commandes multimodales compactes et d'obtenir une interface flexible.

2 Détection et suivi des parties du corps

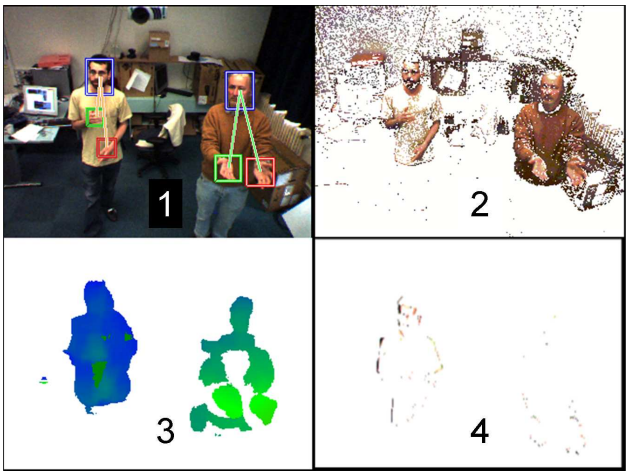

Figure 1 – *(1) Image rectifiée (les cadres représentent les parties du corps). (2) Image teinte chair (25 % des pixels). (3) Image de disparité filtrée. (4) Image du mouvement.*

La détection et le suivi des parties du corps s'appuie sur la détection de teinte chair (figure 1-2), la disparité (figure 1-3) et le mouvement (figure 1-4). La teinte chair est extraite à partir d'un filtre large construit à partir de différent utilisateurs dans différentes conditions d'éclairage. Une disparité fournie par la caméra est filtrée en éliminant les pixels situés à plus de $1.3m$ de la tête de l'une des personnes (après détection du visage). Le mouvement est estimé en soustrayant de l'image courante une image moyenne, adaptée à chaque image de façon à ce qu'une personne immobile s'intègre rapidement au fond. Les algorithmes de détection et de suivi de la tête et des mains sont issus de ceux décrits dans [1].

2.1 Espace corporel et suivi du corps

Après détection par un réseau de neurones, la position 3D de la tête sert de repère pour définir plusieurs zones impliquées dans la détection des mains et l'intentionnalité de l'utilisateur. Des contraintes morphologiques délimitent l'espace de recherche lors de la détection des mains à un sphéroïde centré sur la tête, et de rayon R ($\approx$ 1,3 m). L'espace extérieur à ce sphéroïde est écarté de l'espace de recherche (figure 2-zone 3). De plus, il est raisonnable d'admettre que, lors d'une interaction avec l'écran, l'utilisateur déplace sa main dominante vers l'écran. Ainsi, l'espace de recherche des mains est délimité par une sphère et un plan P, situé à une distance D ($\approx$ 30 cm) devant le plan parallèle à l'écran et passant par la tête. Nous appelons ce volume la zone d'action.

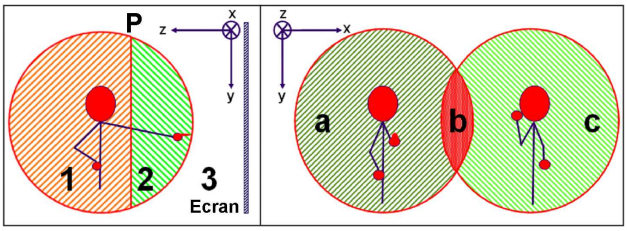

Figure 2 – *Gauche : vue de profil 1 : zone de repos, 2 : zone d'action, 3 : zone hors d'atteinte de la main. Droite : vue de face a : espace privé de l'utilisateur A , b : espace commun à A et B, c : espace privé de l'utilisateur B.*

Chaque utilisateur dispose d'un espace privé (figure 2-espace a et espace c). Si la distance entre les deux utilisateurs est trop faible, les espaces privés s'interpénètrent. L'espace commun (figure 2-espace b) est exclu de la zone de recherche des mains de manière à ne pas affecter une main à la mauvaise personne, lors de la détection.
Pour chacun des deux utilisateurs, la première main détectée est étiquetée main de pointage et la seconde, utilisée pour effectuer des gestes de sélection, est étiquetée main de contrôle. Une main est détectée en tant que zone de teinte chair en mouvement, la plus proche de l'écran, dans la zone d'action. Ainsi, SHIVA, fonctionne aussi bien pour des droitiers que pour des gauchers, sans avoir besoin d'étiquette main gauche ou main droite. Nous faisons en effet l'hypothèse qu'un utilisateur pointera d'abord un objet avec sa main dominante avant d'utiliser son autre main ou la parole pour interagir avec cet objet. La fonction de la main (pointage ou sélection) n'est activée que lorsqu'elle se trouve en zone d'action afin de prendre en compte l'intentionnalité de l'utilisateur et de ne pas déclencher involontairement des sélections incontrôlées. Ces zones et espaces étant référencés par rapport à la tête, ils accompagnent l'utilisateur lors de son déplacement tout en conservant le même comportement. Le suivi s'accommode de fonds complexes (figure 3) et est robuste à la présence de bras nus (figure 3-droite) ou de vêtements de teinte

chair (figure 1-2). Les utilisateurs peuvent être assis ou debout (figure 3-gauche) et se déplacer dans le champ de la caméra. Une fois détectée, une main est suivie même dans l'espace privé de l'autre utilisateur (figures 2, 3-centre et 3-droite), tant que cette main n'est pas automatiquement reconnue comme étant perdue, par exemple lorsqu'elle est masquée par un buste. Si l'une des deux personnes sort du champ de la caméra, une nouvelle personne entrant en scène héritera de l'étiquette de la personne sortie.

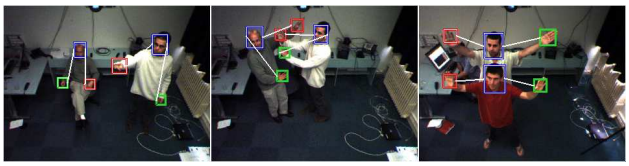

Figure 3 – *Quelques exemples de suivi : Les parties du corps suivies sont représentées par des rectangles (bleus pour la tête, rouge pour la première main et vert pour la seconde main). Les lignes blanches indiquent l'association main-tête. De gauche à droite : (1) Un gaucher assis dans le fond et une personne pointant en avant plan.(2) Les parties du corps sont correctement suivies et affectées à la bonne personne même lorsque les espaces privés se recouvrent.(3) Le suivi est robuste aux avant-bras nus et confère aux utilisateurs une grande liberté de déplacement.*

3 Le système multimodal SHIVA

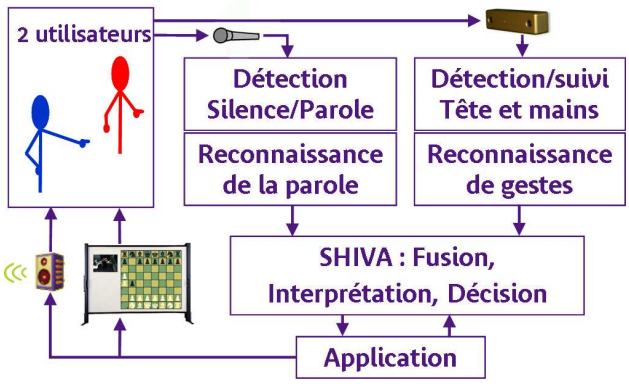

Figure 4 – *Les utilisateurs partagent les dispositifs physiques (écran, caméra, microphone) et les processus de reconnaissance oro-gestuelle de SHIVA.*

Le but du système SHIVA est de permettre aux deux utilisateurs d'effectuer des gestes de pointage et de prononcer des commandes vocales simultanément. Actuellement, SHIVA est testé sur un jeu d'échec, où les utilisateurs interagissent tour à tour au geste et à la voix (figure 5). Le tour de parole permet de n'utiliser qu'un dispositif de prise de son et qu'un processus de reconnaissance de la parole. Pour deux personnes parlant en même temps, on pourrait envisager une solution à base d'antennes acoustiques (en orientant dynamiquement chaque lobe dans la direction connue d'une personne) ou de microphones HF et d'autant de processus de reconnaissance de la parole. Par convention, la première personne qui entre dans le champ de la caméra et est détectée joue les blancs et la seconde les noirs. Un jeu d'échec [16] a été modifié de façon à accepter des commandes générées par le système multimodal SHIVA et à fournir des informations de contexte et des feedbacks adaptés à la nature multimodale des commandes. Une interface homme-machine est une boucle où les utilisateurs sont des entrées (geste et parole dirigés vers les systèmes de reconnaissance) du système et également des sorties, en tant que destinataires des feedbacks audiovisuels (figure 4). Les résultats des modules de reconnaissance de geste et de la parole sont affichés et des animations permettent aux deux joueurs de visualiser sur le grand écran l'interprétation faite par le système de leur commande multimodale, que cette commande soit légale, légale mais ambiguë, illégale voir impossible par exemple en invoquant des pièces qui ne figurent plus sur l'échiquier. Seul le lieu pointé par le joueur dont c'est le tour est représenté sous la forme d'un curseur à l'écran.

Figure 5 – *Gauche : Le joueur à gauche déplace une pièce pendant que l'autre réfléchit. Droite : Le joueur de droite joue et l'autre attend son tour. La caméra et le microphone sont au dessus de l'échiquier.*

3.1 Fusion et déplacement de pièce

Pour interpréter une commande multimodale, l'une des premières étapes consiste à synchroniser les modalités de parole et de geste, en déterminant pour chacune des modalités une référence sémantique commune et à mettre en relation les instants correspondants. Pour une interface gestuelle similaire à Shiva [8], lors d'une interaction avec une carte, 93,7% des gestes sont temporellement alignés avec l'énoncé vocal associé. Chen [3] fait l'hypothèse d'un haut degré de simultanéité entre parole et geste et fusionnent les deux modalités qu'à l'issue des processus de reconnaissance. Stiefelhagen [14] introduit un intervalle de temps de 1 s, avant et après le signal de parole, de façon à prendre en compte l'essentiel de la dynamique gestuelle. Dans notre cas de figure, le processus de pointage est

permanent tant que la main demeure en zone d'action et tous les lieux pointés sont connus en fonction du temps. Le geste de sélection est un événement discret se produisant à l'instant où la main de sélection franchit le plan P. La commande orale est un événement dont le début et la fin sont connus. La fusion geste-parole est donc réalisée dès que le résultat de la reconnaissance de parole est disponible, c'est à dire 240 ms après la fin du signal de parole. Nous faisons également l'hypothèse que geste et parole sont synchrones, c'est à dire que nous considérons les gestes de pointage réalisés pendant l'intervalle de temps de la parole (figure 6-A). Nous étendons cet intervalle d'une durée de 240 ms avant et après la parole, cette durée correspondant au temps nécessaire pour séparer un signal de parole du silence qui l'entoure. Sur un échiquier, si chaque pièce obéit à des règles de déplacement différentes, deux cases suffisent à définir un coup : la position de départ de la pièce concernée et la pièce d'arrivée. Pour déplacer une pièce, un joueur peut pointer et sélectionner successivement les deux cases concernées, fournir l'information uniquement sous la forme d'une commande orale unique du type "pion C2C3" [6] ou encore pointer une case et indiquer oralement la pièce concernée par le déplacement "met la reine". Le premier mode implique deux pointages successifs comme avec une souris, le second fait appel à des commandes surtout connue par des joueurs expérimentés et le dernier associe un seul geste naturel de pointage et une désignation ordinaire de la pièce par son nom. La reconnaissance vocale est sujette à erreurs, notamment sur des énoncés courts, tel "C2C3", erreurs qui peuvent être levées en associant le pointage sur l'une des cases.

La figure 6 présente des gestes de pointage réalisés par un utilisateur effectuant le déplacement d'une pièce, entre deux cases distinctes, sur un échiquier selon trois modalités distinctes. L'amplitude du déplacement est normalisée par la distance entre les deux cases.

La première modalité utilisée pour déplacer une pièce est analogue à la fonction glisser-déposer de la souris ; le pointage est fait avec le geste et la sélection est effectuée à la voix, par l'intermédiaire de commandes vocales "prends"-"lâche". On constate que la parole intervient sur un palier, lorsque l'utilisateur est certain que le curseur est localisé sur la bonne case (figure 6-A). On constate également que le geste de pointage demeure sur le palier, le temps que l'utilisateur perçoive que la commande multimodale a été prise en compte ; à ce moment, les changements de lieu pointés se traduisent par un déplacement de la pièce de l'échiquier visible par l'utilisateur. Le temps nécessaire au déplacement de la pièce est mesuré entre le moment où commence le pointage sur la première case (Début) et le moment où se termine le pointage sur la seconde case (Fin). Sur cet exemple, le temps de déplacement (et donc de pointage) est de 3,52 s.

La seconde modalité (figure 6-B) consiste à synchroniser le geste de pointage avec l'instant où se produit un geste de sélection, réalisé lorsque la seconde main franchit le plan P d'arrière en avant. On constate que ce geste de sélection (indiqué par une verticale) se produit également sur un palier de pointage. La fusion est effectuée en considérant que la case sélectionnée est celle qui correspond au lieu pointé au moment du geste de sélection. La pièce, correspondant à la case sélectionnée, est déplacée puis déposée par l'utilisateur sur une autre case, lorsque la seconde main franchit le plan P d'avant en arrière. Sur cet exemple, le temps de déplacement est de 3,26 s.

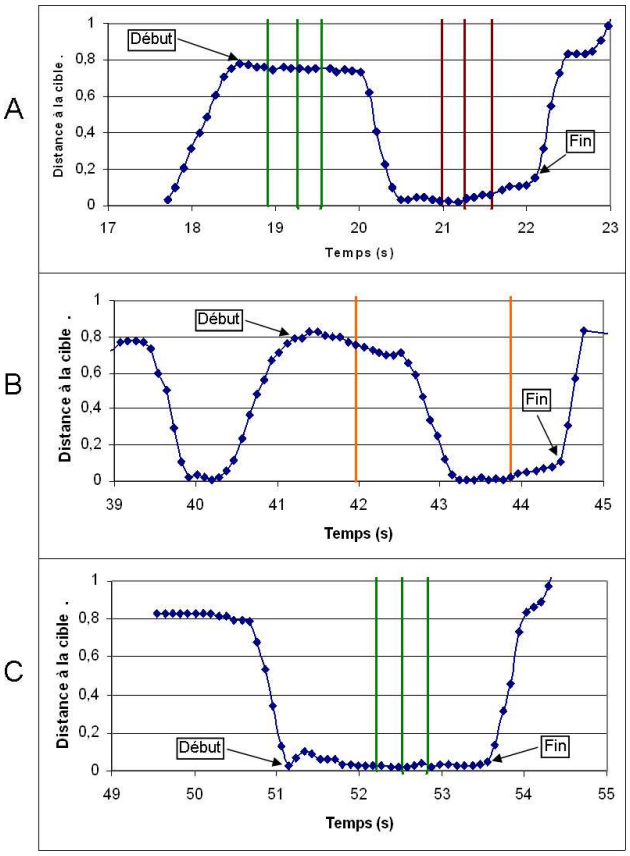

Figure 6 – *Trois façons de déplacer une pièce de l'échiquier. Les temps de déplacement sont calculés entre les instants Début et Fin. (A) La courbe représente les gestes de pointage vers la case de destination et les commandes orales "prends"/"lâche" sont représentées par deux groupes de trois verticales représentant les instants de début et de fin de parole et l'instant où le résultat de la reconnaissance de parole est disponible. (B) : la courbe représente les gestes de pointage et les verticales les instants où se produisent les gestes de sélection et de désélection. (C) Commande multimodale associant gestes de pointage (courbe) et commande orale unique (groupe de trois verticales). Par commodité, les gestes de pointage sont représentés, a posteriori, par la distance entre le lieu courant de pointage et le centre de l'une des deux cases concernées par le déplacement et cette distance est normalisée par la distance entre les deux cases.*

La troisième modalité permettant de déplacer une pièce (figure 6-C) consiste, comme pour la première modalité, à synchroniser le geste de pointage avec l'intervalle de temps correspondant à la commande orale, dont le contenu est suffisant pour effectuer le déplacement sans avoir à préciser par le geste la position de la seconde case. Sur cet exemple, le temps de déplacement est de 2,41 s. Cette modalité, qui n'implique qu'une seule case pointée mais fait appel au contexte de l'application, est réalisée dans un temps plus court que le temps des deux autres modalités qui ne font pas appel au contexte de l'application pour réaliser la fusion oro-gestuelle.

3.2 Fusion et contexte applicatif.

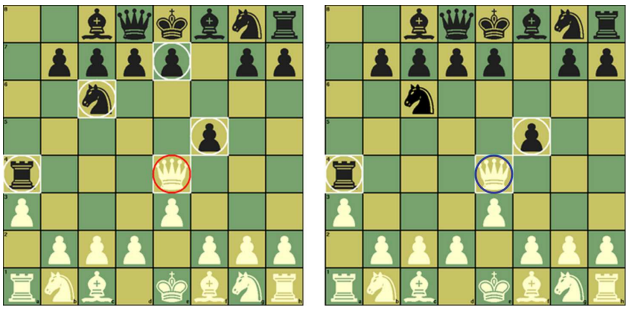

Figure 7 – *Gauche : Quand le joueur qui joue les blancs pointe sa reine blanche (cercle rouge), le contexte correspondant inclut les pièces (cercles blancs) pouvant être prises par la reine : une tour, un cavalier, deux pions. La commande "prend le pion" est ambiguë. Droite : Dans la même configuration, si le joueur qui joue les noirs pointe la reine blanche adverse (cercle bleu), le contexte correspondant inclus ses pièces noires (cercles blancs) qui peuvent prendre la reine blanche : une tour et un pion. La commande orale "prends avec le pion" n'est pas ambiguë.*

En pointant sur une case, un joueur indique l'une des deux cases relatives à un coup. L'information relative à la seconde case peut être spécifiée par la parole. En observant la configuration du jeu, qu'il suppose partagée par le système, le joueur fournit juste l'information qu'il estime suffisante pour compléter l'information manquante. SHIVA doit déterminer quelle est la seconde case concernée, à partir des informations extraites de la reconnaissance de la parole et du contexte du jeu. Le contexte de l'application est principalement lié à la case pointée par le joueur actif selon qu'il joue les blancs ou les noirs (figure 7).
- Lorsque le joueur pointe l'une de ses pièces, le contexte est essentiellement décrit par la liste des pièces adverses prenables (ou les cases qu'elles occupent) par la pièce pointée .
- Quand un joueur pointe une pièce adverse ou une case vide, le contexte est principalement décrit par la liste de ses pièces susceptibles de prendre la pièce adverse pointée ou de se déplacer vers la case vide pointée.
Lorsqu'un joueur prononce le nom d'une pièce du contexte applicatif et si une seule pièce porte ce nom, alors la seconde case est connue et SHIVA effectue le déplacement. Si plusieurs pièces du contexte portent le même nom que le mot prononcé (par exemple les pions de la figure 7-gauche), il y a ambiguïté. SHIVA l'illustre par un déplacement fictif de la reine vers les deux pions (figure 8) avant de revenir à la situation d'origine et d'attendre une nouvelle commande. Quand le joueur ne pointe aucune des cases de l'échiquier, toute l'information nécessaire doit être contenue dans le contexte et dans la parole par exemple sous la forme du nom de deux pièces non ambiguë et d'un verbe tel que "prends". Par exemple l'énoncé oral "reine prends tour" ne fait pas appel au pointage car une seule tour peut être prise par la reine (figure 7-gauche). L'ambiguïté du seul énoncé "reine prends pion" peut être levée en pointant l'un des deux pions.
Le contexte est complété par l'ensemble des coups qui ne respectent pas les règles de déplacement des pièces ou des prises. En effet, rien n'empêche un joueur d'effectuer une commande illégale (par exemple en ne réalisant pas que son roi est en prise). Si le système ne réagit pas à une commande illégale, le joueur peut penser que sa commande est légale mais qu'elle n'a pas été comprise par le système et le joueur réitérera sa commande. En revanche, si le système déplace la pièce du joueur conformément à la commande illégale et replace la pièce à sa position d'origine, le joueur prendra conscience que le système a interprété la commande du joueur mais que le système refuse de la valider.

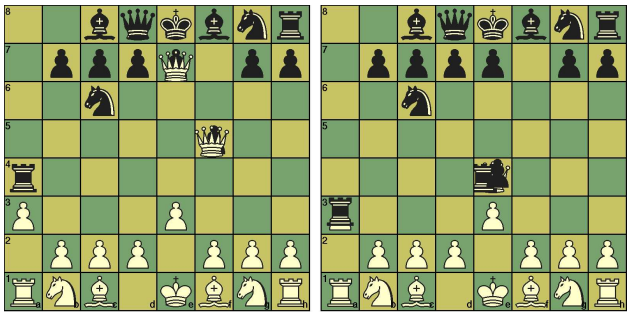

Figure 8 – *Gauche : En pointant la reine blanche, le joueur blanc visualise l'interprétation de la commande orale ambiguë "prends le pion". Droite : Le joueur noir visualise l'ambiguïté de la commande orale "prends" sans pointage. Puis dans les deux cas, la commande est rejetée et la situation d'origine est affichée.*

De la même manière, considérons la situation où un joueur effectue une commande légale, mais que le système l'interprète à tort comme une commande illégale (par exemple en raison d'une erreur de la reconnaissance vocale). En

l'absence d'illustration de la commande interprétée, le joueur ne pourrait déterminer l'origine de la non prise en compte de sa commande. Le système ne peut qu'illustrer la commande qu'il a interprété, à tort ou à raison. Le feedback permet au joueur de comprendre que le système s'est trompé et qu'il a intérêt à réaliser le même but selon une modalité différente.
Enfin le contexte est éventuellement complété par une liste de coups impossibles. En effet, la reconnaissance de la parole peut reconnaître à tort ou à raison une commande vocale impliquant une pièce qui n'existe plus sur l'échiquier. Le système doit savoir qu'une telle pièce n'existe pas, puis matérialiser la commande reconnue en faisant apparaître la pièce inexistante et enfin puis afficher la disposition précédant la commande de façon à ce que l'utilisateur dispose d'un retour lui permettant de comprendre l'interprétation faite par le système de la commande.

4 Conclusion

Nous présentons un système permettant une interaction oro-gestuelle de deux personnes jouant aux échecs sur un grand écran. Les gestes de pointage et de sélection sont obtenus en suivant les parties du corps de deux personnes en quasi temps réel. La précision du suivi confère un pointage précis sur un grand écran. Le système de suivi fonctionne à une cadence de 20 Hz lorsqu'un utilisateur est suivi et à 15 Hz pour deux utilisateurs (Biprocesseur Xeon 3,4 GHz). La faible différence du cout du suivi est encourageante pour le suivi de plus de deux personnes.
Associée à la reconnaissance de la parole, l'interface multimodale oro-gestuelle SHIVA, permet à deux personnes de jouer, tour à tour. La prise en compte du contexte permet de réaliser des commandes plus rapides et plus intuitives mais nécessitent d'adapter le vocabulaire à l'application alors qu'une interface du type souris oro-gestuelle bénéficie d'une interface graphique répandue, ne demande pas de modifications de l'application mais ne permet pas de bénéficier de commandes oro-gestuelles spécifiques.

Références

[1] S. Carbini, L. Delphin-Poulat, L. Perron, O. Bernier, J.E. Viallet, Interaction Multimodale Oro-Gestuelle Personne Libre, Corésa 2005, p. 195-200, Rennes, France, 2005.

[2] T.H. Chang, S. Gong, Tracking multiple people with a multi-camera system, IEEE Workshop on Multi-Object Tracking, p. 19-26, Vancouver, Canada, 2001.

[3] F. Chen, E. Choi, J. Epps, S. Lichman, N. Ruiz, Y. Shi, R. Taib, M. Wu, A study of manual gesture-based selection for the PEMMI multimodal transport management interface, *ICMI (International Conference on Multimodal Interfaces)*, p. 274-281, Trento, Italie, 2005.

[4] N. Checka, K. Wilson, M. Siracusa, T. Darrell, Multiple Person and Speaker Activity Tracking with a Particle Filter, ICASSP, Montréal, Canada, 2004.

[5] D. Demirdjian, T. Darrell, 3-D Articulated Pose Tracking for Untethered Diectic Reference, Proceedings of International Conference on Multimodal Interfaces, p. 267-272, Pittsburgh, Pennsylvanie, 2002.

[6] M. Gabsdil, Combining Acoustic Confidences and Pragmatic Plausibility for Classifying Spoken Chess Move Instructions, Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, p. 27-30, Cambridge, Massachussets, 2004.

[7] R. Kehl and L. Van Gool, Real-time Pointing Gesture Recognition for an Immersive Environment, IEEE International Conference on Automatic Face and Gesture Recognition, p. 577-582, Séoul, Corée, 2004.

[8] S. Kettebekov, R. Sharma, Understanding Gestures in a Multimodal Human Computer Interaction, International Journal of Artificial Intelligence Tools, vol. 9, n. 2, p. 205-223, 2000.

[9] V. Girondel, L. Bonnaud, A. Caplier, A Human Body Analysis System, EURASIP Journal on Applied Signal Processing, vol. 2006, 2006.

[10] A. Micilotta, R. Bowden, View-based Location and Tracking of Body Parts for Visual Interaction, British Machine Vision Conference, p. 849-858, Kingston, Royaume-Uni, 2004.

[11] K. Moustakas, D. Tzovaras, S. Carbini, O Bernier, J.E. Viallet, S. Raidt, M. Mancas, M. Dimiccoli, E. Yagci, S. Balci, E. Ibanez Leon and M.G. Strintzis, "MASTERPIECE : Experiencing Physical Interaction in VR Applications", IEEE Multimedia, p. 92-100, Volume 13, Issue 3, July-September 2006.

[12] E. Polat, M. Yeasin, R. Sharma, A Tracking Framework for Collaborative Human Computer Interaction, International Conference on Multimodal Interfaces, p. 27-32, Pittsburgh, Pennsylvanie, 2002.

[13] K. Sato, J.K. Aggarwal, Tracking and recognizing two-person interactions in outdoor image sequences, Workshop on Multi-Object Tracking, p. 87-94, Vancouver, Canada, 2001.

[14] R. Stiefelhagen, C. Fuegen, P. Gieselmann, H. Holzapfel, K. Nickel, A. Waibel, Natural Human-Robot Interaction using Speech, Gaze and Gestures, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, p 2422-2427, Sendai, Japon, 2004.

[15] Y. Yamamoto, I. Yoda, K. Sakaue, Arm-Pointing Gesture Interface Using Surrounded Stereo Cameras System, ICPR (International Conference on Pattern Recognition), p. 965-970, Cambridge, Royaume-Uni, 2004.

[16] Xboard : http ://www.tim-mann.org/xboard.html.