Mots-clés :
Télé-virtualité, Communications vidéo, Clonage de visages, Spatialisation vidéo.

Réunion virtuelle : aspects du traitement vidéo

par Jean-Luc DUGELAY, *Institut EURECOM*

Le projet TRAIVI de télé-virtualité a pour objectif de créer et gérer des espaces virtuels de réunion. En traitement vidéo, les techniques de clonage de visages et de spatialisation vidéo sont au centre des études.

L'ESSENTIEL

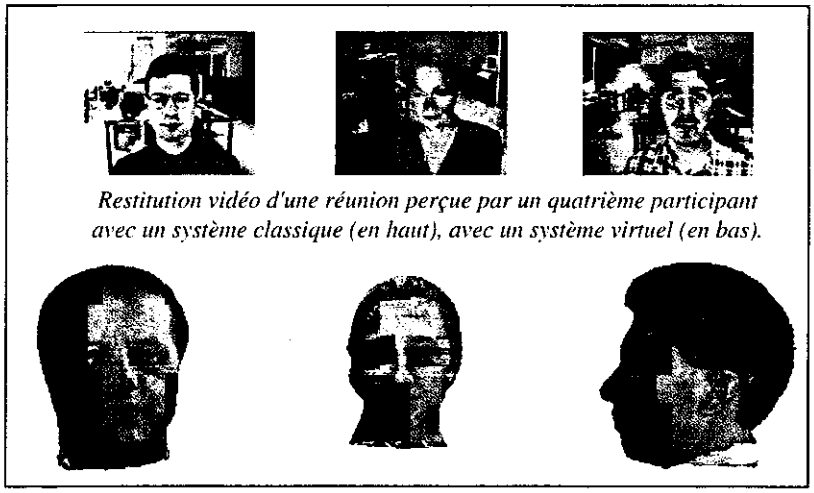
- Le clonage des visages est réalisé par analyse d'images, et utilise un modèle CYBERWARE pour la restitution du locuteur.
- La spatialisation vidéo permet de reconstruire des points de vue fictifs d'un espace de réunion réel à partir d'un ensemble de vues de référence non calibrées.

SYNOPSIS

- Face cloning is performed by image analysis, using a Cyberware model to reproduce the correspondent's features.
- Video spatialization reconstructs fictitious viewpoints in a real-life meeting space from a set of non-calibrated reference views.

1. *INTRODUCTION*

Actuellement, une réunion audio-vidéo à plusieurs participants peut se tenir à distance via des liaisons et des équipements spécialisés. Ces systèmes offrent une qualité de service acceptable pour des réunions limitées à deux participants, ou plus généralement à deux sites. Au-delà, l'environnement proposé par les systèmes actuels devient critique et les communications entre participants deviennent beaucoup plus difficiles qu'elles ne le sont dans le cas d'une réunion réelle (pas de contacts visuels entre les participants, pas d'immersion sonore, point de vue imposé,...). Un système de vignette de chaque participant et/ou un affichage alterné des différents sites n'offrent pas, à ce jour, un service satisfaisant (figure 1). *TRAIVI* est un projet de télé-virtualité qui vise à mettre en place des espaces virtuels de téléconférence via des liaisons bas-débit. Il s'agit d'implanter une structure de communication audio-vidéo entre plusieurs sites distants afin que différentes personnes puissent converser confortablement, en réduisant au maximum l'impression de distance grâce à des outils de réalité virtuelle. Pour atteindre un niveau de réalisme audiovisuel satisfaisant dans la création et la gestion de ces espaces virtuels de réunion, plusieurs techniques de traitement audio et vidéo doivent être parfaitement maîtrisées. Actuellement, des études sont menées en traitement vidéo pour d'une part, cloner les visages des participants (section 2) et pour d'autre part, construire un espace virtuel commun de réunion (section 3) dans lequel seront insérés ultérieurement les clones. Le clonage des visages est réalisé par analyse d'images (sans marqueurs) et utilise un modèle *CYBERWARE* de chaque participant pour la restitution. La spatialisation vidéo permet quant à elle de reconstruire par interpolation d'un ensemble de vues de référence non calibrées des points de vue fictifs de l'espace de réunion. Le modèle de visage de chaque participant ainsi que plusieurs vues de référence de la salle de réunion sont supposés être soit déjà disponibles sur les sites récepteurs, soit préalablement télé-chargés en début de réunion. Afin de limiter la quantité d'informations à transmettre, seuls les paramètres de mise à jour du clone de chaque participant ainsi que ceux caractérisant, à chaque instant, le point de vue à afficher de l'espace de réunion seront transmis au cours de la session. Afin d'assurer une cohérence par rapport à la position de chaque participant autour de la table de réunion et de son centre d'intérêt dans la scène, les points de vue restitués sur chaque site récepteur ne seront pas identiques.



Restitution vidéo d'une réunion perçue par un quatrième participant avec un système classique (en haut), avec un système virtuel (en bas).

1. Systèmes de vidéo-conférence : classique et virtuel.

2. CLONAGE

Le clonage est l'une des techniques-clés pour la création d'espaces virtuels de communication et de télé-présence [1]. Dans le cadre de cette application, le clonage de visage utilise un modèle *CYBERWARE* pour la restitution afin de garantir un niveau acceptable de réalisme. Un modèle *CYBERWARE* est obtenu par le biais d'une acquisition tridimensionnelle d'un visage réel. Chaque participant a donc son propre modèle. Les données sont matérialisées par deux fichiers numériques (figure 2) : le premier contient un ensemble de points 3D représentant la géométrie de la tête, et le second contient les informations de texture associées à ces points. Le télé-asservissement d'un modèle *CYBERWARE* est ici divisé en deux parties : globale (i.e . : mouvements du visage) et locale (expressions du visage). La détection, puis la restitution, des mouvements globaux de la tête sont décrits dans la section 2.1. Différentes approches sont envisagées dans la section 2.2, pour assurer la restitution des expressions faciales, incluant le mouvement de la bouche, des yeux, des paupières, des sourcils,...

2.1 Animation globale

Afin de valider les techniques développées en analyse et synthèse d'images, le schéma suivant de télé-asservissement global d'un modèle *CYBERWARE*, incluant la transmission des paramètres, a été simulé (voir fig. 3) :

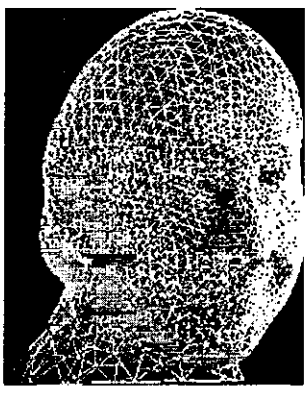
— un individu λ, situé devant sa station de travail sur un site 1 (dit site émetteur), est filmé par une caméra ;

— les images acquises sont analysées afin d'extraire des informations sur la position de la tête ;

— les informations ainsi extraites sont transmises au site distant 2 (dit site récepteur) ;

— le site 2, qui possède localement en base de données le modèle *CYBERWARE* de l'individu λ, interprète les paramètres transmis et restitue en conséquence le visage cloné du participant.

La première étape consiste à séparer la silhouette du locuteur du fond de la scène. Cette opération est réalisée simplement par une différence seuillée entre l'image courante et une image de référence de la scène prise avant l'arrivée du locuteur devant sa station de travail. Cette approche requiert actuellement que le fond de la scène soit statique. Afin de donner l'impression aux participants d'être réunis dans un même espace, l'image de fond associée à chaque site émetteur sera remplacée à la restitution par une image de fond associée à la salle de réunion commune dans laquelle les participants sont censés se trouver. Une fois le visage isolé du fond, celui-ci est à chaque instant englobé par une fenêtre rectangulaire F, à l'aide d'une technique d'histogrammes. Au sein de cette fenêtre, la position des yeux G et D est déterminée par une technique d'appariement dynamique de blocs. Sont enfin définis, l'axe des yeux H et l'axe médian perpendiculaire V. Les



Un modèle « CYBERWARE » est obtenu par le biais d'une acquisition tridimensionnelle d'un visage réel. Les données sont matérialisées par deux fichiers : le premier contient un ensemble de points 3D représentant la géométrie de la tête (à gauche), et le second contient les informations de texture associées à ces points (à droite). L'information géométrique du visage représentée sur cette illustration a été obtenue après triangulation par la méthode proposée par H. Delingette (environ 3000 triangles construits sur une base d'environ 1,5 million de points fournis en sortie par le système d'acquisition).

2. Modèle CYBERWARE.

six degrés de liberté de la tête (translations gauche-droite et haut-bas, translation avant-arrière, rotation gauche-droite, rotation haut-bas et inclinaison de la tête) sont alors respectivement déterminés par la position de *W*, la largeur de *W*, la position de *V* dans *W*, la position de *H* dans *W*, et l'inclinaison du segment inter-oculaire *GD* (figure 3).

Les yeux sont localisés à l'aide de motifs de référence extraits en début de session et d'une technique d'appariement de blocs. L'algorithme de suivi des yeux doit pouvoir s'adapter aux changements 2D photométrique et géométrique dus aux mouvements 3D de la tête du locuteur. Pour ce faire, les motifs de référence sont remis dynamiquement à jour à chaque instant à partir des résultats obtenus aux instants précédents. Cette remise à jour inclut un recentrage et une remise à l'échelle afin que le motif soit stable au cours de la séquence.

Le paragraphe suivant aborde la partie la plus délicate en clonage de visage, à savoir la détection puis la restitution des expressions faciales des participants.

2.2 Animation locale

Plusieurs stratégies, éventuellement complémentaires, sont envisagées pour tout d'abord détecter par analyse d'images les expressions faciales du locuteur, et ensuite les répercuter lors de la phase de synthèse afin d'animer les clones des participants. Certaines expressions telles que le mouvement des yeux et des sourcils peuvent être simulées lors de l'affichage par la seule modification des informations de texture associées aux données géométriques du modèle *CYBERWARE*. A chaque modèle géométrique, il sera associé, en plus du fichier global de texture de référence, plusieurs dictionnaires prédéfinis de textures locales permettant de simuler certaines expressions faciales (figure 4). Ainsi, à chaque instant, en fonction de l'état du locuteur identifié lors de l'étape d'analyse (les yeux vers la gauche ou vers la droite, la bouche ouverte ou fermée,...), des informations nécessaires à l'animation faciale seront à ajouter à celles déjà transmises pour l'asservissement global du modèle *CYBERWARE*. Néanmoins, la quantité globale d'informations à transmettre restera extrêmement faible au regard du volume initial d'informations.

Pour d'autres expressions, il sera de plus nécessaire de compléter la texture du modèle par des portions d'images réelles. Typiquement, si la langue ou les dents d'un participant sont visibles lors de sa locution alors que ces éléments ne figuraient pas dans le modèle *CYBERWARE* de référence, les portions d'images réelles correspondantes devront être segmentées lors de l'étape d'analyse, puis transmises, et enfin mises à l'échelle pour être incrustées dans la texture artificielle du modèle.

La vidéo et l'audio peuvent être tous deux utilisés pour détecter le mouvement de la bouche. Il est possible par analyse d'images, à l'aide de contours actifs, d'évaluer avec précision l'état spatio-temporel des lèvres du locuteur (aire, déformations, périmètres). Ces informations extraites, il s'agira ensuite de les répercuter sur le modèle lors de l'affichage. Pour ce faire, l'ensemble des coordonnées 3D représentant la partie géométrique du modèle *CYBERWARE* sera manipulé à l'aide d'un maillage actif, sur lequel un modèle de déformation du visage suivant le système " Facial Action Coding System " (FACS) pourra ensuite être appliqué. A défaut de restituer les expressions réelles des participants, il sera également tout à fait acceptable, dans le cadre de cette application, d'afficher des expressions cohérentes par rapport aux signaux de parole.

En haut à gauche: le locuteur est devant sa station de travail. La position de sa tête, puis de ses yeux sont automatiquement détectés par analyse d'images.
En bas à gauche: le modèle « CYBERWARE » du locuteur est affiché sur un site distant.

En haut à droite: le locuteur tourne la tête vers sa gauche.
En bas à droite: le modèle « CYBERWARE » du locuteur tourne également à gauche.

Cette simulation tourne, en temps réel (une dizaine d'images/seconde), sur stations SGI.
Les signaux audio sont provisoirement transmis sans traitement particulier. Les images n'étant plus transmises explicitement puisque seuls quelques paramètres de mise à jour du clone le sont, des problèmes de synchronisation audio/vidéo peuvent apparaître.

3. Télé-asservissement global d'un modèle CYBERWARE.

Au centre: la texture originale; A gauche et à droite: modification de la direction du regard par altération de texture.

4. Exemple d'animation locale d'un modèle " CYBERWARE " par altération de texture.

Ces clones devront être ensuite intégrés dans l'espace virtuel de réunion généré par spatialisation vidéo à partir de quelques vues réelles. Cette étape réalisée, chaque participant devra pouvoir visualiser l'espace de réunion sous un angle cohérent par rapport à sa place virtuelle (chacun voyant les autres sauf lui-même) mais également par rapport à son centre d'intérêt dans la scène.

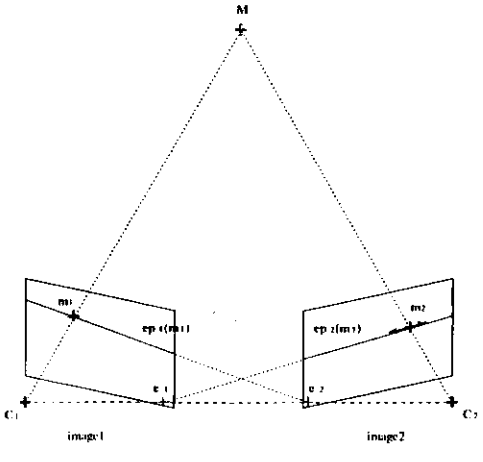
3. *SPATIALISATION VIDÉO*

La spatialisation vidéo a pour objectif de générer et de manipuler un espace complet de réunion, construit à partir de quelques vues de référence [2]. La maîtrise de cette technique permettra à chaque participant de percevoir la salle sous un angle cohérent par rapport à la position qu'il est censé occuper dans cet espace virtuel. C'est précisément ce qui nous différencie des systèmes de visioconférence actuels, qui imposent encore à chaque conférencier un point de vue général sur la scène. Notons qu'aucun modèle artificiel de la salle de réunion n'est créé et qu'au contraire, afin de garantir le réalisme de la scène visualisée, les resynthèses de points de vue de la scène sont produites à partir d'images réelles, sans phase préalable de calibration explicite. Actuellement, la reconstruction d'une vue fictive est réalisée à partir de trois vues de référence. Des extensions sont en cours pour travailler sur un nombre de vues de référence plus important permettant une couverture visuelle complète et précise d'une salle de réunion afin d'y immerger les clones des participants.

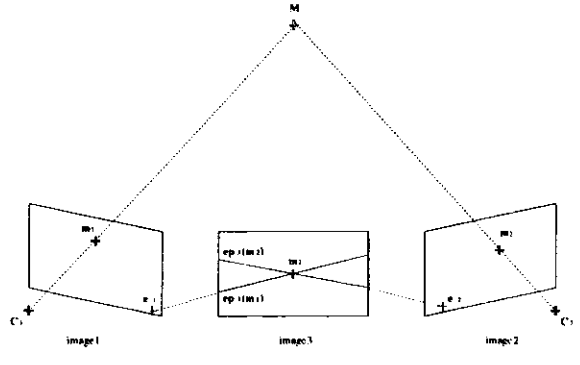
L'algorithme présenté dans cette section constitue une première étape vers l'objectif précédemment défini. L'approche retenue s'appuie sur la notion de géométrie épipolaire bien connue en stéréovision, et plus particulièrement sur les travaux de A. Shashua (figure 5).

Ces travaux, qui ont permis d'établir des relations trilinéaires entre trois vues d'une même scène, permettent de reconstruire l'une des vues à partir des deux autres et de quelques paramètres. L'implantation suivante a été réalisée (figure 6) :

Etape d'analyse. Mise en correspondance de triplets de points dits d'appui afin de calculer les paramètres de pseudo-calibration (quelques dizaines de points caractéristiques). Il est à noter que ces paramètres sont spécifiques à une vue, et sont de plus relatifs aux deux autres vues utili-

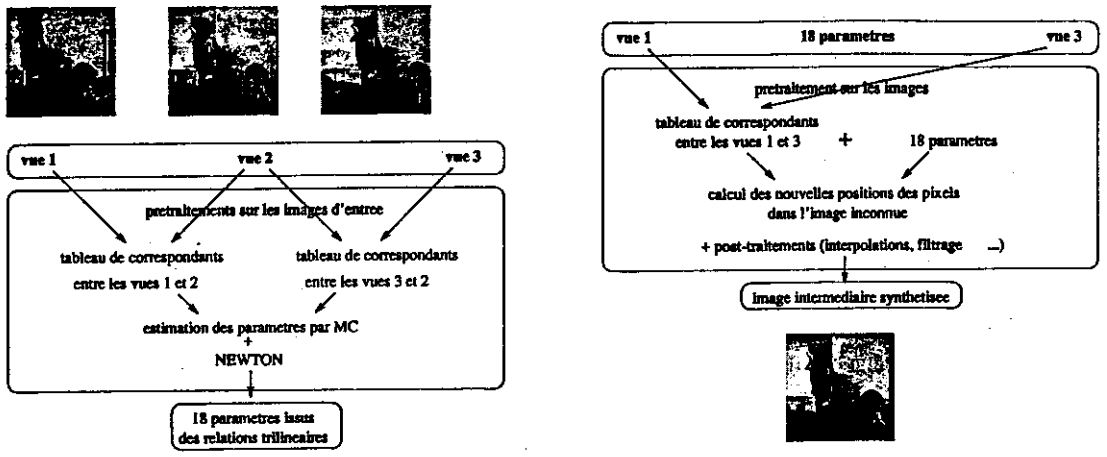


En vision binoculaire, la géométrie épipolaire permet d'associer à un point donné m1 d'une des deux vues, un ensemble de points de l'autre vue qui définit une droite appelée « droite épipolaire », le long de laquelle le point homologue m2 se situe.



Si l'on étend ce principe à un ensemble de 3 vues, il est clair que pour un couple de points homologues (m1,m2), la position du troisième point m3 de la troisième vue se déduira directement comme étant l'intersection les droites épipolaires ep3(m1) et ep3(m2).

5. Principe de la géométrie épipolaire en vision binoculaire et trinoculaire.



L'algorithme de spatialisation vidéo se décompose en deux étapes : une première d'analyse et une seconde de synthèse. Lors de l'étape d'analyse, un jeu de paramètres identifiant la configuration des caméras est estimé. A partir de ces paramètres, l'une des vues peut se reconstruire à partir des vues voisines lors de la synthèse. La manipulation des paramètres entre les étapes d'analyse et de synthèse permet de reconstruire des vues virtuelles.

6. Etapes d'analyse et de synthèse.

sées pour la reconstruction. Ces paramètres sont donc fonction de la configuration des trois caméras.

Etape de synthèse. Mise en correspondance dense des deux vues qui ne sont pas à synthétiser, puis reconstruction de la troisième vue à partir des paramètres dits de calibration implicite (obtenus à l'analyse) et de la mise en correspondance réalisée précédemment.

Afin de restituer une image aussi proche que possible de l'originale, un post-traitement est nécessaire car il existe des conflits dans certaines zones de l'image (i.e. certains points ont plusieurs valeurs de luminance, d'autres aucune,...).

A ce stade, cette approche permet de reconstruire une troisième vue d'une scène à partir de deux autres vues à la condition que ces trois vues existent, y compris la troisième (i.e. la vue à reconstruire). La connaissance de cette troisième vue est actuellement indispensable pour d'une part calculer les paramètres dits de calibration qui interviennent dans les relations trilinéaires (étape d'analyse), et d'autre part valider l'algorithme. Cette limite est la contrepartie du non-passage par une phase explicite de calibration 3D. A ce niveau, bien que cette approche puisse apporter un gain en terme de compression (c'est-à-dire la possibilité de transmettre une vue à l'aide de quelques paramètres seulement, dans la mesure où les vues de référence sont connues du récepteur), elle ne répond pas entièrement aux objectifs fixés ci-dessus : au-delà de l'aspect communication, être en mesure de reconstruire une vue éventuellement inexistante (figure 7). Il est néanmoins possible d'agir à deux niveaux lors de l'étape de synthèse pour reconstruire de nouveaux points de vue :

— en modifiant les valeurs d'un ou plusieurs paramètres dits de calibration,

— en changeant l'une ou les deux vues de référence.

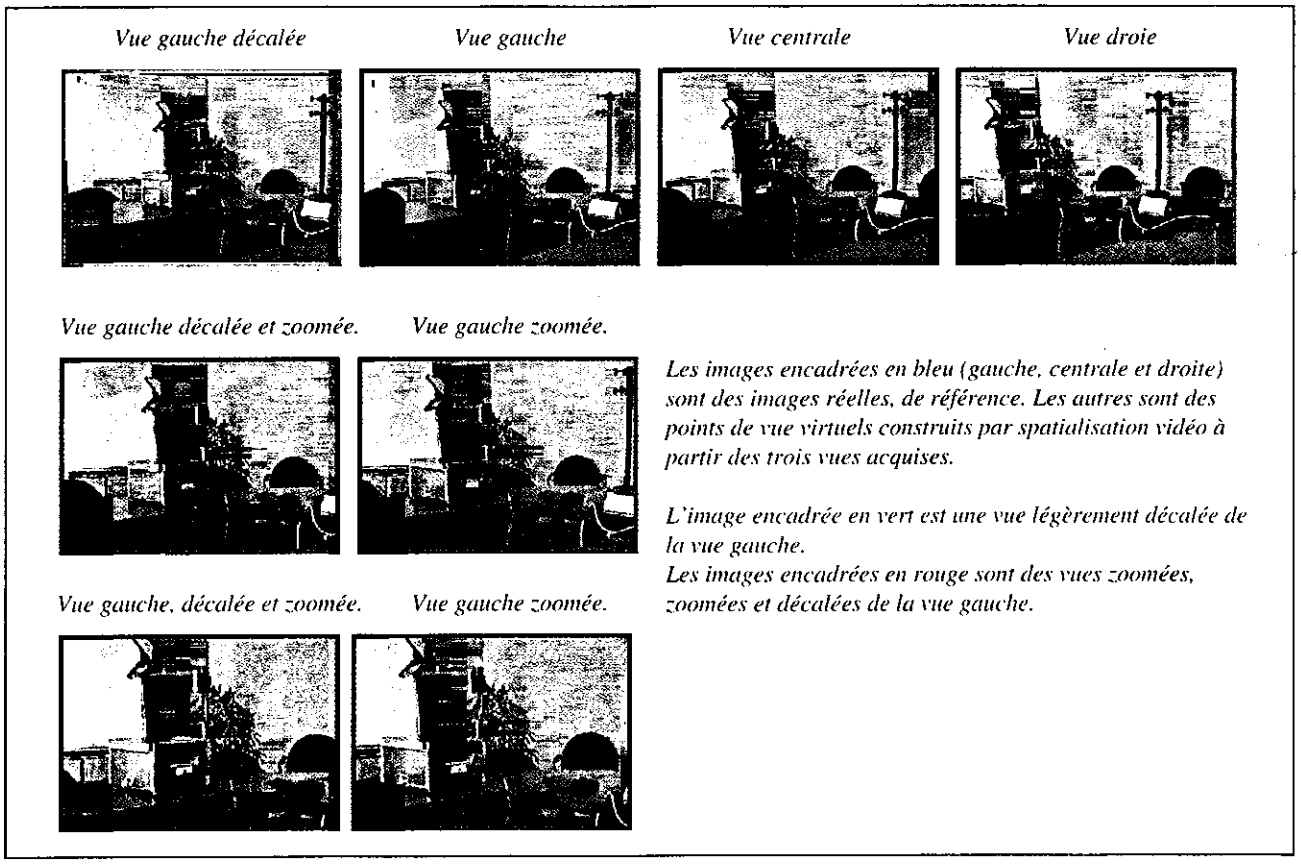
Ces deux possibilités permettent effectivement de reconstruire une vue fictive mais il est encore difficile de maîtriser a priori le point de l'espace qui sera ainsi atteint.

Afin de couvrir entièrement la salle de réunion avec une précision suffisante, il sera nécessaire d'extrapoler l'approche présentée précédemment afin qu'elle soit opérationnelle pour un ensemble de n vues ($n >> 3$). Il s'agira ensuite de calculer pour chaque vue de référence l'ensemble des paramètres de calibration précités (relatifs à un ou plusieurs couples de vues sélectionnés parmi les $n-1$ autres vues). Les paramètres de calibration associés à un point de vue fictif de la scène seront ensuite estimés, par interpolation, à partir des paramètres de calibration associés aux vues de référence. Une fois que ces paramètres auront été calculés, il sera alors possible de reconstruire un point de vue virtuel en choisissant également les deux vues de référence les plus adéquates parmi les n.

Cette technique maîtrisée, il sera possible de restituer, derrière les modèles *CYBERWARE* une image de fond correspondant à la salle de réunion dans laquelle vous êtes censé être, et ce, sous un angle de vue cohérent avec la position que vous êtes également censé occuper d'une part et la direction de votre regard d'autre part.

4. CONCLUSION

TRAIVI est un projet de télé-virtualité. Ce projet vise à mettre en place une structure de télécommunication audio-vidéo entre plusieurs sites distants via des liaisons bas-débit afin que différentes personnes puissent converser confortablement, en réduisant au maximum l'impression de distance grâce à des outils de réalité virtuelle. Les traite-



7. Quelques points de vue virtuels obtenus par spatialisation vidéo à partir d'un triplet d'images.

ments vidéo interviennent à plusieurs niveaux dans ce projet : le clonage et la spatialisation vidéo. Afin de garantir un niveau de réalisme visuel satisfaisant, les images virtuelles sont créées à partir d'une base d'images issues d'acquisitions réelles non calibrées.

Les techniques de clonage de visages et de spatialisation vidéo présentées ici dans le cadre de ce projet de télé-virtualité sont également étudiées pour d'autres applications. La technique de clonage est aussi utilisée pour le clonage des acteurs au cinéma, ou bien encore pour le clonage des présentateurs de télévision. La technique de spatialisation vidéo est également étudiée pour la navigation dans un environnement virtuel dont on ne connaît que quelques points de vue seulement (visite à distance d'un bâtiment ou d'un site en milieu hostile), ou bien encore pour la visualisation, sous un angle de vue défini de manière interactive par le téléspectateur, d'un événement sportif couvert par plusieurs caméras de télévision.

Bibliographie du projet

[1] «A Multi-Site Teleconferencing System using VR Paradigms». S. VALENTE & J.-L. DUGELAY, ECMAST'97, 21-23 May 1997, Milan, Italy.

[2] «Image Reconstruction and Interpolation in Trinocular Vision». J.-L. DUGELAY & K. FINTZEL, 3rd Int. Conf. IMAGE'COM 96, 20-22 May, Bordeaux, France.

L'imagerie virtuelle est de plus en plus étudiée et utilisée dans des domaines aussi variés que les télécommunications, la télé-chirurgie, la robotique, les jeux vidéo... Nous proposons deux serveurs à consulter :

• Club "Réalité Virtuelle" : http://www.iiriam.fr/
Groupe de Travail "Réalité Virtuelle" du GDR PRC CHM
http://www.inria.fr/epidaure/personnel/subsol/GT-RV/gt-rv.html
• ftp zenon.inria.fr/epidaure/gt-rv

Jean-Luc DUGELAY, Docteur de l'Université de Rennes I, est actuellement enseignant chercheur, en charge des activités vidéo du département Multimédia de l'Institut EURECOM, Sophia Antipolis.