Perception et restitution des actions des utilisateurs en interaction distante dans un environnement 3D virtuel

Patrick Horain, José Marques Soares, Dianle Zhou,
David Antonio Gómez Jáuregui et André Bideau

Institut Télécom / Télécom SudParis, 9 rue Charles Fourier, 91011 Evry Cedex, France
Patrick.Horain@Telecom-SudParis.eu

Résumé : Dans une collaboration à distance médiatisée par ordinateur, la communication entre personnes est renforcée par le contact visuel. Toutefois, la vidéoconférence ne permet pas restituer les actions sur les objets partagés. La perception mutuelle des participants et de leurs actions peut être renforcée au moyen d'un environnement virtuel en réseau. Des représentations virtuelles des utilisateurs distants (avatars) et des objets qu'ils partagent y sont rassemblées de sorte que les actions des utilisateurs sur les objets sont utilisées pour animer les avatars qui miment leurs actions. Les gestes de communication sont eux acquis par vision artificielle et restitués virtuellement en temps réel. Un système d'acquisition par vision artificielle des gestes et expressions du visage en 3D peut être développé à partir d'une unique webcam. Ce système à bas débit et bas coût permet d'envisager des applications grand public de téléprésence virtuelle en 3D.

1 Introduction

Les environnements virtuels classiques pour le travail collaboratif distant n'offrent aux utilisateurs qu'une perception mutuelle limitée. Si la vidéoconférence multipoints permet aux utilisateurs de se voir les uns les autres au prix d'une charge relativement importante du réseau de télécommunication, elle ne permet pas une réelle immersion, i. e. la sensation d'être plongé dans un environnement virtuel ou réel, car chaque utilisateur apparaît dans une fenêtre séparée et il est malaisé de percevoir qui est en train d'agir sur les objets partagés (Leung & Chen, 2001).

Pour améliorer la perception de « qui fait quoi » dans une session de travail collaboratif à distance, l'interaction entre les utilisateurs peut être augmentée en les plongeant ainsi que les objets graphiques sur lesquels ils agissent, dans un environnement 3D virtuellement partagé grâce à un réseau de télécommunication.

Les utilisateurs sont représentés dans cet environnement par des avatars animés qui restituent leurs actions et gestes. Percevoir ces gestes aide les autres utilisateurs à focaliser leur attention sur les objets manipulés. Voir directement une représentation d'une action, plutôt qu'observer seulement son résultat, permet de mieux percevoir cette action. En outre, la restitution des gestes des participants dans l'environnement virtuel collaboratif permet une communication non verbale entre utilisateurs (Guye-Vuillème, Capin, Pandzic, Thalman, & Thalman, 1999).

Dans cet article, après avoir défini un certain nombre de termes relatifs aux environnements virtuels collaboratifs, nous cherchons à montrer ce qui est techniquement possible de faire dans ce domaine en décrivant d'abord une plateforme logicielle, que nous avons développée, pour partager une application 2D dans un monde virtuel habité par des avatars représentant les utilisateurs distants et animés pour restituer leurs actions sur l'application. Nous présentons ensuite une interface d'acquisition en 3D des gestes et expressions du visage par vision artificielle en temps réel à partir d'une unique caméra (webcam) : elle permet, quand un utilisateur n'agit pas sur l'application, que son avatar reproduise ses gestes ayant un rôle de communication.

2 Définitions des termes utilisés dans le domaine des environnements virtuels

Un environnement collaboratif virtuel est un outil donnant la possibilité à plusieurs personnes distantes d'évoluer dans un même environnement virtuel pour réaliser un objectif commun. Chaque participant est représenté par un avatar. Les participants peuvent agir entre eux et avec les objets de l'environnement virtuel.

Cet environnement est caractérisé par :

- des objets graphiques partagés tels que la maquette numérique d'un objet technique ou une application informatique…,
- une représentation des participants sous une forme appelée avatars,
- des mécanismes d'interaction.

L'environnement peut être immersif ou non immersif. Un environnement immersif est un environnement virtuel dans lequel le participant se perçoit lui-même comme étant plongé dans l'univers virtuel, où il peut se déplacer et contrôler les gestes de son avatar, et interagir avec les objets de cet environnement.

La communication entre les participants peut être verbale si un canal de communication orale est disponible. Elle peut être complétée par la communication gestuelle qui repose sur les mouvements des avatars représentant les participants, et par la communication par action qui se traduit par la production d'un effet sur un objet appartenant à l'espace virtuel partagé.

3 Collaboration distante dans un environnement virtuel 3D

Pour augmenter la perception de l'activité d'un utilisateur par les autres utilisateurs dans un environnement collaboratif virtuel, une interface juxtaposant deux vues peut être utilisée. L'une présente l'*espace applicatif*, c'est-à-dire l'application partagée sur laquelle les utilisateurs interagissent directement. L'autre affiche un *espace immersif*, c'est-à-dire un monde 3D habité qui rassemble virtuellement les participants et l'application qu'ils partagent.

La plateforme NetICE (Leung & Chen, 2001) proposait déjà de représenter chaque utilisateur par un avatar humanoïde se tenant face à un tableau virtuel où l'application partagée était projetée. Ces avatars peuvent exécuter quelques gestes communicatifs prédéfinis, mais ne permettent pas de restituer les actions des utilisateurs sur l'application. De plus, NetICE n'affiche qu'une vue à la fois, soit celle de l'application partagée, soit celle de l'environnement virtuel 3D avec les avatars des participants. Le tableau applicatif virtuel étant souvent peu lisible dans l'environnement 3D, ce qui gêne la collaboration, NetICE oblige à basculer fréquemment vers la vue de l'application partagée, ce qui entrave alors la perception des autres utilisateurs.

Nous proposons de restituer les actions des utilisateurs sur l'application partagée en les faisant reproduire par leurs avatars respectifs (Figure 1). L'avatar actif, placé devant le tableau, est animé par cinématique inverse de façon à ce que sa main suive sur le tableau la position des événements dans l'interface graphique de l'application. Pour permettre le partage d'une application quelconque, dont l'interface n'est en général pas conçue pour gérer des actions concurrentes d'utilisateurs multiples, un seul utilisateur peut agir à la fois. Les demandes des autres utilisateurs sont mémorisées dans une file d'attente et visualisées dans l'environnement virtuel par la main levée de leurs avatars respectifs.

Notons que les animations des avatars sont calculées localement sur le poste de travail de chaque participant, et donc que seuls les évènements dans l'interface de l'application partagée transitent par le réseau de télécommunication, dont la charge reste très limitée par comparaison avec celle d'un système de vidéoconférence multipoints (Marques Soares, Horain, & Bideau, 2003).

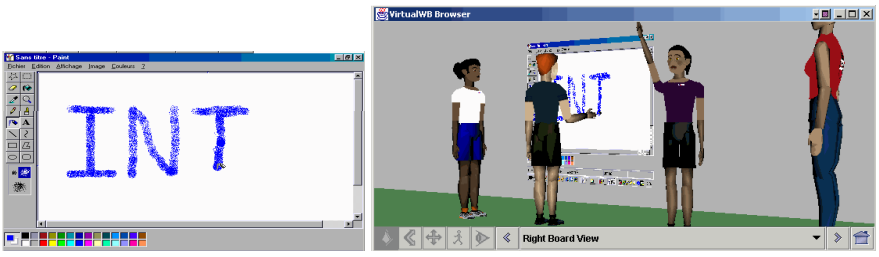

*Figure 1 : Environnement augmenté pour la collaboration distante : à gauche l'*espace applicatif*, à droite l'*espace immersif *rassemblant des représentations des utilisateurs et l'application qu'ils partagent. Des vidéos sont disponibles à l'adresse :* http://www-public.it-sudparis.eu/~horain/MarquesSoares/WB.

Plus qu'un simple canal de communication non verbale, ce système permet de restituer virtuellement à coût de revient raisonnable une scène distante réelle. Il suffit pour cela de vidéo-projeter un écran d'ordinateur affichant l'application partagée et d'acquérir les actions de l'utilisateur sur l'écran au moyen d'une pointe interactive (Dymo). L'espace immersif de notre environnement fournit une restitution proche de la scène réelle (Figure 2). Ce système, constitué d'équipements relativement peu chers et fonctionnant à bas débit réseau, apporte donc à la collaboration distante virtuelle autour d'un tableau réel une forme de téléprésence.

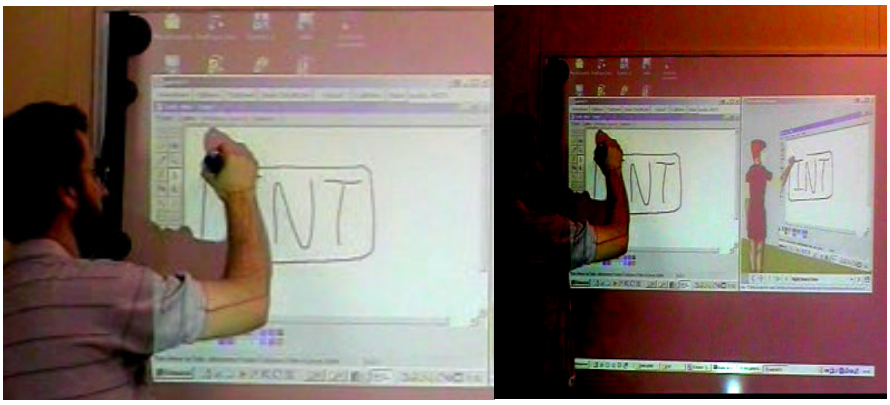

Figure 2 : Restitution virtuelle distante des actions d'un utilisateur sur un tableau augmenté. Vidéo : http://www-public.it-sudparis.eu/~horain/MarquesSoares/WB.

Nous avons complété cet environnement par un canal de voix sur IP et l'avons testé dans un contexte d'enseignement à distance. Un cours enseigné conjointement par 2 enseignants dans 2 salles séparées où les étudiants étaient répartis. Plutôt que d'utiliser 2 caméras en réseau à haut débit, nous avons partagé l'application de présentation des documents dans notre environnement 3D collaboratif. Une large majorité des étudiants se sont déclarés prêts à utiliser à nouveau cet outil (Marques Soares, 2004).

4 Acquisition des gestes en temps réel avec une webcam

Lorsque les utilisateurs ne sont pas actifs sur l'application, il reste nécessaire d'acquérir leurs gestes pour les restituer dans l'espace immersif, c'est-à-dire pour les faire reproduire par l'avatar associé à l'utilisateur. La vision artificielle permet d'acquérir les gestes humains avec un équipement léger, sans entraver la personne observée par des équipements encombrants (Poppe, 2007). La recherche sur ce sujet est stimulée par de nombreuses applications en animation infographique, jeux, interaction homme-machine, vidéosurveillance, interaction médiatisée entre personnes (Moeslund, Hilton, & Kruger, 2006)…

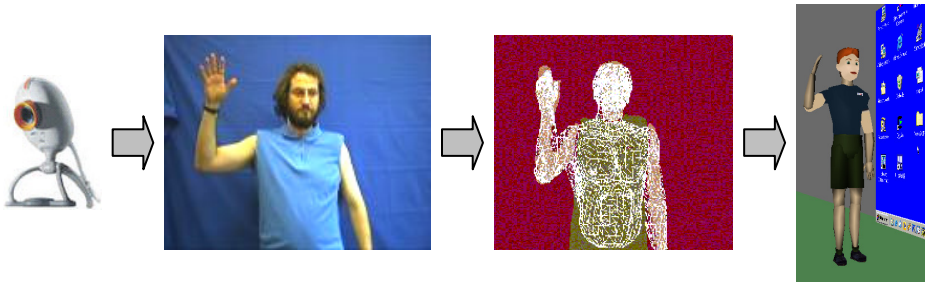

Figure 3 : Acquisition des gestes en temps réel par webcam et restitution virtuelle.

Nous avons développé un système d'acquisition du geste en 3D, en temps réel, sans marqueurs et fonctionnant avec une unique caméra en réseau (webcam). L'approche suivie consiste à recaler un modèle 3D articulé du corps humain sur la vidéo, c'est-à-dire à rechercher pour chaque image l'attitude du modèle qui réalise la meilleure correspondance avec cette image (Figure 3).

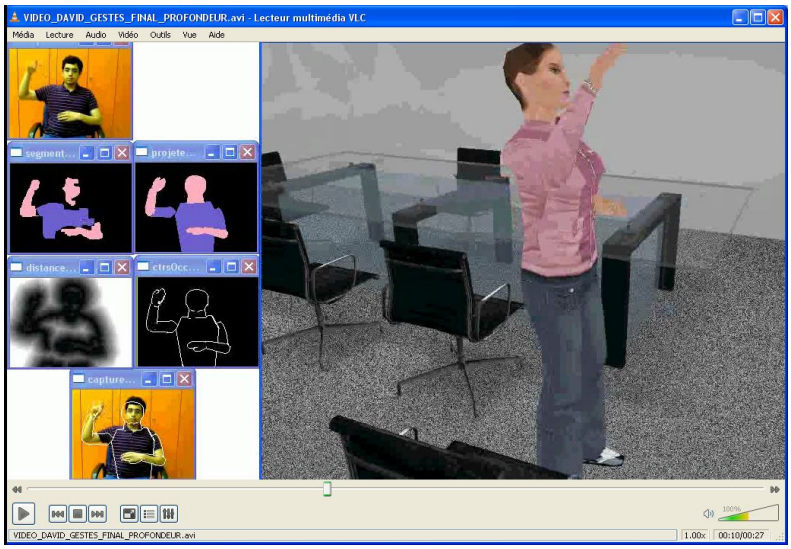

Figure 4: A gauche : en haut, image du flux vidéo ; en dessous, l'image précédente segmentée en classes de couleur, à coté d'une projection du modèle 3D recalé en maximisant le recouvrement entre régions de colorées ; en troisième ligne, la carte des distances aux contours extraits de l'image vidéo, et la projection des contours occultants du modèle dans la pose ajustée par rapprochement des contours ; en bas : les contours occultants du modèle projetés sur l'image initiale. A droite : la pose 3D restituée en temps réel par l'avatar de l'utilisateur, ici observé latéralement par la caméra virtuelle d'un autre participant. L'environnement virtuel utilise ici la plate-forme open source *OpenSpace3D (I-Maginer). Des vidéos sont disponibles en ligne à l'adresse : http://MyBlog3D.com (page* Demos*).*

Pour ce faire est utilisé un modèle 3D de la partie supérieure du corps humain avec 3 paramètres globaux de position et 20 angles aux articulations. Une attitude corporelle est ainsi représentée par un vecteur de 23 paramètres. Pour chaque image vidéo, nous recherchons l'attitude qui lui correspond le mieux. Des caractéristiques de l'image (régions de couleur, contours) sont extraites et appariées avec les caractéristiques du modèle placé dans diverses attitudes candidates. Pour chaque image, l'attitude réalisant le recalage optimal est recherchée itérativement. Dans une première étape, on maximise la superposition des régions de couleur. Le recalage est ensuite précisé en minimisant la distance entre les contours extraits de l'image et les projections des contours occultants du modèle 3D comme le montre la Figure 4 (Gómez, Jauregui & Horain, 2009).

Les paramètres définissant l'attitude optimale issue du recalage sont ensuite envoyés à notre serveur gestionnaire d'événements distribués qui les relaie vers les participants de la session collaborative, pour appliquer cette attitude à l'avatar de l'utilisateur (Figure 5).

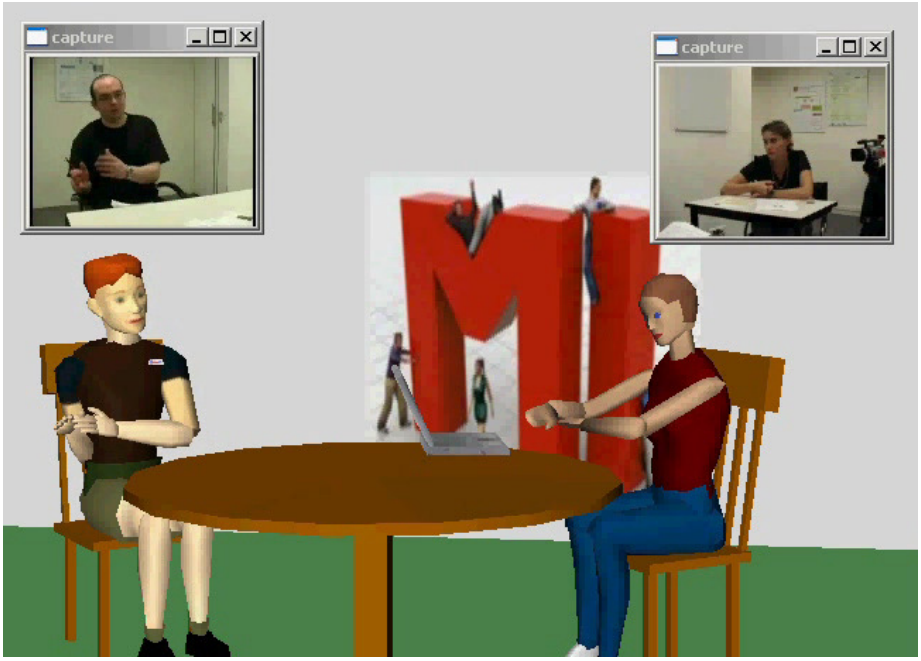

Figure 5 : Acquisition des gestes en temps réel avec une caméra unique pour chaque participant et restitution virtuelle immersive. Vidéo disponible en ligne à l'adresse : http://MyBlog3D.com (page Demos*).*

5 Suivi 3D en temps réel des mouvements du visage et restitution par avatar

La communication entre personnes passe aussi par les expressions du visage. Il faut donc que le visage reste un canal de communication dans l'environnement virtuel. Les expressions du visage de chaque utilisateur doivent être acquises par vision artificielle et restituées par leur avatar. Détecter et suivre le mouvement du visage est une tâche difficile car l'apparence du visage varie tout à la fois avec l'orientation 3D de la tête, les expressions, l'éclairement, les occultations éventuelles.

Nous avons suivi une approche par modèle actif d'apparence (Cootes, Edwards, & Taylor, 2001) qui procède en recalant globalement un modèle du visage de la personne observée sur les images. L'apparence du visage et ses variations en fonction tant de l'expression que de l'orientation de la tête sont apprises. Cette approche a été étendue pour modéliser les variations de la forme 3D du visage (Dornaika & Ahlberg, 2004). Pour chaque image vidéo, les paramètres du modèle de visage sont itérativement ajustés pour minimiser l'écart résiduel entre l'apparence observée et celle générée à partir du modèle. Nous obtenons ainsi les 6 paramètres de pose de la tête ainsi que, en temps réel, 6 paramètres d'expression du visage (Zhou & Horain, 2009) (Figure 6).

Figure 6 : Suivi des expressions du visage en 3D et en temps réel. Une vidéo est disponible en ligne à l'adresse : http://MyBlog3D.com (page Demos*).*

Pour acquérir le visage de nouvelles personnes sans nécessairement devoir procéder au fastidieux apprentissage de leur apparence et de ses variations, nous avons proposé une estimation statistique du modèle de variations d'apparence à partir d'une image de visage. Nous avons montré expérimentalement que cette approche améliorait le suivi du visage de personnes inconnues (Zhou & Horain 2009).

Les paramètres d'animation du visage obtenus sont ensuite encodés et transmis pour être restitués par l'avatar de l'utilisateur (Figure 7).

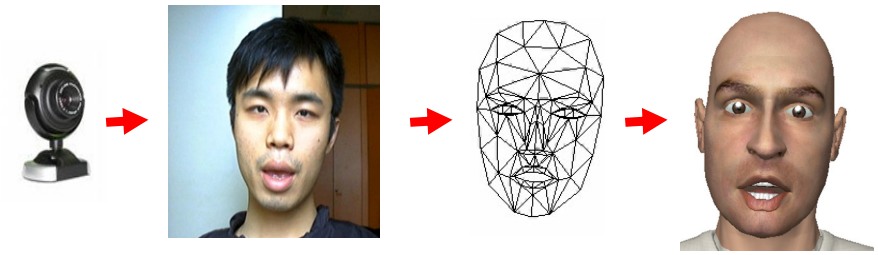

Figure 7 : Suivi 3D des expressions du visage et restitution virtuelle. Le modèle de visage CANDIDE-3 (Ahlberg) est recalé sur les images vidéo, et les expressions sont restituées ici avec l'agent conversationnel GRETA (Pelachaud, C. et al.). Une vidéo est disponible en ligne à l'adresse : http://MyBlog3D.com (page Demos*).*

6 Conclusion et perspectives

Nous avons montré l'état des techniques pour renforcer la collaboration médiatisée en rassemblant des représentations des participants et des objets qu'ils partagent dans un même espace virtuel, en restituant virtuellement les actions des utilisateurs sur ces objets et en instaurant cet espace comme canal de communication visuelle entre utilisateurs. Ces outils sont développés et fonctionnent en temps réel sur des ordinateurs personnels grand public munis d'une simple webcam. Exploitant ces équipements peu coûteux, ils permettent une communication gestuelle compatible avec un réseau de télécommunication à très bas débit dans un environnement virtuel offrant un fort sentiment de téléprésence.

Les interfaces perceptives décrites sont susceptibles d'applications grand public. Les environnements virtuels en réseau tels que *Second Life* (Linden Research, Inc.), qui ont généré bien des discussions voici quelques années, suscitent aujourd'hui beaucoup moins d'attention au profit de réseaux sociaux aux interfaces plus conventionnelles. Alors que l'avatar est l'ambassadeur de chaque utilisateur au sein du monde virtuel, son contrôle reste une tâche complexe et fastidieuse que des interfaces par vision peuvent rendre naturelle. La perception de l'utilisateur par la machine ouvre la voie à de nouveaux modes d'interaction.

Remerciements. José Marques Soares a bénéficié de la bourse CAPES/COFECUB n°266/99-I du gouvernement brésilien.

Références

Ahlberg, J. (n.d.). *CANDIDE - a parameterized face.* Retrieved from http://www.bk.isy.liu.se/candide

Brown University. (n.d.). Retrieved from http://vision.cs.brown.edu/humaneva

Carnegie Mellon University Graphics Lab. (n.d.). *Motion Capture Database*. Retrieved from http://mocap.cs.cmu.edu

Cootes, T., Edwards, G., & Taylor, C. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(6), 681–685. doi:10.1109/34.927467

Dornaika, F., & Ahlberg, J. (2004). Fast and Reliable Active Appearance Model Search for 3D Face Tracking. *IEEE Transactions on Systems, Man, and Cybernetics–Part, 34*, 1838-1853.

Dymo. (s.d.). *Mimio Interactive*. (Athena Global Service, Éditeur) Consulté le août 10, 2010, sur http://mimio.fr

Gómez Jáuregui, D. A., & Horain, P. (2009). Region-based vs. edge-based registration for 3D motion capture by real time monoscopic vision. In A. Gagalowicz, & W. Philips (Eds.), *Computer Vision/Computer Graphics Collaboration Techniques – Proceedings of MIRAGE 2009* (Vol. LNCS 5496, pp. 344–355). Rocquencourt, France: Springer Berlin / Heidelberg. doi:10.1007/978-3-642-01811-4_31

Guye-Vuillème, A., Capin, T. K., Pandzic, I., Thalman, N., & Thalman, D. (1999). Nonverbal Communication Interface for Collaborative Virtual Environments. *Virtual Reality J., 4*, 49-59. doi:10.1007/BF01434994

I-Maginer. (s.d.). Récupéré sur OpenSpace3D: http://www.openspace3d.com

Leung, W. H., & Chen, T. (2001). Creating a Multiuser 3 D Virtual Environment. *IEEE Signal Processing Magazine, 18*(3), 9-16. doi:10.1109/79.924884

Linden Research, Inc. (n.d.). *What is Second Life?* Retrieved from http://secondlife.com/whatis

Marques Soares, J. (2004). *Contribution à la communication gestuelle dans les environnements virtuels collaboratifs.* INT.

Marques Soares, J., Horain, P., & Bideau, A. (2003). Sharing and immersing applications in a 3D virtual inhabited world. *VRIC 2003*, (pp. 27-31). Laval, France.

Moeslund, T., Hilton, A., & Kruger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding, 104*(2-3), 90-126. doi:10.1016/j.cviu.2006.08.002

Poppe, R. (2007). Vision-based human motion analysis: An Overview. *Computer Vision and Image Understanding, 108*(1-2), 4–18. doi:10.1016/j.cviu.2006.10.016

Walsh, A. E., & Bourges-Sevenier, M. (2002). *The Mpeg-4 Jump-Start.* Prentice Hall Professional Technical Reference.

Zhou, D., & Horain, P. (2009). Robust 3D Face Tracking on Multiple Users with Dynamical Active Models. In B. Huet, A. Smeaton, K. Mayer-Patel, & Y. Avrithis (Eds.), *Advances in Multimedia Modeling – Proceedings of the 15th International Multimedia Modeling Conference (MMM2009)* (Vol. LNCS 5371, pp. 74-84). Sophia Antipolis, France: Springer Berlin / Heidelberg. doi:10.1007/978-3-540-92892-8_9