Apparaît dans : *CORESA, 26 et 27 mars 1997, Issy les Moulineaux, France*

CLONAGE DE VISAGES DANS UN CONTEXTE DE TÉLÉRÉUNION

Stéphane Valente & Jean–Luc Dugelay

Institut EURÉCOM
2229, route des Crêtes, B.P. 193
F-06904 Sophia–Antipolis Cedex

Tél.: +33–(0)4.93.00.26.66
Fax: +33–(0)4.93.00.26.27
E–mail: {valente,dugelay}@eurecom.fr
URL. http://www.eurecom.fr/∼image

1 INTRODUCTION

Les systèmes classiques de téléconférence codent en général la redondance spatiale et temporelle d'une séquence vidéo vue comme un signal stochastique. Dans le cas de téléconferences multisites, de tels systèmes conduisent au mieux à des vues comme indiquées sur la figure 1(a) où chaque site est représenté par une imagette différente. Sinon, quand les réseaux n'ont pas la bande passante suffisante pour transmettre plusieurs images, on obtient un affichage alterné des sites pour montrer la personne qui est en train de parler. Dans tous les cas, il est difficile pour les utilisateurs d'avoir l'impression de faire partie d'une vraie réunion.

(a) Vue classique (b) Vue virtuelle

FIG. 1 – *Ce que voit un quatrième participant lors d'une téléconférence multisite.*

À l'inverse, les techniques de compression orientées–objet considèrent les images comme une projection perspective d'objets physiques $3D$, et codent la manière dont les objets sont agencés dans la vue courante. Associé aux techniques de la réalité virtuelle, un tel système peut devenir meilleur qu'un système classique, aussi bien du point de vue du taux de compression que du confort de l'utilisateur. L'idée–clée est de construire un espace virtuel de réunion partagé entre tous les utilisateurs, de synthétiser les points de vue individuels qu'ils auraient lors d'une vraie réunion, et de leur donner la possibilité d'avoir des contacts occulaires entre eux par l'intermédiare de clones $3D$ des participants, en bref, de synthétiser une vue comme celle de la figure 1(b). Pour atteindre un haut niveau de réalisme, d'autres techniques doivent être mises en oeuvre, comme la spatialisation et le multiplexage audio, l'annulation d'écho, la synchronisation audio/vidéo…

1.1 ÉTAT DE L'ART

Un système de téléconférence virtuelle soulève en fait deux problématiques différentes que l'on retrouve dans la littérature en traitements vidéo (en dehors des aspects purement réseaux) : pouvoir reproduire en $3D$ les participants, et construire un espace de réunion pour les immerger.

Depuis quelques années, des recherches ont été menées sur des sujets liés tels que l'analyse et la synthèse de visages humains [1], et le clonage de visages [2]. Historiquement, ces travaux sont partis d'un modèle de visage générique (le plus répandu étant le modèle CANDIDE), mais donnent des résultats peu réalistes [3]. De plus en plus d'équipes travaillent sur l'amélioration du réalisme du modèle : l'INA (l'*Institut National de l'Audiovisuel*) plaque une texture construite à partir de photographies sur le modèle $3D$ [4], Reinders *et al.* adaptent un maillage générique à une personne sur des vues $2D$, et Choi *et al.* obtiennent des expressions hyper–réalistes par déformation de maillages [5] ; nous proposons de prendre la démarche inverse, c'est–à–dire de partir d'un modèle $3D$ non–générique dépendant d'une personne, et de le rendre plus générique pour l'utiliser dans un système automatique.

Pour ce qui est de la construction d'un environnement virtuel, quelques expériences ont déjà eu lieu avec le projet TELEPORT [6] : une scène synthétique est construite par des logiciels de CAD dans le but de prolonger précisément un espace de réunion par un écran de la taille d'un mur.

1.2 LE PROJET "TRAIVI"

Le projet TRAIVI ("TRAItement des images VIrtuelles") a pour objet de mettre en place des espaces virtuels de réunion sur des liaisons à bas–débit avec un haut niveau de réalisme. Nous utilisons pour le clonage vidéo des modèles de visage texturés acquis spécifiquement pour chaque participant. Le projet prévoit également la construction d'environnements virtuels à partir de photographies non–calibrées par *spatialisation vidéo* [7]. Pour une présentation plus détaillée de ces deux aspects, les lecteurs sont invités à se reporter à [8].

Cette communication présente nos premiers résultats en *clonage vidéo* : comment nos modèles sont télécontrolés globalement (leur position et orientation dans l'espace) à la section 2, et localement (leur expression faciale) à la section 3.

2 ANIMATION GLOBALE

L'animation globale consiste à déterminer la position et l'orientation d'un interlocuteur dans l'espace $3D$ par analyse d'image en temps–réel, et à interprèter les paramètres extraits pour synthétiser son clone dans une position cohérente. Cette section décrit un algorithme qui y parvient sans ajouts de marqueurs sur le visage et sans requérir l'intervention de l'utilisateur pour initialiser la procédure.

2.1 PARAMÈTRES SUIVIS

Les paramètres qui nous intéressent sont (voir figure 2) : la silhouette du visage entourée par le rectangle W, la position des yeux L et R, et les axes horizontal H et vertical V

médians des yeux. Les 6 degrés de liberté de la tête sont alors déterminés par :

translations gauche/droite et haut/bas : données par la position de W
translation avant/arrière : par la largeur de W
rotation gauche/droite : par la position de V dans W
rotation haut/bas : par la position de H dans W
la dernière rotation : par l'angle d'inclinaison de la droite (L, R)

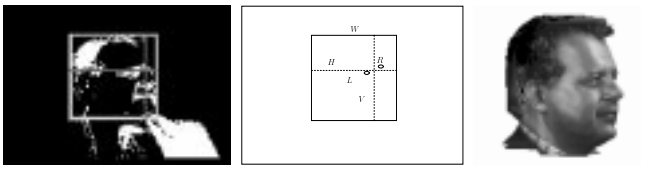
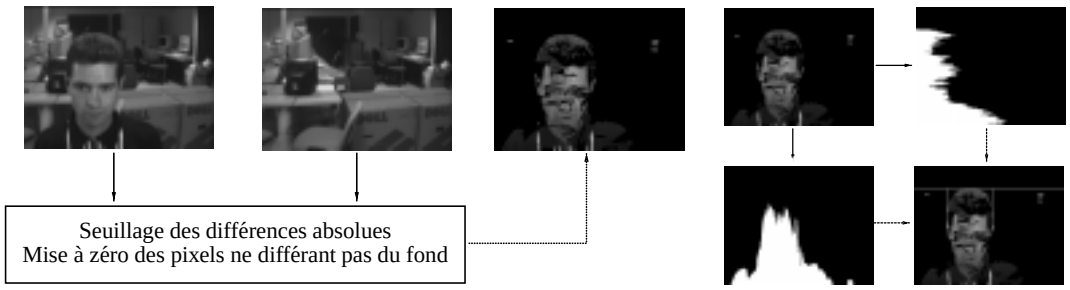


FIG. 2 – *Paramètres d'analyse des mouvements globaux.*

2.2 ALGORITHME D'ANALYSE ET DE SUIVI TEMPOREL

Les paramètres de la figure 2 sont estimés en deux étapes : tout d'abord, le rectangle de la silhouette W est recherché, puis les yeux sont suivis par *template–matching* à l'intérieur de W.

2.2.1 Détermination de la silhouette de la tête



(a) Seuillage de la silhouette. (b) Histogrammes binaires.

FIG. 3 – *Recherche de la silhouette du locuteur.*

Notre algorithme de suivi de la silhouette fait comme hypothèse que le fond derrière l'image du locuteur reste statique durant la session. Une image de référence du fond est alors soustraite à l'image courante (figure 3(a)), puis les histogrammes binaires horizontaux et verticaux sont calculés pour en déduire le haut, la gauche et la droite de la tête (figure 3(b)). Le bas de la tête n'est en pratique pas déterminé puisqu'il n'intervient pas dans l'estimation des paramètres. On peut noter que l'hypothèse du fond statique éliminable par soustraction est satisfaisante pour cette application car elle autorise un algorithme simple et temps–réel. De plus, les clones seront insérés dans un environnement dont le point de vue dépendra de la personne qui regarde la scène, ce qui permet de faire abstraction du fond lors de la phase d'analyse.

2.2.2 Suivi des yeux

Les paramètres H, V, L et R sont obtenus à partir de la position des yeux. L'algorithme de suivi des yeux doit surmonter quatre problèmes distincts :

changements d'échelle : quand l'utilisateur se rapproche ou s'éloigne de la caméra vidéo

rotations $2D$ dans le plan image : quand le visage subit une rotation $3D$ par rapport à l'axe tête/caméra (par exemple lorsque l'utilisateur incline sa tête sur ses épaules face à la caméra)

transformations non–linéaires dans le plan image : quand le visage subit une rotation $3D$ par rapport à un axe autre que celui de la tête/caméra (lorsque l'utilisateur tourne la tête de gauche à droite)

changements d'illumination : quand l'utilisateur bouge en général vis–à–vis des sources lumineuses

On peut résumer les trois dernières difficultés en disant que l'algorithme d'analyse doit pouvoir s'adapter aux variations géométriques et/ou photométriques des motifs des yeux.

Nous traitons le problème de variation temporelle des yeux par un *template–matching dynamique* : à chaque fois que les yeux sont localisés dans l'image courante, les templates sont mis à jour avec les yeux trouvés, et ils s'adaptent ainsi automatiquement à tous les changements. Parallèllement, la taille des template est modifiée suivant la taille de la fenêtre W de manière à ce qu'ils contiennent la même quantité de détails discriminants malgré les changements d'échelle (si les templates initiaux contenaient les yeux et les sourcils, et si l'utilisateur s'éloigne de la caméra, les templates seront mis à jour avec des portions d'images plus petites pour ne pas inclure le nez ou les oreilles qui pourraient "leurrer" les templates par la suite).

Toutefois, les templates dynamiques doivent être mis à jour avec précaution. En effet, si le template courant est mis à jour avec l'image des yeux décalée de un pixel, il est évident que petit–à–petit, les yeux vont glisser de l'intérieur du template jusqu'à disparaître (figure 4). Pour rendre les templates dynamiques plus fiables dans le temps, il est nécessaire, avant de les modifier, de localiser précisément le centre des yeux par une *seconde passe* de template–matching qui utilise cette fois un template de réference du centre des yeux mis à l'échelle (pas à jour) pour s'adapter aux changements d'échelle du visage.

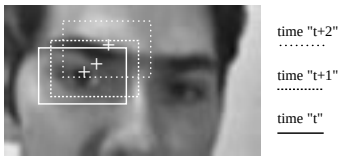


FIG. 4 – *Déviation des templates dynamiques.*

La figure 5 donne quelques exemples du la robustesse du suivi des yeux. Les templates de la première passe contenaient les yeux et les sourcils de l'utilisateur, tandis que les templates de recentrage contenaient uniquement les iris. On notera que grâce à l'inclusion des sourcils, les yeux peuvent être localisés même lorsque l'utilisateur les ferme. Pour plus de détails sur cet algorithme, les lecteurs sont invités à se reporter à [8].

FIG. 5 – *Robustesse des templates dynamiques aux rotations $2D$ et $3D$, aux changements d'échelle et aux yeux fermés.*

3 ANIMATION LOCALE

L'animation globale replace un clone dans l'espace virtuel sans en changer l'expression faciale. Une stratégie d'animation locale est nécéssaire pour reproduire les expressions des différents participants. La section 3.1 traite des spécificités des modèles $3D$ CYBERWARE. La section 3.2 décrit les différentes procédures d'animation possibles. Enfin, la section 3.3 traite des techniques d'analyse vidéo pouvant asservir les animations locales.

3.1 MODÈLES CYBERWARE

Les modèles CYBERWARE sont produits par des scanners cylindriques $3D$, et sont constitués de deux fichiers : le premier décrit la forme géométrique d'un visage par un nuage de points $3D$ (environ 1,5 million), et le second est une texture à appliquer sur le maillage des points.

Ces modèles sont hautement réalistes, mais ne sont valables que pour un individu donné dans une expression figée. De plus, ils ne sont pas optimisés en terme de nombre de points, et ne contiennent aucune information anatomique ou physique (comme un modèle d'os sous–jacents ou de muscles déformables de manière élastique sur l'axe des temps [1]).

Pour réduire la complexité des modèles, nous avons adopté l'approche de Delingette [9] qui transforme le maillage initial cylindrique en maillage triangulaire dont la densité est proportionnée à la surface décrite, avec en tout $1\,400$ points combinés en $2\,800$ triangles (voir figure 6(b)).

3.2 STRATÉGIES DE SYNTHÈSE

Nous avons identifié deux voies différentes pour animer localement le modèle : la première simule une animation en modifiant la texture appliquée sur le maillage, et la seconde modifie directement celui–ci.

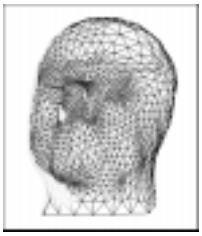
3.2.1 Animation par la texture

Nous avons validé cette méthode en nous attachant aux yeux du modèle : à l'aide de logiciels de retouche d'images, nous avons défini les trois textures de la figure 6(a), et le logiciel de synthèse bascule en temps–réel d'une texture à l'autre afin de modifier la direction du regard.

3.2.2 Animation par le maillage

L'autre solution d'animation locale consiste à animer directement le maillage. Celui de Delingette a été utilisé avec succès dans des simulations chirurgicales en raison de ses possibilités de déformations physiques réalistes [10], et permet également des interpolations de type "splines" conduisant à la simulation de rides d'expressions [11]. Nous travaillons actuellement sur des méthodes visant à contrôler les déformations du maillage suivant le système standardisé FACS [12].

(a) Contrôle du regard par la texture.



(b) Maillage.

(c) Modèle inséré dans un environnement.

FIG. 6 – *Images syntétisées.*

3.2.3 Combinaison

Certaines parties d'une tête humaine, comme la langue et les dents, ne sont pas scannées, et doivent pourtant apparaître quand le clone ouvre la bouche par animation du maillage. On doit alors recoller des images réelles dans la texture CYBERWARE.

Le problème est alors de savoir si l'on utilise des portions d'images extraites lors de l'analyse ou des textures prédéfinies, suivant la bande passante disponible pour la communication du système. D'un côté, se servir d'un dictionnaire de textures signifie référencer un index, et donc n'a quasiment aucune incidence sur les besoins en bande passante. De l'autre côté, envoyer des portions d'images "live" demande plus de bande passante, et surtout les besoins sont difficilement prédictibles, car l'envoi peut se produire à n'importe quel moment lors de la session. Choi a décrit un mécanisme basé sur l'orthonormalisation de Graham–Schmidt pour diminuer la bande passante nécessaire à l'envoi d'une nouvelle texture en considérant celles qui ont déjà été transmises, mais sans résoudre la question de la prédictabilité [5]. En outre, le "copier/coller" d'images $2D$ en temps–réel peut s'avérer délicat à réaliser sur un modèle visualisé sous un point de vue différent, les difficultés étant ici d'opérer lors du collage la bonne transformation $2D \rightarrow 2D_{cylindrical}$ et d'homogénéiser les caractéristiques photométriques entre les images "live" et la texture CYBERWARE.

La solution des dictionnaires de texture prédéfinies (éventuellement construit à partir d'images de sessions typiques) semble être la solution la plus efficace en l'état actuel de nos travaux.

3.3 ÉTUDES PROSPECTIVES

Une fois que la politique de synthèse sera figée dans notre plateforme de simulation pour chaque caractéristique faciale, il faudra mettre en oeuvre des techniques d'analyse locale. Il est important de noter que les techniques d'analyse dépendent des solutions choisies pour la synthèse, et que des éléments faciaux différents peuvent requérir des techniques d'animation différentes.

Pour l'animation par la texture, le *template–matching* est ce qu'il y a de plus simple pour mesurer la similarité entre une texture prédéfinie et l'image courante. Si besoin est, la mesure de similarité peut être rendue invariante à l'illumination, à l'échelle et aux rotations [13].

L'animation du maillage par analyse d'images est une tâche beaucoup plus difficile. La technique la plus courante dans la littérature est l'utilisation de *contours actifs* (*snakes*) [1]. Nous aimerions toutefois nous orienter vers une solution plus novatrice : puisque les modèles CYBERWARE offrent un haut niveau de réalisme à travers le lien précis qu'ils font entre l'image (la texture) et le maillage, il est possible de réaliser une coopération analyse/synthèse indirecte en entraînant des *eigenfeatures*. Elles ont été largement utilisées pour faire de la reconnaissance de visages de par leur capacité à représenter de manière compacte un espace complexe en calculant un jeu de vecteurs orthogonaux dans les sous–espaces d'énergie maximale [14], ou représentant des orientations particulières [15]. La difficulté majeure pour calculer des eigenfeatures optimales est d'établir une base de données calibrée. Les exemples d'entraînement doivent tous présenter la même échelle et le même éclairement, ce qui est loin d'être acquis si des images réelles sont utilisées. Par contre, un modèle CYBERWARE sera idéal pour synthétiser des images calibrées d'entraînement en faisant varier les expressions faciales générées par animation du maillage ou de la texture. De cette manière, la base d'eigenfeatures calculée sera optimale non seulement du point de vue des conditions d'entraînement, mais aussi du point de vue de la décomposition des expressions faciales par rapport aux paramètres qui contrôlent l'animation du maillage.

Et finalement, dans le cas où le visage d'un participant n'est pas totalement visible par une caméra, il sera également acceptable d'appliquer au maillage de la bouche une déformation réaliste basée sur l'analyse du signal audio [16].

4 REMARQUES CONCLUANTES

Nous avons présenté les premiers résultats obtenus en *clonage vidéo* pour le projet TRAIVI. Le but est de contrôler des interfaces $3D$ représentant des personnes réelles dans un environnement virtuel. Nous avons abordé les spécificités des modèles CYBERWARE, et décrit comment ils peuvent être animés globalement et localement indépendamment du point de vue utilisé lors de leur synthèse. Tandis que la procédure d'animation globale est validée, nous travaillons actuellement sur les animations locales.

Nous avons réalisé un prototype logiciel qui télécontrôle les mouvements globaux d'un ou de plusieurs modèles à travers des sockets UNIX. Le logiciel d'analyse tourne sur une "SGI Indy" à environ 10 images/sec avec des trames de taille 208×160 en niveau

de gris, et est surtout limité par la vitesse d'acquisition de la carte vidéo de la machine. Le logiciel de synthèse foncionne sur une station "SGI High Impact" avec des modèles d'environ $3\ 000$ triangles, plaquage de textures, et une image de fond dépendant de la personne qui visualise la scène virtuelle (voir figure 6(c)).

Références

[1] 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.

[2] A. Saulnier, M.-L. Viaud, and D. Geldreich. Real–time facial analysis and synthesis chain. In *International Workshop on Automatic Face— and Gesture— Recognition*, pages 86–91, Zurich, Switzerland, 1995.

[3] I. A. Essa, T. Darrell, and A. Pentland. Tracking facial motion. In *IEEE Workshop on Nonrigid and Articulate Motion*, Austin, Texas, November 1994.

[4] Institut National de l'Audiovisuel. Televirtuality project: Cloning and real-time animation system. URL http://www.ina.fr/INA/Recherche/TV.

[5] C. S. Choi, K. Aizawa, H. Harashima, and T. Takebe. Analysis and synthesis of facial image sequences in model–based image coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 4(3):257–275, June 1994.

[6] GMD's Digital Media Lab. TelePort: The Communication Wall. URL http://viswiz.gmd.de/DML/cwall/cwall.html.

[7] K. Fintzel and J.-L. Dugelay. Spatialisation vidéo. In *CORESA*, Grenoble, France, 1996.

[8] S. Valente and J.-L. Dugelay. A multi–site teleconferencing system using VR paradigms. In *Ecmast*, Trieste, Italy, 1997.

[9] H. Delingette. *Modélisation, Déformation et Reconnaissance d'Objets Tridimensionnelles à l'aide de Maillages Simplexes*. PhD thesis, Ecole Centrale de Paris, Châtenay–Malabry, France, 1994.

[10] H. Delingette, G. Subsol, S. Cotin, and J. Pignon. A craniofacial surgery simulation testbed. Research Report RR–2199, INRIA, 1994. URL http://www.inria.fr/rapports/sophia/RR–2199.html.

[11] M.-L. Viaud. *Animation Faciale avec Rides d'Expression, Vieillissement et Parole*. PhD thesis, Université de Paris XI–Orsay, Orsay, France, 1992.

[12] P. Ekman and W. V. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Palo Alto, California, 1977.

[13] G. S. Cox and G. de Jager. Template matching with invariance. In *Proceedings of the Fourth South African Workshop on Pattern Recognition*, pages 152–156, 1993. URL http://dip1.ee.uct.ac.za/papers/cox93.html.

[14] A. Pentland, B. Moghaddam, and T. Starner. View–based and modular eigenspaces for face recognition. In *International Conference on Computer Vision and Pattern Recognition*, June 1994.

[15] T. Darell, B. Moghaddam, and A. Pentland. Active face tracking and pose estimation in an interactive room. Technical Report 356, M.I.T. Media Laboratory Perceptual Computing Group, 1996.

[16] R. R. Rao and T. Chen. Cross–modal prediction in audio–visual communication. In *International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA, May 1996.