VERS LA TELECONFERENCE VIRTUELLE

Jean-Luc Dugelay[‡], Katia Fintzel[‡⌇], Stephane Valente[‡]
Philippe Dubois[†] & Hervé Delingette[⌇]

[‡]Institut EURECOM, Département Communications Multimédia
2229, route des Crêtes, B.P. 193
F-06904 Sophia–Antipolis Cedex
Tél.: +33–(0)4.93.00.26.26
Fax: +33–(0)4.93.00.26.27
E–mail: {dugelay}@eurecom.fr
URL. http://www.eurecom.fr/∼image

[⌇]Espri Concept
Les Taissounières HB2, BP 277,
F-06905 Sophia–Antipolis Cedex

[†]Institut EURECOM, Département Communications d'Entreprise

[⌇]INRIA, Projet Epidaure
2004, route des Lucioles, B.P. 93
F-06902 Sophia–Antipolis Cedex

Résumé

Dans ce papier, nous proposons des outils d'imagerie virtuelle (tels que le clonage de visages et la spatialisation vidéo) qui peuvent être utilisés pour définir de nouveaux systèmes de vidéoconférence offrant plus de "confort d'utilisation" que les systèmes actuels, malgré des liaisons très bas-débit. Ce nouveau concept repose sur la métaphore d'une salle de réunion virtuelle où les utilisateurs pourront choisir leur place.

En particulier, nous proposons des algorithmes de clonage vidéo pour représenter les participants par l'intermédiaire de modèles synthétiques 3D de leur visage, visualisables sous des points de vue différents de celui de la caméra qui analyse les mouvements d'un interlocuteur.

Par ailleurs, le réalisme de la salle de réunion virtuelle est renforcé par des techniques de spatialisation vidéo, qui a pour but de créer des points de vue inédits à partir d'images statiques non-calibrées d'une salle de réunion existante.

1 INTRODUCTION

Après une brève description des systèmes de téléconférence couramment utilisés dans la section 2, nous introduisons de nouveaux outils de traitement d'image pour les télécommunications dans la section 3.
Le but est de créer un système de téléconférence virtuelle, qui offre une meilleure qualité de services (QoS) que les systèmes conventionnels, malgré l'utilisation de liaisons très bas-débit (Internet, communications mobiles).
Dans la dernière section, plus prospective, nous étudions les possibilités d'intégrer ce type de systèmes de téléconférence virtuelle aux stations de travail, réseaux et interfaces standards tels que les PC, Internet et VRML.

2 SYSTEMES DE VIDEOCONFERENCE

2.1 RNIS, réseau adapté aux téléconférences

Initialement, les systèmes de téléconférence ont été conçus pour une utilisation via des réseaux RNIS (Réseau Numérique à Intégration de Services). RNIS a rapidement semblé être un bon support pour ce type de services étant donné qu'il est lui-même basé sur le système téléphonique (POTS: Plain Old Telephone System), d'où une large couverture, une bande passante et des paramètres de QoS garantis. D'autant plus que la granularité d'une bande passante constituée de canaux à 64 kbits/s (ou 56 kbits/s en Amérique du Nord) est suffisante pour assurer le transport de flux audio et vidéo d'une qualité acceptable. Afin d'obtenir un haut niveau d'interopérabilité, l'UIT a développé une série de standards, comme la norme H320 [1] incluant toutes les définitions requises pour la création d'un système de vidéoconférence. Dans la lignée de H320, on trouve les normes H263 [2], pour la compression vidéo, G711 pour le codage audio et quelques autres standards pour définir entre autres les processus de contrôle et de multiplexage.

Du fait de l'effort significatif de standardisation, nous pouvons trouver sur le marché de nombreux produits développés suivant ces standards et RNIS est sans doute devenu le moyen le plus efficace d'organiser une vidéoconférence point à point.

2.1.1 Un mauvais support de diffusion

Cependant, RNIS est très restreint en ce qui concerne le nombre de sites participant à une vidéoconférence commune. Comme la bande passante et les facilités de diffusion de RNIS sont très limitées, l'utilisation d'un pont est nécessaire pour assurer la distribution des flux audio et vidéo aux différents sites. Un pont est un site maître qui reçoit tous les flux audio et vidéo et qui les redistribue aux différents sites. Cette solution coûteuse et non paramétrable apparaît comme une sérieuse limitation de RNIS dans le cas de communications multipoints.

2.2 Internet, réseau pour les téléconférences?

Ces dernières années, Internet n'a cessé de grandir et est devenu le plus important réseau au monde. Au départ Internet était dédié au transport de données sans contraintes de temps-réel. Avec l'émergence des supports multimédia, il devient tentant d'utiliser Internet pour le transfert de flux audio et vidéo.

2.2.1 Un manque de garantie de QoS

Durant les cinq dernières années, la communauté de chercheurs Internet s'est particulièrement intéressée au développement d'outils et de solutions pour le transfert de données audio et vidéo via des réseaux à très bas-débit.

En réalité, l'inconvénient majeur d'Internet est l'absence de méchanismes de réservation de QoS, particulièrement en ce qui concerne la bande passante requise pour l'acheminement de flux audio et vidéo en temps réel. Lorsque le réseau est surchargé, la vidéo reçue via Internet est en général de mauvaise qualité. Plusieurs algorithmes [3] essaient de résoudre ce problème, mais sans succès d'autant que le réseau est de plus en plus surchargé (la charge du réseau augmente aujourd'hui plus rapidement que son "renforcement").

2.2.2 Un bon support de diffusion

Avec son extension multidiffusion (multicast) [4], le protocole Internet (IP) fournit maintenant un support pour les communications multipoints, simplifiant considérablement la configuration et le développement des applications de vidéoconférence multipoints. Aujourd'hui, on trouve plusieurs outils permettant d'organiser des vidéoconférences multipoints via Internet comme vic ou vat [5].

2.3 Limites des systèmes de vidéoconférence

2.3.1 D'un point de vue technique

Les systèmes de vidéoconférence RNIS courants limitent le nombre de participants à une session de téléréunion. Les outils Internet qui supportent correctement les communications multipoints ne produisent pas une QoS suffisante. De plus la capacité de la bande passante dépend de la charge du réseau et est souvent insuffisante pour supporter des flux multimédia sans algorithme de compression.

2.3.2 D'un point de vue ergonomique

Les systèmes de téléconférence utilisant la technologie Internet aussi bien que ceux utilisant RNIS sont basés sur la vidéo. Ils ne produisent pas un réel sentiment de présence pour les raisons suivantes:

– ils n'offrent qu'une vue 2D des participants, avec aucune cohérence des positions relatives de chaque participant,

– ils n'offrent pas d'environnement commun (par exemple les arrières-plans d'une même pièce).

C'est pourquoi nous proposons un nouveau concept de téléconférence virtuelle utilisant des outils de traitement d'image comme le *clonage de visage* et la *spatialisation vidéo*.

3 OUTILS D'IMAGERIE VIRTUELLE

3.1 Clonage

Dans le contexte de notre projet de télé-virtualité, l'intérêt du clonage vidéo est le suivant:

- fournir aux utilisateurs une représentation 3D réaliste des autres participants, qui peuvent être visualisés sous n'importe quel angle de vue, suivant la position initiale que chaque personne veut ou est sensée occuper dans la salle de réunion virtuelle;
- éviter de transmettre des images via le réseau et n'envoyer qu'une représentation compacte sous forme de paramètres permettant l'animation du modèle et nécessitant une bande passante aussi faible que possible.

3.1.1 Travaux précédents

Des algorithmes de clonage vidéo sont utilisés couramment pour l'animation globale d'un modèle (correspondant à la position et à l'orientation de l'intervenant dans l'espace 3D), ainsi que pour son animation locale (révélant ses expressions faciales courantes). Dans la littérature, on trouve plusieurs références concernant le clonage vidéo, comme [6, 7, 8, 9], mais la plupart d'entre-elles (exceptée [6]) considèrent que l'intervenant regarde la caméra et restreignent les mouvements globaux du visage. D'autre part, on peut déplorer le manque de réalisme des modèles faciaux utilisés dans [6, 8, 9], produits "à la main" ils ne représentent pas réellement le sujet. Ces modèles artificiels sont néanmoins largement utilisés car du fait de leur géométrie relativement simplifiée, leur animation et leur manipulation sont aisées. Seuls Terzopoulos et Waters [10] ont commencé à utiliser de façon générique des modèles CYBERWARE, réalistes et mono-locuteurs [7], qu'ils ont retravaillés pour les rendre manipulables par un algorithme général.

3.1.2 Premiers résultats

Nous pensons que les recherches présentées dans la section 3.1.1 ne sont pas applicables dans le cas d'un système de téléconférence virtuelle, car aucune d'entre elles ne remplit à la fois toutes les conditions suivantes: les algorithmes doivent être temps-réel, les résultats visuels obtenus doivent être très réalistes, aucun marqueur de couleur doit être utilisé sur le visage des intervenants et enfin les utilisateurs doivent avoir la possibilité de bouger raisonnablement la tête devant la caméra.

Afin d'obtenir un niveau de réalisme satisfaisant, nous proposons d'utiliser des modèles CYBERWARE retravaillés [11] pour représenter les participants. Nous avons développé un démonstrateur [12] de suivi des mouvements globaux, basé sur un algorithme de block–matching différentiel de régions 2D et sur un filtre de Kalman permettant d'estimer la position et l'orientation du modèle 3D à partir de l'observation des positions 2D des régions d'intérêt.

Grâce au réalisme des modèles synthétiques, une boucle d'asservissement basée sur des images synthétiques avec compensation d'illumination 3D permet de rendre le suivi plus robuste sans avoir recours à des marqueurs artificiels mettant en valeur les points caractéristiques du visage du locuteur.

FIG. 1 – *Résultats de l'algorithme de suivi vidéo (colonne gauche) et rendus synthétiques (colonne droite) — Le modèle CYBERWARE de J.–L. Dugelay nous a été fourni par le laboratoire LUAP de l'Université Paris 7.*

Comme le montre la figure 1, le système peut suivre des rotations très larges hors du plan image. Pour étendre les aspects du clonage vidéo au système de téléconférence, nous étudions actuellement un algorithme d'animation des traits du modèle par combinaison de déformations linéaires du maillage et d'animation de texture.

Le clone d'un participant est synthétisé sous l'angle de vue de la caméra (figure 1), mais il est bien sûr possible de modifier les paramètres de synthèse afin d'immerger le clone sous un point de vue et sous un éclairage différents dans un espace de réunion virtuel géré par *Spatialisation Vidéo*.

3.2 Spatialisation vidéo

Le deuxième aspect du traitement d'image que nous étudions dans le but de créer un espace de conférence virtuel est la *spatialisation vidéo* pour le contrôle des images de fond de la scène, représentée uniquement par des vues 2D non calibrées sans modèle CAD 3D. Un tel processus doit pouvoir nous offrir la possibilité de visualiser la salle de réunion en question depuis n'importe où et dans n'importe quelle direction, au lieu d'imposer un point de vue unique pour chaque site participant, comme le font les systèmes de téléconférence actuels. Cependant, il semble impossible, en termes d'acquisition d'une part et de liaisons bas-débit d'autre part, de créer dans un premier temps puis de transmettre toutes les vues nécessaires de la scène. Notre travail utilise donc la trilinéarité combinée au plaquage de texture pour compresser les données à transmettre et accroître l'information en créant des points de vues inédits.

Pour ce faire, nous nous appuyons sur une première méthode de base permettant la reconstruction d'une vue existante à partir de deux vues voisines (résumée en section 3.2.1). Cette méthode de reconstruction constitue aujourd'hui une étape de validation de l'utili-

sation de la trilinéarité, que nous avons ensuite étendue à la synthèse de vues inexistantes (section 3.2.2) afin de couvrir plus largement l'espace virtuel.

3.2.1 Régénération de vues réelles

Par extension des concepts de stéréovision à trois vues perspectives de la même scène, nous définissons les *tenseurs trilinéaires*, qui peuvent être exprimés en termes algébriques par quatre systèmes trilinéaires modélisés pour la première fois par A. Shashua [13]. En utilisant l'une de ces formes trilinéaires, nous pouvons reconstruire une vue existante à partir de deux vues voisines par l'algorithme suivant (voir fig 2):

- une phase d'analyse, utilisant les correspondants dans trois vues originales non calibrées permet d'obtenir une estimation des dix-huit paramètres d'une forme trilinéaire (pour plus de détails concernant la définition des paramètres trilinéaires voir [14]);

- une phase de synthèse, utilisant les correspondants des deux images externes et les paramètres précédemment estimés, permet de reconstruire l'image centrale.

Cette technique initialement "orientée pixel" nécessitant des mises en correspondances denses tant au niveau de la phase d'analyse que de la phase de synthèse, a été modifiée pour en faire une méthode "orientée maillage". Dans ce cas, à la suite de la phase d'analyse, un maillage basé sur les points d'intérêt des images est associé à chaque image originale. La séquence de trois images initiales est alors remplacée par une texture et trois maillages plus ou moins denses suivant la complexité de la scène. La phase de synthèse n'utilisent plus alors que les noeuds des maillages des deux images originales externes et les dix-huit nombres flottants préalablement estimés pour reconstruire le maillage associé à l'image centrale, avant plaquage de texture de référence.

L'information initiale concernant les deux vues externes à transmettre est donc réduite à une texture complète accompagnée de deux maillages téléchargés par avance. En termes de données, un jeu de dix-huit nombres flottants remplace une vue complète (la vue centrale dans notre cas), mais il faut évidemment procéder à la reconstruction de cette vue.

Etant en mesure de reconstruire une vue à partir de ses voisines, nous nous intéressons maintenant à la synthèse de vue inexistantes.

3.2.2 Synthèse de vues virtuelles

Des extensions possibles de la méthode de reconstruction ont été étudiées afin de créer des points de vue virtuels à partir d'un jeu de vues initiales [15]: simulant un changement de distance focale ou une transformation géométrique 3D de la caméra relative au point de vue à reconstruire. En appliquant directement sur les paramètres trilinéaires les modifications algébriques simulants des changements de paramètres intrinsèques (distance focale) ou extrinsèques (position et orientation 3D) de la caméra relative à la reconstruction, nous pouvons générer de nouvelles vues. Seule l'étape de synthèse de la méthode est nécessaire après avoir modifié le vecteur de 18 paramètres (voir fig 2). Quelques résultats visuels sont présentés figure 3.

Ceci est particulièrement intéressant pour notre application. Effectivement, après une phase de téléchargement de quelques vues réelles non calibrées de l'espace de réunion et

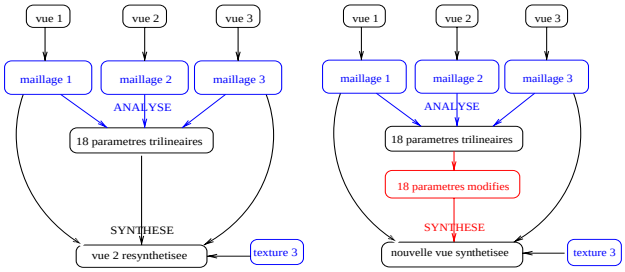

FIG. 2 – *Méthodes de synthèse de vues réelles et virtuelles.*

une pré-estimation des vecteurs de paramètres trilinéaires correspondants, nous sommes capables par calculs algébriques de créer, pour chaque site indépendamment des autres, de nouveaux points de vue cohérents pour chaque participant, basés sur ses paramètres de mouvement (rotation et translation globales de sa tête et de ses yeux) et sur sa position virtuelle dans la salle de réunion.

4 Remarques concluantes

L'imagerie virtuelle offre de nouvelles perspectives en ce qui concerne les systèmes de téléconférence, qui utilisent des liaisons très bas-débit [16]. Après une phase préliminaire de téléchargement (i.e. transmission des modèles CYBERWARE des participants et de plusieurs points de vue de la salle de réunion), de tels systèmes peuvent offrir plus de liberté d'interaction qu'une réunion classique; par exemple, les participants peuvent choisir leur point de vue par rapport à leur place virtuelle dans la salle de réunion ou par rapport à leurs centres d'intérêt.

Dans l'état actuel du projet, des prototypes d'algorithmes de clonage et de spatialisation ont été développés sur des stations de travail SGI utilisant la librairie graphique OpenGL. Les exemples de la figure 1 sont extraits d'une séquence vidéo d'une durée de 30 secondes acquise à la cadence de 10 images de résolution 320×242 par seconde. La bande passante nécessaire (hors téléchargement) pour transmettre les paramètres globaux au visualiseur de scènes virtuelles est 6 paramètres/image $\times$ 2 octets/paramètre $=$ 12 octets/image.

La prochaine phase du projet concerne l'intégration de tels outils dans un système réseau standard et dans un browser WEB sur Internet. Deux solutions sont envisagées: la première consiste à implanter un module de téléconférence virtuelle parmi les outils MBone multicast existants [17], la difficulté à prévoir venant ici de la disponibilité et des performances de la librairie OpenGL sur différentes plate-formes matérielles; la seconde serait l'utilisation des algorithmes de synthèse via le langage VRML [18], le challenge ici étant de contourner les problèmes posés par la communication multi-points temps-réel entre les différents participants via Internet.

Enfin, l'essor des multimédias mobiles pourrait offrir à l'avenir un nouvel espace d'utilisation de ce projet, en distinguant les deux modes suivants:

– le téléchargement via des réseaux fixes,

– le mode de fonctionnement en séance sur des terminaux multimédias via des réseaux mobiles.

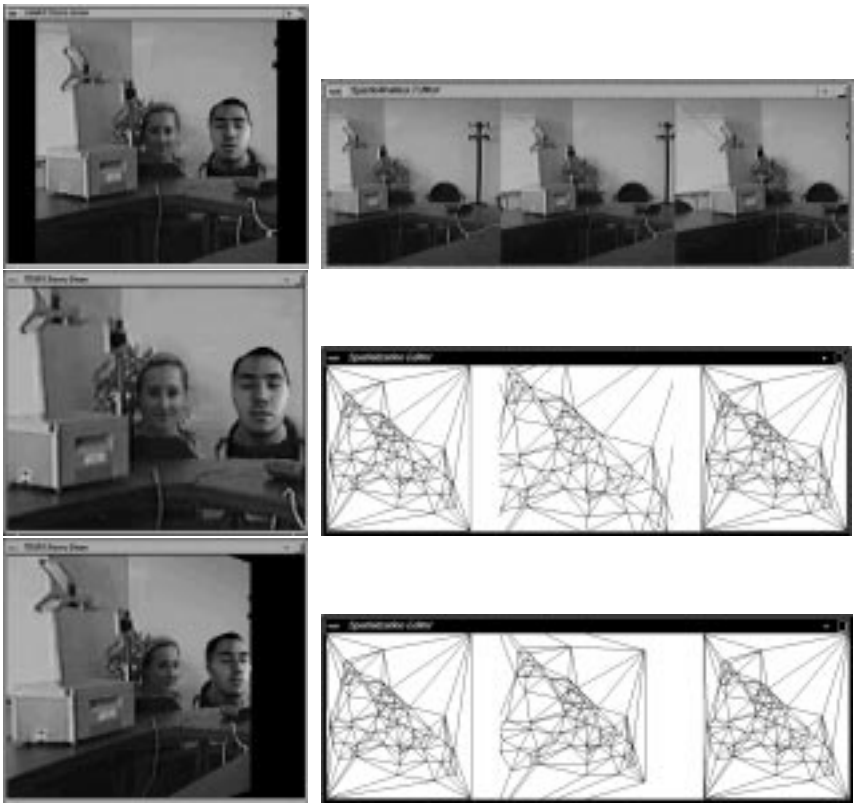

FIG. 3 – *Nouveaux points de vue synthétisés aprés un changement de distance focale ou une rotation de caméra. Les maillages des images fond correspondants sont présentés à droite.*

Références

[1] IUT. Narrow-band visual telephone systems and terminal equipment, March 1996.

[2] IUT. Video codec for audiovisual services at p x 64 kbit/s, March 1993.

[3] T Talley & F.D Smith K Jeffay, D.L Stone. Adaptive, best effort delivery of audio and video across packet-switched networks. In *3rnd Intl. Workshop on Network and OS Support for Digital Audio and Video*, SanDiego, CA, November 1992.

[4] IETF. Host extensions for IP multicasting, November 1988. rfc1112.

[5] S MacCanne V Jackobson. vat. Technical report, Lawrence Laboratory, University of California, Berkley, CA.

[6] P.-E. Chaut, A. Sadeghin, A. Saulnier, and M.-L. Viaud. Création et animation de clones. In *Imagina — Méta–mondes/Metaverses*, pages 244–257, Monaco, Février 1997.

[7] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6), June 1993.

[8] I. S. Pandzic, P. Kalra, and N. Magnenat Thalmann. Real time facial interaction. *Displays*, 15(3), 1995. *Butterworth — Heinemann*.

[9] I. A. Essa, S. Basu, T. Darrell, and A. Pentland. Modeling, tracking, and interactive animation of faces and heads using input from video. In *Computer Animation'96 Conference*, Geneva, Switzerland, June 1996.

[10] CYBERWARE Home Page. URL http://www.cyberware.com.

[11] H. Delingette. *Modélisation, Déformation et Reconnaissance d'Objets Tridimensionnelles à l'aide de Maillages Simplexes*. PhD thesis, Ecole Centrale de Paris, July 1994.

[12] Cyberware Model Analysis. URL http://www.eurecom.fr/~image/KFINTZEL/TRAIVI/traivi.html.

[13] A Shashua. On geometric and algebraic aspect of 3D affine and projective structures from perspective 2D views. In A Zisserman & D Forsyth eds J-L Mundy, editor, *Applications of Invariance in Computer Vision*. Second European Workshop Invariants, Ponta Delagada, Azores, October 1993.

[14] S Avidan & A Shashua. Tensorial transfer: representation of N>3 views of 3D scenes. In *ARPA Image Understanding Workshop*, Palm Springs, CA USA, February 1996.

[15] K Fintzel & J-L Dugelay. Spatialisation vidéo. In *Proc. CORESA'96 Conf.*, CNET Grenoble, France, Février 1996.

[16] J. Ohya, Y. Kitamura, F. Kishino, and N. Terashima. Virtual space teleconferencing: Real-time reproduction of tridimensional human images. *Journal of Visual Communication and Image Representation*, 6(1):1–25, March 1995.

[17] MBONE (or IP Multicast) Information Web. URL http://www.mbone.com.

[18] VRML. URL http://vrml.sgi.com.