Clones 3D pour communication audio et vidéo

Frédéric Elisei

UNIVERSITÉ JOSEPH FOURIER-GRENOBLE 1
SCIENCES ET GÉOGRAPHIE
U.F.R. D'INFORMATIQUE ET DE MATHÉMATIQUES APPLIQUÉES

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Discipline : Informatique

arrêtés ministériels du 5 juillet 1984 et du 30 mars 1992

Présentée et soutenue publiquement
par

Frédéric ELISEI

Le 15 Novembre 1999

CLONES 3D POUR COMMUNICATION AUDIO ET VIDÉO

effectuée sous la direction de

Jacques LEMORDANT

au sein du laboratoire GRAVIR-IMAG. UMR 5527.

COMPOSITION DU JURY :

Claude	PUECH	Président
Bernard	PÉROCHE	Rapporteur
Peter	SANDER	Rapporteur
Olivier	AVARO	Examinateur
André	GILLOIRE	Examinateur
Jacques	LEMORDANT	Directeur

Table des matières

Table des figures

Introduction

Notre société nous offre de nombreux modes de communication. Depuis l'invention de l'art, du langage et de l'écriture, le génie humain n'a eu de cesse d'inventer et d'utiliser de nouveaux moyens de partager son savoir et ses sentiments. Délaissant parfois la proximité physique et le contact, contournant certains impératifs de distance et de temps, ces outils sont désormais si nombreux qu'il ne serait pas possible de les énumérer tous. À titre d'exemple, citons quand même le téléphone, son répondeur, la télévision et le fax.

Une définition de la communication

Il s'agit de *la prise de connaissance d'informations, d'opinions ou de sentiments auprès d'autrui.* C'est la réception d'un message, sans présumer de la forme de celui-ci.

De nombreux types de communication

On peut opposer différentes spécificités, parmi les exemples précédemment cités, en les comparant à ce mode de communication primordiale où les interlocuteurs sont tout simplement en présence, dans un même lieu.

Certaines communications nécessitent **un contact**. C'est le cas d'une accolade ou d'une poignée de main, que cherchent à propager certaines expériences de réalité virtuelle.

D'autres outils informent de **la présence** et des occupations de personnes distantes (Boards, ICQ ou les MediaSpaces).

On peut aussi classifier des communications qui se restreignent au **canal oral** (la radio ou le téléphone), ou **transmettent une image** (comme le fax) ou s'appuient sur **du texte** (le talk d'Unix). Il peut aussi s'agir **d'un message sans début ni fin** (cas de la radio ou de la télévision).

Certaines communications se font **en mode différé**, permettant aux utilisateurs de converser ou d'accéder à l'information à leur rythme (rediffusions, messagerie ou répondeur téléphonique) ou bien parce que le délai est inhérent à la méthode employée (courrier postal, envoi d'objets).

D'autres sont **interactives**, permettant de réagir ou de répondre à des intervenants, une caractéristique qui nécessite **une communication bidirectionnelle** plutôt qu'**à sens unique** (émission radiophonique ou télévisée).

Les moyens de communication diffèrent aussi selon **le destinataire visé**: il peut être ciblé (par exemple par un numéro de téléphone) ou le message peut s'adresser au plus grand

nombre (affiches, mailing de masse). Le moyen utilisé conditionne aussi **la discrétion obtenue** (mégaphone, enveloppe scellée et adressée, ou message crypté). Ainsi, la prise de connaissance d'un message est **plus ou moins volontaire**.

Pour des raisons techniques ou sociologiques, la communication peut-être **régulée** du fait d'un arbitre ou de contraintes technologiques. Par exemple les participants peuvent se voir imposer des temps de parole, une hiérarchie ou un maximum d'un seul orateur à la fois (radio-amateurs sur un seul canal).

Enfin, **l'information communiquée et sa fidélité** diffèrent selon le média. Toute la gamme des sentiments ne se transmettra pas forcément avec la même fidélité et facilité selon qu'elle est véhiculée par de la vidéo, seulement par du son ou qu'elle doit être décrite par des mots.

En définitive, *la communication n'est pas forcément réciproque ou volontaire, et le risque est plus ou moins grand que le message soit mal interprété*. Il est donc souhaitable de choisir *le bon outil pour la bonne tâche*, en étant conscient de ses imperfections et de ce qui est ou n'est pas transmis, à défaut d'être perçu.

Problématique initiale

On a choisi de s'intéresser à **un outil de communication qui permettrait de réunir plusieurs personnes distantes, travaillant sur un projet commun par exemple**. Quand un tel groupe ne peut pas se réunir dans un même lieu pour tenir la discussion nécessaire, quelles sont les solutions possibles?

une conférence téléphonique sur un réseau voix ou informatique, permet de se parler à plusieurs. Seules la voix et les intonations sont transmises et l'on ne voit pas les intervenants, ni pour les identifier, ni pour jauger leurs expressions. Difficile dès lors, avec plus de trois participants, de différencier sur un écouteur monophonique lesquels parlent en même temps.

une téléconférence profite d'un canal vidéo dédié, par exemple pour émettre vers plusieurs salles de spectateurs. Le canal de retour (parfois seulement sonore) ne permet généralement pas aux spectateurs de se transformer en participants à part entière. Cette solution est donc plutôt utilisée pour la communication d'entreprise ou la télé-éducation que pour des réunions de travail.

une vidéo-conférence connecte entre eux plusieurs participants, par des flux audios et vidéos croisés, sans hiérarchie ni dissymétrie, permettant une communication distante visuellement attrayante. L'attente de clients cherchant plutôt un outil de travail est souvent déçue par la qualité des images (petite taille et faible précision) et leur débit (quelques images par secondes constituent-elles une vidéo?). Même dans les meilleures conditions (réseau ATM dédié par exemple), le nombre des participants peut dépasser les capacités des machines ou du réseau, au point de rendre ces meetings difficilement utilisables, sinon pénibles et frustrants.

Idéalement, on souhaiterait un système de communication :

- utilisable entre personnes distantes,
- temps-réel, sans notion mesurable ou gênante de différé, c'est-à-dire sans que ce soit préjudiciable à une communication «naturelle» ; c'est un problème constaté lors d'interviews par satellite par exemple : le retard induit par la transmission entre la fin d'une question et la réception de sa réponse est généralement mis à profit pour relancer la question, sans se rendre compte que la réponse, en cours de transit, sera couverte ou coupée. Lors de cette expérience relativiste, chacun découvre quelques secondes plus tard qu'il a parlé «en même temps» que l'autre.
- totalement interactif, sans hiérarchie ni protocole, laissant libre cours à une communication collaborative naturelle,
- muni d'un support audio minimal pour la parole, par exemple celui du téléphone,
- doté d'un support vidéo suffisant pour la reconnaissance des visages d'autrui et d'un minimum d'expressions, les plus réelles et personnelles possible,
- qu'il rappelle une expérience assez proche d'une vraie réunion, avec une certaine notion d'un espace commun,
- avec une architecture et des performances suffisantes pour des communications entre plus de deux personnes.

Est-ce là simplement le profil d'un vidéophone idéal à plusieurs ? Comme on va le voir à l'occasion de l'état de l'art, les points précédents inciteraient à la fois à retenir et à écarter les solutions du tout vidéo, mais aussi celles de la synthèse 3D et de l'animation de clones ressemblants.

Démarche

La première partie de ce document propose un bref tour d'horizon de différentes approches qui semblent se prêter au type de communication envisagé. En comparant les méthodologies basées 3D et celles basées vidéo, on cherchera dans quelle mesure chacune correspond ou échoue face aux contraintes précédemment listées, soit d'un point de vue pratique, soit d'un point de vue théorique.

En conséquence, la seconde partie définira plus précisément la problématique retenue en fonction des difficultés pratiques ou théoriques identifiées. On fixera en particulier les contraintes minimales et les possibles libertés qui ont prévalues pour ce travail, en se plaçant face à diverses opportunités sociologiques ou commerciales notamment.

On proposera ensuite une solution et des choix aux divers problèmes rencontrés, avant de présenter le prototype construit, ses résultats puis nos conclusions.

Tour d'horizon

Introduction

Dans cette partie consacrée à l'état de l'art, on s'intéressera particulièrement aux nombreux domaines qui ont un lien avec les notions de vidéo-conférence et de communication assistée par ordinateur. Par contre, les problèmes très vastes liés aux réseaux et protocoles, par exemple pour garantir les délais ou gérer une dégradation acceptable des baisses de performances ou erreurs de transmission, ne seront pas expressément abordés ici.

Historiquement, ce sont les méthodes de **transmission et compression d'images et de vidéo** qui ont généré les premiers espoirs de vidéo-téléphone ; le chapitre 1 ne détaillera pas explicitement les classes de standards et d'algorithmes qui restent utilisés ou vont l'être, mais rappellera plutôt leurs forces ou leurs faiblesses, en terme de qualité et de coûts notamment. On verra donc en quoi ces techniques ne sont pas idéales pour la classe de communication envisagée, et pourquoi on peut être tenté de les écarter au profit de méthodes orientées 3D.

Dérivés d'un savoir faire en animation, les travaux autour des **modèles 3D pour le visage** sont aujourd'hui très nombreux. Après un rappel sur les principes de modélisation 3D et de rendu, le chapitre 2 exposera les techniques d'acquisition qui permettent de capturer des modèles de visages existants, et discutera la ressemblance obtenue.

On verra ensuite, au chapitre 3, quelles sont les techniques que les animateurs utilisent traditionnellement pour animer les visages et synthétiser des expressions.

Parce que ces méthodes d'animation des visages synthétiques ne forment au mieux qu'une partie d'une application de communication, c'est à leur asservissement synchrone à un locuteur du monde réel qu'on s'intéressera particulièrement, lors du chapitre 4. Dans ce cadre analyse/synthèse, on s'interrogera à nouveau sur la fidélité et la ressemblance, obtenues ou qui pourraient être atteintes, car ces méthodes adaptées présentent des difficultés et des résultats potentiels qui leur sont propres.

Enfin, le chapitre 5 sera consacré aux **approches vidéo-réalistes**, auxquelles cette thèse contribue pour partie. Combinant des images naturelles de la vidéo avec différentes techniques de synthèse 2D ou 3D, ce sont le plus souvent des méthodes hybrides ou basées images.

Avant de conclure ce tour d'horizon, et à titre de réflexion, quelques exemples de projets proposant des **espaces de communication** seront évoqués dans le chapitre 6.

Chapitre 1

Le modèle du vidéophone

On dispose aujourd'hui de nombreuses techniques et standards pour les transmissions et la compression, dont certains sont dédiés à l'image, au son ou à la vidéo, de sorte que ces données, tellement présentes dans notre environnement, inondent aussi le paysage informatique.

Ce chapitre ne dressera pas un catalogue exhaustif des méthodes de transmission ou de compression, mais cherchera plutôt à catégoriser leurs caractéristiques et spécificités, en rapport avec leur application à la communication entre les personnes. On verra alors pourquoi le concept de vidéophone ne s'étend pas parfaitement à plus de deux personnes.

1.1 La richesse de la vidéo

Avec le cinéma puis la télévision, la civilisation de l'image s'est enrichie du mouvement, ou plutôt de son illusion, lorsque ces images se succèdent suffisamment rapidement pour paraître animées. Alors que, comme le prouve un arrêt sur image, les images reçues sur un poste sont bruitées et de faible résolution (chromatique en particulier), une scène vidéo (du réseau hertzien par exemple) dégage pourtant une fidélité et un naturel qui servent encore d'étalon au grand-public pour mesurer la qualité de leurs équivalents informatiques ou digitaux. Le rafraîchissement par images bruitées et approximatives n'empêche pas le cerveau, bien au contraire, de reconstruire une image plus précise, basée sur son expérience du réel.

Dans le cadre de la communication, l'image se révèle riche d'informations utiles, que l'on soit un simple spectateur, un intervenant ou les deux tour à tour :

si le locuteur est visible :

- on peut le reconnaître, ou apprendre à le reconnaître, selon le cas,
- on perçoit ses expressions ou les émotions qui accompagnent ses paroles, ainsi que tous ses gestes,
- la reconnaissance de la parole est facilitée, particulièrement en environnement bruité lorsque les lèvres, les dents et la langue sont visibles.

si l'auditoire est visible :

- on peut voir l'attention et les réactions de l'auditoire, et donc moduler le discours de façon interactive, en accélérant ou en se répétant par exemple,
- on peut authentifier les destinataires, même s'ils ne parlent pas.

Ainsi, Arlésienne maintes fois annoncée qui enrichirait l'expérience du téléphone, le vidéophone est toujours souhaité – à prix abordable – dans les foyers. S'il a été plusieurs fois déployé sur des réseaux expérimentaux, par exemple câblés en fibre optique jusque chez les abonnés, il est technologiquement bien plus réalisable maintenant que lorsqu'on l'envisageait comme un outil analogique, grâce aux applications de la théorie de l'information, l'évolution de la puissance des machines et des techniques de traitement du signal.

1.2 La vidéo digitale

Pour l'image comme pour la vidéo, on peut caractériser quelques critères significatifs pour sa représentation digitale et sa compression :

- **le taux de compression**, qui est lié au gain acquis sur la bande passante nécessaire pour faire transiter le message compressé par rapport au message intégral,
- **le taux de fidélité**, selon que le flux décompressé restituera exactement l'image ou la vidéo initiale, ou que l'on autorise des altérations, parfois parce qu'elles ne sont que peu ou pas visibles, le plus souvent parce que c'est une façon d'améliorer le taux de compression,
- **le coût du compresseur** qui conditionnera la possible utilisation d'une technique donnée pour la compression temps-réel d'un message en direct,
- **le coût du décompresseur**, lorsque la technique n'est pas symétrique, demandant par exemple moins de calculs pour reconstruire le message que lorsqu'on a cherché à le compresser.
- **le type du message original**. Les images ou vidéo de dessins, schémas ou lettrages (sous-titres et légendes par exemple) diffèrent beaucoup de celles obtenues en capturant des scènes naturelles. Ces dernières sont plutôt bruitées et comportent des détails riches en texture, avec peu de ces contours très appuyés qui sont plus caractéristiques de la première catégorie. En conséquence, les algorithmes ne se comportent généralement pas de la même façon selon le contenu du message original, et certains ont été développés spécifiquement pour leur adéquation à l'un ou l'autre type seulement.

1.3 La communication par vidéo

1.3.1 Survol des techniques et évolutions

Il y a encore quelques années, le frein à la communication vidéo se situait au niveau de la taille des données nécessaires pour encoder et faire transiter le message vidéo, en analogique le plus souvent, de sorte qu'un réseau dédié (satellitaire ou hertzien par exemple) était nécessaire.

Avec le codage digital, il devenait possible, par exemple en compressant chacune des images d'un film, de conserver des fragments de vidéo sur des périphériques de stockage communs, ou de réaliser des expériences de dialogue sur des réseaux informatiques réservés.

Avec des méthodes spécialisées dans le codage de la vidéo, comme MPEG-1 puis MPEG-2 il est devenu possible de stocker des films entiers, sur quelques CD par exemple. Mais le coût du compresseur et le délai introduit ne permettaient pas de les détourner utilement pour communiquer à distance.

Maintenant, on dispose de standards spécialement développés pour de telles applications. Comme GSM pour le son, les normes H-261 et H-263 par exemple savent s'astreindre aux contraintes matérielles spécifiques, comme de limiter le débit maximum, de minimiser les délais dans les codeurs et les décodeurs, ou de permettre des implémentations logicielles efficaces.

Avec de nouveaux codeurs orientés maillage ou région [BT97, LLS98, MFL98] par exemple ou des méthodes basées sur des analyses statistiques, on peut générer de la vidéo d'un débit encore inférieur et viser l'utilisation du réseau téléphonique existant. Commercialement et sur le Web, de nombreux formats propriétaires se répandent en autant de *plug-ins*. Sur toutes les plate-formes, on dispose désormais de logiciels de vidéo-conférence [VC], offrant parfois aussi une interface de partage de documents.

1.3.2 La communication à plusieurs

Ainsi, les réseaux informatiques et les techniques de compression rendent possibles différentes expériences de communication, avec une qualité et une souplesse variables, mais qui ira à n'en pas douter en s'améliorant.

Pour un dialogue (deux personnes), le choix du **mode vis-à-vis** semble s'imposer naturellement : chaque personne voit l'autre sur son écran. Mais quel mode de représentation faut-il adopter lorsque les participants sont nombreux ?

- **mode** «**locuteur privilégié**» : si une seule personne est autorisée à parler, comme c'est le cas lors d'une téléconférence, seule son image est retransmise vers tous les sites. Cela économise la bande passante, ou permet de la réserver pour envoyer cette vidéo avec le maximum de résolution et de précision. Cette solution n'est cependant pas satisfaisante puisqu'on ne voit pas les réactions de l'auditoire (importantes au moins pour le locuteur) ni ceux qui lui coupent (plus ou moins poliment) la parole.

- **mode** «**album**», où les différentes vidéos sont juxtaposées pour occuper l'écran. Si les participants sont très nombreux, il va falloir réduire chaque vidéo (ou la plupart) à la taille d'imagettes, tout juste utiles à savoir qui est qui ou qui parle. À ce stade, les détails seront de moins en moins visibles, ce qui signifie que la bande passante – multipliée ou répartie suivant le nombre de participants – est finalement bien mal utilisée, puisque les expressions sont perdues. Enfin, une machine qui devra recevoir et afficher de nombreux flux différents risque de voir ses performances se dégrader (les accélérateurs graphiques n'étant pas forcément conçus pour décompresser simultanément plusieurs vidéos) au point d'handicaper sa mission d'émission du flux vidéo capturé localement.

Dans tous les cas, les participants n'ont pas l'impression de partager un **espace commun** :

- ni au sens réalité virtuelle, puisque chaque participant apparaît derrière un écran qui le sépare des autres,
- ni dans la représentation des autres, puisqu'ils n'apparaissent pas «ensemble» à l'écran, semblant regarder le spectateur derrière l'écran plutôt que la personne à laquelle ils s'adressent.

La figure 1.1 illustre le problème qui apparaît si on veut créer un espace unique qui ressemble à une réunion, à partir des images des intervenants : cette image les représente parfaitement de face (cas A), mais ne pourra servir à les représenter de coté (cas B) que si l'on évite des cas plus extrêmes (C et D). Sans faire appel à des techniques différentes (3D implicite ou explicite), qui seront introduites dans la suite de ce tour d'horizon, on ne pourra pas produire une image composite comme celle de la figure 1.1.

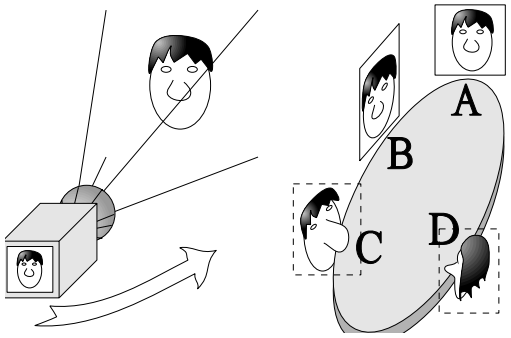


FIG. 1.1 – *Les limites de l'utilisation de l'image*

1.4 Conclusion

On voit donc qu'un outil qui serait parfait pour une communication en vis-à-vis, façon vidéophone, s'adaptera d'autant plus mal que le nombre de participants sera supérieur à deux.

Cependant, la vidéo est fidèle et permet de reconnaître personnes et expressions, de voir et d'entendre simultanément le message et son porteur.

L'évolution des techniques de compression de la vidéo illustre aussi qu'il faut oser altérer le message, perdre de façon contrôlée certains détails, si possible les moins perçus ou les moins importants. Indépendamment, on gagne à envisager le contenu de l'image à un niveau plus structuré (lignes ou régions) ou plus abstrait (zones en déplacement) que celui des pixels. En tirant partie d'une connaissance a priori de ce que l'on doit représenter, on peut même à l'extrême substituer à l'image celle d'un modèle ad-hoc, piloté par quelques paramètres. Ainsi, comme on s'attend à des images de visages/bustes, on pourrait les approximer et les coder par des clones 3D, dès lors que l'on dispose de techniques pour les créer et les animer. Ce sont ces techniques que présentent les trois prochains chapitres.

Chapitre 2

Capture de modèles 3D pour les visages

La déferlante d'animations et d'effets spéciaux en images de synthèse, coursée depuis par l'industrie du jeu vidéo, n'a probablement pas permis au grand public d'échapper à la rencontre avec des personnages 3D. Certaines séquences de Morphing, les animateurs virtuels d'émissions télévisées ou encore la production de longs dessins animés prouvent la réussite, ne serait-ce que commerciale, de ces techniques, dont certaines sont liées aux modèles 3D de visages.

Penchons nous sur ce qui constitue un modèle 3D et fait sa spécificité pour les visages, avant de montrer comment ces structures peuvent être construites et sont généralement animées.

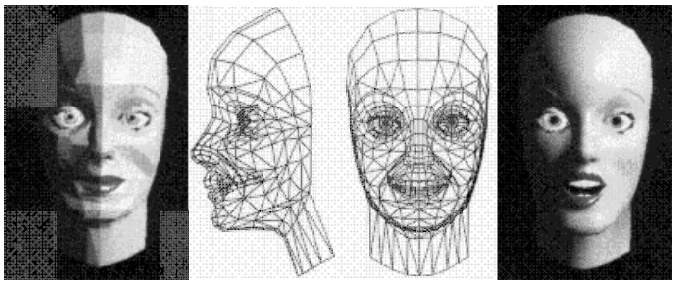


FIG. 2.1 – *Un des premiers modèles de visage [Par82]*

2.1 Notions de modèle 3D

Le modèle 3D est une structure de données qui représente un objet en permettant d'en calculer une multitude de vues, selon différentes conditions (au minimum en faisant varier les points de vue possibles, mais parfois aussi l'éclairage). Cette définition particulièrement peu restrictive inclut par exemple la représentation d'une sphère par un point et son rayon.

En pratique, et du fait de l'existant dans les machines disposant d'accélérateurs graphiques, le modèle 3D est souvent une liste de triangles ou de polygones. Pour ces formats, on dispose d'une grande souplesse à l'utilisation (affichage, transformations, interactions 3D...) et de nombreux outils de saisie ou d'édition. Malheureusement, les polygones ne constituent pas la meilleure des représentations pour les sphères ou les surfaces courbées, comme les visages. Ainsi, il faut beaucoup de triangles pour que le modèle et ses contours externes n'apparaissent pas trop anguleux.

En colorant uniformément chacun des triangles, on obtiendrait un rendu assez peu réaliste, qui rappelle plutôt l'aspect d'un mannequin, mat ou au teint cireux. Pour un rendu photo-réaliste, où la ressemblance dépasse celle de la forme en approchant l'image du vrai modèle, l'usage (reconnu par les accélérateurs graphiques) est d'employer une texture, sorte de papier-peint élastique qui colorera chacun des points des triangles.

Un modèle 3D qui approxime un visage peut être utilisé pour personnifier un individu réel (dans un univers virtuel par exemple). On parle alors généralement **d'avatar**. Si cet avatar est suffisamment ressemblant (par exemple à un acteur ou une personnalité), on pourra dire qu'il s'agit **d'un clone**. S'il est texturé plutôt que peint, il permettra d'obtenir un rendu **photo-réaliste**.

2.2 Création de modèles 3D ressemblants

Aux débuts de l'animation, il fallait créer des modèles de toute pièce, directement depuis les logiciels de modélisation. Les ressources graphiques assez faibles limitant la complexité, donc la ressemblance, il s'agissait plus souvent de créer du neuf ou de l'artistique.

Avec la disponibilité des premiers capteurs, par exemple les stylets 3D, il devenait possible à un manipulateur appliqué de numériser avec précision divers objets. Ainsi, en partant d'un buste réel, soit sculpté soit obtenu par moulage d'un vrai visage, on pouvait – de façon fastidieuse – construire un modèle 3D d'une forme évoquant celle d'une personne ou d'une célébrité donnée.

Le réalisme de la forme n'empêchait généralement pas que le rendu rappelle plutôt l'apparence d'un mannequin : en affectant au modèle des couleurs par zones (chair, cheveux, lèvres, iris...), on ne rend pas compte de la complexité des phénomènes lumineux à la surface de la peau (selon que la zone est plus ou moins irriguée ou constituée de tissus adipeux), ni de celle produite par l'ensemble de la chevelure.

Si l'on a progressé dans la construction de modèles de synthèse plus convaincants [KK89, HK93, YS90], il reste beaucoup moins coûteux d'utiliser une texture pour rendre l'aspect de la peau ou de la texture des cheveux, en capturant les spécifités de la personne (grains de beauté, cicatrice, pilosités faciales...).

Plus automatiques, les scanners lasers ont rendu possible la construction en quelques secondes de modèles aussi volumineux que détaillés (plusieurs milliers de points, avec des détails de l'ordre du millimètre sur les modèles haut de gamme). Du fait du principe de l'acquisition par balayage, certaines zones cachées ou trop diffuses, tels les cheveux ou l'arrière des oreilles ne sont

généralement pas correctement capturées. Comme ces scanners combinent souvent une deuxième caméra, cela permet de digitaliser aussi les données de réflectance (couleur et luminosité) de chacun des points du visage sous forme d'une texture, utile à un rendu photo-réaliste.

Récemment, de nombreuses techniques de reconstruction utilisant une ou plusieurs caméras vidéos sont apparues ou ont été adaptées pour le cas des visages. Certaines demandent deux ou trois caméras calibrées, d'autres travaillent à partir d'un flux vidéo non calibré, ou encore à partir d'une paire d'images (de coté et de front), voire d'une vue unique.

Après une nécessaire discussion sur la notion de ressemblance, on pourra comparer ces méthodes de création. En effet, les méthodologies diffèrent selon l'exigence de ressemblance qu'elles ont adoptée, et se retrouvent donc inégalement adaptées pour une application de télé-présence ou de télé-communication 3D.

2.2.1 Ressemblance, mimétisme et fidélité

Lorsqu'on décide de construire le modèle d'une personne, cherche-t-on seulement à être le plus proche possible de son apparence lorsqu'elle tenait la pose ? Cette **ressemblance statique** n'est pas suffisante dans le cadre d'une communication. On souhaite par exemple que les lèvres soient synchronisées avec la parole, et que le visage s'anime de façon vivante, par ses yeux ou ses attitudes par exemple. La palette de ces comportements est large, depuis des mouvements quasi-symboliques d'automates (mouvements de la mâchoire et clignements des yeux pseudo-aléatoires) en passant par des mouvements physiologiquement corrects mais impersonnels, jusqu'à la fidélité totale au message facial original.

FIG. 2.2 – *De célèbres acteurs du MIRALab [Mir]*

Si les expressions faciales, l'articulation voire la voix ne sont pas du tout celles du modèle vivant, tout en respectant une apparence humaine, on parlera de **mimétisme physiologique**, ou d'avatars ressemblants. Le choix a été fait pour ce document de réserver le terme de **clone**

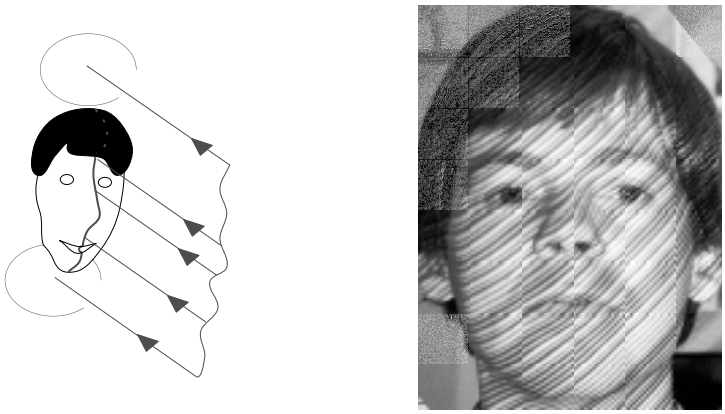
dynamique à une reconstruction 3D fidèle dans ses articulations et mimiques à *un couple message/messager.*

2.2.2 Acquisitions par capteurs 3D

Puisqu'on veut construire un modèle 3D d'un visage, l'approche qui consiste à utiliser un capteur 3D semble assez naturelle.

À l'aide d'un stylet 3D, on pouvait capturer des coordonnées de points à la surface d'un objet, mais pour modéliser une personne réelle (ou un acteur disparu), il fallait d'abord passer par l'intermédiaire d'un buste, réalisé par un artiste ou par moulage. Par ce procédé, la forme 3D est effectivement capturée, mais on n'a aucune information sur le comportement optique de la surface (la pigmentation locale de la peau par exemple). Il faudrait donc éventuellement compléter cette modélisation à l'aide de plusieurs photos, pour créer la texture.

Des services spécialisés dans la capture de modèles 3D vivants – par exemple pour les effets les plus complexes de Morphing – sont ensuite apparus, principalement en utilisant un faisceau laser plan et deux caméras (une pour capturer la trace du faisceau et l'autre pour capturer l'apparence sous forme de texture cylindrique), en rotation pendant quelques secondes autour du modèle [Cyb].



Dans les deux cas, des faisceaux de lumière trahissent le relief de la surface interceptée. À gauche, c'est un faisceau laser plan qui est en rotation relative par rapport au modèle, et permet en quelques secondes la reconstruction d'un modèle complet. Sur l'image de droite [BA99], la projection d'un motif spécialement choisi pour être non répétitif et éviter les ambiguités permet la capture en une seule image, mais la reconstruction du visage sera partielle.

FIG. 2.3 – *Deux méthodes d'acquisition par trace lumineuse*

Parce que certains points, les cheveux ou l'arrière des oreilles le plus souvent, n'ont pas été capturés, le modèle doit généralement être retouché. Indépendamment de ces déficiences, le maillage (ou le nombre de points) obtenu est trop grand, et il faut donc remplacer le modèle par

une version épurée et en extrapolant les données manquantes. Finalement, il n'est pas rare de ne se servir de ces points 3D que pour déformer un modèle 3D générique, sur lequel un travail fastidieux et manuel a déjà été fait (optimiser le nombre de polygones, étiqueter des zones du visage ou attacher des réseaux de muscles virtuels par exemple). Cette opération est presque toujours réalisée pour faciliter l'animation synthétique, comme on le détaillera dans le chapitre suivant.

FIG. 2.4 – *Un même modèle 3D adapté à deux visages différents [BT]*

Les méthodes utilisées pour adapter un clone générique aux données mesurées sont très variables, mais s'imposent toutes de spécifier une déformation qui respectera la topologie initiale. Elles diffèrent généralement selon qu'elles imposent la transformation exacte de certains points, ou l'optimisent plus globalement.

Certains travaux rapportent l'utilisation de l'imagerie médicale comme l'IRM [TH98], mais il s'agit à priori de tirer partie de données accessibles par Internet [VH], utiles pour créer un modèle générique. Plus classiquement, les modèles génériques sont obtenus comme une moyenne de plusieurs clones qui présentent le même maillage, par exemple une base de données de visages obtenus avec une même technique de balayage cylindrique.

2.2.3 Création à partir d'images non-calibrées

Pour s'épargner de coûteux capteurs 3D, il est possible d'utiliser de simples images (photos digitalisées, ou obtenues à l'aide d'un appareil numérique par exemple) du visage que l'on souhaite modéliser. Cela peut aussi permettre de cloner des personnages disparus.

L'approche la plus classique consiste à établir, plus ou moins automatiquement, des correspondances entre diverses images, par exemple entre une vue de face et une vue de coté (ou deux vues de coté quand – à raison – on ne fait pas l'hypothèse de symétrie du visage). Ce faisant, on a précisé une mesure 3D de certaines paramètres, typiquement des données anthropométriques, comme la taille et l'écartement des yeux, la limite du front ou du menton... Un modèle 3D générique peut alors être déformé, pour se conformer à ces mesures et minimiser des contraintes plus globales de distance. En changeant de forme, mais pas de topologie, le modèle générique approche le modèle en gardant nombre de ses propriétés initiales (son paramètrage pour l'animation par exemple).

Par exemple, des chercheurs de NTT [ASW93] utilisent une segmentation basée sur la couleur pour trouver la zone des cheveux (supposés noirs) puis les régions des yeux, de la bouche et des narines. À l'aide d'un filtrage, des points de la ligne du menton sont aussi détectés sur la vue de face, et estiment la taille du modèle avec la limite supérieure des cheveux. Les pixels du profil sont eux aussi extraits de la vue de coté et servent à trouver la déformation d'un profil générique composé de 50 segments, par minimisation d'une fonction de coût. Pour créer la texture, les deux vues (face et profil) sont combinées, dans des proportions liées à la position sur le modèle 3D générique (pré-étalonné pour cette tâche). Ainsi, ils construisent un modèle assez ressemblant, constitué de 800 polygones, à partir de deux images seulement.

Au MIRALab, ce sont aussi deux images (une de face et une de coté) qui sont utilisées pour adapter leur modèle générique d'animation (qui inclut des dents sous la bouche, et de nombreux pseudo-muscles, dont l'action est visible lors de l'articulation de messages parlés synthétiques par exemple). Les profils, la limite des cheveux et de nombreux points positionnés autour des yeux, du nez et du menton sont connus sur le modéle générique, ce qui permet de les projeter de face et de profil.

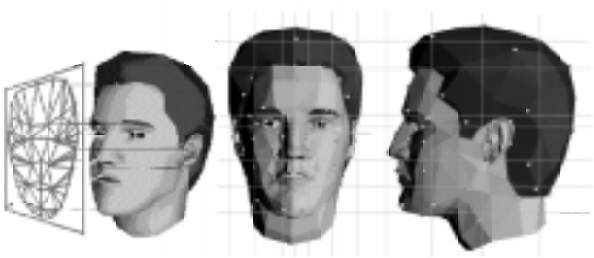


FIG. 2.5 – *Exemples de points de calibration d'un modèle générique [LMT98]*

Ainsi initialisés, ces éléments sont recherchés et localisés sur les images de la personne réelle par des contours actifs contraints (*structured snakes*) ou avec une aide humaine pour les ajuster dans les cas difficiles. Localisés sur une paire d'image, ces points définissent donc des correspondances 3D entre le modèle canonique et la personne réelle. La déformation de tous les points du modèle, (par DFFD, *Dirichlet free-form deformation* [MTKE98, LMT98], qui étend la déformation ponctuelle au reste de l'espace), permet d'obtenir un modèle adapté, compatible avec le modèle générique et toutes ses procédures et modalités d'animation. Une texture personnelle peut aussi être créée, en fusionnant progressivement les images de références, autour d'une ligne fixée sur le modèle générique et qui passe notamment par l'extrémité extérieure des yeux.

Le même paradigme d'adaptation de modèle générique est employé dans [PHSS98]. Par contre, ils procèdent manuellement et avec plus d'images de référence (par exemple cinq), dans le but de créer des textures de haute qualité. Une détermination des paramètres de caméra et un premier ajustement du modèle générique sont réalisés à partir de la mise en correspondance

manuelle de quelques points (par exemple 13) entre les images et le modèle. En rajoutant plus de correspondances (une centaine par exemple), l'opérateur pourra affiner l'adaptation du modèle. Comme alternative à une texture cylindrique unique, les auteurs proposent de faire le rendu texturé directement à partir des images de définition, en combinant les deux plus proches du point de vue de la synthèse. Quoique plus coûteux, ce procédé a déjà fait ses preuves [DTM96] et limite la perte de détails (qui n'étaient pas en correspondance du fait de l'imprécision du modèle 3D). Il permet aussi de texturer avec plus de précision celles qui sont mal définies sur une vue cylindrique (les zones qui vues de face sont orientées de coté, comme les ailes du nez) et même les zones «cachées» par la projection cylindrique (comme l'arrière des oreilles).

À l'EPFL, Pascal Fua [Fua98] a proposé de reconstruire des modèles 3D à partir de séquences vidéos non calibrées, sous l'hypothèse d'un mouvement rigide (la personne ne parle pas) et d'une focale assez stable. Après une estimation des paramètres et du déplacement de la caméra, grâce à cinq points initialisés manuellement, des cartes de disparités peuvent être établies pour des images consécutives. Comme le nuage de points 3D qui en dérive est présumé très bruité et ne forme pas un modèle surfacique, il cherche à estimer des déformations (de plus en plus détaillées, depuis un ajustement global et en pondérant les points selon qu'ils semblent atypiques ou non) qui seront appliquées à un modèle générique. Pour rendre le modèle plus ressemblant, des contraintes supplémentaires peuvent être données manuellement, sous forme de correspondances entre des points/segments du modèle générique et l'image initiale, pour augmenter la précision des profils ou de la délimitation front/cheveux par exemple.

2.2.4 Extrapolations statistiques

À partir d'une seule image, [BV99] construit un modèle 3D probable, relativement à une base de donnée de visages 3D (forme et texture), où les cheveux n'apparaissent plus. Cette base sert en fait à construire un modèle générique paramétré qui représente un espace de visages (3D et texture) de grande dimensionnalité et qui inclut tous les membres de la base. Cette paramétrisation permet de générer de nouveaux visages, par exemple pour passer continuement *(Morphing)* entre des visages de l'espace, ou d'amplifier certaines propriétés (âge, aspect féminin ou masculin...) pour peu qu'on ait étiqueté des axes de l'espace qui les caractérisent. Mais c'est aussi un espace de représentation, qu'ils proposent d'utiliser pour créer un modèle 3D à partir d'une vue (ou plus) et de l'estimation des paramètres de caméra.

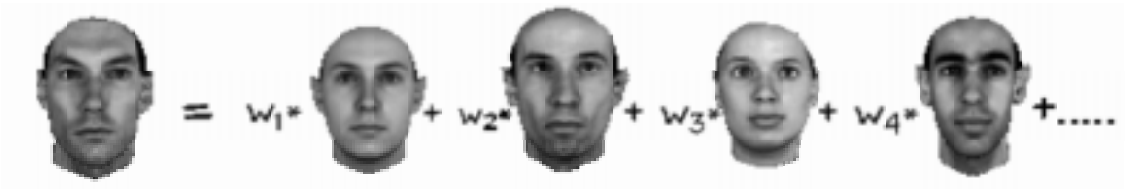


FIG. 2.6 – *Paramétrisation pour une synthèse ou une représentation photo-réaliste*

Pour construire une approximation globale du modèle, ils demandent à un opérateur de construire interactivement une première estimation en taille et position. Puis, leur méthode affine la correspondance en effectuant une descente de gradient, pour minimiser l'erreur observée.

2.3 Conclusion

Les méthodes se sont dirigées vers beaucoup plus d'automatisme, et capturent généralement la géométrie et la texture. De nombreuses techniques à base d'images ou utilisant des caméras ont été développées pour éviter des capteurs 3D spécialisés, dont l'offre s'est pourtant considérablement diversifiée, avec la commercialisation d'appareils portables ou de techniques concurrentes [INS, C3D], et de techniques plus rapides qui permettraient de capturer des expressions animées à la volée.

Dans presque tous les cas, pour créer des modèles 3D qui soient directement animables et pour régulariser les erreurs que pourraient introduire les capteurs, on cherche plutôt à modifier automatiquement un modèle générique qu'à créer directement un modèle exact mais incomplet qu'il faudrait retoucher. En conséquence, le volume des cheveux ou la position des oreilles ne sont pas toujours respectés, et **la ressemblance «statique» du clone final n'est plus forcément aussi parfaite** qu'en capture directe et complète.

Alors que les caméras, mais aussi les cartes 3D, sont devenues bien plus courantes dans les environnements informatiques, les clones restent encore rares. Il faut en effet remarquer qu'il n'y a pas encore de «toolkit» (logiciel clef-en-main, sans matériel supplémentaire) – en distribution libre ou commerciale – qui permettrait à chacun de créer facilement son clone, comme c'est le cas pour les photos panoramiques par exemple.

Chapitre 3

Animation des modèles 3D des visages

On a déjà souligné qu'il fallait animer les visages synthétiques, pour qu'ils paraissent vivants ou dans un souci de ressemblance. Lorsqu'il n'est plus statique, le visage 3D présentera des apparences multiples, qu'il faut pouvoir choisir, c'est-à-dire contrôler avec des paramètres.

La première étape étant de paramétrer ces visages 3D, il n'est donc pas inutile de savoir ce qui constitue une expression réelle, et comment elles sont générées ou perçues par notre cerveau.

3.1 Cause et analyse des expressions

Les expressions réelles se matérialisent à la surface du visage, où d'autres pourront les interpréter. Pour comprendre comment nos expressions réelles sont créées ou perçues, il faut avoir recours à d'autres domaines du savoir, comme l'histologie ou la psychologie.

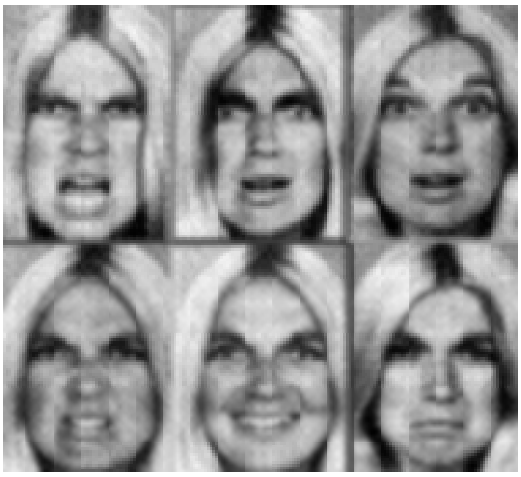


FIG. 3.1 – *Quelques expressions universelles d'après [EF]*

L'étude des tissus révèle un empilement de différentes couches, plus ou moins élastiques (mais pas de façon linéaire) et dont l'épaisseur varie selon l'endroit. On trouve aussi des muscles, reliés aux os du crâne ou à la mâchoire et à différents points du visage. Selon le mode de fonctionnement de ces muscles (sphincter ou linéaires par exemple), les contractions provoquent des mouvements différents, qui peuvent se propager à travers les tissus adipeux, l'épiderme et le derme jusqu'à apparaître visibles, mais transformés, à la surface du visage.

En tant que psychologues, étudiant la communication non-verbale, Ekman et Friesen [EF] se sont intéressés aux muscles faciaux, dans la mesure où ils ont un effet sur les expressions et donc jouent un rôle dans la communication interpersonnelle. Ils dressent un catalogue des mouvements faciaux du point de vue d'un observateur, en définissant des *action units (AU)* qui expriment le déplacement, causé par un muscle ou par un ensemble de muscles quand leur action n'est pas différentiable par la simple observation visuelle. Leur système, appelé FACS (*Facial Action Coding System*), explicite 46 AU en précisant comment les lire visuellement, le ou les muscles qui provoquent ce déplacement, et comment un utilisateur entraîné peut réussir à reproduire ce seul mouvement.

Cet «alphabet» des mouvements faciaux permet alors de lire et de représenter objectivement la façon dont tout visage s'anime dans le temps, par exemple pour générer ou répertorier le dictionnaire des expressions, dont au moins six seraient universellement reconnaissables, dans toutes les cultures.

Comme on va le voir dans les sections suivantes, ces deux considérations, biomécaniques et perceptives, ont aussi influencé le domaine de l'animation faciale.

3.2 L'évolution de l'animation faciale

Les méthodes classiques de l'animation du dessin animé sont basées sur la notion de «*keyframe*», des points clefs du mouvement ou de la déformation, sous forme de couples (date, positions), entre lesquels l'animation est suffisamment régulière pour pouvoir être générée de façon plus mécanique, par interpolation.

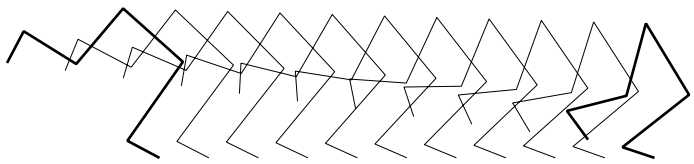


FIG. 3.2 – *Le keyframing : une interpolation temporelle entre des positions clefs*

Les premiers visages 3D, comptant très peu de polygones, allaient donc être animés en réutilisant ce paradigme : on pouvait produire une animation complète par l'interpolation des positions des quelques points à leurs extrêmes. Cette localisation, par exemple des coins de la bouche, pouvait être réalisée avec une tablette graphique et deux photos orthogonales.

Avec l'augmentation de complexité des modèles, pour éviter les erreurs et assurer un résultat de qualité constante en un temps raisonnable, il fallait faciliter et automatiser la tâche de déformation du modèle. Aussi sont apparues des méthodes hiérarchiques, avec des abstractions progressives regroupant les points en zones par exemple, et définissant des actions sémantiques, comme «sourire» ou «cligner de l'œil». On peut classifier ces approches suivant deux tendances, selon qu'elles s'intéressent directement aux effets (les points du modèle 3D doivent être déplacés), ou qu'elles simulent par la physique ses causes (par exemple avec des muscles virtuels).

Toutes ces techniques prennent donc le parti de créer l'animation uniquement au niveau du modèle de forme, c'est à dire sans même toucher à la texture. On verra pourtant à l'occasion du chapitre 5 que pour créer **des images de visages animés** d'autres techniques sont utilisables, dont certaines n'utilisent même pas de modèle 3D. Pour l'heure, ce sont seulement les approches «classiques» où la paramétrisation influence seulement les positions des points du modèle 3D qui vont être abordées.

3.3 Animation par paramétrisation directe

Si on pourra générer de nombreuses images par interpolation de plusieurs situations clefs, il n'en reste pas moins indispensable de spécifier la position de tous les points du modèle, ce qui est d'autant plus fastidieux que ce dernier est précis. Pourtant, nombre de ces points sont immobiles, ou bougent dans un mouvement d'ensemble de façon prévisible, parce qu'ils sont plus ou moins solidaires ou liés par une action qu'on pourrait décrire sémantiquement. C'est cette dépendance qui est à la source des modèles paramétrés: la donnée de quelques paramètres, avec une sémantique assez claire, permet de générer automatiquement les positions de tous les points du modèle. Ainsi, on pourra faire du *keyframing* directement sur les paramètres. C'est à cette attente que répondent les modèles paramétrés pour l'animation du visage.

3.3.1 Le modèle de Parke

Pour automatiser ses premiers résultats d'animation par *keyframes*, Parke a développé un modèle procédural qui contrôle avec quelques paramètres la construction d'un représentant: selon les zones du visage et le paramètre mis en jeu, des points seront crées ou subiront une interpolation, une rotation, une mise-à-l'échelle ou un déplacement. Ainsi, les yeux, les paupières ou la bouche pourront être animés en faisant varier un petit nombre de valeurs de contrôle. D'autres paramètres liés à des mesures du visage servent à déformer le modèle, par exemple pour simuler sa croissance ou le faire ressembler (dans sa forme seulement) à une personne donnée.

C'est ce modèle qui est représenté sur la figure 2.1, page 13, selon différents angles de vues et méthodes de rendu.

3.3.2 Le modèle CANDIDE

Plus fortement influencé par les *Action Units*, un autre modèle paramétré est très utilisé, c'est le modèle CANDIDE [Ryd87], développé à l'université de Linköping.

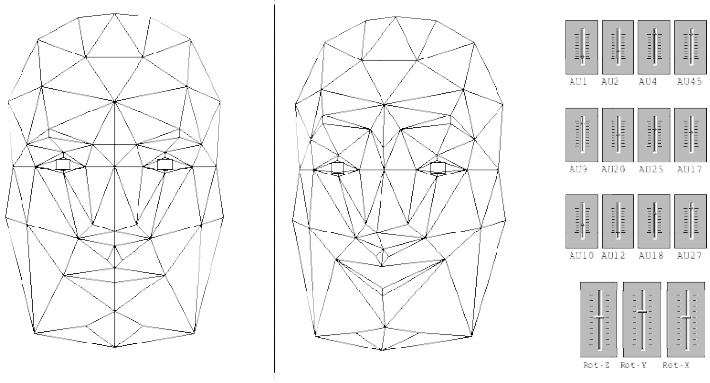


FIG. 3.3 – *Deux vues du modèle CANDIDE d'après [For97]*

Comme il est peu complexe, il est facilement animable, même sans machine puissante, par exemple en JAVA [For97]. Parce qu'il introduit un minimum de triangles et peu de paramètres, ce modèle reste encore très utilisé : tel quel lorsqu'on cherche à estimer le mouvement ou la position d'un visage dans une vidéo, ou raffiné par un maillage plus fin et moins anguleux, par exemple des B-splines triangulaires dans [TEGK97].

3.3.3 Le modèle de MPEG-4

Dans la vaste entreprise de normalisation connue sous le nom de MPEG-4 [MPE99], en plus de techniques de compression «classiques» des signaux audio et vidéo, de nombreuses approches de codage «hybride» ont été proposées [SNH97, SNH], qui permettront à tout décodeur compatible de synthétiser une image ou un son à partir de données moins volumineuses que le signal naturel compressé. C'est notamment le cas pour les images de visages, qu'il est prévu de pouvoir transmettre à l'aide de quelques paramètres qui contrôlent un modèle 3D.

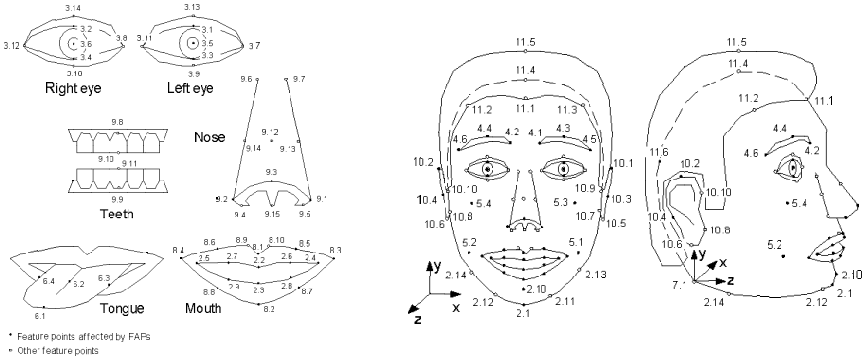


FIG. 3.4 – *Points caractéristiques et FAP pour un visage 3D selon MPEG-4*

Ici aussi, les paramètres sont découpés en deux classes : les FDP (*Facial Definition Parameters*) et les FAP (*Facial Animation Parameters*), qui vont permettre :

- d'encoder les mouvements globaux du visage dans la scène (déplacements et rotations),
- de coder des expressions faciales, comme avec les AU, mais avec bien plus de détails (par exemple, pour la langue).
- de synchroniser des lèvres avec la parole (naturelle ou synthétique),
- d'adapter le modèle 3D du décodeur à une morphologie faciale particulière (FDP).

La norme ne spécifie pas comment obtenir ou générer ces paramètres (quelle sorte de caméra les capture si l'on veut compresser une vidéo ?), mais précise comment ils transitent jusqu'au décodeur, en même temps que les autres informations nécessaires au décodeur (l'usage d'une texture par exemple, ou les canaux sonores), et définit leur sémantique pour l'utilisation.

À priori, ce standard permettrait des reconstructions plus précises qu'avec les précédents modèles, pour fournir des services présentant une interface humaine, ou pour compresser des flux vidéo. Aussi, et du fait des opportunités commerciales, nombre de projets se réorientent avec pour but la création de contenus qui soient compatibles avec le codage MPEG-4.

3.4 Animation par simulation de muscles

On sait reconnaître (par des algorithmes ou par la vision humaine) un sourire au mouvement caractéristique des extrémités de la bouche. Certaines expériences [Bas79] ont même montré que le cerveau savait décoder les expressions à partir du seul mouvement de quelques marqueurs (par exemple fluorescents) sur un visage invisible. Mais pour la synthèse, il faut savoir générer des mouvements d'ensemble : si l'animateur déplace le coin de la bouche, d'autres parties doivent suivre le mouvement.

Cette cohérence de la peau, et aussi son élasticité ou son volume par exemple, viennent de sa constitution sous forme de muscles et tissus connectés de façon plus ou moins élastique au squelette. Aussi, certaines modélisations se sont fortement inspirés de la réalité anatomique, ou ont cherché à reproduire sa mécanique, en introduisant des comportements physiques (cinématique et dynamique).

3.4.1 Les modèles «physiques»

Waters [Wat87], puis Terzopoulos et enfin Lee [LTW93, LTW95] ont contribué à développer des modèles successifs de plus en plus performants pour créer des animations faciales de haute qualité.

Avec leurs modèles, l'animation découle de la contraction de muscles synthétiques, hérités d'un modèle générique déformé. Dans les évolutions de leur travaux, le modèle générique est plus ou moins complexe, par le nombre de couches physiques qui sont simulées par réseaux de masses-ressorts interconnectés (figure 3.5), qui généreront la dynamique de la couche supérieure, celle

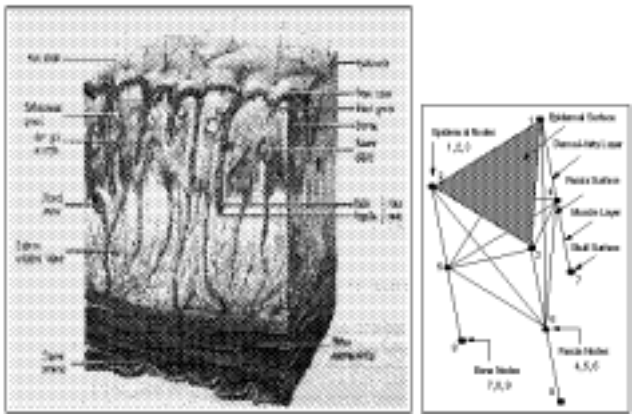


FIG. 3.5 – *Les tissus de la peau et une possible modélisation d'après [LTW93]*

qui est visible sur le modèle 3D affiché. Dans les versions les plus sophistiquées, des contraintes supplémentaires sont ajoutées pour que les tissus ne pénètrent pas la couche du crâne et que leur volume restitue un comportement d'incompressibilité. Dans tous les cas, le maillage du modèle de simulation est donc soumis à de nombreuses forces, dont la résultante, calculée selon les lois de la physique, sera de déformer la couche qui représente le visage d'une façon bien plus fine et naturelle qu'avec les modèles paramétriques, puisqu'une force musculaire va voir ses effets se propager «physiquement».

FIG. 3.6 – *Animation «physique» d'un modèle générique adapté [LTW95]*

3.4.2 Modèles avec pseudo-muscles

Plus vieux que les précédents, le modèle de Platt et Badler [PB81] ne comportait qu'une couche de points reliés élastiquement, celle qui était visible. Le concept de muscles, des ressorts liés à des points fixes représentant les os permettait d'activer le mouvement du maillage visible et de modéliser les expressions. Antérieur et plus simple que le modèle précédent, il peut générer des expressions moins réalistes et moins physiques, notamment pour des forces trop grandes.

D'autres formalismes expriment l'action visible des vrais muscles par un ensemble de pseudo-muscles. En nombre différent, et placés autrement que ceux des modèle FACS ou physiologiques, ces paramétrisations ont été utilisées par leurs auteurs pour divers clones parlants notamment, qui peuvent être pilotés en imbriquant des niveaux d'abstraction de plus en plus élevés. Ce sont les *Abstract Muscle Action procedures* [MTPT88] et les *Minimal Perceptible Actions* [KMMTT91].

3.4.3 Animation par influences de zones

Par rapport aux paramétrisations directes, l'approche de simulation de l'ensemble muscles + tissus conduit aux déplacements simultanés de nombreux points, restituant l'impression de zones d'influence pour chaque muscle.

Certaines méthodes proposent d'appliquer directement diverses déformations sur des zones de la surface, pour approcher à moindre coût les résultats de la simulation. Selon la technique de déformation utilisée, on va, par le réglage d'une pseudo-force, entraîner ou repousser plusieurs points de la surface par le déplacement d'un point de contrôle qui influe sur ses voisins proches (comme sur la figure 3.8) ou par la déformation d'un volume englobant (figure 3.7).

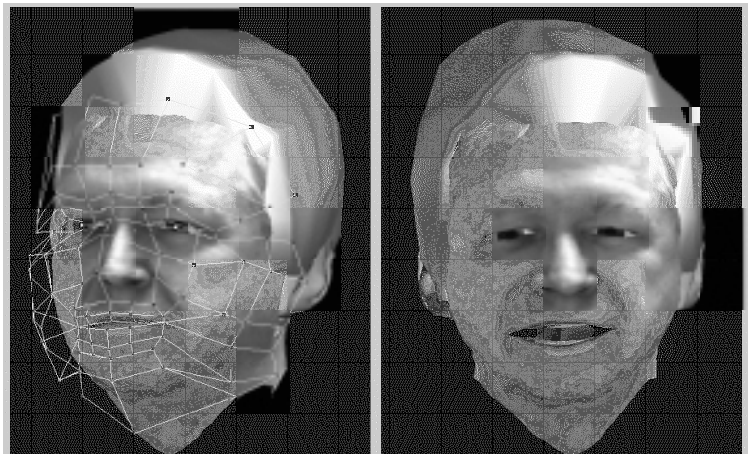


FIG. 3.7 – *Exemple de déformation par des volumes de Bézier [TH98]*

Si l'on peut obtenir de bons résultats avec cette classe de méthodes, il faut souligner la difficulté du placement et de l'utilisation des éléments de contrôle. Lorsque l'amplitude des déformations n'est pas limitée (ou contrebalancée physiquement), on peut aussi obtenir des résultats excessifs (intersection de surfaces) ou exagérés et non naturels (volumes ou tailles visiblement non-constants). Cela peut être utile pour des effets spéciaux, des caricatures ou les personnages stylisés de dessins-animés, mais n'est pas souhaitable pour des clones ressemblants.

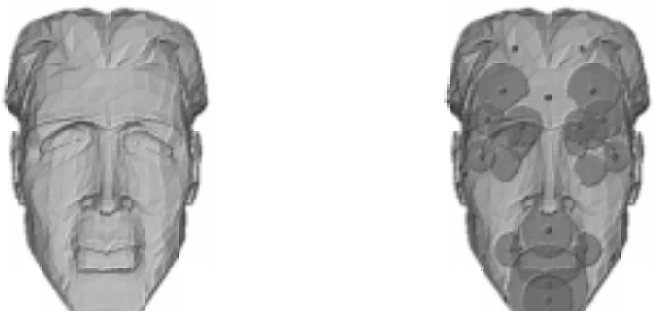


FIG. 3.8 – *Quelques points de contrôle et leurs zones d'influence radiale [Sol98]*

3.4.4 Les automates parlants

Un autre utilisation de l'animation faciale est liée à la délivrance d'un message sonore, sous forme d'automate parlant. Face à un client qui utiliserait un écran tactile, un clavier ou un micro couplé à une reconnaissance vocale, un distributeur de billets pourrait proposer une interface simili-humaine, avec un visage qui articulerait les questions ou les réponses. Dans un tel scénario applicatif, les réponses ont été pré-enregistrées et la ressemblance du modèle ou de la voix sont jugées moins importantes que leur esthétique ; la communication, si elle semble interactive, est alors plus convenue que vraiment bidirectionnelle.

De très nombreuses implémentations existent, qui génèrent leurs visèmes (analogues visuels des phonèmes [BLM92]) à partir d'une piste sonore renseignée[1] [ECG97] parfois synthétisée d'après un fichier texte [WL94] («décompression» *text-to-speech*, comme dans MPEG-4) et parfois en enrichissant l'avatar d'expressions et attitudes faciales créées par des agents réagissant à leur audience (les passants ou le client visible par la caméra d'une borne interactive).

Dans le cas particulier où l'on s'adresse à des personnes sourdes, par exemple pour l'enseignement, la visibilité et le déplacement des dents et de la langue doivent être traités avec une plus grande rigueur.

3.5 Conclusion

Ainsi, le talent et les méthodes des animateurs sont et restent très riches en terme de degrés de libertés donc d'expressivité, par exemple pour caricaturer, inventer ou mélanger diverses attitudes faciales. Indiscutablement, on sait créer et animer de tels modèles à un niveau d'anthropomorphisme très élevé, avec un automatisme de plus en plus développé, par exemple pour la synchronisation des lèvres avec la parole.

1. pour synchroniser les transitions, le plus souvent en analysant des tranches temporelles du signal. Ces *buffers* doivent être suffisamment longs pour permettre une «prédiction» précise, mais introduisent donc des décalages avec l'image, qu'on compense en retardant le son. En tenant compte de ce délai, l'utilisation en mode «téléphone enrichi» est compromise.

FIG. 3.9 – *August, un exemple d'agent interactif parlant [KTH]*

Mais dans le cadre d'une communication par clones interposés, il ne s'agit plus seulement de synthèse ou d'animation, mais avant tout d'un problème de **reproduction** d'une certaine réalité : l'image d'une personne et son message (audio et facial dans notre cas) qu'on doit représenter à distance, en les compressant. Dans une approche évolutive, il est logique d'essayer d'ajuster tous les paramètres qui contrôlent les modèles de visages précédents pour ressembler le plus possible à l'image ou aux expressions du locuteur. Croisant souvent la piste de la compression orientée objet, c'est cette **reproduction** des expressions des télécommunicants, sous forme d'analyse de leur image ou de leurs expressions par la synthèse d'un modèle 3D asservi que le chapitre suivant va analyser.

Chapitre 4

Analyse/synthèse de clones 3D

Le chapitre précédent a montré qu'il était possible de synthétiser des modèles 3D animés de visages qui semblent assez naturels et vivants, par exemple pour l'industrie du cinéma. On a vu qu'on pouvait les créer pour qu'ils ressemblent (forme et texture) à une personne réelle, généralement en modifiant un modèle générique.

Cependant, dans l'optique d'une conférence à distance, on souhaite qu'ils adoptent plus que l'apparence de leur original. Le message d'un participant mérite d'être reproduit le plus fidèlement possible, dans ses intonations sonores comme dans ses mimiques faciales. C'est ce que ferait un visiophone et c'est ce que l'on souhaiterait d'un codage par la 3D (mais sans bloquer l'angle de restitution à celui de la prise de vue).

Ce chapitre va donc s'intéresser aux techniques qui permettent de «piloter» un clone pour qu'il articule et affiche les expressions du locuteur qu'il représente. Cependant, parce que la conférence 3D à distance n'est pas encore une réalité, on va aussi et surtout analyser le cas d'applications connexes pour lesquelles les contraintes du temps-réel ou du visio-réalisme sont abandonnées.

4.1 Les applications connexes

Quelles sont les possibles utilisations du pilotage d'un clone par une personne réelle?

- pour la génération d'animations : plutôt que d'animer des modèles faciaux en manipulant des points du modèle ou des abstractions de plus haut niveau, il peut être utile d'utiliser un acteur, professionnel ou non, qui pilotera directement un modèle (qui ne lui ressemble pas, voire n'est pas humain). Ainsi, des animaux en 3D jouent le rôle de présentateurs d'émissions, contrôlés par un nouveau type de marionnettistes.
- pour des interfaces homme-machines humanisées : dans le cadre de la norme MPEG-4, pour des CD éducatifs ou ludiques et des bornes commerciales qui souhaiteraient présenter un message pré-enregistré mais en 3D.
- pour la compression bas-débit de flux vidéos, notamment dans le cadre de MPEG-4/SNHC : si l'on enlève la contrainte du temps réel, il est déjà possible d'encoder des scènes du style

visage/buste dans le strict respect du standard.

4.2 Les codages 3D basé objet et basé modèle

C'est l'idée que la 3D devrait pouvoir approcher n'importe quel objet, donc leur image, qui est à la base des codages basé objet et basé modèle. Quand on l'applique au cas de la visioconférence, on sait que c'est un visage (ou un buste) qui devra être encodé. Mais selon le scénario, on disposera d'un modèle exact de la personne qui parle dans la scène, ou seulement d'un objet générique qu'il faudra adapter et raffiner.

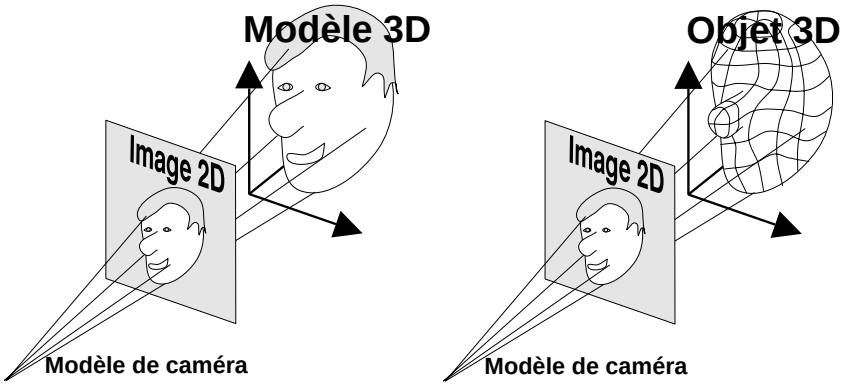


FIG. 4.1 – *Codages basé modèle et basé objet*

Dans ces deux cas, le codage basé modèle et le codage basé objet, un décodeur pourrait reconstruire le flux vidéo initial, son approximation ou une vue sous un autre angle en effectuant le rendu du modèle 3D, mis-à-jour dans ses déplacements et transformations selon ceux repérés dans l'image par le codeur (qui peut choisir de ne pas tout répercuter).

En plus de l'analyse des paramètres d'animation (par exemple les FAP de MPEG-4) pendant toute la durée de la communication (ou du message vidéo s'il est pré-enregistré), il y a aussi une phase d'initialisation de l'analyse, qui peut soit :

- avoir lieu sur les toutes premières images (de la conférence, ou juste avant d'émettre) avec ou sans la participation volontaire du locuteur (rester immobile, cligner des yeux ou se placer de coté...) et ne plus être remise en cause,

- se poursuivre et s'affiner pendant toute la prestation, de sorte que l'objet qui est initialement incomplet ou imprécis, devienne un modèle de plus en plus précis.

Si l'on sait réaliser un tel codage avec peu de paramètres, mais assez pour avoir une bonne qualité, les possibilités de compression à bas ou très bas débits sont réelles. Selon les paramètres et le modèle interne, connu du codeur comme du décodeur, l'espace de reconstruction sera un maillage 2D texturé, un modèle 3D ressemblant ou ne reproduira sur des avatars (qui peuvent être ressemblants) que des unités sémantiques, comme les expressions.

Dans le cas des visages 3D, MPEG-4 propose un tel cadre mais pour la décompression (FDP et FAP), c'est-à-dire sans spécifier comment on peut capturer ces paramètres. Il faut donc trouver une technique de capture qui garantisse des paramètres tels que la synthèse reconstruite approchera suffisamment l'image reconstruite et/ou le message facial initial.

4.2.1 Le paradigme de l'analyse-synthèse

Puisque l'on sait déjà synthétiser des visages 3D synthétiques, la tentation est grande de vouloir s'en servir pour encoder des images réelles ou animer des avatars en temps réel. Ainsi, une fois que le modèle est connu du décodeur (parce qu'il a été transmis, complètement ou sous forme de déformation d'un modèle générique), il suffirait de transmettre les paramètres de déplacement ou de déformation du modèle pour décrire les modifications de l'image en termes de transformations de la scène virtuelle. Mais pour cela, il faut être capable de trouver l'évolution temporelle de ces paramètres, en ne voyant que leur influence sur les images. C'est ce problème, inverse de celui de la synthèse, qu'on appelle analyse, et qui pose le plus problème pour la matérialisation de la conférence 3D.

La difficulté de l'analyse

En pratique, pour un visage, il faut estimer sa position et son orientation, et trouver les positions ou formes des éléments caractéristiques du visage (bouche, sourcils...) qui devront être mimés par le modèle 3D et sont non rigides. Selon les techniques, le suivi peut être fait en 3D ou en 2D, et l'estimation du mouvement global précède ou suit celle des caractéristiques. Selon la nature du modèle interne de représentation, on a un codage modèle, un codage objet ou un codage sémantique si on ne s'intéresse qu'aux expressions par exemple.

Mais qu'est ce qui rend le suivi des éléments du visages si difficile (pour une machine) ? En vision pour la robotique, on peut souvent chercher des points de contraste et des arrêtes rectilignes, par exemple dans des environnements statiques qui comportent des murs et du mobilier. Dans le cas d'un visage, de telles primitives n'apparaîtront pas. Peu de détails sur le visage présentent les bonnes qualités [ST93] pour matérialiser sûrement les mouvements locaux et globaux. Il faudrait des points qui :

- présentent une texture riche, différentiable de leur environnement, avec une orientation non ambiguë,
- correspondent à des éléments réels à la surface du visage : Il ne faudrait pas suivre des «fantômes» comme des ombres ou des limbes (le contour apparent qui se crée aux points où la vue est rasante), qui ne sont pas des éléments physiques du visage.
- soient répartis sur l'ensemble du visage, en restant visibles le plus souvent possible, pour nous renseigner sur les mouvements globaux et locaux.

De tels points sont peu nombreux sur un visage sans marqueurs. Classiquement, les techniques de suivi peuvent s'appliquer au visage au niveau :

- des narines, si la caméra est placée sous le moniteur par exemple,

- du centre des yeux, parce qu'ils sont très texturés. Il faut cependant prévoir les clignements des yeux. Des lunettes, capables de générer de soudains éclats lumineux, compliquent grandement la tâche,
- des extrémités (externes voire internes) des sourcils,
- des coins de la bouche, parce que ces points offrent un aspect pointu caractéristique. Lorsqu'on parle et que la bouche s'arrondit, leur suivi devient souvent plus difficile, et il est plus robuste de suivre toute la bouche avec un modèle global.

Seules les deux premières entités sont rigides. En comparaison d'une grille de calibration pour stéréovision par exemple, on ne dispose donc en général pas de beaucoup de points réputés sûrs pour estimer le mouvement global et le séparer des mouvements locaux des expressions.

Dans beaucoup d'approches d'analyse, on trouvera la plupart de ces différentes phases :

- à l'initialisation, localiser le ou les visages présents dans une image,
- à l'initialisation, établir les correspondances avec le modèle interne (en localisant les points caractéristiques, les zones significatives du suivi ou en affinant le placement d'un modèle 3D),
- pendant la communication, faire évoluer les paramètres visibles du suivi pour s'adapter aux mouvements locaux et/ou globaux dans l'image,
- pendant la communication, et s'ils sont différents, effectuer la traduction, précisément ou par des règles empiriques, des paramètres de suivi 2D vers les paramètres internes (3D ou 2D) du modèle qui sert à faire le rendu et constitue la représentation.

De nombreux travaux proposent une solution pour l'une ou l'autre de ces étapes, mais rares sont les approches qui proposent une chaîne complète, qui permette de juger de l'applicabilité, en matière de robustesse notamment. Comme ces phases peuvent – voire devraient – être traitées par des méthodes différentes, c'est une difficulté supplémentaire que de rassembler les diverses compétences pour tous ces domaines, très actifs actuellement.

Pour illustrer toutes les classifications qui viennent d'être énoncées, on va tout d'abord s'intéresser aux systèmes d'analyse pour l'animation qui permettent de piloter des avatars (ou des clones ressemblants) en temps-réel, par exemple pour réaliser des présentateurs virtuels dans un cadre interactif.

4.2.2 L'analyse temps-réel pour l'animation

Ce sont probablement les systèmes de capture pour présentateurs virtuels ou pour la réalité virtuelle qui sont le plus répandus. À l'aide de capteurs (magnétiques ou optiques par exemple) plus ou moins pénibles à porter, on peut mesurer le déplacement de divers points du corps du manipulateur, par exemple pour calculer la flexion des articulations. Pour le visage, un harnachement spécifique ou quelques marqueurs optiques sont suffisants pour animer plus ou

moins grossièrement des avatars selon quelques degrés de liberté : ouverture de la bouche et mouvements des sourcils le plus souvent.

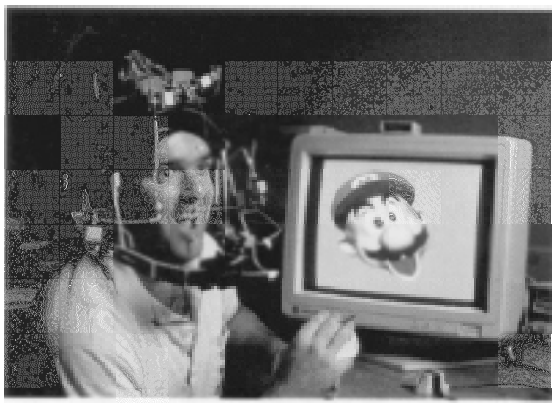


FIG. 4.2 – *Un des premiers systèmes commerciaux d'animation [VAS]*

C'est le rôle de l'acteur que d'amplifier ses mimiques faciales pour que les capteurs, dont les mesures doivent être filtrées pour diminuer les incertitudes, génèrent une reconstruction qui soit satisfaisante à l'écran. Parfois, les données capturées peuvent être éditées manuellement et aider un animateur à construire une séquence de meilleure qualité technique ou artistique.

Comme ces systèmes commerciaux très invasifs ne se placent manifestement pas dans le cadre d'une communication naturelle, on va donc les délaisser pour examiner plutôt les recherches et expériences qui correspondent à la problématique posée. Dans un premier temps, on va illustrer les approches qui procèdent par analyse 2D sur l'image, toujours pour reproduire les expressions.

Quelques exemples de systèmes d'analyse 2D/synthèse

Vu l'absence de marqueurs naturels, certains auteurs demandent à leurs cobayes de porter quelques marqueurs ou du maquillage, pour obtenir une information plus dense sur les mouvements locaux et globaux, ou tout simplement rendre leur capture plus robuste.

Par exemple, Williams [Wil90] a proposé d'utiliser une vingtaine de marqueurs fluorescents, ce qui permet de les suivre facilement avec un éclairage adapté, tel que le reste du visage soit quasiment invisible. Chacun des spots contrôle une région à la surface du clone, sous forme d'une fonction radiale qui lors de ses déplacements influencera les points proches en fonction décroissante de la distance (*warp kernel*), avec une portée qui dépend de l'étendue de la zone faciale d'ancrage.

À l'INA, un système d'analyse/synthèse [SVG95] reproduit les mouvements du menton, de la bouche (soulignée par un rouge-à-lèvres dans certains cas pour augmenter la robustesse) et des sourcils, ainsi que les orientations de la tête du locuteur, placé face à une caméra non calibrée. Un

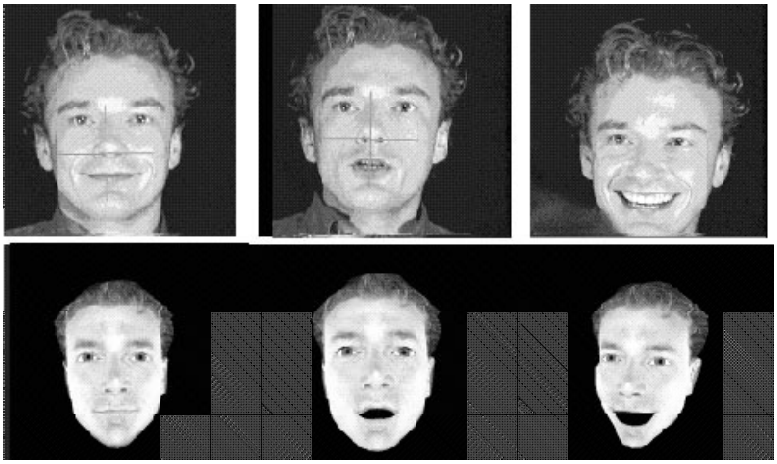


FIG. 4.3 – *Le système du projet Télévirtualité [INA]*

modèle du locuteur a été construit à partir de deux vues, et va servir de base pour transmettre une téléprésence, à l'aide d'une dizaine de paramètres, comme la taille de la bouche ou la distance œil/sourcil. Ces valeurs sont estimées, après le suivi 2D par segmentation et *snakes*, d'après des règles empiriques, qui conduisent le clone à ouvrir la bouche ou bouger les sourcils de concert avec son modèle (figure 4.3).

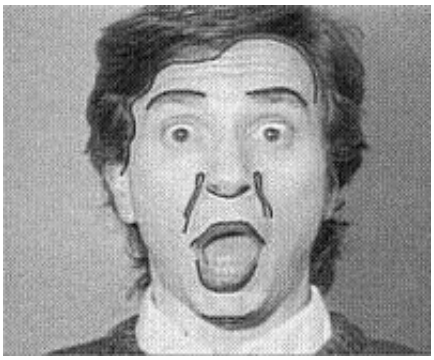


FIG. 4.4 – *Marquage facial pour faciliter le suivi de l'animation [TW93]*

Terzopoulos et Waters ont proposé de piloter un de leurs modèles à base de muscles à partir d'images non calibrées. Ils utilisent un maquillage du visage (comme sur la figure 4.4), notamment entre le creux des joues et l'aile du nez, pour matérialiser des déformations qu'ils supposent dues à 6 muscles indépendants. Chaque ligne d'intérêt est suivie (avec des *snakes*), débouchant sur 11 points de confiance (milieux et extrêmes des sourcils, coins de la bouche, bas du menton et position des joues).

Des règles arbitraires déduisent de ces mouvements, considérés à tour de rôle, des contractions quantitatives pour les seuls six muscles qui sont modélisés.

FIG. 4.5 – *Estimation des contractions des muscles selon un flot optique par régions [Ess95]*

Avec la volonté de créer un système automatique et non ambigu de mesure de l'évolution faciale (et des expressions), Essa [Ess95] propose un nouveau codage baptisé *FACS+* puisqu'il étend le modèle d'Ekman. Le principe consiste à s'intéresser au flot optique mesuré entre deux images successives, mais va quantifier les mouvements par régions. Chacune de ses régions est censée être influencée de façon représentative par un muscle ou ensemble de muscles (Cf. figure 4.5). Ces mesures servent à l'auteur à quantifier le déroulement temporel de certaines expressions et à reconnaître les expressions classiques. Il propose aussi de les reproduire sur un modèle 3D avec muscles comparable à ceux de Waters. En terme de synthèse, les résultats obtenus souffrent, comme pour l'approche précédente, de ce que les actions des muscles, principalement au niveau de la bouche, ne sont en fait pas si facilement différentiables.

Pour différents projets, dont VL-Net [CPN+97], les auteurs mesurent l'évolution de différentes caractéristiques du visage en mouvement après une initialisation manuelle d'un *soft mask*. Ce dernier a aussi servi à enregistrer des modèles de couleur de la peau et des cheveux, qui seront utilisés lors du suivi, avec presque autant de stratégies de recherche qu'il y a de caractéristiques suivies (position horizontale des iris, ouverture des yeux, distance entre les sourcils, Cf. figure 4.6). Sans être forcément la plus robuste, c'est de loin, parmi toutes les approches de ce début de chapitre, celle qui propose le plus de détails pour l'animation reconstruite.

Conclusion des approches précédentes

À l'aide de modèles de couleur, de *snakes* ou de flot optique, il est possible de suivre en temps réel quelques caractéristiques 2D de la surface du visage. Une phase de transformations empiriques doit par contre suivre pour les traduire en des caractéristiques à peu près équivalentes

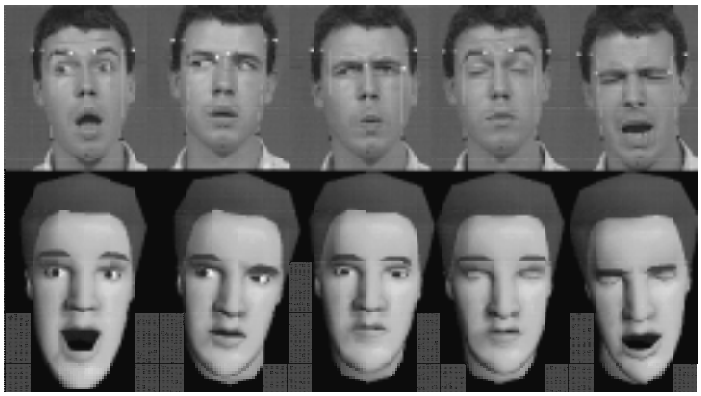


FIG. 4.6 – *Commande d'un avatar dans VL-Net [MTKP95, CPN+97]*

sur le modèle 3D (qui peut être aussi bien un avatar qu'un clone ressemblant). Par principe, la «reconstruction» n'a pas vocation à être fidèle à un niveau vidéo-réaliste, mais plutôt aux grandes expressions classiques, et à une certaine synchronisation des lèvres et de la parole. À raison, certains auteurs parlent de codage basé sémantique, puisque le petit nombre de paramètres qui sont reçus du coté de la reconstruction sont généralement des FAP ou des AU [EG97a].

Lorsque le modèle de reconstruction n'est pas photo-réaliste, cela n'aurait pas grand sens d'estimer la différence pixel-à-pixel entre l'image originale et l'image reconstruite sous le même angle. Par contre, cette mesure est utilisable voir exploitée directement pour les techniques qui vont maintenant être évoquées, et qui peuvent à priori s'avérer vidéo-réalistes.

4.2.3 Le paradigme de l'analyse par synthèse

Lorsque l'on possède un modèle 3D censé représenter la scène de façon suffisament fidèle, il est possible de l'utiliser pour générer une image à partir des paramètres courants, et faire évoluer ces paramètres pour minimiser une certaine erreur avec l'image réelle. Comme dans un filtre de Kalmann, on peut espérer prédire plus précisément ce qui est observé dans les prochaines images réelles ou ce qui se passe dans l'image courante si l'on peut utiliser un modèle adapté.

En cherchant à optimiser les paramètres déjà connus (qui peuvent avoir été spécifiés manuellement pour l'initialisation), cette approche peut conduire à rendre le suivi plus robuste, si le modèle est suffisamment expressif pour représenter la scène et qu'on sait corriger ou faire abstraction des différences avec l'image réelle, comme l'arrière plan ou l'éclairage.

Par contre, du fait de la taille de l'espace des paramètres et selon la méthode adoptée pour trouver la paramétrisation, les temps de calculs pour chaque image peuvent être éloignés du temps réel de plusieurs ordres de grandeur. Lorsque les paramètres sont utilisés directement, il faut avoir recours à des techniques de minimisation pour trouver ceux qui induiront les transformations voulues dans l'écran. Les méthodes les plus efficaces cherchent à **créer des contraintes linéaires, sur un très grand nombre de points de la scène,** pour ensuite estimer aux moindres carrés les nouvelles valeurs des paramètres.

En 1993, Koch [Koc93] propose un large panorama des problèmes et des principes qui seraient applicables pour le visiophone, avec une boucle de rétroaction pour coder des scènes de bustes : chaque rendu du modèle 3D est comparé à l'image réelle et conduit à la modification des paramètres de ce rendu. Lorsque la distance est acceptable, on a obtenu un jeu de paramètres qui permet d'approximer l'image ou de la coder si on la complémente par une image de résidu. Il propose de détecter le buste de la personne par différentes techniques de segmentation (soustraire le fond, détecter de zones en mouvement cohérent, estimer la profondeur avec plusieurs caméras), puis d'y adapter un modèle de buste générique, et d'extraire la texture initiale.

En concluant sur la difficulté d'estimer indépendamment les mouvements et les déformations (puisque tous deux génèrent des mouvements dans l'image), quand on n'a qu'un modèle générique des objets ou personnes observés et qu'ils sont non rigides, cet article, sans être le premier, a placé le cadre formel et la problématique de nombreux travaux même très actuels.

Intérêts par l'exemple de l'analyse par synthèse

Si le modèle est représentatif, on peut générer des hypothèses et des éléments qui aideront la recherche et le suivi robustes du modèle. On peut par exemple prévoir les occlusions des éléments mobiles, l'apparence de fenêtres de texture autour des éléments caractéristiques du visage (pour faire un suivi par une méthode de corrélation du style *pattern matching* [LCH97]) ou un masque pour décorreler l'influence du fond pour les zones (comme le nez) qui s'y détachent partiellement [VDD98].

FIG. 4.7 – *Un exemple de suivi d'un visage rigide par analyse/synthèse [CLH97]*

On peut aussi s'en servir pour prédire le mouvement 2D apparent (en linéarisant les rotations par de petits déplacements tangents comme dans [EG98]) et le mettre en correspondance avec un flot optique.

FIG. 4.8 – *Reconstructions par un modèle personnel animé [EG98]*

4.2.4 Les perspectives d'utilisation de codages basés objet ou modèle

Plusieurs méthodes résolvent le suivi (approché 2D [Bir97], approché 3D [MPB99] ou par recalage précis) lorsque le visage est supposé rigide (la personne vue ne parle pas). Lorsque la mâchoire est mobile, ces méthodes vont à priori induire dans le modèle reconstruit des déformations globales, mais pas toujours suivant la direction voulue, de sorte que le paramètrage (u,v) de la surface ne sera même plus en correspondance rigide avec les yeux. Ces techniques doivent donc plutôt être réservées à l'estimation de certains paramètres et à la génération d'hypothèses, mais ne peuvent pas réaliser une chaîne complète d'analyse.

Un cadre formel très générique d'adaptation hiérarchique pour tous les codages objets est proposé par Reinders [Rei95] : il permet de générer des hypothèses de niveau de plus en plus abstrait, en liant par exemple avec des contraintes géométriques les positions relatives possibles (sous forme de modèle élastique à l'écran, et de graphe structurel pour les connaissances à priori).

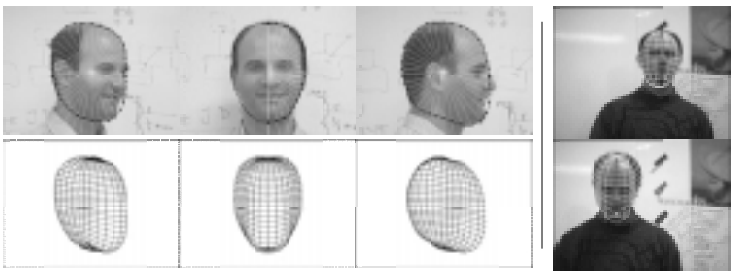


FIG. 4.9 – *Un modèle générique coûteux mais précis pour le suivi des visages [MPB99]*

De nombreuses techniques proposent donc des briques d'une solution complète de reconstruction : détection du visage, estimation de son orientation, adaptation/création d'un modèle à ses caractéristiques, suivi des caractéristiques ou flot optique hiérarchique, pour un suivi rigide ou seulement pour les composantes. Rares encore sont celles qui réalisent complètement un suivi non-rigide d'un visage, et une telle réalisation en temps réel et sans qu'un modèle personnel doive être disponible avant la conférence n'existe pas encore à notre connaissance.

Ici plus qu'ailleurs, les problèmes de l'analyse restent à résoudre dans des cas plus généraux, et en temps-réel si l'on souhaite pouvoir les utiliser non plus pour encoder des vidéos pré-enregistrées, mais les interventions en temps-réel de personnes distantes.

4.3 Conclusion

C'est incontestablement une tâche très délicate que d'animer les modèles 3D pour créer des clones dynamiques qui soient vidéo-réalistes. Le nombre des paramètres utilisés en synthèse pure doit souvent être revu à la baisse lorsqu'on est asservi à une image de référence, ce qui pénalise la richesse et la qualité de la reconstruction dynamique et la rend assez peu fidèle.

Dans le cadre du temps réel, plusieurs prototypes, parfois assez anciens, capturent et peuvent transmettre les expressions franches : après détection d'un sourire appuyé, un large sourire sera

synthétisé. Mais généralement la fidélité n'est pas parfaite : si l'on parle du bout des lèvres ou que l'expression est ambiguë (grimace, demi-sourire...) **on n'est pas certain de ce qui sera transmis et vu**. Sans un rendu local de contrôle, sorte de miroir imparfait, on ne peut pas se rendre compte de ce que les autres verront ni de quand il y a «trahison». Il n'y a en fait communication que d'une certaine abstraction d'expression (par exemple les 6 expressions universelles), du fait de l'interprétation qui s'est insérée au moment de l'analyse/synthèse, à cause du faible nombre des paramètres estimés (donc génératifs) ou de leur précision relative.

Avec les méthodes plus coûteuses qui font du recalage de modèle 3D et/ou de l'analyse par synthèse, sans la contrainte du temps réel, on va trouver de plus en plus de systèmes pour faire du codage basé objet 3D ou compresser un flux multimédia. Les plus complexes des algorithmes d'analyse déjà existants fournissent des résultats prometteurs : avec des temps de calcul de l'ordre de 10 secondes par image sur des stations de travail, on peut ainsi compresser des vidéos réelles, avec une qualité et des taux de compression correspondant à un haut niveau d'exigence. L'évolution très rapide de ce domaine est incontestablement liée à MPEG-4. En offrant une plate-forme commune et ambitieuse, ce standard stimule la recherche, la concurrence et la comparaison de résultats. En spécifiant le codage, le transport et le décompresseur, il n'a pas préjugé du rythme des progrès qualitatifs qui seront accomplis par les compresseurs et a su être suffisamment ouvert pour que l'incorporation de nombreux travaux courants ou futurs s'avère possible. En plus de se révéler payante, en terme de résultats sur le marché du *broadcast* digital avec de nouveaux services et une nouvelle génération de programmes, cette norme verra probablement à terme l'arrivée de codeurs temps-réels pour la compression vidéo-réaliste d'interlocuteurs 3D distants.

Il n'empêche qu'il reste délicat de rebondir maintenant dans le cadre de la conférence vidéo-réaliste à distance sans délai, d'un point de vue pratique voire théorique avec les approches actuelles des codeurs. L'analyse/synthèse 3D (pour les visages, les mains, ou des corps complets) reste un domaine de recherche récent, qui a bifurqué depuis plusieurs spécialités, et n'est actuellement pas caractérisé par la simplicité de ses solutions, ni par son vidéo-réalisme.

À défaut de visages vidéo-réalistes pilotés en temps-réel sous la forme d'une animation 3D classique par déformation de maillage, le chapitre qui suit va donc se tourner vers d'autres techniques, des techniques hybrides qui se proposent d'imiter la qualité de la vidéo.

Chapitre 5

Transformation et production de visages vidéo-réalistes

Plusieurs travaux très différents existent, qui codent ou représentent **les images des visages** de façon vidéo-réaliste, c'est à dire cherchant à leurrer un spectateur en approchant la qualité de la vidéo animée.

Certaines méthodes photo-réalistes de déformation d'images, pour le *Morphing*, la caricature ou l'animation [OTO+87, Yau88, BN92, LW94b], sont manifestement du nombre dès lors qu'on les utilise pour produire un flux continu, mais demandent une intervention humaine, avec une technicité ou un talent artistique. Ici, **on va se restreindre aux techniques automatiques** qui proposent de transformer des vues réelles (par exemple un visage vu par une caméra) pour synthétiser des images sous un angle de vue différent. On va bien sur chercher des méthodes qui résolvent aussi le problème dual de l'analyse, et sont donc à priori des candidats possibles pour réaliser le type de communication convoité.

Qu'elles ajoutent ou non une information 3D, mesurée où connue à priori, toutes les méthodes que l'on va retenir dépassent les limitations de la vidéo pure, puisqu'elles donnent l'illusion au spectateur qu'on a changé l'angle de vue.

Quoique différentes, elles forment une classe de techniques «hybrides» en ce sens qu'elles combinent généralement des informations de provenances diverses pour reconstruire des images avec plus de degrés de liberté, quelque part entre le 2D et la 3D.

Ce chapitre regroupe ici les expériences les plus représentatives en les classant selon leur méthodologie, analysant leurs points forts et leurs faiblesses.

5.1 Interpolations et extrapolations dans des espaces de petite dimension

Par un fondu enchaîné entre deux photos proches, on peut créer une courte séquence vidéo de bonne qualité. En généralisant cette interpolation pour un plus grand nombre d'images de référence, on opère sur des espaces linéaires, qui peuvent être de très grande dimension (autant que de pixels au maximum).

Avec des outils supplémentaires, comme l'analyse en composantes principales (PCA) ou en composantes indépendantes, on affine et rationalise leur utilisation en se restreignant à des sous-espaces de plus petite taille. On peut facilement faire des projections pour trouver les paramètres qui représentent ou approximent un objet, c'est à dire faire de la reconnaissance [TP94]. Toutes ces méthodes ont donc été largement étudiées et utilisées, par exemple pour l'identification des visages, ou le codage très bas débit d'images de visages vus de face, notamment pour les visages parlants [EP98] ou des concepts de *Video-mail*.

FIG. 5.1 – *Un exemple d'Eigenfaces [Rom]*

D'autres approches de combinaison non-linéaire des images ont aussi été proposées, avec par exemple des réseaux de neurones comme estimateurs pour retrouver les paramètres nécessaires à encoder une image donnée [BSP93, EP96].

5.1.1 Interpolation de modèles personnels

Plutôt que sur des images, on peut aussi travailler sur des modèles 3D (avec texture, pour être photo puis vidéo-réaliste). L'interpolation peut alors être réalisée entre les points du modèle (trivialement quand ils sont en correspondance par construction, ou par *Morphing* 3D sinon). C'est la méthode utilisée dans [GGW+98], à partir de 6 à 8 clones statiques d'une personne donnée pour créer diverses expressions (en 3D). Pour augmenter les combinaisons possibles, une paramètrisation par régions indépendantes est utilisée, comme sur la figure 5.2.

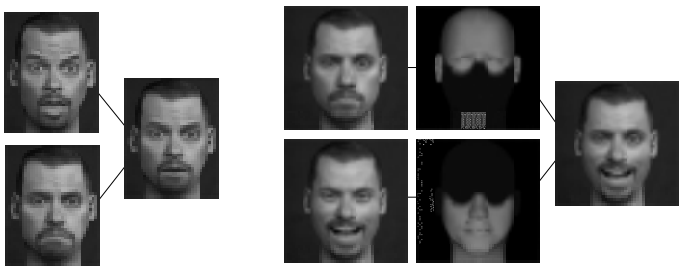


FIG. 5.2 – *Synthèse par combinaison de modèles 3D statiques d'après [GGW+98]*

L'analyse ne peut bien sûr pas être faite par une méthode linéaire. À défaut, les auteurs proposent une minimisation [PSS99] qui, à partir d'une initialisation manuelle arrive à des résultats convaincants quand l'image originale n'est pas trop éloignée du sous-espace codé, comme sur la figure 5.3.

FIG. 5.3 – *Exemples de reconstructions par interpolations de modèles d'après [PSS99]*

C'est une question ouverte que d'évaluer les attitudes (en un nombre minimal) qu'une personne donnée devrait enregistrer (en 3D) pour pouvoir utiliser une telle technique, dans un cadre d'analyse/synthèse fidèle (à défaut de téléconférence, puisqu'on est loin du temps réel et que des initialisations manuelles sont nécessaires).

5.1.2 Extrapolation à partir de modèles génériques

On a déjà évoqué les résultats de [BV99], où est créé un espace générique de modèles morphables (pour la face seulement) qui peut fournir (après estimation non-linéaire) un représentant pour toute image donnée en entrée. Il n'est pas très clair si l'utilisation de cette méthode avec un flux continu d'images en entrée donnerait en sortie (mais pas en temps-réel...) un flux de modèles 3D texturés qui soit suffisamment continu pour être utilisable. Pour cela, leur base de donnée devrait contenir un large ensemble d'expressions et de personnes. Le risque est néanmoins grand d'obtenir une image ou un modèle reconstruit dont la ressemblance fluctue quand l'original sourit ou se tourne (ou que la lumière change, comme cela est rapporté pour l'approche précédente) : après tout, il en est de même de la projection d'un mouvement circulaire, qui n'est pas toujours une ellipse presque circulaire si on l'observe dans un sous-espace.

5.1.3 Conclusion

Les sous-espaces (linéaires ou pas) d'images et de modèles peuvent permettre d'analyser (facilement quand on est dans le cas linéaire), représenter et donc synthétiser des flux d'images ou de modèles 3D, automatiquement à partir d'images animées et avec une qualité vidéo-réaliste. Si la théorie ne limite pas la dimensionnalité du sous-espace utilisé, il faut bien en pratique rester dans des limites raisonnables, pour l'occupation mémoire et le temps de calcul (surtout dans les cas non linéaires où une minimisation doit être effectuée). Lorsque les images/modèles

de la base doivent correspondre à la personne à modéliser, la contrainte ne concerne plus le matériel : on ne peut pas demander au locuteur de créer sa base personnelle en exécutant plus de quelques expressions ou postures. Une idée possible [VD99a, VD99b], pour avoir quand même un espace personnel de grande dimensionnalité, donc représentatif, consisterait à synthétiser ces représentants depuis un modèle synthétique déformable. La difficulté tient alors en plusieurs points :

- quels sont les déformations à appliquer ? C'est un problème de synthèse et de paramètrisation qu'on a déjà évoqué au chapitre 3.
- quels représentants générer, pour effectuer une bonne couverture de l'espace utile en un temps acceptable ? Un échantillonnage linéaire des paramètres n'est bien sûr pas le meilleur.
- comment faire pour que ces images synthétiques soient suffisamment proches (et donc utilisables pour l'analyse) des images réelles qu'elles remplacent ? On peut voir cela comme un problème de paramètrage (rajouter des degrés de libertés au modèle pour corriger l'illumination et quantités d'autres paramètres), mais cette approche ne fait que rendre le problème pratique plus insoluble. Dans le cas original, l'analyse envisagée utilisait un flux optique, qui s'accomode donc par construction des différences entre les deux mondes (réel et synthétisé) pour l'analyse.

Ces travaux particulièrement récents et ambitieux méritent donc d'être surveillés.

À l'heure actuelle donc, quoique vidéo-réalistes, les animations de visages obtenues dans la littérature par interpolation, extrapolation ou par l'exemple exhibent assez rapidement un caractère répétitif, qui reste apparent même si l'on utilise un découplage des zones actives selon les régions du visage, du fait des contraintes pratiques sur les dimensions des espaces manipulés.

Si le mélange d'images fournit seulement l'illusion de la vidéo, mais pas sa richesse, ne faut-il pas considérer que ce que l'on observe est modélisable comme de la vidéo sur support 3D, et effectuer une telle analyse/synthèse ?

5.2 Vidéo sur supports 3D

Si l'on connaît très précisément (pour chaque pixel) la profondeur d'un point sur l'image, on peut à priori réaliser un codage de la scène sur deux canaux découplés : l'un qui regroupe les informations 3D, et l'autre celles de la texture, telle qu'elle serait vue en faisant abstraction de son support (par exemple, une vue cylindrique). En plus de permettre de changer de point de vue à la reconstruction, cette séparation des flux fait espérer de bons résultats pour la compression. En effet, cette texture ne devrait à priori pas changer si le modèle se contente de se déplacer face à la caméra, et on peut donc envisager de la compresser et la transmettre sous forme de mises à jour lorsqu'elle s'anime. Le modèle 3D ne subit lui aussi que peu de transformations, dont certaines comme les rotations s'expriment avec un nombre très faible de paramètres. Au final, il n'est pas déraisonnable d'espérer encoder efficacement toutes ces modifications apparentes dans l'image, en les considérant comme une vidéo animée sur un support 3D en mouvement et en déformation.

Reste à définir de quoi est constituée l'information 3D, et si elle est connue suffisamment précisément à priori ou doit être capturée complètement et en temps réel. Enfin, l'information 3D peut être connue pour toute l'image, ou seulement pour certains objets de la scène.

5.2.1 Bas reliefs

À partir de deux ou trois caméras calibrées ou en correspondance, on peut estimer les profondeurs relatives ou globales de divers éléments de la scène. On dispose en fait d'un bas-relief – généralement assez bruité, dans l'espace comme dans sa cohérence temporelle – qui permet par exemple de différencier le sujet des éléments de l'arrière plan (qu'on peut ainsi éliminer, remplacer ou ne transmettre qu'une seule fois). L'image devient une surface, avec des discontinuités. On peut par exemple approcher ce bas-relief par un maillage, qu'on transmettra et qui servira au moment du rendu comme support pour la texture. Ainsi, on peut changer le point de vue et/ou proposer un rendu en stéréovision (par des lunettes synchronisant l'œil gauche avec les trames paires du moniteur, ou des écrans auto-stéréoscopiques [Pan] par exemple.)

FIG. 5.4 – *Restitution stéréoscopique d'un locuteur distant dans PANORAMA [Pan]*

L'expérience est assez riche pour l'utilisateur, qui peut vraiment se déplacer et apprécier des points de vue différents, sans lunettes. Bien sûr, seule la partie qui était visible par les caméras est en fait texturée, ce qui limite les angles de vues naturels. Autre point négatif, l'information supplémentaire (la profondeur relative, ou les images sources) doit être transmise aux sites distants (avec un codage adapté aux discontinuités de la carte de profondeur), donc avec un volume de données à priori plus important que pour une seule des images.

5.2.2 Recalage d'un modèle 3D de visage

Pour les deux expérimentations qui vont être rapportées ici, les auteurs **possédaient le modèle exact de la personne qui se trouvait devant la caméra**, grâce à un scanner 3D. Par rapport à la situation précédente, on ne dispose donc plus de l'information de profondeur directement pour tous les points de l'image, mais on peut envisager de la retrouver, au moins pour les points renseignés par le modèle, s'il est mis en correspondance avec l'image.

Si c'est le cas, on va pouvoir extraire la texture de chacune des images observées, pour construire le flux animé de la texture. Une fois cette tâche accomplie, on peut rejouer la vidéo

sur le clone en changeant l'angle de vue ou compresser les mises à jour dans cette texture. Cette dernière approche est à priori très efficace en terme de compression car si l'on a correctement extrait la texture, seuls subsistent des animations locales à la surface (et des changements, d'assez basse fréquence, dus à la lumière ou aux ombres). Si par contre les postures et déformations du modèle 3D utilisées pour extraire la texture ne sont que de piètres estimations, la texture ne se compressera plus aussi bien et la reconstruction risque même d'être très grossière. C'est pourquoi les méthodes classiques de recalage de modèle que l'on avait évoquées à la fin du chapitre 4 ne sont pas adaptées à cette difficile tâche.

Les deux approches qu'on va détailler ici [TEGK97, GGW+98] varient surtout par les conditions dans lesquelles posture et déformation du modèle sont estimées.

Recalage par minimisation d'erreur

Partant d'une séquence de quelques (9 seulement) images non-calibrées, les auteurs de cette première expérimentation [TEGK97] ont cherché à recaler le modèle, disponible à l'avance, sur l'image de la caméra (dont les paramètres intrinsèques sont supposés connus). Pour cela, les positions sur le modèle comme dans l'image de quelques points caractéristiques (extrémités des yeux et de la bouche, narines...) doivent être fournies, et conduisent à une estimation de la rotation et de la translation du modèle qui le ferait coïncider avec l'image réelle (par un calcul analogue à celui d'une calibration, mais pour le modèle). L'information de texture de l'image approchée par le modèle peut alors être extraite, par projection inverse de l'image vers la surface du modèle. Comme prévu, les auteurs notent que cette texture est partielle (seulement ce qui était visible), avec des erreurs manifestes aux limites (comme sur la figure 11.2, page 104). En mettant à jour la texture globale qui sert à la reconstruction, plusieurs problèmes se posent à eux : en copiant la partie interne de la texture partielle (sans les bordures), des défauts de raccord sont visibles à cause de la variation de l'illumination selon l'angle de vue et des imprécisions du recalage, suite au faible nombre de correspondances. Si la texture est intégrée progressivement (en la mélangeant à 50% avec la précédente valeur), les détails de l'animation et des contours apparaissent plus flous. Un panachage des deux méthodes est donc utilisé, selon que la vue vidéo est ou non de face.

Sans proposer une méthode automatique et achevée, les auteurs illustrent donc le concept et les problèmes associés. Leurs résultats montrent les inadéquations du positionnement reconstruit, pour les oreilles et autour du renfoncement des orbites oculaires en particulier.

On va voir que l'approche suivante met en œuvre une débauche de moyens pour accomplir la même tâche, mais avec une qualité quasi-parfaite.

Mesure dense des déformations

On a déjà souligné que le manque de points de contraste naturels et physiques à la surface de la peau handicapait les algorithmes de suivi. On a aussi vu que certains auteurs y remédiaient par des marqueurs colorés ou un maquillage facial.

Pour l'expérimentation rapportée ici [GGW+98], ce sont près de 200 marqueurs de différentes couleurs qui sont fixés sur un visage. Avec un éloignement maximal pour ceux qui présentent la même couleur, ces marqueurs sont répartis sur toute la partie chair du visage, autour des caractéristiques faciales, et permettront de matérialiser visiblement les déformations de l'original, pour les reproduire sur son modèle. Pour cela, six caméras de studio, calibrées, fixent la zone frontale de l'acteur. En plus d'un éclairage de studio (pour obtenir un signal couleur fiable), des lampes fluorescentes assurent que les spots maximiseront leur visibilité sous une grande gamme d'angles de vue.

Enfin, l'acteur, qui a été scanné avec ses marqueurs, est placé face aux caméras. Son visage est maintenu calé pour diminuer les mouvements globaux, et la posture du modèle est adaptée pour correspondre aux vues filmées. Une fois modélisées pour chaque caméra les apparences des marqueurs et leurs emplacements initiaux, l'enregistrement peut alors commencer. À l'aide des mouvements des marqueurs observés précisément en 3D grâce aux caméras calibrées, les déformations du modèle sont capturées et peuvent être utilisées pour capturer une texture animée «corrigée», décorrélée de sa forme. Par une combinaison d'interpolations et de filtres non-linéaires, l'image des marqueurs et leurs reflets peuvent être supprimés et remplacés par des fragments de texture de peau «neutre».

Après tous ces calculs, ils disposent d'un «film 3D», sous forme d'un clone 3D et d'une texture sans marqueurs, tous deux animés. La reconstruction est de très grande qualité, et capture par exemple toutes les rides faciales.

Clairement, leur environnement n'est pas celui d'une communication à distance. Dans des conditions de studio, avec beaucoup de matériel spécialisé et d'interventions manuelles, leur démonstrateur poursuivait manifestement un autre but : montrer que la texture sur la 3D permettait un vidéo-réalisme jamais atteint pour la 3D, et que dans le cadre de MPEG-4 on pouvait obtenir un flux fortement compressé et de haute-qualité.

5.3 Conclusion

Ce chapitre nous montre par l'exemple que pour synthétiser un flux vidéo de haute qualité, il n'est pas nécessaire de tout vouloir faire en 3D. Au contraire, pour créer une succession d'images qui soit vidéo-réaliste, il est tentant et parfois plus facile de partir de plusieurs images ou modèles réels (photo-réalistes), pour les combiner. Les problèmes de dimension des sous-espaces manipulés (pour l'analyse) sont cependant un problème, et grèvent la part d'expressivité qui est retranscrite dans les images synthétisées.

L'autre piste dégagée pour le vidéo-réalisme consiste à réutiliser des flux digitalisés pour les déformer par un support 3D. Si sa forme n'est pas adaptée (plan, ellipsoïdes, cylindres ou modèles génériques) on n'obtiendra pas un résultat utilisable. Par contre, les plus ambitieuses de ces méthodes utilisent un modèle précis de la personne observée (et de la caméra). Avec ce modèle 3D animable «classique», la texture vidéo du visage peut être extraite, et réutilisée pour un rendu animé dans d'autres conditions (changement de l'éclairage, ajout de tatouages ou d'objets font partie des applications déjà démontrées). Très impressionnants, les résultats obtenus ne doivent cependant pas faire oublier les conditions très particulières de leur obtention.

Pratiquement, on n'a donc pas trouvé dans la littérature de technique vidéo-réaliste qui soit parfaitement adaptée à notre cadre de vidéo-conférence en vraie 3D sans délai : selon les travaux, le domaine synthétisable ne couvre pas toutes les expressions ou bien l'analyse demande des temps de calculs plus que prohibitifs, ou a seulement été simulée pour tester l'expressivité de la synthèse.

Chapitre 6

Exemples d'espaces de communication

On ne listera pas ici toutes les tentatives de communautés virtuelles rapportées dans la littérature. Nombreuses sont les expériences qui ne se placent dans notre perspective, mais certaines, par leurs échecs ou leurs succès, apportent des enseignements sur des problèmes qui n'ont pas encore été abordés dans ce tour d'horizon. En particulier, quelques approches qu'on va évoquer dans ce chapitre ont introduit des concepts novateurs et dignes d'intérêt pour la téléprésence, ou sensibilisent à des écueils dont il faudrait avoir conscience.

6.1 Modalités et interfaces de communication

Il n'est plus à démontrer qu'il ne suffit pas d'inventer un (mauvais) outil technologique pour qu'il soit adopté. Plutôt que de créer le besoin et un succès marketing, on peut s'intéresser aux besoins des utilisateurs et adapter ou créer des outils spécifiques.

Le problème peut résider dans l'interface homme-machine (lorsque l'on présente trop de réglages à l'utilisateur, ou que la présentation des informations est inadaptée par exemple) ou bien quand le média choisi est trop restreint, ne permettant pas une communication adaptée.

À ce titre, les *mediaspaces* [DB92, BHI93, CBCC98] sont particulièrement dignes d'intérêt, de même que certaines études menées sur leurs utilisations [OTP, Rou97].

6.2 Matérialisation de la présence

Comment donner l'impression que des personnes distantes sont présentes près ou autour de nous, plutôt que juste sur un même écran d'où proviennent toutes les voix ? Quels sont les degrés de liberté qu'on peut offrir à l'utilisateur, et comment les contrôlera-t-il ?

6.2.1 Individualisation des rendus audio et vidéo

Dans le cadre de l'*Ontario Telepresence Project* et du sous-projet Hydra [Bux92], chaque personne distante est représentée sur un poste de restitution différent. De petite taille, incorporant

un écran LCD et un haut-parleur, chaque unité peut être placée indépendamment des autres, comme sur la figure 6.1. Parce que les présences (image et son) sont distinctement réparties dans l'espace, l'expérience est plus réelle. Pour changer l'agencement de la réunion, il suffit de déplacer ces objets, sans interface informatique puisqu'ils appartiennent au monde réel.

FIG. 6.1 – *Des postes individuels pour restituer les présences audiovisuelles [Bux92]*

6.2.2 Les espaces de communication audio

Il est possible de créer des espaces uniquement sonores de communication [ACK94, AHMS97, TL99]. Certains utilisent des techniques de spatialisation du son [SPA, USA] pour enrichir l'espace sonore, par exemple en simulant la perception de positions pour les locuteurs. Selon les cas, la configuration de l'espace sonore est imposée aux auditeurs, ou peut être contrôlée (consoles de mixage, interfaces graphiques plus ou moins symboliques ou capteurs et objets du monde réel dans l'expérience *SomeWire* [SHSW99]).

6.2.3 Murs de téléprésence

Puisque lorsqu'on «téléporte» l'image des participants, on s'aperçoit qu'ils ne sont pas dans un espace commun, certaines approches [MK92, EG97b] ont pris le parti de mettre tout en œuvre pour que plusieurs espaces distants semblent suffisamment confondus pour inspirer l'impression d'une unité. Cela passe par une spécification très précise des salles et de leur équipement : l'éclairage et le mobilier par exemple sont tels que les conditions locales donnent l'impression de se raccorder avec les images des sites distants, représentées sur des écrans en vraie grandeur (comme sur la figure 6.2). L'aspect sonore n'est lui non plus pas négligé : avec des placements contraints pour les micros et les participants, on dispose de l'information pour pouvoir restituer une impression de la provenance sonore qui soit en accord avec l'image : avec par exemple un haut-parleur «relié» à chacun des micros distants et situé à l'emplacement normalisé pour l'image du locuteur, ou par des solutions d'holophonie [NEG98]. De tels environnements proposent aussi une annulation sophistiquée de l'écho (les différents chemins possibles sont nombreux lorsque tous les micros sont actifs, pour ne pas restreindre les prises de parole).

Ces coûteuses salles spécialisées exigent aussi des liaisons dédiées (par exemple ATM), avec un haut-débit pour des images de qualité (elles sont affichées en taille réelle) et les nombreux canaux sonores nécessaires.

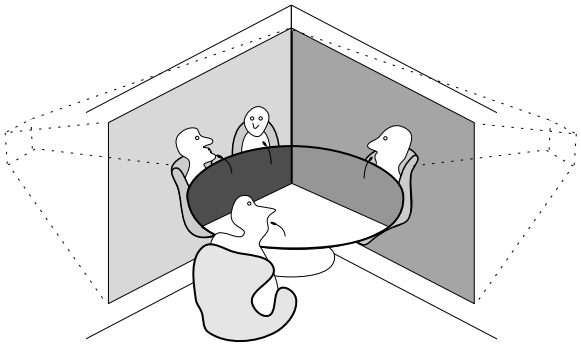


FIG. 6.2 – *Une résolution matérielle «parfaite» de la téléprésence*

6.3 Le débat réel diffère des débats reconstruits

Selon les applications de téléprésence, l'espace transmis est plus ou moins différent de l'espace original, mais pas seulement par déficience de la technique : ce peut être un choix délibéré, par exemple pour respecter le désir d'intimité [CBCC98].

Dans le cas des représentations 3D pilotées, toutes les approches ne font pas les même choix : si le locuteur tourne la tête dans la réalité (pour regarder hors de son écran), faut il relayer ce mouvement sur son clone et risquer de donner l'impression qu'il regarde son voisin virtuel? Faut-il aussi corriger la direction du regard synthétique? Le placement des interlocuteurs doit-il nécessairement être calqué sur une situation réelle (en créant une table virtuelle par exemple)? La vue de la scène doit elle être la même pour tous les participants, ou chacun peut-il placer ses vis-à-vis et choisir pour son poste de visualisation une vue individualisée?

Il n'y a pas de réponse unique et franche à toutes ces questions. Chaque alternative génère un outil différent, qu'il serait souhaitable de pouvoir tester en vraie grandeur pour le qualifier. Rendre tout paramètrable, en léguant de trop nombreux contrôles à l'utilisateur n'est pas une solution qui lui permettra de se sentir intégré à la conférence. À défaut, il faut se fixer une application particulière et imposer les réglages ou un scénario.

6.3.1 Virtual Party

L'expérience de *Virtual Party* [HCS96] propose à des intervenants distants de piloter des représentants dans un monde partagé, une sorte de télé-cocktail. Ils peuvent donner des ordres à leurs avatars, comme de se rendre au bar, d'aller discuter avec une personne et de regarder quelqu'un ou quelque chose.

L'originalité de ce monde virtuel est qu'il s'insère dans une architecture de *Virtual Cinematographer* qui se propose de le filmer (au sens paramétriser et générer un flux d'images synthétiques) de façon complètement automatique. Cela est réalisé en construisant un modèle comportemental hiérarchique ad-hoc qui intègre différents comportements préétablis et «câblés» sous forme de modules ou d'idiomes dont les exécutions seront cascadées comme autant d'appels de procédures imbriquées. Ainsi, des caméras spécialisées dans certaines tâches (filmer deux personnes de face,

ou en contre-champ) seront activées par un module spécialisé dans les dialogues ou discussions à trois, lui même un représentant du concept de discussion (dont un exemple de comportement alternatif serait le suivi de personnages en mouvement).

Les transitions sont faites de façon déterministe, en réponse à des évènements prévus lors de la création des modules : une personne commence à se déplacer, ou parle depuis exactement 10 secondes. De nombreuses règles ou recettes issues de la pratique cinématographique sont intégrées de façon empirique au niveau de chaque module ou idiome, avec l'espoir qu'ils se combinent à plus haut niveau comme prévu[1]. En pratique, le résultat au niveau global est totalement déterministe, mais pas forcément aisé à régler (comme avec des réseaux de neurones). Si on est assuré d'avoir un déroulement donné (par exemple que les même enchaînements rejaillissent continuellement), il n'est pas facile de le modifier ou de garantir que c'est le meilleur.

6.4 Conclusion

Ce court chapitre a illustré quelques modalités supplémentaires pour la communication, notamment le rendu sonore et l'agencement visuel. Il oppose aussi différentes approches pour le contrôle des présences et souligne l'importance d'un scénario applicatif.

En abordant ces problèmes, on a présenté plusieurs des pistes que l'on va suivre pour construire une nouvelle solution au problème de communication initialement formulé, une fois celui-ci de nouveau précisé.

1. grâce notamment à un concept d'exceptions, qui n'est pas très élégant car en autorisant l'interconnexion entre des modules différents, il brise leur indépendance : les modules des caméras doivent par exemple savoir qu'un mouvement pourrait intéresser des modules de suivi, et «s'arranger» pour provoquer une cascade de retours de procédure qui précède l'activation de leurs concurrents, les modules de suivi...

Chapitre 7

Conclusions du tour d'horizon

Même en négligeant tout ce qui concerne l'architecture réseau et les protocoles, les techniques et pistes de recherche qui touchent de près ou de loin **au contenu et à la forme** d'un service de communication à distance sont donc nombreuses, et font appel à des domaines de compétence très variés : techniques de compression, vision 2D, vision 3D, optimisation non-linéaire, synthèse d'images, analyse statistique ou biomécanique et anthropométrie par exemple, voire psychologie ou histologie.

Le constat relatif aux techniques vidéo doit tenir compte de l'application envisagée. Un vidéophone pour deux personnes ne peut que voir le jour, et satisfera probablement aux attentes du grand-public : voir la personne avec qui l'on dialogue et ses sourires. Son extension à plusieurs personnes ne le transformerait cependant pas en un outil très adapté à une communication de travail.

Indépendamment, l'évolution des approches de compression nous montre qu'il ne faut pas hésiter à faire appel à des approximations fortes ni à considérer la vidéo à un plus haut niveau d'abstraction que celui des pixels.

C'est ainsi que le codage 3D offre un fort potentiel pour représenter les visages, même si leur animation reste un point délicat, particulièrement du point de vue de la fidélité lorsqu'elle est pilotée par un interlocuteur réel. Par contre, la liberté de recomposition, par exemple dans des scènes virtuelles, est totale et très intéressante pour les applications de télécommunication avec plusieurs participants et plusieurs sites.

Entre la communication basée tout-vidéo et la solution de l'animation tout-3D émergent aussi quelques techniques hybrides, cherchant à concilier plusieurs de leurs avantages respectifs :

la vidéo est pleine d'une information riche et animée que notre cerveau exploite naturellement, pour reconnaître les locuteurs et leurs expressions. En ce sens, elle joue le rôle d'une **fenêtre qui abolirait la distance sans altérer le message de la communication,**

la 3D résout bon nombre des problèmes pratiques, notamment lorsqu'elle se propose de diminuer les flux. Elle est aussi le point d'entrée vers les réalités virtuelles ou augmentées, par exemple en permettant de **regrouper les participants distants dans un espace partagé**, s'astreignant des échelles et angles de vue.

L'impact, par leur qualité visuelle inhabituelle dans le monde du graphique, de certains résultats de synthèse récents [BCS97, TEGK97, GGW+98, PHSS98, BV99] confirme s'il en était besoin que la texture et son animation sont très efficaces, sinon primordiales, pour **conserver le réalisme plutôt que le créer** : dans un cadre d'animation qui ne serait pas de la synthèse pure, où l'on dispose d'une source riche, photo-réaliste ou vidéo-réaliste, sa réutilisation peut s'avérer payante.

Mais comment combiner ces briques de base que sont la 3D et la vidéo pour construire une communication de plus grande échelle que celle d'un vidéophone ? Ne manque-t'il pas des concepts de plus haut niveau pour permettre à plusieurs personnes de partager une expérience visuelle et sonore suffisamment naturelle pour autoriser un travail en groupe de discussion ?

Les contributions de la thèse

Vers une solution hybride

Initialement, on a listé les contraintes qu'on attendait d'un système de communication idéal, dans le cadre d'une discussion à distance. Reprenons ces points, en marquant par un encadré ceux qui posent problème, parce qu'ils ne sont pas «simplement» réglés par l'utilisation d'un réseau :

(a) utilisable entre personnes distantes,

(b) temps-réel, sans notion mesurable ou gênante de différé, c'est-à-dire sans que ce soit préjudiciable à une communication «naturelle»,

(c) totalement interactif, sans hiérarchie ni protocole, laissant libre cours à une communication collaborative naturelle,

(d) muni d'un support audio minimal pour la parole, par exemple celui du téléphone,

(e) doté d'un support vidéo suffisant pour la reconnaissance des visages d'autrui et d'un minimum d'expressions, les plus réelles et personnelles possible,

(f) une expérience assez proche d'une vraie réunion, avec une certaine notion d'un espace commun,

(g) avec une architecture et des performances suffisantes pour des communications entre plus de deux personnes.

Suite à l'analyse des techniques vidéos présentée dans l'état de l'art au chapitre 1, rappelons en quoi la vidéo pure n'est pas compatible avec les deux derniers points :

- parce que les possibilités de transformation de la vidéo reçue sont limitées à des opérations 2D ou 2D et demi, on ne peut pas utiliser les images reçues pour montrer les participants de coté et de dos, dans une vue combinée, comme s'ils étaient dans une même pièce qu'observerait le spectateur.

- parce que les flux vidéos – plus ou moins volumineux selon la technique de compression et la précision/expressivité sacrifiée – sont multipliés avec le nombre de participants, ils peuvent rapidement devenir un goulot d'étranglement pour le réseau ou les machines chargées de l'acquisition et l'affichage.

L'approche tout-3D, si elle diminue les flux et autorise une composition virtuelle, n'assure pas la fidélité de la vidéo qu'exige le point (e) : une scène vidéo d'un visage dégage une ressemblance et un naturel qui ne sont pas encore atteints par son approximation sous forme d'un clone 3D piloté selon les méthodes des animateurs d'image de synthèse.

Peut-on dès maintenant exiger tout coder en 3D dans le cadre des communications qui sont envisagées dans ce document ? N'existe-t-il pas une solution qui mélangerait de façon satisfaisante 3D et prises de vues réelles ?

Audio, 3D et vidéo

Une nouvelle solution – hybride, parce que 3D et vidéo y sont combinées pour représenter et animer les participants – sera proposée dans le chapitre 11. On y exposera les principes et problèmes de toute approche de ce type, avant de proposer une résolution pratique sous certaines contraintes et de présenter les résultats du prototype d'évaluation construit.

Comment orchestrer ces représentations pour ne pas contredire les points (c) et (g) ? On propose au chapitre 9 une gestion de l'espace virtuel – visuel 3D et sonore spatialisé – sous la forme d'un régisseur qui filmerait le débat. En choisissant selon les événements un sous-ensemble plus au moins grand des participants, et donc des images plus ou moins détaillées, il ménage ainsi les points (e) et (f), selon les vues produites.

Enfin, pour minimiser la dégradation des performances avec le nombre d'intervenants, le chapitre 8 propose un algorithme de rendu rapide, spécialisé dans les visages.

Ainsi architecturée, notre approche va consister en fait à construire pour l'écran et le canal sonore la perception d'un débat filmé en temps réel sur un plateau de télévision virtuel. À la fois spectateur de ce programme produit pour lui à la volée, chaque participant y est aussi «acteur», audible et visible sur le plateau par une représentation mélangeant vidéo et modèles 3D de son visage ou de son buste.

Chapitre 8

Structure et algorithme pour un rendu rapide de visages synthétiques

Le rendu par polygones et textures est une technique classique et maîtrisée, qui donne lieu à des implémentations efficaces, aussi bien sur les stations graphiques dédiées que sous forme de cartes 3D accélératrices pour les compatibles PC. Cependant, leur utilisation éventuelle dans le cadre de la vidéo-conférence peut soulever quelques remarques :

- la généralisation de services de vidéo-conférence pourrait se faire avec la distribution ou la location de terminaux à usage générique, par exemple par le fournisseur d'accès ou de services. Ces terminaux sont censés être moins chers que les PC, et se voudraient aussi utilisables qu'un magnétoscope, même sans compétence informatique de ses utilisateurs. C'est par exemple le cas de certains *Network Computer* (NC) ou de *SetTopBox* (STB), qui ne disposent pas d'accélérateurs 3D, voire de coprocesseur flottant. Or les algorithmes génériques de 3D pour la manipulation et le rendu ne sont plus du tout aussi efficaces s'ils ne sont implémentés qu'en logiciel.

- la présence de nombreux clones, donc de nombreux polygones et textures peut devenir un frein même sur du matériel de performance moyenne. Par exemple, le manque de mémoire texture peut se traduire par une surcharge du bus, puisqu'il faut – à chaque nouvelle image de chacun des visages – retransférer la texture –différente– depuis la mémoire centrale. La vitesse de rendu peut alors dégénérer et dépendre plus du nombre de primitives que de leur impact utile à l'écran.

- à priori, plus les visages seront nombreux à l'écran, plus ils seront affichés en petite taille, et n'auront pas besoin d'être définis avec une grande précision. Au contraire, lorsqu'un visage sera vu seul, en gros plan, le besoin de fidélité impose de restituer les courbes naturelles du visage, par exemple tous les profils visibles. Avec des primitives polygonales, qui ont plutôt tendance à générer des arêtes, cela nécessite de multiplier le nombre de triangles. Le compromis précision vs taille et vitesse peut être délicat à optimiser lorsque les primitives sont aussi mal adaptées.

– puisque les objets qu'on veut afficher sont des visages auxquels on n'appliquera pas tous les effets, déformations ou opérations envisageables pour des objets 3D quelconques, on pourrait se contenter d'un sous-ensemble d'opérations de manipulation et d'affichage, et d'optimiser la structure et son utilisation en conséquence.

Cette thèse propose donc une nouvelle structure pour la description et le rendu de visages synthétiques. Elle a l'avantage de pouvoir convenir à des machines modestes, sans cartes accélératrices, et offre une vitesse de rendu qui dépend suffisamment de la taille d'affichage, de sorte que la multiplication des visages (d'autant plus petits qu'ils sont nombreux) dans une même scène se fait sans trop de pénalités. Pour cela, il n'est pas inutile de commencer par s'interroger sur la nature de la 3D.

8.1 Création de l'illusion 3D

Qu'est ce que la 3D ? Voici ce qu'écrivait il y a près de 40 ans un professionnel du dessin artistique [Hog92], dans un contexte non informatique :

> « *Les artistes de toutes les générations ont cherché des voies aptes à faire apparaître la surface de l'image plate d'une manière plus plastique, d'y pénétrer et de la mettre plus en valeur, de la faire avancer ou de la faire reculer. Les principes d'illusion, d'ombre et de lumière pour créer la forme, la perspective linéaire du dessin et la perspective de la figure avaient valeur de grandes découvertes dans la question de la décomposition de la surface bidimensionnelle dans l'espace en profondeur.* ***La vue humaine, l'œil, ne peut pas percevoir la profondeur. La troisième dimension est un facteur de perception du jugement de l'expérience****, développé à partir du contact physique et du mouvement corporel dans le monde objectif des choses. Si nous pouvions voir la profondeur de l'espace comme une réalité tridimensionnelle, il serait alors possible de voir un objet à la fois d'en haut et d'en bas, de coté et de derrière tout comme une main qui fait l'expérience de la profondeur de l'espace quand elle tient une balle. [...] la main n'y croit pas quand elle s'étend pour toucher la forme photographique* ».

Sans motivation artistique mais avec celle de la fidélité dans un contexte informatique, ce que l'on veut faire est bien de recréer l'illusion 3D à la surface d'un écran, pour leurrer l'œil et les processus cérébraux de la vision et de la reconnaissance.

8.2 L'illusion d'un visage

En fait, plus que le visage, c'est son image qui nous intéresse, et de ne pas briser l'illusion que plusieurs images correspondent à une même personne. Mais puisque deux images d'un même visage ne sont pas identiques, qu'est ce qui fait qu'on les apparie ? On peut répondre à cette question en se demandant quelles sont les causes des différences observées entre deux images. Il peut s'agir :

– d'une transformation dans le visage, comme un oeil qui cligne, des lèvres qui articulent. Cette déformation partielle fait que **le visage s'anime**.

– d'un changement d'éclairage ou de couleur de teint (comme pâlir ou rougir), qui se matérialisent à la surface de la peau. Le visage **change d'aspect en surface**.

– d'un changement de point de vue (donc d'échelle), qui fait que les parties visibles ne sont plus les mêmes, ou que les formes apparentes semblent changer plus ou moins profondément. Il en est de même si **le visage tourne ou se déplace**.

Supposons dans un premier temps (jusqu'au chapitre 11), que le visage ne se déforme pas, ne s'anime pas. De tels visages statiques et rigides seraient alors un équivalent informatique de bustes sculptés. Pour une meilleure ressemblance, on les dotera cependant d'une coloration de peau, comme sur la figure 8.1, où le buste «peint» semble bien plus naturel.

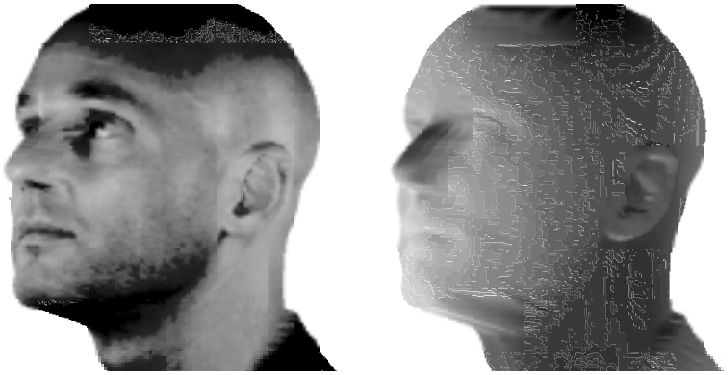


FIG. 8.1 – *Images de bustes, avec et sans pigmentation*

Pour d'un visage produire plusieurs images, avec entre elles une cohérence qui fait qu'un observateur peut mentalement reconstruire et reconnaître «l'objet» original, on va les mémoriser sous la forme d'un modèle 3D ad hoc, spécifique à un visage donné.

FIG. 8.2 – *Une texture cylindrique et la forme associée*

Ayant pour l'instant écarté l'animation, il y a donc deux types d'informations à mémoriser : une information de forme, en tant que support 3D qui occupe de l'espace, et une information de couleur, qui s'affiche en tant que surface. C'est pourquoi on retient une structure double : une texture cylindrique et une forme en fil de fer du visage, comme l'illustre la figure 8.2.

La texture cylindrique collecte une vue à 360°, de sorte que chaque bande verticale semble vue de face. La forme en fil de fer est issue de la déformation d'un cylindre, pour s'approcher le plus possible du visage voulu.

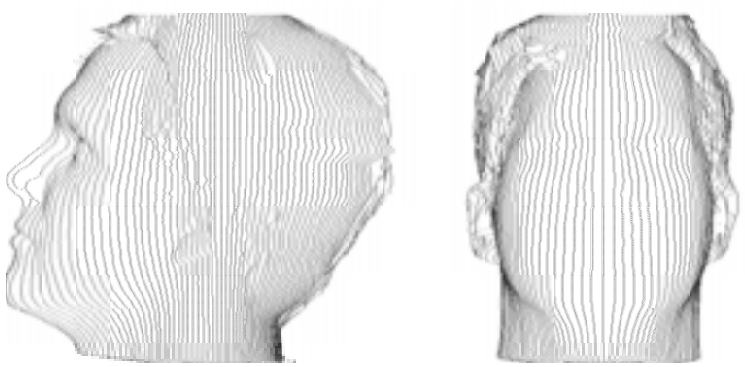


FIG. 8.3 – *Images d'une structure en fil de fer*

FIG. 8.4 – *Exemples de rendu «forme + texture» sous différents angles*

Il est facile de montrer que ces deux informations sont bien complémentaires : Ainsi, sur les vues d'un modèle fil de fer, comme en figure 8.3, on ne perçoit la forme du visage que sur les cotés. La zone centrale semble sans relief, sans qu'on puisse la différencier de l'image d'un cylindre.

Inversement, la texture seule, même sur une forme simple comme un cylindre, n'est pas convaincante : si la bande centrale de la texture de la figure 8.2, par ses variations de lumière et notre habitude rappelle naturellement une photo, un rapide coup d'oeil horizontal suffit pour

percevoir que la périphérie n'a pas une forme apparente compatible avec notre modèle perceptif des visages (pas de profil tranché qui rappellerait une oreille ou un nez).

En combinant les deux structures, on peut obtenir une déformation de la texture qui épousera les formes du modèle en enrichissant les zones dont la déformation n'apparaît pas visiblement. La forme est directement perçue sur les bords tandis qu'au centre elle peut être interprétée d'après les variations de teintes, comme sur la figure 8.4.

8.3 Encodage de la structure 3D retenue

On choisit de discrétiser le cylindre déformé et la texture associée à cet axe par un empilement de h couches, comme sur la figure 8.5.

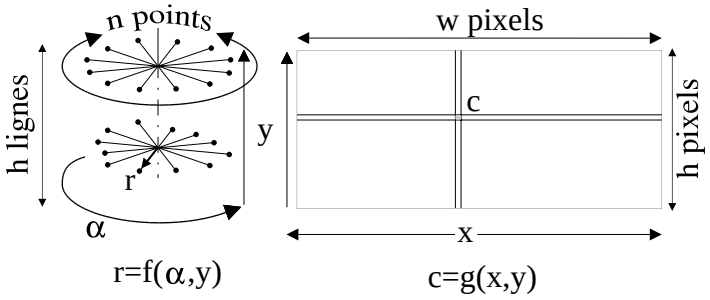


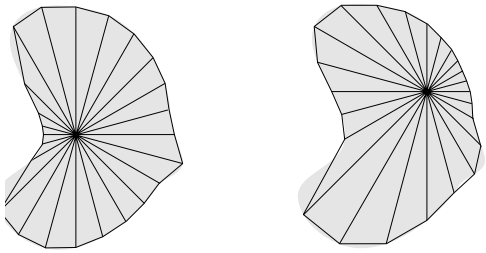
FIG. 8.5 – *Encodage et paramètres de la double structure*

Pour **la texture**, ce sont h lignes de w pixels chacune. Elle peut donc être encodée suivant n'importe quel format d'image, couleur ou monochrome, par exemple le JPEG (Cf. Annexe B) puisqu'il compresse efficacement et qu'on peut s'autoriser de légères altérations (peu ou pas visibles sans zoom, selon les options de compression) de l'image originale.

Pour **la forme**, chacune des h couches comportera exactement n points, répartis à angle régulier, soit $\frac{2\pi}{n}$, par rapport à l'axe haut-bas du cylindre. Les points relatifs à un même angle s'empilent donc pour construire l'un des n profils de h lignes.

En pratique, on a choisi de stocker les points par leur distance à l'axe du cylindre, sous forme d'octets (donc de 0 à 255, à un facteur multiplicatif près). Les valeurs négatives compliqueraient le problème du rendu, elles sont donc interdites. Sur des objets massifs comme une tête, cela n'est pas vraiment une contrainte, puisqu'en translatant l'axe central vers l'intérieur de la tête, on évite généralement toute valeur négative et l'échantillonnage angulaire se repartira mieux sur la surface (Cf. fig 8.6).

La forme est donc un tableau rectangulaire ($n \times h$) d'octets, qu'il est possible d'interpréter comme une image en niveaux de gris. On peut alors la compresser assez efficacement à l'aide de formats comme GIF ou PNG (Cf. Annexe B). Le format JPEG n'est pas souhaitable car il introduirait des artefacts dans la forme du modèle.



La qualité de l'approximation de la forme grisée dépend de la position de l'axe. De fortes courbures et des changements de concavité, plus rares pour un visage, sont plus difficiles à approximer. Un nez peut cependant être capturé assez fidèlement s'il est aligné avec la distribution angulaire des échantillons. Ainsi, l'approximation de gauche préserve mieux la petite proéminence de droite.

FIG. 8.6 – *Erreurs d'approximation selon la courbure et la position de l'axe*

	Figure 8.1	Figure 8.3	Figure 8.2
Dimensions de la texture	512×450	512×450	512×200
Dimensions du modèle	256×450	512×450	16×200
Taille de la texture en JPG	25 Ko	22 Ko	15 Ko
Taille du modèle en PNG	6 Ko	5 Ko	1 Ko

FIG. 8.7 – *Dimensions et tailles compressées de quelques modèles*

À titre d'exemple, la figure 8.7 présente les dimensions et tailles des modèles utilisés dans ce chapitre. Les modèles très précis (figures 8.1 et 8.3) ont été convertis depuis des modèles polygonaux CyberWare du domaine public, tandis que celui des figures 8.2 et 8.4 – d'une précision suffisante pour un rendu sur un téléviseur – a été obtenu à l'aide d'une caméra non calibrée.

On pourrait assouplir le codage en transmettant pour chaque point la position complète dans son plan horizontal (distance et angle, ou paire de coordonnées), pour permettre un échantillonnage adaptatif de la surface, qui concentrerait la précision dans les zones les moins lisses du visage, mais ce supplément d'information ajouterait un coût important à l'encodage. En pratique, il vaut mieux utiliser un échantillonnage régulier deux fois plus précis par exemple.

L'avantage de ce codage est d'être simple et très compact. Il privilégie avant tout la précision des profils, qui sont un élément clef de la ressemblance et de la reconnaissance. D'ailleurs, si l'on pouvait regarder le modèle de dessus, des discontinuités éventuelles entre les couches horizontales pourraient apparaître. Elles n'apparaîtront pas si l'on se contente de tourner autour du visage synthétique, restriction qui n'est pas forcément très gênante dans notre application et qu'on justifiera plus en détail par la suite. Dans notre cadre, le format choisi va permettre de proposer des algorithmes de rendu rapides.

8.4 Rendu des points du modèle

Cette section présente le contexte et les notations indispensables pour la production d'images à partir du modèle précédent, en se limitant à considérer ses points. C'est seulement dans les deux sections suivantes (8.5 et 8.6) qu'on construira des rendus plus complexes, en altérant l'interprétation du modèle selon le réalisme et les fonctionnalités voulues.

8.4.1 Modèle de caméra et projection d'un point

On utilisera la modélisation de caméra perspective classique, avec un centre optique placé derrière un écran virtuel plan (de façon à ne pas inverser l'image).

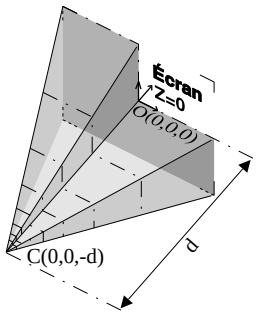


FIG. 8.8 – *Modèle de caméra perspective*

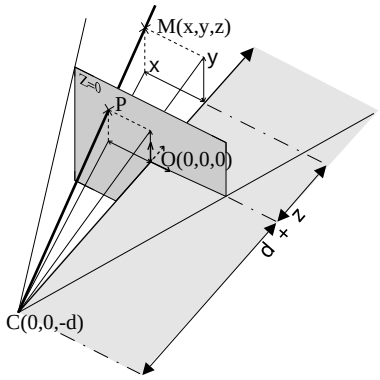


FIG. 8.9 – *Image d'un point de l'espace*

Avec les notations de la figure 8.8, plaçons nous dans le repère orthonormé $(O, \vec{i}, \vec{j}, \vec{k})$ qui est tel que le plan de l'écran E est confondu avec $(O, \vec{i}, \vec{j})$, et où le centre optique de la caméra est placé en $C(0, 0, -d)$.

Alors, comme sur la figure 8.9, tout point $M(x, y, z)$ du demi-espace observable $(z > 0)$ a son image non-inversée à l'intersection du rayon lumineux CM avec l'écran virtuel E, soit en $P\left(x\frac{d}{d+z}, y\frac{d}{d+z}, 0\right)$, par exemple en remarquant les triangles semblables.

En projetant ainsi tous les points du modèle cylindrique, on en obtiendrait bien sûr une image, mais sous la forme d'un nuage de points peu lisible, puisqu'on n'a pas différencié de points qui seraient visibles ou invisibles, selon qu'ils seraient par exemple «tournés» vers la caméra ou lui feraient face.

8.5 Algorithme de rendu avec points cachés

Si l'on étend notre modèle fil de fer à une interprétation surfacique, on héritera de points cachés : la présence d'un fragment de surface plus près de la caméra rendra certains points invisibles (à peu près la moitié arrière de la tête, car des proéminences comme le nez ou les oreilles peuvent cacher une portion de la joue ou rester visibles de profil). Si seules sont visibles les extrémités des segments (les points du modèle) qui ne sont pas masquées, on obtiendra des

vues fil-de-fer du type de celles déjà présentées. Cette représentation plus lisible du visage et plus conforme à l'expérience sera facilement étendue au cas texturé, dès la prochaine section.

8.5.1 Restriction des mouvements de caméra

On va limiter les mouvements de caméras que l'on autorise. Ainsi, la dimensionnalité du problème de visibilité sera réduite, et un algorithme de rendu plus efficace pourra être proposé.

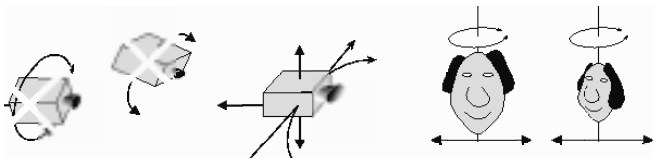


FIG. 8.10 – *Degrés de libertés de la caméra relativement au modèle*

On impose que les axes haut/bas du modèle et de la caméra restent colinéaires. Les placements relatifs permettent toujours de tourner autour du visage (1 degré de liberté) et de s'éloigner ou se rapprocher, comme le montre la figure 8.10. On ne peut donc plus voir les visages s'incliner, ni sur le coté, ni vers le haut ou le bas. Avec cette restriction, un clone pourrait donc faire «non» de la tête, mais pas opiner du chef. Cela n'est pas très gênant puisque dans un contexte où une personne n'est pas sûre d'être vue ou regardée par les autres, ce genre de communication visuelle ne devrait pas être utilisé.

8.5.2 Hypothèse de perspective faible

L'hypothèse de la perspective faible est presque aussi fréquemment utilisée que le modèle classique de la caméra perspective qu'on a introduit à l'occasion du rendu point-par-point de la structure. Commençons par remarquer que les points situés dans chaque plan $z = constante$ subissent une même transformation, composée d'une projection orthogonale et d'une mise à l'échelle : comme un point $M(x, y, z)$ se visualise en $P\left(x\frac{d}{d+z}, y\frac{d}{d+z}, 0\right)$, le rapport $\frac{d}{d+z}$ est effectivement constant pour tout $z = constante$ donné.

Qu'en est-il de points situés à proximité d'un plan $z = z_0$ donné? Le facteur d'échelle $\frac{d}{d+z}$ est dans certains cas très proche de $\frac{d}{d+z_0}$. L'erreur exacte vaut :

$$\begin{aligned} \frac{d}{d+z} - \frac{d}{d+z_0} &= \frac{d^2 z_0 - d^2 z}{(d+z)(d+z_0)} \\ &= \frac{d^2(z - z_0)}{(d+z)(d+z_0)} \end{aligned}$$

Pour une distance focale d constante, l'erreur de position dans l'image sera d'autant plus faible que z et z_0 seront proches et que z et z_0 seront grands devant la focale.
On appliquera l'hypothèse de perspective faible pour le rendu des points de chaque visage[1]. Si

1. qui ne sont jamais zoomés au point de ne pas apparaître en entier dans l'image.

$z = z_0$ est le plan qui contient l'axe du clone, tout point $M(x, y, z)$ du modèle sera visualisé en $P\left(x\frac{d}{d+z_0}, y\frac{d}{d+z_0}, 0\right)$, c'est à dire que tous les points d'un même visage subiront une projection orthogonale et une mise à l'échelle qui ne dépendra que de la position de l'axe. Ainsi, le calcul de la transformation de chaque point pourra être fait plus efficacement, et l'on gardera l'impression de perspective, lorsque le visage se déplacera dans la scène, ou du fait de sa taille relative par rapport à des clones plus proches ou plus éloignés dans la scène.

8.5.3 Notion de segments

On considère que le modèle est en fait la réunion de segments horizontaux qui relient chaque point $P\left(\alpha = k\frac{2\pi}{n}, y\right)$ du modèle à son voisin de droite $P\left(\alpha + \frac{2\pi}{n}, y\right)$ et à son voisin de gauche $P\left(\alpha - \frac{2\pi}{n}, y\right)$. Le modèle se présente donc comme l'empilement de n polygones horizontaux (qui feraient tous un pixel de hauteur). Ainsi, l'intersection d'un modèle avec un plan horizontal passant par l'un de ses points est un polygone non-croisé qui joint tous les points de même altitude, comme sur la figure 8.11. On suppose que seules les extrémités (les points du modèle) sont visualisées (pas les segments), pour reproduire des vues fil-de-fer analogues à celles de la figure 8.3.

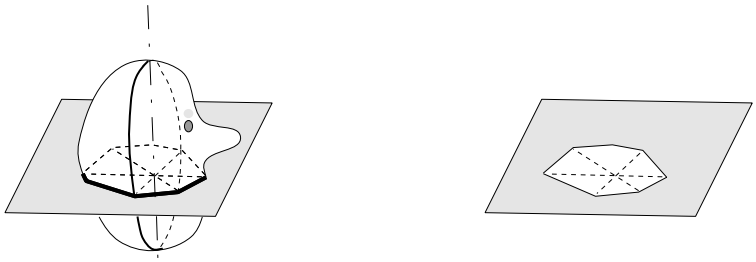


FIG. 8.11 – *Segments du modèle*

8.5.4 Ordonnancement du rendu

Puisqu'on n'incline pas la caméra, chacun des polygones horizontaux qui forme le modèle se projettera dans un segment sur une ligne horizontale de l'écran. Aux considérations d'anti-aliasing près, le rendu de chaque plan horizontal du modèle (chaque contribution à une ligne de l'écran) peut donc se faire indépendamment. Par exemple, pour un angle de vue donné, on va pouvoir tracer le modèle de haut en bas, en obtenant ligne après ligne les contributions à l'image, comme sur la figure 8.12.

Après avoir **décomposé le problème verticalement**, il reste donc à effectuer efficacement le rendu d'un de ces polygones, en détectant les points cachés.

Si le modèle est face à la caméra, son axe centré sur l'axe optique, un point qui n'est jamais caché est celui qui tombe sur cet axe optique, ou en est le plus près en terme de déviation angulaire. Comme l'échantillonnage du modèle est régulier, il suffit de connaître le degré de

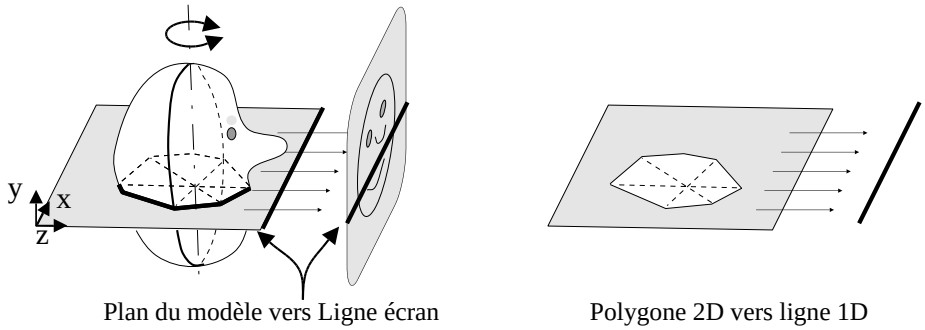


FIG. 8.12 – *Liens entre lignes d'écrans et polygones horizontaux du modèle*

rotation du modèle autour de son axe pour identifier ce point, sans que la forme du modèle n'intervienne : par exemple, quand on tourne le modèle de 30° sur la droite, c'est le point du modèle situé 30° sur la gauche qui passe dans l'axe, de face et toujours visible (grâce à la contrainte de positivité des distances à l'axe).

Dans le modèle de perspective faible qu'on utilise, cela reste vrai même si le clone n'est pas centré : son image est exactement translatée, sans que les angles vus ne soient modifiés (en perspectif pur, ils sembleraient tourner faiblement mais à des vitesses relatives liées à leur profondeur).

On va donc effectuer le tracé en partant de ce point frontal, et en cherchant quels sont les points qui apparaîtront à sa droite (respectivement à sa gauche). Lorsqu'on suit l'arc en s'éloignant de ce point frontal, la condition de visibilité est de ne pas être masqué par les segments précédents, donc de dépasser leurs extrémités. Pour être visible, un point doit apparaître plus excentré (plus à droite ou plus à gauche respectivement) que le plus excentré des points qui le précèdent vers le front. La figure 8.13 montre l'ordre d'affichage pour le demi-modèle de droite, ainsi que le cas d'un point caché.

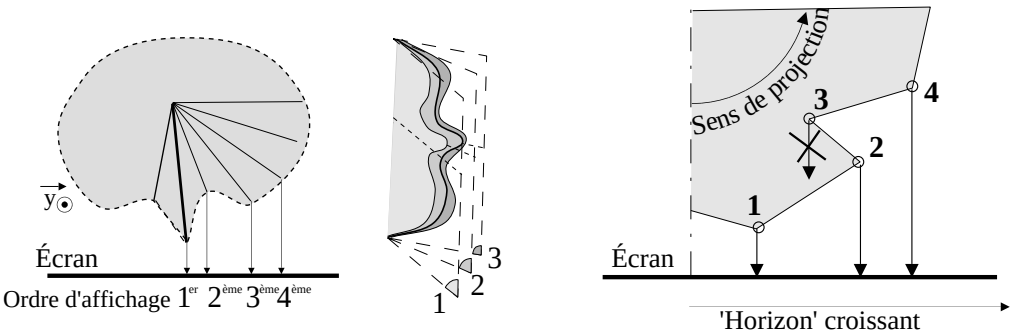


FIG. 8.13 – *Principe d'affichage du demi-coté droit*

8.5.5 Principe de l'implémentation

On implémente cette propriété à l'aide d'un «horizon flottant» : cette variable est initialisée à chaque ligne d'écran avec l'abscisse où se projette le point frontal du modèle. Elle est comparée à l'abscisse de la projection de chacun des autres points du modèle, énumérés pour moitié en tournant dans le sens positif. Lorsque cette abscisse est supérieure, c'est que ce point est visible ; le point est affiché et son abscisse devient la nouvelle valeur de l'horizon. On procède de même pour la moitié gauche du modèle, mais en tournant dans le sens indirect depuis le point frontal. Ce sont des abscisses minimales qui caractérisent les points visibles et commandent la mise à jour de l'horizon flottant, toujours initialisé d'après l'abscisse de la projection du point frontal du modèle.

Prenons le cas de droite de la figure 8.13 : l'horizon flottant h est initialisé à x_1, abscisse de la projection du point 1, qui est visible et affiché. Comme $x_2 > h$, le point 2 est visible et affiché, tandis que h vaudra désormais x_2. Le point 3 n'est pas visible, car $x_3 < h$, donc h reste égal à x_2. Le point 4, visible car $x_4 > h$ est affiché et modifie h...

On peut remarquer qu'on ne se sert pas de la profondeur de chaque point. Cette information n'est pas visible dans l'image, et ne sert pas dans l'hypothèse de perspective faible (la profondeur de l'axe est tenue pour représentative pour tout l'objet).

8.5.6 Interprétation de l'algorithme

L'algorithme présenté peut être vu comme une variante de «l'horizon flottant», une méthode développée il y a de nombreuses années pour effectuer efficacement le rendu de terrain (ou plus généralement d'une fonction de deux variables), de l'avant vers l'arrière [FvDFH95]. Par opposition à l'algorithme du peintre où tous les éléments sont tracés d'arrière en avant, recouvrant éventuellement des parties tracées précédemment, on cherche explicitement les fragments visibles pour ne tracer qu'eux, minimisant le nombre des écritures aux seuls pixels qui contribuent à l'image finale.
Dans le cas d'un visage, notre support n'est pas la base plane du paysage, mais un cylindre. Au lieu de balayer le modèle d'avant en arrière, on le balaie de l'intérieur vers les extérieurs.

8.6 Algorithme de rendu texturé

8.6.1 Notion de segments de texture

Par construction, la texture cylindrique associe un point coloré (pixel) à chaque ligne et angle. Les angles de la forme $k\frac{\pi}{n}$ qui échantillonnent les points du modèle fil de fer correspondent sur la texture à des lignes verticales régulièrement espacées de $\frac{w}{n}$ pixels. Cette association est visible sur la figure 8.14.

À chaque segment horizontal de points du modèle est donc associé un segment horizontal de texture – toujours de $\frac{w}{n}$ pixels de largeur – celui situé à la même altitude et délimité par les angles associés.

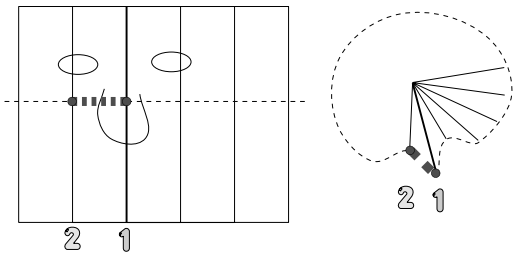


FIG. 8.14 – *Définition des segments de texture*

8.6.2 Principe de l'implémentation

Le segment de texture précédemment défini devra donc remplir à l'écran l'espace (horizontal) entre la projection de ses deux extrémités, calculées comme lors des rendus précédents, à partir des 2 points du modèle auxquels il est associé. Si la plupart du temps le segment de texture est visible en même temps que ses extrémités, ce n'est pas toujours le cas. En effet, le segment peut n'être visible que partiellement – du coté extérieur – dès que son extrémité intérieure est cachée, comme pour CD sur la figure 8.15. On résout ce cas particulier comme le cas général en copiant toujours le segment depuis l'extérieur, et en s'arrêtant une fois l'horizon atteint.

Initialement, $h = x_1$. Comme le point 2 est visible, on commence la copie du segment de texture associé aux points 1 et 2 en partant de x_2. Ses $\frac{w}{n}$pixels de largeur, devraient être «étalés» sur $x_2 - x_1$ pixels de large, et seront effectivement tous visibles, puisque h valait x_1. On met à jour h, qui vaut désormais x_2, et on détecte que 3 n'est pas visible. Il n'y a donc pas de texture à copier, et h reste égal à x_2. Comme x_4 dépasse l'horizon ($x_4 > h$ où $h = x_2$), le point 4 est visible. On commence donc la copie du segment de texture depuis x_4. Ses $\frac{w}{n}$ pixels de large, devraient être «étalés» sur $x_4 - x_3$ pixels de large, mais seuls auront été copiés (sont visibles et seront visualisés) ceux entre x_4 et $h = x_2$.

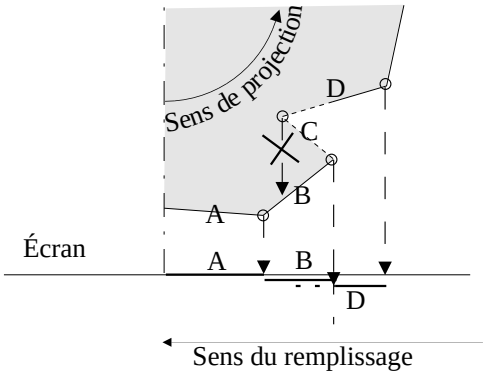


FIG. 8.15 – *Visibilité partielle d'un segment de texture*

8.6.3 Vitesse et complexité relative de l'algorithme

En la comparant à un modèle générique à base de polygones, on pourrait qualifier notre structure de «pré-rasterisée», parce qu'on a réalisé une partie de la transformation **triangle vers écran**, cassant les triangles en segments horizontaux («rasters»). L'écran, le modèle et sa texture partagent un alignement forcé. Ainsi, nos primitives (segments horizontaux, texturés ou non) ont été pré-triées selon la ligne d'écran qu'elles pourraient influencer, diminuant la dimensionnalité du problème de la projection : avec les restrictions sur les mouvements de caméra, on ne peut pas perturber beaucoup le lieu de projection de chacune des primitives. Il n'y a plus besoin de les trier (algorithme du peintre) ou de les comparer de façon semi-globale et indirecte (par un espace mémoire «volumineux» comme un Z-buffer[2]) : une seule passe d'un parcours ordonné permet de résoudre le problème de visibilité à la volée en ne mémorisant qu'une valeur par ligne. On affiche tous les pixels nécessaires, et seulement ceux là, obtenant la même image que celle du modèle polygonal associé, mais par un algorithme moins gourmand.

Du fait de ces mêmes contraintes d'angles de vue, la texture ne subit plus de rotation et sera accédée plus efficacement à travers le cache d'un processeur : dans le cas de segments de texture triangulaires, la localité 2D peut être perturbée par un cache de taille trop petite[3] pour le problème ; c'est un défaut de cache par capacité [LW94a, VL97] (qui ne serait pas arrivé si le cache avait une taille plus grande ou infinie), alors que la localité 1D – donc moins volumineuse – des segments se prêtera mieux à des tailles/niveaux de cache inférieurs.

8.7 Application : 3D finger

Il n'est pas rare de croiser sur le Web des pages personnelles qui incluent une photographie de l'auteur. Par une telle image, on peut espérer être reconnu par quelqu'un qu'on n'a croisé qu'une fois ou qu'on va être amené à rencontrer. Une seule image, fut-elle de face ou de profil, n'est cependant pas suffisante pour illustrer à la fois des informations comme la taille et l'empâtement d'un nez. Plutôt que de recourir à des vues multiples «façon identité judiciaire», il est possible de placer sur le Web un modèle 3D qui pourra être examiné par les visiteurs. L'utilisation d'un format 3D polygonal classique pour cette application peut poser quelques problèmes :

- si l'on ne préjuge pas de la plate-forme et du logiciel utilisé, on s'expose à des risques d'incompatibilité. En choisissant plutôt VRLM, 3D Studio ou Inventor, on privilégierait l'un des navigateurs ou systèmes d'exploitation en handicapant ou négligeant les autres... Les bibliothèques d'objets 3D n'ont pas tranché et proposent généralement chaque fichier dans de nombreux formats.

- performance et délai de chargement seront pénalisants si on utilise un modèle suffisamment précis pour bien modéliser les profils caractéristiques,

- on se heurte à la multiplication des *plug-in* ou à une mauvaise intégration avec le logiciel (il faut sauver puis visualiser le document dans une fenêtre et/ou une application séparée).

2. qu'il faut prendre le temps d'effacer régulièrement dans une implémentation logicielle.

3. même dans les architectures où des caches de niveau 2 ou 3 prennent le relais, il est toujours souhaitable de travailler sur un cache de niveau inférieur, plus rapide mais plus petit, idéalement au niveau 1.

L'utilisation du format proposé ici présente les caractéristiques suivantes :

– les restrictions de rotation n'empêchent pas de visualiser les profils, qui sont un élément clé de la reconnaissance,

– la compacité du modèle 3D est excellente, puisque sa structure cylindrique concentre le codage de l'information essentiellement au niveau des profils verticaux,

– l'algorithme est simple et permet une implémentation rapide et concise [IE] en Java, donc accessible à tout Browser Web qui supporte ce langage.

FIG. 8.16 – *Applet 3D-finger sur le Web*

Au final, le coût total en terme de données transitant sur le réseau est du même ordre que celui d'une image simple de qualité (Cf. tableau 8.7, page 66) et n'handicape donc pas cette utilisation [Eli97].

8.8 Conclusion

Dans ce chapitre, on a proposé une structure simple et compacte qui permet d'encoder spécifiquement des visages, en concentrant l'information sur la texture et les profils. En ayant restreint les conditions de vue à une rotation axiale et aux changements d'échelle, on a pu proposer un algorithme de rendu texturé qui reste performant sur les architectures (matérielles ou logicielles) simples ou dégradées, par exemple des machines grand-public sans accélérateur graphique ni processeur flottant mais aussi pour les machines plus puissantes, dans le cadre du Web et de machines virtuelles multi-plate-formes. Cette propriété est illustrée par l'existence d'une implémentation sous forme d'applet Java pour les navigateurs [Eli97, IE].

8.8 Conclusion

Avec cet algorithme rapide, on est techniquement capable de combiner sur une même image les visages 3D de différentes personnes en temps réel (25 images par seconde pour 4 clones sur un NC/STB). Comment se servir de cette possibilité dans le cadre d'une téléconférence ? Tous les participants doivent ils être visibles ? C'est ce dont traite le chapitre qui suit.

FIG. 8.17 – *Une vue avec plusieurs clones à différentes échelles*

Chapitre 9

Gestion automatique de l'espace virtuel

Comment utiliser et agencer la représentation des participants ? Lorsqu'on n'a qu'un seul interlocuteur, comme avec un visiophone, toute l'image peut lui être dédiée. Mais quelle visualisation adopter lorsqu'on dépasse le dialogue, en permettant à plus de deux personnes de se connecter, que ce soit pour participer ou comme spectateur ? Plusieurs alternatives sont possibles, dont la plus classique consiste à diviser l'écran en surfaces – égales ou non – pour afficher l'image de chacun, vu de face. Cette approche présente cependant plusieurs défauts :

- plus il y a de participants, plus l'imagette les représentant sera imprécise et petite, au point de ne plus être forcément très utile : les expressions deviennent moins visibles, les visages sont moins reconnaissables et il devient difficile de regarder plusieurs personnes à la fois.

- cette juxtaposition en lignes/colonnes ne correspond pas à la réalité, où les gens se parlent le plus souvent de face. Ici, deux personnes en cours de dialogue seront vues par toutes les autres comme étant côte-à-côte, comme si elles se répondaient sans se regarder, en s'ignorant.

Il s'agit pourtant d'une solution communément employée lors de la confrontation télévisée de personnes présentes sur des sites éloignés, car elle ne pose pas de problème technique (il «suffit» de composer les deux signaux vidéos dans une même image). Elle ne privilégie l'image d'aucun des intervenants et permet de voir les expressions et gestes du locuteur en même temps que les réactions de son interlocuteur.

Si par contre le débat se déroule sur un plateau unique, doté de plusieurs caméras, l'usage courant est d'employer un régisseur. Parmi les différents points de vue filmés, ce professionnel choisira à la volée la vue qui sera diffusée aux téléspectateurs. Son travail consiste aussi à piloter les cameramen pour faire en sorte que les images disponibles suivent ou précèdent l'action et restent intéressantes, par exemple en proposant des points de vue différents, par exemple complémentaires si de nombreuses personnes sont présentes. Contrairement à un travail de montage, qui se réaliserait entre le moment du tournage et celui de sa diffusion, la difficulté tient dans le fait que les choix doivent être faits en temps-réel, au risque de commettre des

erreurs (par exemple, ne pas avoir montré un geste d'acquiescement ou de dénégation important pour la compréhension, ou avoir commuté sur une personne au moment où elle arrête de parler, de sorte qu'on ne pourra pas montrer immédiatement celle qui prend la parole...)

9.1 Vers un débat virtuel

Dans notre cadre, où il peut y avoir plusieurs interlocuteurs possibles, dont on espère qu'ils ne parlent pas tous en même temps, on va proposer le service d'une **régie automatisée**. On cherchera à façonner un rendu intéressant du débat, notamment en alternant les vues : voir qui parle, qui est présent ou observer les réactions de l'interlocuteur à qui l'on répond sont des exemples d'informations utiles aux participants comme aux spectateurs. Dans un tel débat virtuel, les visages peuvent tourner, se regarder comme s'ils étaient dans un espace commun. De plus, lorsque la réunion intègre un grand nombre de participants, on peut choisir de n'en montrer que deux ou trois à l'écran à un moment donné, pour optimiser l'utilisation de la surface de l'écran, en proposant des vues détaillées et expressives des intervenants.

9.1.1 Les contraintes et libertés du virtuel

Contrairement à un plateau réel, le virtuel permet de s'affranchir de nombreuses contraintes :

- on n'a pas **l'unité de lieu**, mais on peut la construire et composer à l'écran la scène voulue, par exemple en plaçant les modèles 3D des intervenants en cercle autour d'une table virtuelle,
- on dispose de **caméras virtuelles**, permettant d'obtenir une image du modèle 3D de n'importe quelle personne sous l'angle voulu[1],
- on dispose **d'autant de caméras que nécessaire**, pour obtenir des vues de groupe ou des images individualisées, sans jamais qu'une de ces caméras présente dans le champ d'une autre ne finisse visible à l'image : virtuelles, elles sont transparentes.
- on peut **créer ou déplacer une caméra instantanément**, pour réagir aux événements ou à l'arrivée d'un nouvel intervenant.

Tout en profitant de ces atouts, il est cependant plus que raisonnable de ne pas bouleverser la forme du débat. En effet, comme les films, ceux-ci obéissent à un code de règles, plus ou moins exprimées ou perçues consciemment par le public : par exemple, on n'utilisera pas pour un débat de ralenti ou de flash-back, qui sont plutôt réservés aux événements sportifs ou aux fictions. Inversement, l'alternance de champs ou de contre-champs est bien adaptée aux scènes de dialogues. La forme ne doit pas rendre le fond incompréhensible, mais peut au contraire amener une compréhension plus facile au spectateur.

Prenons un exemple d'une telle règle filmique, **la règle des** 180° **:** selon que la caméra est placée de part ou d'autre d'une ligne imaginaire qui relie deux personnes, leur placement relatif à

1. si l'on réalise le rendu avec un moteur 3D classique, ou avec des restrictions d'angle si l'on utilise le rendu rapide proposé au chapitre 8. Mais sur le principe, la gestion de l'espace virtuel s'applique à tout moteur de rendu.

l'écran (qui apparaît à droite de qui) n'est pas le même. Si la caméra (ou une autre vue) franchit cette ligne, le placement relatif à l'écran s'inversera. Pour conserver une unité apparente, il est d'usage de respecter cette règle des 180° en n'enchaînant que des vues où la caméra est placée du même coté de la ligne d'action. Inversement, si l'on brise cette continuité, on suggère à priori qu'une rupture spatiale ou temporelle a eu lieu.

Il s'agit donc d'une règle pragmatique, tellement employée que même les spectateurs non-initiés qui ignorent son existence décodent inconsciemment pour se construire une représentation mentale de la scène filmée. En pratique, on n'a donc à priori pas intérêt à placer trop de caméras, notamment de part et d'autre de la scène. On peut par exemple réunir les participants sur la moitié seulement de la périphérie d'une table virtuelle, face à la majorité des caméras.

9.1.2 Comportements visuels typiques pour un débat

Voici quelques uns des comportements à reproduire dans l'image synthétique du débat :

- une présentation des participants, par une vue d'ensemble ou une alternance de vues,
- rendre visible une personne qui prend la parole, soit sur un plan d'ensemble soit en plan plus rapproché,
- filmer, en alternant champ et contre-champ, deux personnes qui se répondent, surtout si leurs interventions sont courtes.
- zoomer sur une personne qui parle depuis un certain temps, pour donner l'impression qu'on s'en rapproche,
- ne pas rester sur une personne qui parle depuis trop longtemps, pour diversifier l'image, par exemple en montrant le public ou les autres participants.

Ces règles de comportement sont bien sûr contradictoires, mais peuvent être enchaînées dans le temps, en alternant vues et caméras. La façon dont on séquencera les plans constitue le scénario du débat, et joue fortement sur sa présentation et son intelligibilité.

9.1.3 Comportements sonores

Du fait de l'exploitation relativement récente de formats sonores multicanaux (*Dolby Digital, DTS, Dolby Surround* et *Prologic...*), et plutôt dans des cadres privilégiés (salles de cinéma équipées, *Home Cinema* avec 5 enceintes ou plus), il n'y pas vraiment d'usage, et encore moins d'habitude des auditeurs. On ne peut donc pas vraiment recenser de codes établis pour le son, dans les débats, et s'en inspirer.

Il semble cependant évident qu'on ne doive pas traiter le son comme l'image des participants, qui est parfois masquée : dans le cadre d'une conférence, aucune intervention verbale ne doit être censurée (au risque d'encourir une cacophonie fidèle à celle qui règne effectivement sur le débat) et on ne se permettra donc pas de couper certains canaux sonores (par exemple en modulant à l'excès les volumes selon les distances).

Il n'empêche qu'il faut créer un espace sonore pour que les participants distants puissent être entendus, de la même façon qu'on restitue un espace visuel pour rendre visibles les clones. Parce qu'un simple mixage monophonique des sources audios provenant des micros réels ne permet qu'une restitution pauvre et une intelligibilité limitée, on va plutôt considérer l'espace sonore comme étant lui aussi 3D. Le but est que l'auditeur puisse appréhender finement ce qui se passe dans la scène, dont on sait qu'elle est plus large que celle visualisée à l'écran. Grâce aux techniques de spatialisation du son, on peut envisager de recréer l'impression de la position relative des locuteurs sur l'écran ou en dehors de celui-ci, ou la perception de l'accoustique de la salle (réverbération, présence des aigus ou graves...).

S'il est donc souhaitable de tirer partie du canal sonore, il a cependant fallu imaginer des nouveaux codes adaptés à cet exercice. Parce qu'il était important de les tester et de les faire évoluer d'un point de vue appliqué, cela n'a réellement été possible qu'une fois construit le prototype qui sera présenté dans le chapitre 10.

9.2 Scène virtuelle

On définit un lieu «informatique», pour la visualisation (sonore et visuelle) du débat. En terme d'expérience pour l'utilisateur, cela reste un lieu distant relayé sur son écran, et qui donne l'impression d'être partagé avec les autres participants, puisqu'on y discute en temps-réel.

Cet espace virtuel commun intègre aussi bien **les représentants des intervenants**, que **les capteurs virtuels** qui permettront de les voir et les entendre. Il nous a semblé suffisant d'avoir un placement 2D, c'est à dire où tous les intervenants sont à la même hauteur, permettant aux caméras de les filmer directement, sans vue plongeante par exemple. On peut alors donner une représentation complète de tous ces éléments par une simple vue de dessus (Cf. figure 9.1).

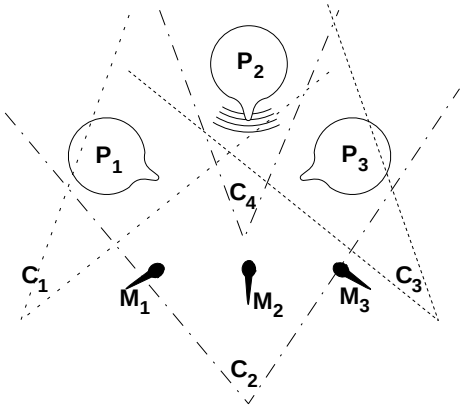


FIG. 9.1 – *Exemple de studio quasi-minimal pour trois participants virtuels*

9.2.1 Modèle des participants

C'est seulement lorsqu'ils sont vus par une caméra virtuelle que les représentants des participants apparaîtront sous forme de clones. Dans la scène virtuelle, ces objets sont caractérisés par leurs positions, orientations et tailles.

9.2.2 Modèle des microphones

À priori, les microphones pourraient être représentés dans la scène par leurs positions et orientations. L'analogie avec le réel pourrait pousser à intégrer des paramètres internes, comme la sensibilité ou la sélectivité, mais il ne faut pas oublier qu'on s'astreint à ne pas modifier l'intelligibilité de toute participation. **Ces microphones représentent en fait des conditions d'écoute, qu'on synthétisera d'après les prises de son réelles**. Leur modélisation exacte dépend donc du rendu sonore que l'on sera capable de réaliser, et sera détaillée au chapitre 10.

9.2.3 Modèle de caméra

On doit modéliser une caméra, si possible dans un formalisme où on pourra lui faire exécuter avec peu de paramètres des mouvements naturels pour le spectateur, c'est à dire du point de vue de l'image capturée. C'est ce que propose le paramétrage de la figure 9.2.

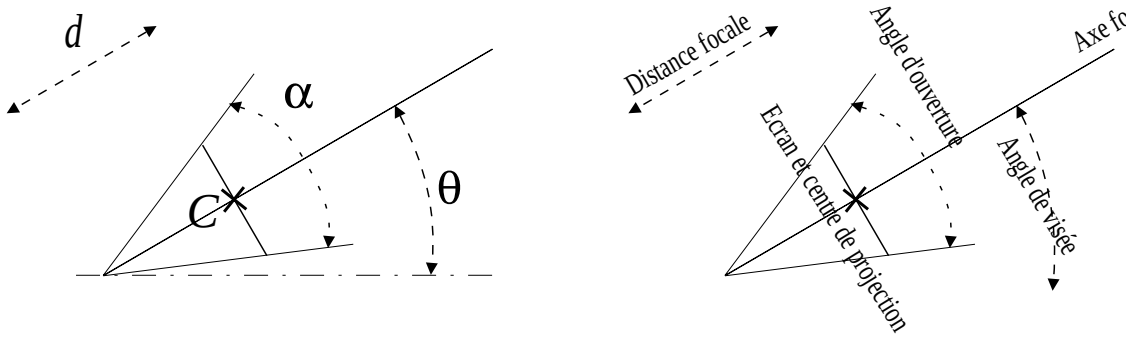


FIG. 9.2 – *Paramètres de pilotage de la caméra*

Avec le formalisme précédent, on réalise facilement :

- **un changement de focale**, pour ouvrir le champ ou donner l'impression qu'on se concentre sur un locuteur, en faisant varier d ou l'angle d'ouverture α,
- **un balayage en rotation**, en modifiant continûment l'orientation θ, par exemple pour montrer les participants, ou recentrer sur une personne située plus à droite ou plus à gauche et qui aurait pris la parole,
- **un travelling**, par la translation du centre C, pour balayer les personnes présentes ou garder un personnage mobile dans le cadre, en utilisant aussi la rotation.

9.2.4 Exemple de scène virtuelle

La figure 9.1 présentait un exemple de placement initial, avec trois personnages, quatre caméras et trois micros. Une telle configuration non symétrique n'est bien sûr pas adaptée à tous les débats avec trois interlocuteurs. Ici, la seule façon de montrer les deux participants à l'extrémité serait d'activer la caméra grand angle. On imagine donc qu'on est plutôt dans la situation où la personne centrale mène l'interview en deux parties, avec l'un puis l'autre de ses invités, ce que confirme l'existence de la caméra C_4. Dans un débat non-virtuel, C_4 apparaîtrait dans le champ de la caméra C_2, aussi utiliserait-on plutôt 3 caméras seulement, en demandant à C_2 de faire des zooms lorsque P_2 prend la parole. On risquerait alors de ne pas pouvoir réagir instantanément à une prise de parole simultanée des deux interlocuteurs.

Mais dans l'état actuel de la description de notre modèle, les caméras sont immobiles, et on ne choisit pas quelle vue sera «diffusée» (synthétisée et affichée à l'écran).

9.3 Principe de l'orchestration

On va faire en sorte que les caméras et l'image affichée sur les postes des conférenciers reproduisent certains scénarios, par exemple en se déplaçant ou en zoomant, et réagissent à ce qui se passe dans le débat.

9.3.1 Les évènements

Différents évènements sont susceptibles d'intervenir lors d'une conférence, comme l'arrivée ou le départ de participants, la mise à disposition d'un document ou une prise de parole.

C'est ce dernier type d'évènements – prise de parole et arrêt – qu'on va utiliser dans la suite pour illustrer les principes de l'orchestration.

On verra dans le chapitre 10 (Réalisation d'un prototype) comment on peut, pratiquement, détecter automatiquement ces évènements. Pour la suite, on suppose que l'on sait à tout moment si une personne est ou n'est pas en train de parler (selon que le dernier évènement reçu est une prise de parole ou un arrêt).

9.3.2 Modèle de scénario

On distingue deux niveaux de scénarios :

- au niveau global, en ce qui concerne la sélection de la vue (caméra), selon des critères préétablis de choix et d'alternances. C'est le **régisseur virtuel**,

- au niveau de **chaque caméra**, selon qu'elle se contente de suivre un interlocuteur privilégié, ou est spécialisée dans des transitions plan large/plan rapproché ou pivote pour suivre les prises de parole. À tout moment, chaque caméra applique son scénario, faisant de son mieux pour être celle qui sera sélectionnée. Elles **sont en concurrence continue, et seule la caméra élue calculera une image virtuelle de la scène.**

On va piloter les caméras indirectement, en définissant des **types de caméras**, conçues pour reconnaître des comportements recensés à l'avance, s'attribuant une note selon qu'elles jugent que l'image qu'elles observent réalise plus ou moins cette hypothèse. Ces caméras sont aussi capables d'ajuster leur paramètres internes (zoom, orientation...) pour réagir dans le temps aux événements qu'elles observent, en cherchant à maximiser la note qu'elles se donnent.

Pondérées, ces notes seront utilisées par le régisseur automatique pour choisir la vue qui sera diffusée. Le régisseur automatique possède lui aussi des comportements préétablis, pour réagir dans le temps à des scénarios répertoriés, ou créer des enchaînements types. Par exemple, il lui incombe d'éviter les changements de vue intempestifs, mais aussi d'en introduire lorsque le débit d'évènements baisse, pour maintenir l'attention des spectateurs.

Le régisseur «publie» certains de ses résultats, de sorte qu'ils sont accessibles aux caméras pour leur évaluation. Ainsi, un jeton unique permet à chacune des caméras de savoir si elle est (ou n'est pas) la caméra sélectionnée pour faire la diffusion. Il relaye aussi un état de visibilité de chaque participant, qui est en fait publié par la caméra sélectionnée.

9.4 Autogestion des caméras virtuelles

Illustrons tout d'abord le fait que les critères de notation dépendent de la situation rencontrée, même pour les paramètres les plus simples :

- **le nombre de participants visibles :** s'il est élevé, la caméra propose une image qui est intéressante comme vue d'ensemble, mais ne permet pas forcément de juger des expressions du locuteur ou d'auditeurs privilégiés (celui qui avait posé la question et provoqué la réponse en cours).

- **un visage visible de face** est avantageux s'il s'agit d'un locuteur. Lors d'un «ping-pong» verbal, il vaut mieux deux vues de coté, ou une alternance de champs et de contre-champs.

Puisqu'une discussion générique n'est pas possible, on va aborder la gestion des caméras, type par type, par l'exemple des quelques classes indispensables qui ont été créées.

9.4.1 Types de caméras

Toutes les caméras virtuelles sont des objets qui héritent des paramètres physiques de la visualisation (focale, angle de visée...) mais chaque sous-classe implémente une stratégie différente pour se noter et maximiser l'intérêt de ce qu'elle observe.

On peut à priori créer autant de types qu'on est capable d'isoler de comportements intéressants, et leur réalisation sous forme d'automates d'états finis, avec des transitions déterministes ou probabilistes ne pose pas de problème particulier. Il est bien sûr souhaitable de modéliser des comportements physiques, par exemple des changements de focale avec des vitesses raisonnables et des déplacements continus.

Caméra grand-champ

Cette caméra est complètement immobile (elle est une valeur refuge lorsque le débat s'anime trop, ou pour ouvrir la diffusion). La note qu'elle s'attribue dépend du nombre de personnes visibles, s'ils sont à peu près de face (mais la contrainte est assez faible, puisqu'on souhaite avoir plusieurs personnes) et/ou en train de parler ainsi que de leur taille apparente à l'écran.

Caméra individualisée

Son but est de ne pas perdre la vue de face d'une personne donnée (elle n'est pas spécifiée explicitement, c'est celle qui sera la mieux centrée lors de l'initialisation). Une caméra de ce type effectue deux types de mouvements : des rotations pour centrer parfaitement sa cible (seulement au début si celle-ci ne se déplace pas) et des changements de sa focale pour se rapprocher du locuteur lorsque celui-ci parle/est visible depuis longtemps. Selon son ouverture initiale, elle peut ressembler à une caméra grand-champ, mais saura zoomer sur le personnage central. Lorsqu'elle est en plan rapproché et perd le jeton, elle ne repasse pas immédiatement en plan large avant quelques secondes. Ainsi, elle a une chance de regagner le jeton et de proposer la même vue, montrant les réactions ou de la suite du discours. Ne sachant pas pourquoi, en termes sémantiques, elle a perdu le jeton (une autre caméra couvrait une intervention, ou parce que la vue fixe s'éternisait), il n'y a aucune raison qu'elle ne continue pas à faire son suivi au mieux. D'une façon générale, le gain ou la perte du jeton sont un non-événement.

Caméra champ/contre-champ

Elle n'est pas très différente de la caméra individualisée, mais sa cible est constituée de deux personnes. Dans son critère de notation, il est considéré comme positif de voir de dos quelqu'un qui est intervenu récemment (ou parle encore), si la personne vue de face a la parole. Elle se note défavorablement si aucune des personnes qu'elle peut montrer n'a un état de visibilité à 1, ce qui favorisera sa prise de jeton sur une transition en continuité.

Caméra de balayage

Il s'agit d'une variante de la caméra grand-champ qui s'autorise des balayages assez lents en rotation. Initialisée par exemple sur trois personnes (toujours de façon non explicite), son but est de pivoter pour maximiser le nombre de personnes qui parlent ou ont parlé récemment. Elle générera donc de bonnes notes dans un débat assez calme où la parole passe de voisins en voisins.

9.5 Modélisation d'un régisseur automatique

Il est lui aussi réalisé à l'aide d'un automate d'états finis, et dispose d'une mémoire pour assurer la cohérence de son comportement.

L'une de ses premières tâches est de relativiser les notes que se donnent les caméras :

- **dans l'absolu**, parce que des caméras de types différents n'établissent pas forcément des notes qui sont directement comparables. Par exemple, une classe de caméra peut générer des notes en moyenne plus hautes ou plus basses qu'une autre classe. En surveillant la répartition statistique de chaque classe/caméra, il possède les données pour réajuster l'intérêt que s'attribuent les caméras.
- **dans le temps**, parce qu'une vue intéressante ne l'est plus forcément après avoir occupé l'écran quelques secondes. Ainsi, lorsqu'une caméra est «élue», le régisseur prévoit une pénalité qui fait que sa note baisse continûment, assurant qu'une autre vue va l'évincer,
- **dans la forme**, car il ne faudrait pas enchaîner des changements de prises de vues avec un trop fort rythme, en attendant qu'une caméra spécialisée dans les plans larges se fasse connaître par exemple. Dans ce but, le régisseur maintient une variable correspondant à la durée minimale de présence d'une vue. Ce délai de garde minimal est susceptible d'être rallongé jusqu'à dix secondes ou plus si les deux ou trois derniers enchaînements étaient trop rapprochés, et redescend à une ou deux secondes lorsqu'il n'a pas servi depuis quelque temps. Dans le cas où l'on a repositionné le délai de garde au maximum, on impose la commutation vers la mieux notée des caméras du type «plan d'ensemble».

La vue qui obtient la meilleur note pondérée est donc choisie, si le délai de garde minimum est écoulé. C'est désormais la caméra associée qui gérera la production de l'image de la scène, au moins pour la nouvelle valeur du délai de garde, et jusqu'à ce qu'un nouvel événement favorise la notation d'une autre caméra ou que le handicap qui croît avec sa durée de présence ne lui fasse perdre la première place.

En pratique, le régisseur dépouille donc les notes que s'attribuent les caméras et agit comme un «zappeur» automatisé. Respectant des règles préétablies, mais soumis à un flux aléatoire d'évènements, il génère une alternance de vue suffisamment complexe pour ne pas sembler répétitive et maintenir plus longtemps l'attention du spectateur. Lorsque surviennent des évènements importants, par exemple une prise de parole, les probabilités restent fortes pour qu'une vue du locuteur, individualisée plutôt que dans un groupe, apparaisse quasi-immédiatement, sauf débat trop agité.

9.6 Critiques et perspectives

Par rapport à un régisseur humain, on dispose de quelques handicaps :

- il n'est pas possible de savoir qu'une phrase ou une intervention est finie, puisqu'on ne comprend pas le sens du discours. Ainsi, on risque de changer la vue vers une personne qui va s'arrêter de parler, ce que ne ferait pas un bon régisseur.
- lorsqu'une personne en nomme une autre, un régisseur humain peut préparer ou activer la transition, en pilotant une caméra vers une vue d'ensemble ou un rapprochement. Sans intelligence du discours, le régisseur automatique ne montrera probablement pas la personne interpellée avant qu'elle ne commence à répondre.

Notre modèle ne réalise pas :

- d'initialisation automatique et intelligente de la scène, en terme de placement des participants ou des caméras. En fonction du nombre de participants, on se contente de choisir une organisation type, avec des caméras pré-positionnées et de choisir l'une des vues d'ensemble comme caméra initiale.

- de commande directe des caméras par le régisseur. Quoiqu'autonomes, elles sont conscientes des choix qu'il a réalisés, par exemple d'être la vue sélectionnée (utile pour s'interdire des raccourcis de déplacement non physiques). Par ce biais, on handicape la notation des caméras qui montrent une vue équivalente, et l'on dope les vues complémentaires : les participants apparaissant de face dans la vue sélectionnée seraient moins bien notés que d'habitude dans les autres vues par exemple. Mais tout cela n'est pas une stratégie où les transitions se préparent à l'avance en pilotant les caméras pour créer une situation ou être prêt si elle survient, seulement une heuristique au coup par coup.

- de comportement sur les modèles. On pourrait penser qu'un clone qui ne parle pas puisse orienter sa tête vers le nouvel interlocuteur (s'il est unique). Cela anime la scène et fournit un indice visuel sur localisation de la personne qui prend la parole.

Les deux derniers points mériteraient sûrement d'être analysés dans un protocole de test, avec l'aide de professionnels de la télévision, de psychologues ou d'ergonomes, plutôt que de continuer à être évalués à titre personnel. Le développement d'autres types de caméras intelligentes devrait lui aussi passer par cette étape.

En comparaison avec le *Virtual Cinematographer* ([HCS96], résumé page 53), l'approche de régisseur automatique présentée ici est très différente en ce que :

- on ne cherche pas à tout moment à déplacer les caméras ou les participants pour forcer un cadrage préétabli ; les bons (ou moins bons) cadrages sont la conséquence des positions initiales, puisque les intervenants ne bougent pas. En ce sens, leur approche pourrait servir ici à initialiser plus automatiquement les placements.

- on découple bien mieux les caméras et le régisseur : au niveau de l'écriture des modules caméras, puisqu'ils sont écrits sans avoir à connaître ceux avec lesquels ils seront en concurrence ; on les découple aussi au niveau de l'utilisation puisqu'on peut rajouter toute caméra dans la scène, sans rien avoir à modifier ou reprogrammer : le régisseur l'utilisera si la vue qu'elle propose rend le débat plus intéressant.

9.6.1 Vues synthétiques et vues réelles

On a proposé des exemples de notation des vues disponibles depuis les caméras virtuelles. Mais il est aussi possible de construire d'autres types d'objets jouant le même rôle qu'une caméra, c'est à dire se notant et diffusant une représentation (animée) de la scène, mais où la vue ne serait pas synthétique. De la vidéo, ou même un document statique (schéma, page web...) sont envisageables, pour peu qu'on puisse les noter.

Par exemple, chaque machine équipée d'une caméra réelle pourrait envoyer sa note sur le réseau et ne broadcaster la vidéo (sur un réseau comme le *Backbone*) que lorsqu'elle a été élue par le régisseur automatique. C'est directement la machine reliée à la caméra qui évalue l'intérêt de ce qu'elle observe : une piste son révélant qu'il est en train de parler, le nombre de pixels de teinte chair qui sont détectés, ou la norme de différence entre deux images successives pourraient être des indicateurs de ce que la personne parle, en suivant la conférence, sans trop bouger (juste les lèvres et les expressions faciales).

On disposerait alors d'une vidéo très détaillée de chaque participant (utile pour apprécier fidèlement en plein écran les expressions du locuteur, ou lire sur les lèvres en bénéficiant de l'image de la langue et des dents) sans payer le prix de flux vidéo multiples, avec des raccords utilisant les vues synthétiques lorsque le besoin d'un plan d'ensemble ou non disponible se fait sentir.

9.6.2 Réseau de caméras

Sur un vrai plateau, le nombre de caméras est limité, mais elles s'adaptent en changeant de comportement. Confronté à un trop grand nombre de caméras, les choix du régisseur peuvent être moins bons (trop de notes trop proches de vues trop équivalentes). Il semble donc souhaitable d'avoir, pour des conférences de plus de cinq personnes par exemple, une gestion moins quantitative des caméras, avec un nombre maximum restreint, et qui évite de créer toutes les caméras individualisées, filmant les couples, les trios etc.

Une approche possible serait d'autoriser les caméras à se dupliquer puis à muter. Chaque instance de caméra peut s'évaluer suivant les critères des autres classes, et si la note obtenue est intéressante (pendant plusieurs images), décider de se dupliquer pour augmenter le pool disponible. La caméra clone adopterait alors le nouveau comportement, donnant une chance au régisseur de la sélectionner. Les instances de caméras non-utilisées finiraient par être détruites (et laisser une place libre pour une nouvelle duplication). Pour éviter de sacrifier la couverture de l'ensemble de la scène, les caméras initiales dont la caméra globale devraient probablement être persistantes. Pour éviter la prolifération de caméras équivalentes (regardant toutes plus ou moins le même sous-ensemble de personnes), il semble souhaitable d'introduire une distance sur l'espace des caméras, et donc un codage, probablement plutôt sur ce qu'elles observent que sur leurs paramètres, de sorte que le problème ressemblerait à un problème de quantisation/partitionnement (comme celui de réduction de palette des images).

9.6.3 Conclusion

Sans aucune des nombreuses extensions proposées ou évoquées dans cette conclusion, la gestion automatique de l'espace virtuel telle qu'on l'a proposée est parfaitement fonctionnelle. Elle constitue d'ailleurs une brique capitale – en ce sens qu'elle transforme profondément le module de rendu des clones en une expérience bien plus intéressante pour l'utilisateur ou le spectateur – du prototype de communication qui a été construit et est exposé dans le chapitre qui suit.

Chapitre 10

Un prototype de communication

Il semblait capital de pouvoir tester le type de communication qui était possible avec les outils et principes développés. Mais l'expérience doit rester interactive pour correspondre à la communication voulue. En ce sens, une simulation ou un rendu en différé n'en sont pas représentatifs. Pire, on ne justifie plus la perte de qualité, par rapport aux images de synthèse ou aux films d'animation qu'on a l'habitude de voir, ni le montage automatique des vues. En revanche, comme on vise une application d'évaluation, on s'est autorisé divers compromis matériels.

On souhaite donc que de vrais intervenants puissent se couper la parole, et réagir au plus tôt à ce qu'ils entendent ou voient. Ceci est réalisable à l'aide de stations connectées, mais plus facilement avec un réseau performant[1]. Chacune des machines correspond à un intervenant et jouera le double rôle de serveur et de client de communication vis-à-vis des autres. Elle échantillonnera le son et les interventions de l'utilisateur local pour les envoyer sur le réseau. Lorsqu'elle dispose d'une caméra, elle pourrait aussi participer en émettant de la vidéo aux autres postes, comme on le verra au chapitre 11. En réception, chaque machine effectue la synthèse et le rendu en fonction des événements détectés, et éventuellement selon les préférences de l'utilisateur (s'il choisit de jouer le rôle du régisseur).

Le prototype, réalisé sur un *Network Computer*[2] comporte notamment :

- la gestion des caméras virtuelles et du régisseur automatique présentés au chapitre 9,
- le rendu de clones 3D, avec l'algorithme rapide proposé au chapitre 8,
- un module de communication réseau, basé sur les *sockets*, qui reçoit les paquets sonores émis par les autres participants,
- un module de rendu sonore spatialisé, qui synthétise l'impression d'une direction de provenance (localisation) pour quatre flux audio monophoniques indépendants,

1. ATM ou un Éthernet local par exemple. Ce premier compromis autorise le *broadcast* du son entre les clients, garantit dans la pratique des délais raisonnables aux transmissions sans qu'il soit nécessaire de recourir à des protocoles élaborés (des familles RTP ou GSM par exemple) pour gérer la compression et la robustesse aux perte de paquets.

2. pas de carte graphique 3D, ni de coprocesseur flottant ou de DSP, ni de mémoire-cache de niveau 2.

10.1 Gestion du son

Les problèmes du rendu et de l'orchestration des vues ont déjà été traités dans les deux chapitres précédents. Les nouveaux aspects introduits lors du prototype sont donc tous liés au son. Une fois digitalisé et transmis par IP vers les correspondants, le message sonore va devoir être traité pour s'intégrer à l'espace virtuel et être restitué de façon utile, en aidant les participants à vivre et comprendre cette communication à distance.

10.1.1 Détection visuelle de la parole

Lors du développement incrémental du prototype, une première utilisation de type «téléphone augmenté» pour deux participants a très vite pu être testé : sans régisseur automatique, un seul clone était affiché, avec un rendu monophonique de la piste sonore reçue par IP depuis l'autre machine. Dans cette configuration, avec un message sonore en même temps qu'une vue fixe sur un visage statique, il ne devrait pas y avoir d'ambiguïté sur l'identité de la personne qui parle. Pourtant, on a l'impression que le message est immatériel, qu'il n'est pas associé à la personne que l'on voit car sur l'image ses lèvres sont immobiles.

Dans de telles conditions, il ne servirait donc à rien de vouloir réunir plusieurs participants virtuels sur la même image, ni même de spatialiser le son : **il faut un indice visuel de la personne qui parle**, comme quand les lèvres bougent (dans la réalité ou sur une vidéo).

Pour animer les lèvres de façon suffisamment synchronisée, on va utiliser le signal audio. Si on note $A[t]$ son amplitude (discrétisée dans le temps, parce qu'on a un signal échantillonné tous les Δt), l'énergie moyenne du signal, pour n échantillons successifs vaudra :

$$E(n, t_0) = \frac{1}{n} \sum_{t=t_0}^{t=t_0+n.\Delta t} (A[t])^2 \tag{10.1}$$

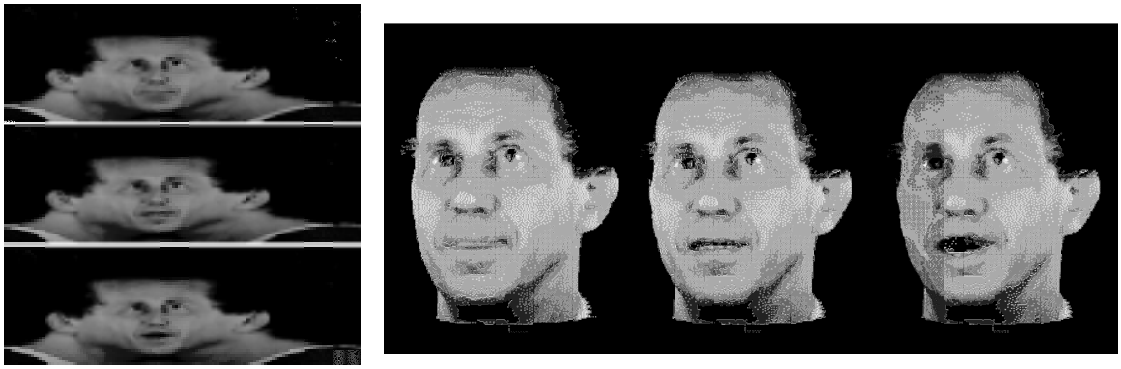
C'est en effectuant ce calcul avec tous les échantillons audios reçus depuis la dernière image calculée qu'on estime quantitativement l'activité sonore de chacun des participants, ce qui va permettre d'indiquer visuellement qui est en train de parler.

Approche vidéo-réaliste

Une première approche a été de vouloir recréer une image de lèvres animées (doublage visuel). On a utilisé trois versions de la texture de chaque clone, pour que la bouche apparaisse respectivement fermée, entre-ouverte et ouverte, comme par exemple sur la figure 10.1. Si l'on utilise l'énergie (avec des seuils) pour choisir la texture qui apparaîtra, le mouvement semblera très mécanique. Il vaut mieux, toujours selon la valeur de l'énergie, choisir deux des textures en les mélangeant progressivement.

Quoique visuellement plaisante, cette solution pose plusieurs problèmes de fidélité :

- on peut avoir utilisé un logiciel de retouche ou avoir incrusté des bouches génériques pour obtenir ces trois textures,
- les mouvements construits ne sont pas ceux de la personne originale, ou pas ceux qu'elles a articulés au moment où elle a prononcé ce message. Même s'ils ont l'air physiologiquement



Avec trois textures différentes, obtenues lors de l'acquisition du clone ou comme ici, par retouche d'images, on peut créer une animation du clone statique et donner l'indice visuel qu'il est en train de parler.

FIG. 10.1 – *Trois textures pour animer la bouche*

réalistes, ils sont trop impersonnels. Jamais un correspondant ne semblera sourire par exemple.

Le spectateur, à qui l'on offre une représentation de trop bonne qualité (entre photo et vidéo-réalisme), sans transmettre l'humanité de la personne qui est représentée est donc trompé. Il risque alors de se faire une idée fausse d'une personne qu'il ne connaissait pas, ou de trouver son humeur inhabituellement différente.

Approche symbolique

Finalement, plutôt que cet artifice ou une méthode encore plus sophistiquée de clone parlant[3] on a choisi de **montrer** visiblement que l'animation des lèvres n'était que symbolique. De même qu'un dessin animé n'utilise qu'un petit nombre de postures pour les lèvres (et qui ne sont pas vidéo-réalistes, ce qui contribue à leur acceptation), on choisit d'introduire des indices visuels de la parole qui soient indubitablement artificiels : c'est un rectangle noir, plus large que haut, qu'on incruste sur la texture du clone, recouvrant d'autant plus la bouche que l'énergie sonore est grande.

Ainsi, sans leurrer le spectateur, on indique visuellement qu'une personne est en train de parler. Avec un prototype qui permettait d'avoir plusieurs interlocuteurs distants – toujours sans son spatialisé et avec une seule caméra fixe en grand angle – on a pu vérifier que l'image et le son étaient suffisamment corrélés (puisque l'énergie pilote l'animation) pour qu'il soit généralement possible d'identifier sans les confondre les clones de deux personnes qui parlaient en même temps.

3. dont on ne renie pas l'intérêt, ne serait-ce que pour des avatars synthétiques, dans des bornes de vente par exemple ou pour des trucages cinématographiques [BCS97].

10.1.2 Détection des évènements de prise de parole

De même que le spectateur doit savoir qui parle, le régisseur présenté lors du chapitre précédent a besoin des évènements de prise de parole. Pour cela, on va utiliser la mesure d'énergie sonore réalisée par l'équation 10.1. En plus de bruits de fonds naturels (paroles lointaines), les environnements informatiques sont souvent très bruyants (disques durs et ventilateurs) et il est souhaitable de ne pas confondre ces bruits de fond, pas toujours très réguliers, avec la voix du locuteur, si possible sans introduire de modèle coûteux.

La détection des évènements de prise de parole sera faite de façon robuste en utilisant un cycle d'hystérésis avec deux états, pour coder si la personne est ou n'est pas en train de parler. Les transitions ne se font pas suivant des conditions symétriques :

- quand on est dans l'état «inactif», il suffit que l'énergie dépasse un seuil minimal S_2 pour qu'on passe dans l'état «actif».
- quand on est dans l'état «actif», il faut que l'énergie sonore reste suffisamment longtemps en dessous d'un seuil $S_1 < S_2$ pour rejoindre l'état «inactif».

Ainsi, comme sur la figure 10.2, on détecte avec un délai quasiment nul les prises de paroles, tandis que les fins de parole sont légèrement retardées. Notons que ce dernier point n'est pas gênant dans notre application :

- puisque le son est toujours transmis et restitué, on ne brise pas la continuité de l'ambiance sonore, et on n'empêche personne d'être entendu,
- on risque moins de changer de vue en croyant que la prise de parole était finie, alors qu'il ne s'agit que d'un silence temporaire,
- si un participant a détecté plus vite que la machine que la prise de parole était terminée (d'après le sens ou l'intonation de la phrase par exemple), et qu'il intervient, créant un nouvel évènement sonore, une caméra a toutes les chances de le montrer immédiatement.

Il n'est donc pas utile de déterminer précisément le seuil du bruit (il est suffisant que S_1 le dépasse franchement), de sorte que ce réglage n'a généralement pas besoin d'être réajusté d'une séance à l'autre. En pratique, des délais de l'ordre d'une seconde et demi pour t_1 semblaient adaptés.

10.1.3 Notions d'espace sonore scénarisé

On souhaite apporter aux spectateurs/auditeurs une plus grande richesse et une meilleure facilité en terme de compréhension. Grâce à des altérations de ce son pour l'enrichir d'une information de provenance qui sera perçue et interprétée par l'auditeur, il pourrait par exemple appréhender plus rapidement le nombre de locuteurs simultanés ou bien associer plus facilement les positions, les voix et les visages ou encore comprendre les interventions hors-champs... En ce sens, le son apportera bien une nouvelle dimension au spectateur.

Cependant, il n'y a pas de code ou d'usage aussi bien établis que pour la gestion des caméras, et il ne semblait à priori pas évident que telle association son/image devait être privilégiée à

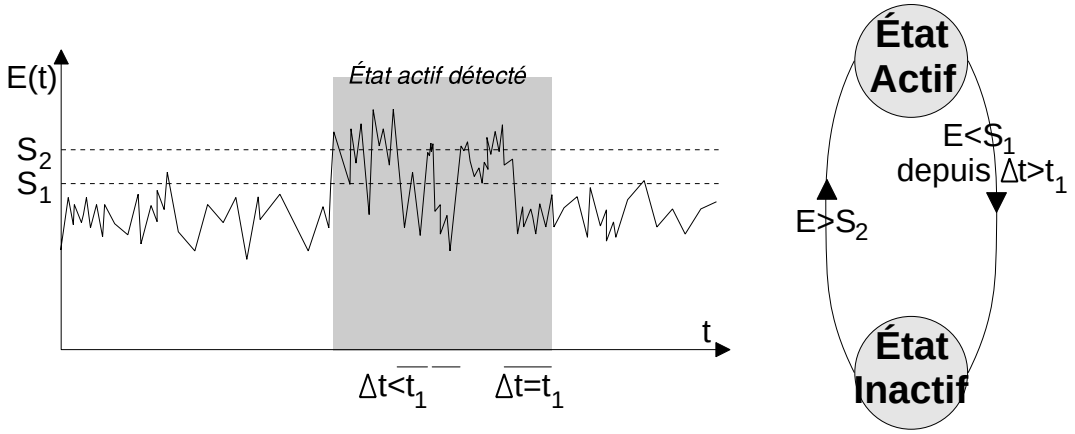


FIG. 10.2 – *Détection des débuts et fins de parole*

d'autres. C'est donc l'un des intérêts de ce prototype que d'avoir pu en tester plusieurs et juger comment, en tant qu'auditeur, on comprend ou cherche à interpréter ces nouveaux codes audiovisuels.

Ce sont ces diverses stratégies de lien entre le canal sonore et l'affichage que cette section va discuter, une fois explicitées les modalités de la restitution du son aux participants.

10.1.4 Cadre de la restitution du son

Du fait du coût en terme de temps de calcul, on ne pouvait pas choisir un spatialisateur riche d'effets. Avec l'hypothèse d'une machine peu puissante et peu chère du style NC ou STB, les quelques cartes sonores dites «3D» qui font leur apparition et implémentent en matériel les algorithmes de convolution et de mixage nécessaires, sont elles-aussi exclues. De toute façon, les effets qui ne troubleraient pas l'intelligibilité sont en fait assez limités : échos et altération des volumes sont à proscrire, car ils pourraient favoriser ou déservir certains locuteurs. Sans oreillette pour le retour sonore, il ne serait d'ailleurs pas possible de se rendre compte qu'on n'est pas entendu des autres comme on le présumait.

On a donc choisi de se restreindre à une **localisation** des sources sonores, c'est à dire faire en sorte qu'on perçoive une direction de provenance du son. Ce résultat peut être atteint, dans certaines limites, en utilisant un matériel simple et courant : deux haut-parleurs habituellement utilisés pour la stéréophonie, avec un auditeur censé être placé à égale distance des deux.

Un modèle perceptif sommaire de propagation et de modification du son dans l'air jusqu'aux oreilles de l'auditeur est présenté dans l'Annexe C. Il propose de restituer sur haut-parleurs [4] un

4. mais pas pour l'écoute au casque ! Pour ce cas précis, il faudrait intégrer les autres modifications qui ont lieu entre la source et leur perception. La plupart varient en fonction de l'auditeur, parce qu'elles interviennent au niveau de ses épaules ou du pavillon des oreilles par exemple. Même si on résout cette difficulté (en mesurant ou approchant ces paramètres individuels, appelés HRTF) le casque n'est pas aussi plaisant qu'il peut sembler: certes il élimine les risques d'échos ou de retours avec la prise de son, mais lorsque l'auditeur tourne ou remue

déphasage et une différence de volume relatifs pour les canaux gauche et droit de ce système de restitution, en fonction de l'angle de perception qu'on veut simuler.

Sans correspondre parfaitement aux «réalités» physiques et physiologiques, ce modèle sera suffisant pour donner une impression de position (au sens direction) des sources sonores, et ce qui est plus important, **des impressions de positions relatives et de mouvements** éventuels (sources tournant pour se rapprocher de l'une ou l'autre oreille ou être perçues de face).

10.1.5 Rappel sur les codes vidéos

Le code classique pour l'image est que la caméra se substitue aux yeux du spectateur. Certes, ce n'est pas lui qui décide de ce qu'il va voir, mais il comprend ce qui se passe, ce à quoi il assiste. Ainsi, lorsque la continuité visuelle se rompt, on diagnostiquera – plus ou moins consciemment d'après le contexte narratif, les objets et personnes vues dans le champ ou encore la (dis)continuité de la bande son – qu'il s'agit d'un changement de point de vue ou de lieu, et si l'unité de temps est préservée.

Dans le cadre d'un débat en temps réel, la continuité temporelle ne sera jamais brisée. Dès lors, on assiste à des changements de point de vue, qui nous seront imposés par le régisseur (réel ou virtuel), et sauf fréquence trop grande, cela n'est pas gênant, voire ne sera pas remarqué consciemment.

Pourtant, si l'on devait faire abstraction de la notion de caméra et qu'on se force à chercher ce qui pourrait instantanément substituer tout ce que l'on percevait dans notre champ visuel, il faudrait probablement conclure que nous avons été téléportés (ce qui n'est pas à proprement parler une expérience commune[5]...).

Il ne fait cependant aucun doute que ce code télévisuel n'est pas interprété comme tel, parce qu'il est parfaitement intégré par les spectateurs.

10.1.6 Micros et caméras confondus

Dès lors, il semble tentant de généraliser ce code au cas du son : la caméra ne sera pas seulement les yeux du spectateurs, elle sera aussi ses oreilles. En termes physiques, cela signifie que le micro (virtuel) qui enregistre le son doit être placé et orienté le plus identiquement possible à la caméra. D'un point de vue symétrique, celui du rendu, la source sonore reste «centrée» autour de l'image.

Avec cette hypothèse, une personne qui est vue au centre de l'écran sera entendue au même endroit. Si l'un bruit parvient de la gauche de l'écran, c'est qu'il provient de la gauche de ce qui est visible (provenant donc de la droite du locuteur s'il est vu de face...).

Lorsqu'une vue différente amènera un autre locuteur au centre de l'image, il sera à son tour entendu au centre de l'écran.

---

la tête, il «emporte» la scène audio avec lui (sauf capteur d'orientation et correction à la volée...), ce qui brise l'illusion que les sources étaient attachées au moniteur et à son image

5. on peut toutefois se demander si la nécessité physiologique de cligner des yeux, de façon inconsciente même lorsqu'on est en mouvement rapide, n'explique pas qu'on tolère ces ruptures de champs de vision.

Ainsi décrit, cela semble cohérent et agréable. S'il n'y avait qu'une seule caméra qui pivote et se déplace dans la scène, l'expérience serait bien analogue à celle d'un auditeur/spectateur tournant sa tête, c'est-à-dire ses yeux et ses oreilles simultanément. Le problème vient de ce que l'on passe de temps en temps d'une vue de caméra à une autre. De même qu'il y a une discontinuité visuelle induite à l'écran, il y a une discontinuité de position dans la restitution du son (héritée de celle de la prise de son virtuelle).

La discontinuité visuelle est généralement perceptible sans effort : les personnages vus ne sont plus les mêmes, ou plus à la même place. Au pire, on se repère d'après le décor. Pour le son, deux cas se présentent :

- il n'y avait pas de son dont on puisse repérer le changement de position (faible différence de localisation ou pas de son). Ce n'est que plus tard qu'on se rendra peut-être compte du changement de point d'audition.

- quelqu'un parlait et la suite de son message semble provenir d'ailleurs, comme s'il s'était instantanément téléporté. Cette sensation étrange est en pratique assez stressante, car on ne comprend initialement pas ce qui se passe, comme si le cerveau hésitait à valider l'hypothèse que c'est un nouveau locuteur. En fait, le délai de prise de conscience semble ne pas être le même pour la discontinuité visuelle et la rupture auditive, ce qui n'aide pas à percevoir la corrélation entre les deux événements. Il n'est pas clair que l'on pourrait s'habituer à ce stress, en faisant l'effort d'apprendre à corréler les deux informations sans perdre la concentration sur le discours et son sens. Il ne serait pas étonnant qu'à défaut on prenne l'habitude de ne plus tenir compte de cette information initialement voulue comme supplémentaire...

Contrairement à l'image pour laquelle le paysage visuel constitue une information qui renseignera du changement, le son pose problème. On pourrait bien sûr songer à rajouter des sources ponctuelles d'ambiance pour créer un décor sonore, mais cette solution n'est pas idéale dans le cadre d'un débat où l'intelligibilité doit primer.

10.1.7 Micros statiques

Une solution pour éviter les changements de «points de vue» auditifs serait de n'utiliser qu'une seule direction de prise de son durant tout le débat, donc sans aucun lien avec quelque caméra que ce soit : telle personne serait toujours perçue au centre, telle autre à 30° sur la droite... Effectivement, une telle solution élimine l'impression de téléportation auditive et restitue une information supplémentaire sur la localisation des intervenants – et donc leur identification – qu'ils soient ou non dans le champ. L'auditeur peut se construire une image de la scène en quelques instants : c'est donc à la fois moins d'ambiguïté, mais aussi moins d'informations, puisque si l'on ne regarde pas l'écran, on ne sait pas que le point de vue a changé. On perd donc la possibilité de rattraper l'attention d'un auditeur qui n'était plus spectateur.

L'inconvénient majeur vient de ce qu'on a privilégié une direction particulière et décorrélé complètement l'image et le son. Ainsi on peut observer une vue de face d'une personne qu'on entend pleine gauche ou pleine droite... Cela ne correspond pas du tout à une expérience

de communication physique et réelle. Au point que les mouvements de lèvres vus, pourtant parfaitement synchronisés, persistent sembler ne pas correspondre du tout au son entendu. Le spectateur-auditeur a l'impression d'entendre une personne et d'en voir parler (sans entendre de voix) une autre.

De plus, la différence prolongée de volume sonore entre les deux oreilles est suffisamment désagréable pour provoquer une gêne allant jusqu'aux migraines, ce qui condamne ce principe même pour des émissions radiophoniques en stéréo.

10.1.8 Micros mobiles

Dans la première approche, voix et image (des lèvres) étaient toujours corrélées, mais la discontinuité de position du son était stressante. Pourtant, c'était elle qui apportait le supplément d'information nécessaire pour comprendre ou entendre les changements de caméra, et les positions des interlocuteurs.

L'idée est de remplacer la discontinuité de position par une discontinuité sur les vitesses : entre les deux directions de prise de son, on va prendre le temps (1,5 seconde dans les tests réalisés, mais cela semble durer moins longtemps) de passer par toutes les positions (directions) intermédiaires. Pour que cet effet soit perçu, on effectue ce déplacement avec une vitesse initiale grande (accélération infinie) qui décroît linéairement, jusqu'à s'annuler une fois atteinte la direction finale. Ainsi, en moins d'une seconde[6] on retrouve la corrélation son/image,

Plus que la stratégie d'interpolation, c'est sa durée et la dureté de l'impulsion initiale qui semblent faire l'efficacité de ce principe de «sources glissantes».

10.1.9 Conclusion sur le son

Pour survivre à un monde léthal, le cerveau humain a su évoluer en développant des réflexes vitaux. En ce sens, vision et audition sont complémentaires, qui nous gardent de dangers dans plusieurs directions et modalités : pour savoir où focaliser le regard, ou imaginer ce qui se passe hors du champ de vision.

En rajoutant des indices sonores assez simples, sous forme de localisation des sources monophoniques, on ne perturbe pas la compréhension des messages vocaux (le volume global reste sensiblement le même), on renforce l'aspect immersif d'une façon naturelle et on augmente les évènements qui peuvent entretenir ou regagner l'intérêt du participant. Appuyant et complémentant l'aspect graphique, l'approche des sources glissantes n'introduit pas d'artefacts (ni discontinuités de position, ni voix désincarnée loin des lèvres) ou de code qui soit compliqué à appréhender. L'auditeur peut donc rester concentré sur le message, tout en profitant d'une nouvelle dimension.

Il y a probablement beaucoup de recherches possibles dans ce domaine, pour établir un code plus évolué, si l'on utilise des techniques de spatialisations plus avancées notamment. Par exemple, un zoom de caméra (qui peut donner l'impression qu'on s'approche, alors que la caméra n'a pas bougé) n'est pas traduit auditivement. Faudrait-il rendre les autres sons moins localisés

6. à cause du caractère non linéaire de la perception, et parce que la différenciation angulaire est plus ou moins précise, de l'ordre de 7° sur l'horizon d'après la littérature.

(plus ambiants), simuler des chuchotements (à volume constant, en jouant sur la composition fréquentielle) ? Avec d'autres dispositifs de rendu, on peut simuler une plus grande gamme de directions, par exemple depuis le haut (la voix d'un administrateur, ou pour annoncer les thèmes, la fin de la réunion ou l'arrivée de participants ?). Clairement, pour ces nouvelles codifications, plusieurs tests en vraie grandeur et avec des professionnels sont probablement nécessaires.

10.2 Résultats des tests de communication

Ce prototype a été initialement utilisé avec un scénario fixe et reproductible (les messages d'un débat de plusieurs intervenants ont été pré-enregistrés, et peuvent donc, en provoquant les mêmes évènements sonores, être utilisés à volonté pour provoquer un débat filmé identique ou différent) ce qui a permis de mener à bien l'expérimentation sur la cohérence audio et visuelle et les types de caméra. On a aussi pu vérifier que l'on n'obtenait pas toujours le même film du débat, en faisant se répéter en boucle les évènements du débat pré-enregistré, ce qui est un gage de richesse donc d'intérêt pour l'utilisateur.

Depuis sa mise au point, ce prototype a été utilisé plusieurs fois en situation réelle de communication (mais sur site local), avec trois ou quatre interlocuteurs. Par rapport à un téléphone, il est bien plus simple de mener et de suivre une discussion à plusieurs, grâce notamment aux indices auditifs et visuels, et le complément graphique aide à rester impliqué dans le débat (ne serait-ce que parce qu'il offre un point où poser son regard).

Ce prototype incorporait aussi un contrôle manuel des caméras, au clavier pour choisir la caméra active, et à la souris pour l'orienter et la déplacer. Comme prévu, il n'est pas très agréable pour un participant de l'utiliser en même temps qu'il communique, malgré sa relative simplicité. Dans l'ensemble, les choix du régisseur automatique n'étaient pas critiqués (comme prévu, le régisseur fait parfois une transition juste avant la fin d'une intervention et manquait donc le début d'une autre intervention) et n'avaient donc pas besoin d'être corrigés.

10.3 Perspectives et conclusion

Du fait du placement des clones, l'espace écran n'est que partiellement occupé, plutôt en largeur qu'en hauteur, sous forme d'un bandeau (Cf. figure 8.17, page 75). On peut imaginer mettre à profit la surface restante pour proposer d'autres informations, ou d'autres services. Elle pourrait servir à la visualisation de documents ou de pages Web, ou être une fenêtre vers un autre espace commun, comme un collecticiel.

Pour une communication sonore ininterrompue (pas de connexion ni de déconnexion, juste un bouton pour couper son micro et régler le volume par exemple), le bandeau seul occupe une place moins importante, et peut donc s'intégrer parmi d'autres fenêtres dans l'environnement graphique, être un élément d'un *Mediaspace*.

Dans le cas de réunions avec beaucoup plus de participants, on pourrait peut-être multiplier les «bandeaux» de vue des participants (l'orchestrateur choisirait les deux ou trois «meilleures» caméras), ou rajouter un bandeau spécial qui aide le participant à se localiser : un barillet où

toutes les personnes sont représentées, et qui tourne (sans discontinuité de position, comme pour l'angle de prise de son) lorsque la vue de caméra change.

Avec ce prototype, on a répondu à presque toutes nos exigences initiales, dans des conditions suffisamment réelles pour que le résultat garde sa valeur et puisse être extrapolé au cas de communications non locales : avec un codage et une transmission du son qui ne rajouterait pas de délai notable, on réunit bien plusieurs participants distants, en leur proposant un environnement riche et assez immersif qui leur permette de communiquer.

Le point important qu'il reste à résoudre concerne le vidéo-réalisme : il faudrait remplacer les clones statiques (aux lèvres près) par des clones animés qui ressemblent plus aux participants en se comportant comme eux, dans leurs expressions ou leurs articulations notamment. C'est la perspective que propose le chapitre suivant.

Chapitre 11

Animation de visages par la texture

Dans les chapitres précédents, on a montré comment représenter un clone et l'utiliser pour construire un débat virtuel avec plusieurs participants et autant de sources sonores. On a vu que pour lever l'ambiguïté du locuteur un indice visuel de la provenance du son était indispensable, par exemple une animation symbolique des lèvres basée sur l'énergie du signal sonore.

Il ne fait cependant aucun doute que le locuteur a en réalité articulé son message d'une façon plus personnelle, en le soulignant de sourires ou de moues qui lui sont propres, et très différentes de tout ce qu'on pourrait synthétiser artificiellement. Il est important de proposer une communication plus riche, plus fidèle, donc de meilleure qualité dès que la situation le permet, lorsqu'**on dispose d'une caméra pointée vers le locuteur**. Plus généralement, c'est l'ensemble des expressions ou mimiques du visage qu'il faudrait transmettre, comme le font le cinéma ou la télévision en utilisant de 24 à 30 images par secondes.

On voudrait profiter à la fois de la fluidité, de la richesse et de la précision d'une telle vidéo, sans renoncer aux libertés du virtuel et de la 3D. Aussi, plutôt que d'animer l'image du clone en déformant uniquement le modèle 3D au cours du temps, comme cela est souvent pratiqué, on a choisi de suivre une piste hybride qui mélange vidéo et modèle 3D. En modifiant uniquement la couleur des points de la texture surfacique, on va animer l'image du modèle sans modifier sa base. Au lieu d'opérations de déplacement des points 3D, on ne fera que des retouches graphiques 2D dans la texture.

11.1 Quelques écueils possibles

En choisissant une approche hybride, qui combinerait à la fois une information vidéo avec une structure 3D, on s'expose à quelques risques :

- avec une simple caméra, non calibrée, **on ne dispose pas d'une vue 3D du locuteur**. Les différentes parties du visage sont déformées par la perspective, et se déplacent dans l'écran de façon complexe, en se déformant ou changeant radicalement d'aspect. Elles peuvent même se masquer partiellement.

- le **coût de la vidéo** en terme de transmission sur le réseau est très vite important, même quand il se fait au détriment de la qualité (précision des images et/ou vitesse du rafraîchissement) donc au détriment de la fidélité.

- **une personne et son modèle 3D ne sont pas identiques**, parce qu'ils n'ont pas été créés dans les mêmes conditions ni par les mêmes capteurs et que l'apparence de la personne réelle a pu évoluer depuis la date de la capture (pilosité, bronzage...). On peut s'attendre à une variation des couleurs due à la lumière, et à ce que le volume des cheveux n'apparaisse jamais coiffé identiquement. La présence de lunettes, bijoux ou maquillage est aussi une complication supplémentaire. Enfin, les clones qui sont obtenus par déformation de modèles génériques ne sont pas toujours les plus ressemblants, soit parce qu'ils optimisent une allure globale plutôt qu'une correspondance parfaite, soit parce que des éléments «standardisés» remplacent les oreilles, la chevelure ou la dentition.

- **la rigidité est capitale** entre les points de la vidéo et le modèle 3D. S'ils oscillent à la surface du modèle ou pire que l'image de zones proéminentes (nez, oreilles, limite des cheveux) n'est pas recalée au bon endroit, l'illusion de rigidité est brisée. Lorsqu'on cherche à estimer beaucoup de paramètres dont certains sont non linéaires (liés à la perspective par exemple) pour recaler un modèle observé qui n'est en fait pas rigide du tout[1], on s'expose bien sûr à des problèmes de stabilité et de robustesse. Chaque fois que la déformation appliquée à l'image pour la réaligner en tout point de la forme présente un défaut de précision spatiale ou de cohérence dans le temps, on peut obtenir des images synthétiques très dégradées, parfois grotesques. Avec la contrainte du temps réel, l'alignement correct est une exigence particulièrement délicate à maintenir, notamment pour la forme du nez, des oreilles, ou aux limites du volume des cheveux.

- **l'image du locuteur est incomplète**. Bien sûr on ne voit pas simultanément tous les cotés du visage, mais pas forcément non plus la zone animée du visage. Même s'il est censé être filmé dans de bonnes conditions par une caméra placée près de l'écran où il regarde la conférence, le locuteur risque de détourner la tête, de la bouger trop vite ou hors du cadre.

- **l'image caméra est imprécise et bruitée**, par rapport à celle de caméras de studio sous un bon éclairage. Le rafraîchissement du moniteur ou des tubes fluorescents s'ajoutent souvent au bruit perçu en se réfléchissant sur le visage ou d'éventuelles lunettes. La composante de couleur du signal est de plus assez peu précise. Les transitions fortes (bleu vers rose par exemple) s'étalent souvent horizontalement, du fait de mauvaises réponses impulsionnelles dans la chaîne d'acquisition.

Pour simplifier ou contourner une partie de ces difficultés, on va opter pour une approche «par morceaux».

1. lorsque l'on parle par exemple, la bouche, le menton et de nombreux muscles sont mobiles à l'image comme dans la structure physiologique.

11.2 Une approche par morceaux

Certains travaux emploient toute l'information vidéo du visage pour calculer la vue synthétique, soit en la plaquant (après une projection inverse estimée, lorsqu'elle existe) sur un modèle 3D pour effectuer un rendu classique, soit par des techniques de calcul direct (morphing de point de vue, ou en ayant des informations de profondeur, mesurées ou estimées).

Pour ce travail, on a préféré s'intéresser à des zones localisées de la vidéo. Parce qu'elles concentrent la plus grande part des expressions et de l'information de synchronisation avec la parole, les **trois zones retenues sont la bouche et les yeux (avec les sourcils)**. Le principe consisterait à faire apparaître ces éléments sur la texture du clone 3D pour l'animer. Il y a donc au préalable une phase où il faut trouver ces zones sur la source vidéo, comme le symbolise la figure 11.1.

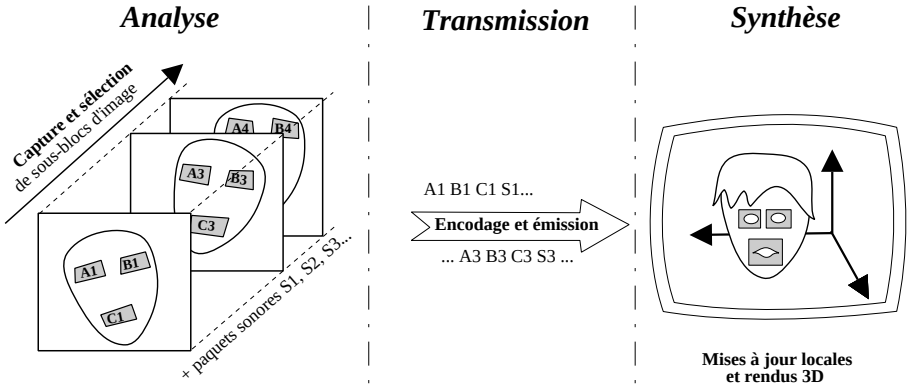


FIG. 11.1 – *Approche hybride par morceaux pour la communication*

11.2.1 Avantages de l'approche par morceaux

En optant pour une approche par morceaux, on bénéficie des avantages suivants :

- la bande passante vidéo nécessaire n'est plus liée au codage de tout un visage (ou toute une image, avec un fond), mais seulement des yeux et de la bouche. On utilisera donc moins de bande passante, ou bien on disposera de sa totalité pour ces éléments clefs, donc avec plus de précision et de résolution.

- on réduit les problèmes de visibilité partielle, les yeux et la bouche restant plus souvent visibles dans de bonne conditions que l'ensemble de l'avant du visage (par exemple, les ailes du nez apparaissent plus détaillées quand on les voit de coté, dont pas en même temps que les yeux et la bouche).

- on évite la plupart des divergences de positionnement entre la forme et son image, puisque la rigidité n'est pas remise en cause pour toutes les parties statiques. Dans l'approche par morceaux, si les yeux ou la bouche sont décalés, sautillants ou même absents parce le suivi

les a perdus, le nez, les oreilles et les cheveux resteront cohérents voire reconnaissables, puisqu'on garde ceux du modèle statique.

- on concentre les problèmes de raccords autour des yeux et de la bouche plutôt qu'entre ce qui était visible ou invisible depuis la caméra. Les problèmes d'aliasing et de bruit de la vidéo pour ces zones limites du visage vues de façon rasante ne nous concernent plus, puisqu'on n'utilisera pas ces fragments vidéo.

- on garde la liberté totale du tout-virtuel sur les angles de vue de l'image construite[2], notamment de pouvoir se placer de coté ou de derrière, où le modèle statique est défini (contrairement au cas du Morphing de point de vue ou d'un bas-relief tiré de quelques caméras frontales),

11.3 Les problèmes à résoudre

Comme souvent dans les approches hybrides, on peut séparer le problème en deux parties :

- **l'analyse** des images qui constituent la source vidéo. Ici, il va s'agir de détecter et délimiter les trois régions d'intérêt sur le flux vidéo en provenance de la caméra,

- **la synthèse** de la texture, nécessaire pour donner au clone son coté animé et ressemblant. Elle dépend des données de l'analyse pour l'intégration sur le modèle statique de texture et précède la synthèse 3D.

Bien sûr, dans le cadre de la communication à distance, ces deux étapes sont séparées par un encodage et une émission puis la réception et le décodage des données (y compris sonores), comme sur la figure 11.1.

11.3.1 Les défis de l'analyse

Idéalement, il faudrait être capable :

- de trouver la position des yeux et de la bouche,

- d'affiner ces positions pour estimer les tailles et orientations,

- d'inverser et de compenser les déformations apparentes des yeux et de la bouche (dues à la forme du visage et à l'effet de perspective), pour se rapprocher d'un élément de texture cylindrique bien aligné,

- de corriger l'illumination et de masquer les changements d'éclairage.

Le premier point est résolu par de nombreux algorithmes, dont certains – généralement parce qu'ils en encapsulent plusieurs aux caractéristiques différentes – sont des estimateurs très robustes, par exemple du centre d'une zone. Il semble que les démonstrateurs les plus fiables

2. le locuteur lui doit toujours rester suffisamment face à la caméra pour être visible. Mais on pourrait avoir plusieurs caméras...

soient ceux qui en parallèle utilisent plusieurs techniques pour lesquelles ils disposent d'indices de confiance, servant à choisir ou combiner les résultats concurrents. Il y a bien sûr un surcoût de charge CPU qui est incompatible avec des machines peu puissantes (NC ou STB).

Le second point semble lié au premier, mais est plus problématique. En effet, les impératifs de précision sont cruciaux dans notre application. L'apparente rigidité entre les zones animées et le support statique est à ce prix. En pratique, des variations de la taille estimée ou un décalage en position d'un demi-pixel suffiront à donner une impression de flottement ou d'oscillation[3]. Les meilleurs résultats de *model fitting*, qui replacent un modèle 3D sur une séquence d'images, prouvent qu'une telle précision n'est pas impossible[4], mais elle n'est pas à notre connaissance atteinte en temps réel ni même à un ordre de grandeur l'approchant.

Toujours dans ce cadre du temps réel, l'approximation classique faite par exemple dans les domaines de la reconnaissance et de la vision pour résoudre la troisième difficulté consiste à supposer qu'un *shearing* (changement d'échelles suivant deux axes non orthogonaux et de normes différentes) est acceptable pour chaque zone rigide, suffisamment plane et où l'effet de perspective reste spatialement comparable.

Concernant le dernier point, de nombreuses techniques sont proposées, par exemple pour la reconnaissance des visages [GKB98, HB96, Hal95] ou la recherche d'illuminants [LM97]. Que ce soit par un histogramme ou des modélisations plus complexes travaillant directement sur le spectre, le résultat final est une fonction (ou une table) qui permet de transformer les couleurs de la vidéo en celles de la texture, ou l'inverse. Pour produire un résultat plus naturel, il est souhaitable d'effectuer les deux transformations «à moitié», c'est à dire corriger la texture et la vidéo pour qu'elles présentent toutes deux les mêmes teintes, intermédiaires à celles initialement capturées.

Dans un premier temps, et jusqu'à la construction du prototype proposé dans la section 11.5, on supposera que le problème de l'analyse est réglé, c'est à dire que l'on dispose du flux vidéo (corrigé en forme, orientation et couleur) des trois zones de texture cylindrique retenues, à savoir autour des yeux et de la bouche. La figure 11.2 montre un exemple de ces zones, telles qu'elles seraient renseignées depuis une vue sous un angle de vue extrême.

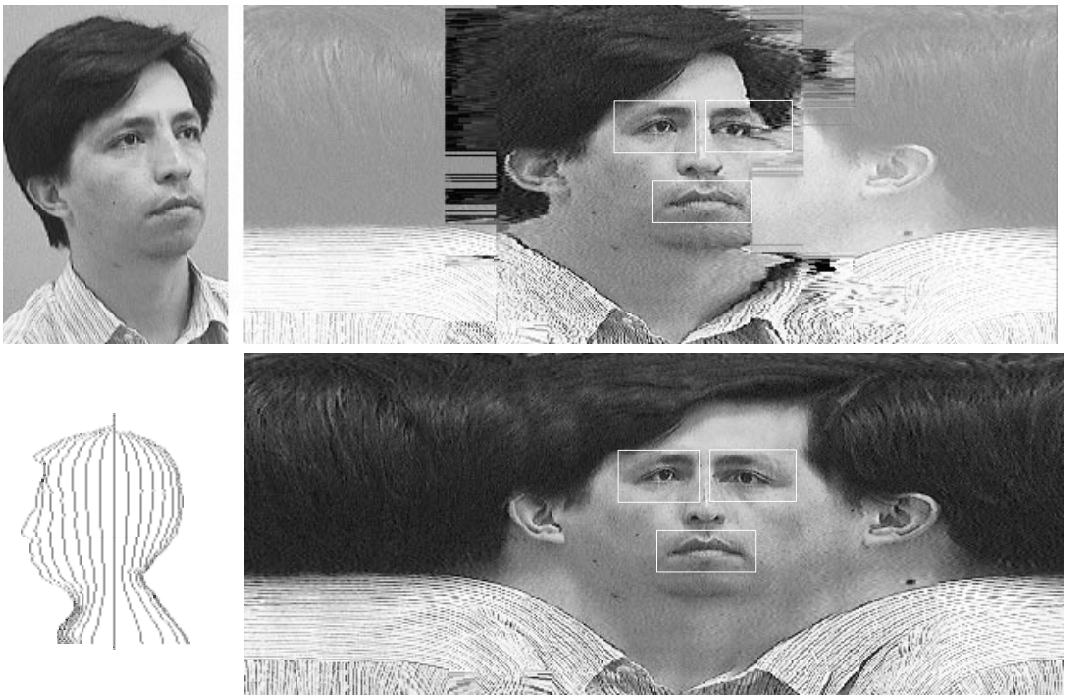
11.3.2 Les buts de la synthèse de la texture

Une fois que l'on dispose des fragments sources, il faut les intégrer à la texture cylindrique, en prenant soin :

- d'intégrer les éléments **à la bonne position et à la bonne échelle**, pour produire un résultat fidèle, et ce de façon automatique, à un débit visuel compatible avec la voix,
- d'insérer les éléments animés **sans introduire de frontière artificielle** qui soit visible, ce qui exclut bien sûr de copier directement les zones de la vidéo,
- de **ne pas montrer les éléments statiques**, par exemple une bouche de plus grande taille que celle observée sur la vidéo,

3. qui seront d'autant plus repérables que, synthétisé sans bruit sur sa position, le clone est immobile en l'absence de mouvements de la caméra virtuelle.

4. au moins pour des modèles rigides.



Partant d'une vue réelle, approchée par le modèle 3D (ici par 105° de rotation sur la droite, puis 10° dans le plan de l'image), on a effectué une projection inverse approchée vers la texture. Pour comparaison, la texture complète apparaît en filigrannes et au dessous. On peut aisément voir les imperfections et incomplétudes, notamment aux limites, dont certaines sont dues à l'angle excessif de la prise de vue.

FIG. 11.2 – *Projection inverse et zones d'intérêt de la texture cylindrique*

– de **ne pas écraser d'éléments de la texture statique**, comme une mèche de cheveux qui descendrait vers les sourcils.

Ce sont ces points qui vont être discutés en détails dans la section suivante.

11.4 Synthèse de la texture

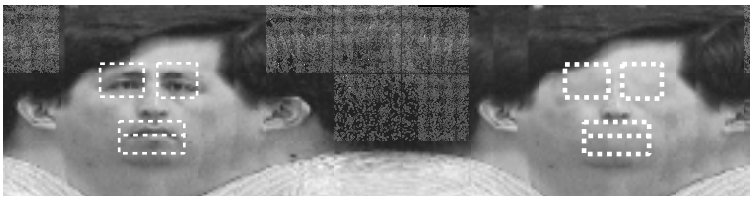
Cette section présente la démarche qui a été retenue pour mener à bien cette tâche : intégrer les yeux, les sourcils et la bouche sur la texture cylindrique, lorsque ces zones sont fournies par l'analyse.

11.4.1 Principe des zones de destination

Puisqu'on veut incruster des images animées des yeux et de la bouche, il est souhaitable de savoir où les analogues statiques de ces éléments se trouvent sur la texture cylindrique originale

du clone. Mais comme il ne faudrait pas que ceux qui seront incrustés pour l'animation interfèrent avec les éléments originaux, on décide de faire disparaître (masquer) le contenu statique.

Ces deux phases, la **définition des zones de destination** et la **neutralisation de la texture** sont des phases d'initialisation, qui doivent être réalisées une unique fois pour chaque clone, par exemple lors de la capture des données statiques. La figure 11.3 montre une texture cylindrique pour laquelle ces deux opérations sont visualisées.



Des rectangles blancs ont été surimposés pour visualiser la délimitation des zones d'intérêt dans l'espace de la texture. Dans la version de droite, ces zones de la texture ont été remplacées par un motif de teinte chair, pour faciliter et rendre plus naturelle l'incrustation qui sera faite des éléments tirés de la vidéo.

FIG. 11.3 – *Les zones d'intérêt sur la texture cylindrique*

11.4.2 Principe d'une copie progressive

Il n'est pas possible de copier entièrement chacune des zones de la vidéo vers les zones rectangulaires de la texture : la correction de couleur ne garantit qu'une teinte globale comparable, pas une correction exacte qui permettrait une incrustation directe. Le pourtour de la zone rectangulaire n'a aucune chance de correspondre parfaitement à son environnement dans la destination.

Or la vision humaine est particulièrement sensible aux discontinuités (de l'intensité lumineuse). Pour incorporer plus discrètement les nouveaux éléments dans leur environnement/fond, on va – comme pour nombre de trucages photo ou de montages – utiliser un masque progressif (*alpha channel*) pour piloter les opérations de copie.
Si pour chaque point (x, y) de l'image, on note $B[x, y]$ les pixels du fond, $S[x, y]$ les pixels de la source à incruster et $M[x, y] \in [0, 1]$ le taux de transparence voulu, la composition donne :

$$C[x, y] = S[x, y].M[x, y] + B[x, y].(1 - M[x, y]) = B[x, y] + M[x, y].(S[x, y] - B[x, y])$$

On reformule ainsi la tâche en terme de **création des masques progressifs** qui serviront lors des opérations de copie.

11.4.3 Intégration des yeux et des sourcils

Les sourcils sont caractérisés par une grande mobilité, plus forte en leur centre mais essentiellement verticale, d'où une forme très changeante. Comme les poils sont des éléments trop petits

à l'échelle où ils interviennent dans l'image, ils apparaissent donc «filtrés» par les capteurs de la caméra : le masquage partiel de la chair par les sourcils génère donc un nuancier des teintes extrêmes de la chair et des sourcils (mono-teinte ou poivre-et-sel).

Les yeux constituent une zone très animée du visage (clignements et direction du regard) mais sont fixes du point de vue de leur position et de leur taille dans la texture cylindrique. Cette zone présente donc des pixels d'apparence très variable : la couleur de l'iris comme la présence de reflets et de cils peuvent contribuer à une texture très colorée et imprévisible.

Pour résumer, **la zone des yeux est stable dans sa forme et sa position, mais assez variable au niveau des pixels tandis que ceux des sourcils ont une apparence plus prévisible, mais avec une forme et une position indéterminées**.

À cause de ces différences, on pourrait être tenté de proposer deux techniques pour capturer ces éléments, et réaliser deux copies. Cependant, lorsque l'on fronce les sourcils par exemple, la frontière entre ces zones n'est pas très nette, et on peut craindre qu'apparaissent dans la destination des artefacts disgracieux lorsqu'elles se recouvrent (une copie écrase l'autre, ou la rend plus pâle/plus sombre à la jointure).

En conséquence, il faudrait **ne faire qu'une copie, mais qui intégrerait les deux éléments à la fois**, même s'ils ont été détectés par des critères différents.

11.4.4 Masque pour l'œil

Comme la position et la taille de l'œil sont invariables, on va utiliser un masque elliptique aux contours progressifs. Ce masque est statique, c'est à dire qu'il ne changera pas pendant toute la conférence.

Si $2 \times w_{œil}$ et $2 \times h_{œil}$ sont la taille de cet élément (par exemple l'œil gauche) dans la texture cylindrique, une expression possible pour générer un masque elliptique progressif, sur le domaine $[-w_{oeil}, w_{œil}] \times [-h_{œil}, h_{œil}]$ est :

$$M_{œil}[x, y] = 1 - Max\left\{1, 3.Min\left(0, (x/w_{œil})^2 + (y/h_{œil})^2 - .6\right)\right\}$$

Le masque ainsi créé présente une zone elliptique centrale où la transparence vaut exactement 1, avec une bordure régulière où elle décroît assez rapidement. Proche du bord, la transparence vaut 0.

11.4.5 Masque pour le sourcil

L'étiquetage de la texture cylindrique nous fournit l'emplacement de deux zones englobantes, censées recouvrir les yeux et les sourcils. On dispose de l'image animée de ces zones, mais sans savoir exactement où apparaissent des pixels des sourcils (puisqu'ils se déforment). Pour les localiser, on va utiliser deux modèles de couleur, étalonnés au début de chaque séance (pour corriger l'effet de l'illumination naturelle/artificielle) en pointant sur des zones de l'image réelle correspondant à :

– des sourcils, ou des cheveux, là où ils sont les plus sombres,

– de la peau, par exemple sur le front, au dessus du milieu des sourcils, mais pas sur un reflet blanchâtre.

C'est à partir de ces distributions gaussiennes (de la teinte, mais aussi de la luminosité, car les sourcils peuvent être suffisamment sombres pour que leur teinte ne soit plus représentative) que l'on va construire le masque qui servira à recopier préférentiellement les points du sourcil, en les mélangeant avec la peau de la texture statique neutralisée.

Le masque des sourcils est dynamique, puisqu'il se construit en chaque point en fonction du pixel de la source $S[x, y]$ où il sera utilisé. On peut donc formaliser sa construction comme suit :

$$M_{sourcils}[x, y] = F\left(S[x, y]\right)$$

où $F(r, g, b)$ évalue quantitativement (entre 0 et 1) si le pixel est plus ou moins près du modèle de couleur du sourcil, relativement au modèle de la peau. En pratique, le lieu des points de l'espace de couleur correspondant à une transparence donnée est une pseudo-ellipse dont les deux foyers sont les échantillons moyens[5]. De son paramètre k, on tire notre transparence, qui vaut 1 sur une plage de pixels proches de ceux du modèle du sourcil, et décroît rapidement vers 0.

11.4.6 Masque composite

On sait que les deux masques précédents concernent des zones spatiales qui peuvent se recouvrir. Pour créer le masque combiné, on utilise :

$$M_{composite}[x, y] = Max\left(M_{\text{œil}}[x, y], M_{sourcil}[x, y]\right)$$

ainsi un pixel sera visible selon la modalité où il a été le plus détecté (œil ou sourcil). Comme l'opérateur de maximum n'introduit pas de discontinuité dans les zones de recouvrement, on réalisera bien une copie conjointe satisfaisante en utilisant ce masque.

11.4.7 Synthèse : Intégration de la bouche

La bouche constitue une zone très animée dont la forme et la taille sont particulièrement changeantes : un sourire déploie toute sa longueur, tandis qu'une bouche en cœur la recentre tout en rondeur. Parce que les lèvres sont très mobiles avec la parole, elles peuvent à tout moment réfléchir différemment la lumière et révéler les dents ou la langue.

L'approche retenue a été d'utiliser le même masque elliptique progressif que pour l'œil, mais avec une largeur et une hauteur correspondant à celles de la bouche dans le fragment vidéo.

11.5 Une implémentation pour l'évaluation

On a déjà souligné l'importance qu'il y avait à ce que de telles techniques interactives puissent être expérimentées en temps-réel pour être évaluées. C'est dans cette même optique que l'on s'est à nouveau autorisé des choix qui ne sont pas ceux d'une réalisation commerciale,

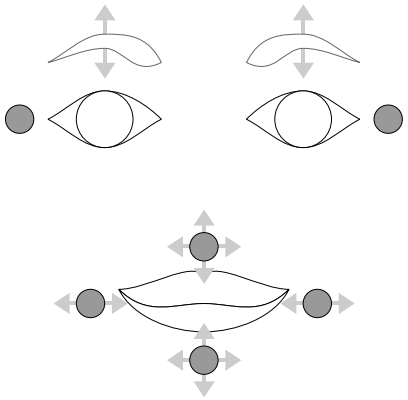
5. la distribution statistique des deux modèles n'est à priori pas la même, elle joue donc sur le calcul de la distance : $d^2_{peau}(M).d^2_{sourcil}(M) = k^2$.

tout spécialement pour le module de suivi. Pour ne pas surcharger la machine (qui fait aussi l'acquisition et les rendus graphique/sonore), on s'est contenté d'un suivi minimal, qui exige quelques marqueurs colorés sur le visage de l'utilisateur.

La priorité ne portait pas non plus sur une optimisation de la couche réseau : la bande passante utilisée n'est pas minimisée en compressant les fragments de vidéo, et l'assurance de leur délivrance dans les temps exigerait normalement un protocole adéquat. Là encore, c'est l'hypothèse d'un réseau performant (Intranet ou ATM) qui simplifie la réalisation du démonstrateur sans sacrifier les résultats.

11.5.1 Principe du suivi

On oblige chaque participant vidéo à porter des marqueurs (verts ou bleus) sur le visage, comme sur la figure 11.4, pour faciliter la tâche de vision 2D/suivi. Bien sûr, tout autre module de suivi pourrait être utilisé : une caméra fixée à un casque présente l'avantage de rester constamment face au visage, sans percevoir de déformations ou de rotations, mais serait encore moins agréable à porter, et inadaptée pour regarder un écran... Idéalement, les marqueurs devraient être rendus caducs par des algorithmes de vision à la fois rapides, précis et robustes.



Les zones et les marqueurs ont des amplitudes et directions de mobilité différentes : la bouche est polymorphe, les sourcils sont assez mobiles.

FIG. 11.4 – *Positionnement des six marqueurs sur le visage*

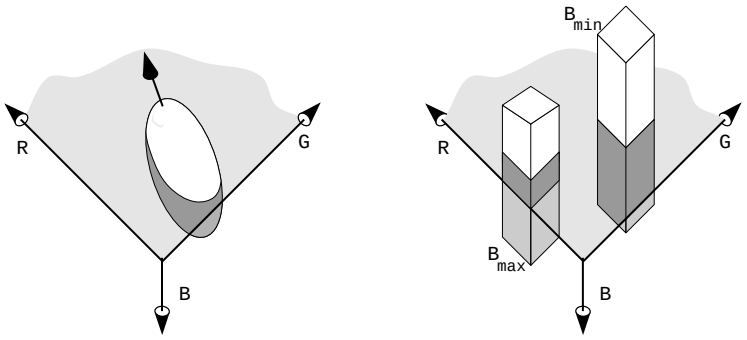
C'est avant tout pour la simplicité d'implémentation et le faible coût en temps de calcul que cette solution avec marqueurs a été retenue, permettant l'évaluation des techniques proposées pour l'incrustation.

Les marqueurs sont au nombre de six, positionnés de façon à cerner les futurs éléments vidéos. La bouche en monopolise quatre, placés horizontalement et verticalement. Les deux derniers encadrent les yeux, comme l'illustre la figure 11.4.

11.5.2 Classification des pixels d'un marqueur

Verts ou bleus, les marqueurs devraient être facilement différenciables de la peau d'après leur couleur. Pour classifier les points «marqueurs», on s'arme d'un modèle de leur apparence

(au niveau de chaque pixel, par la distribution des couleurs possibles, cf. Annexe A). En plus de la couleur (ou teinte si on fait abstraction de la luminance) moyenne, une distance et un seuil servent à les caractériser. Pour être efficace lorsque l'on veut décider si un pixel doit être classifié comme «marqueur», on utilise une table, comme sur la figure 11.5. Celle ci est pré-affectée mais peut être réinitialisée avec une autre distribution gaussienne en pointant quelques représentants à l'image.



Pour un seuil et une distance de Mahlanobis donnés, les points de l'espace RGB à classifier comme «marqueurs» forment un ellipsoïde. En pratique, comme les pixels de l'image sont déjà à valeurs discrètes, une table sert à accélérer la décision. Plutôt qu'un index complet $[R, G, B] \rightarrow$ oui/non, une double table $[R, G] \rightarrow (B_{min}, B_{max})$ minimise l'occupation dans le cache.

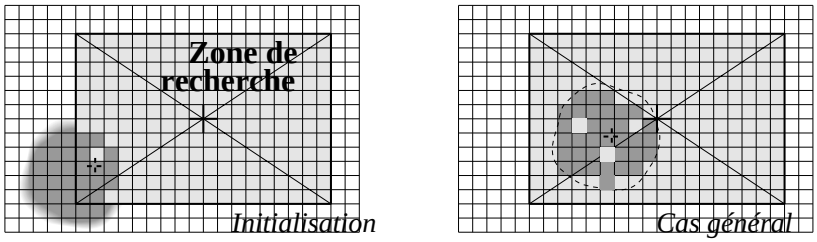
FIG. 11.5 – *Espace des «marqueurs» : ellipsoïde exact et approximation tabulée*

En pratique, avec une lumière normale (faible pour une caméra), on observe un bruit important sur les images vidéos capturées par des caméras CCD. Les mêmes pixels ne présentent vraiment pas la même couleur dans plusieurs images consécutives, et les points des marqueurs ne seront pas toujours faciles à classifier comme tels à chaque image.

Plutôt que d'adopter des techniques de filtrage temporel ou d'utiliser des opérateurs morphologiques, on pourrait penser à relaxer le modèle de couleur verte pour qu'il soit moins sélectif. En pratique, ce ne sont pas les points du marqueur, mais son centre que l'on va en fait estimer et suivre. Comme chaque marqueur occupe plusieurs pixels visibles, le centre des pixels classifiés comme «marqueurs» peut être suffisamment stable pour être utilisé directement, sans filtre de Kalmann notamment, si l'on prend quelques précautions : en restreignant les points reconnus comme marqueurs, on perd la détection de certains points verts mais sans jamais inclure de points de la peau. L'ensemble des points classifiés marqueurs est souvent un anneau (le reflet blanchâtre n'est pas reconnu), ce qui ne change pas le centre (alors que si on incluait des points de la peau, le centre migrerait vers celui des points de la peau).

11.5.3 Réalisation du suivi d'un marqueur

Le suivi est réalisé par un algorithme itératif : une fenêtre de recherche rectangulaire se recentre à chaque trame sur le barycentre des points qu'elle a classifiés «marqueurs», comme schématisé sur la figure 11.6.



Dans la fenêtre de recherche centrée sur la croix noire, tous les pixels sont classifiés «marqueurs» (gris foncé) ou «non-marqueurs»(gris pâle). Le barycentre des pixels «marqueurs», matérialisé par la croix pointillée devient le centre de la fenêtre de recherche pour la prochaine itération. Dans le cas général, sauf vitesse de déplacement trop rapide du marqueur ou à l'initialisation comme à gauche, le futur centre de la fenêtre estime celui du marqueur après une seule passe.

FIG. 11.6 – *Principe du suivi d'un marqueur*

La taille de la fenêtre correspond au déplacement maximum qui pourra être suivi, et conditionne par son aire le coût de la recherche. Pour un fonctionnement normal, il faut prévoir une fenêtre suffisamment grande pour inclure le rayon du marqueur en plus de son possible déplacement (apparent dans l'image) selon toutes les directions. Pour être sûr de ne pas avoir estimé qu'une partie d'un marqueur, comme à l'initialisation ou lors des déplacements juste repérables, il est possible d'appliquer le recentrage deux fois par image acquise.

Il n'y a pas de problème pour choisir une taille suffisante pour la fenêtre au niveau des yeux, puisqu'il n'y a que très peu de marqueurs dans l'image du visage. Au niveau de la bouche, une fenêtre haute est particulièrement souhaitable pour le marqueur inférieur, car il est très mobile. Pour ne pas risquer la fusion avec le marqueur supérieur, la technique suivante a été utilisée avec succès : quand un pixel est reconnu comme marqueur, sa valeur originale dans la source est écrasée (à la deuxième passe, et par une valeur spéciale), de sorte qu'il ne sera plus reconnu comme marqueur par une autre fenêtre de recherche. Si la fenêtre inférieure est balayée en dernier, elle ne risque plus d'être influencée par les marqueurs précédents, et peut donc être plus haute sans risque de confusion.

11.5.4 Modèle des déformations

On ne va pas chercher à corriger l'apparence des yeux selon toutes les déformations induites par le modèle 3D. Comme dans de nombreux travaux de suivi pour la reconnaissance, on fait l'hypothèse que la zone autour des yeux et de la bouche est quasi-plane (ou vue suffisamment de face pour que la projection induise peu de déformations non affines). Sous cette hypothèse, ces morceaux de la texture du modèle subissent chacun une transformation vers l'écran que l'on approxime par un *shearing*. Il suffit donc de spécifier la déformation inverse (elle aussi un *shearing*) pour obtenir la mise à jour approchée de la texture.

11.5.5 Définition des zones d'intérêt

Les deux marqueurs horizontaux, appelés X et Y sur la figure 11.7, servent de référence, non seulement pour l'horizontalité mais aussi pour dériver la verticalité supposée (dans le cas de faibles rotations de la tête) : on fera comme si l'orthogonalité était préservée par la projection.

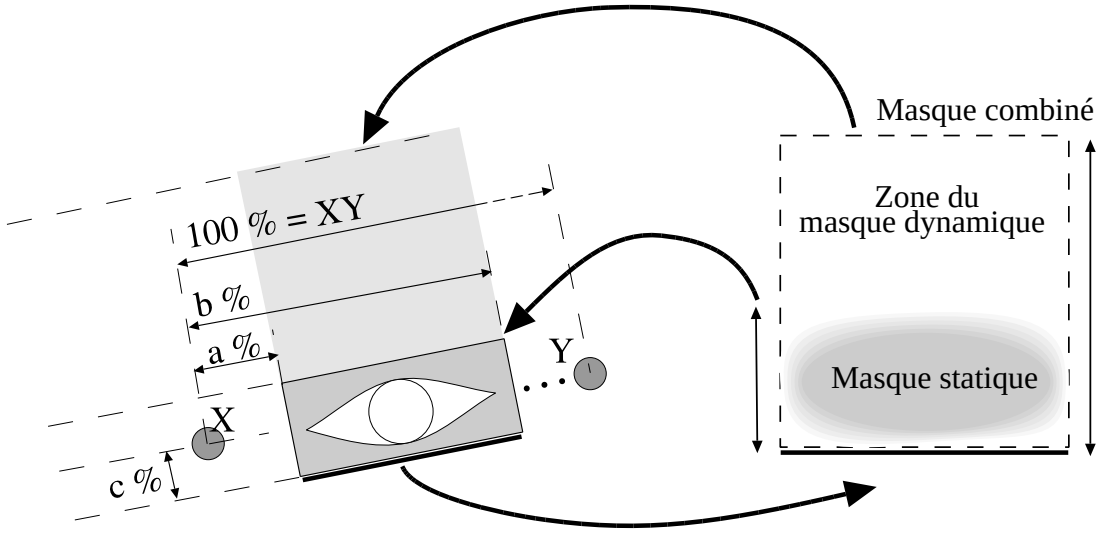


FIG. 11.7 – *Positionnement en proportions des yeux dans la vidéo*

On ne peut pas espérer que ces marqueurs soient toujours positionnés exactement de la même façon à chaque utilisation d'un clone, il faut donc passer par une phase de réglage – manuelle dans cette implémentation – où l'on ajuste (par rapport aux réglages mémorisés lors d'une précédente utilisation) les proportions relatives qui cadrent les yeux et les sourcils. Ainsi, $a\%$ et $b\%$ servent à préciser la position horizontale[6] (de l'œil gauche dans la figure), tandis que $c\%$ délimite le bas de l'œil.

Ces positions sont **relatives et proportionnelles**, de sorte que lorsque les marqueurs se déplaceront dans l'image (parce que le locuteur bougera par rapport à la caméra), on puisse toujours **retrouver les éléments de texture associés, en position mais aussi en taille**, puisque les hauteurs pour l'œil ou la zone qui inclut le sourcil peuvent être automatiquement déduites des proportions de la zone dans la texture cylindrique, en appliquant un ratio qui dépend du modèle : si dans la vidéo les pixels sont carrés (ou rectangulaires) et font tous la même taille, ce n'est pas le cas dans la texture cylindrique, où la largeur d'un pixel correspond en fait à une portion d'arc. En pratique, on approxime la «largeur» d'un pixel de la texture cylindrique en utilisant le rayon moyen du modèle dans la zone de l'œil par $\frac{2\pi}{w} r_{moyen}$ où w est la largeur de la texture cylindrique (qui représente un tour complet).

En plus de préciser les zones d'intérêt, ces paramètres vont aussi servir à définir l'échelle de la correspondance avec les éléments analogues de la texture cylindrique.

6. au sens de la référence XY ou des contributions dans la texture cylindrique.

11.5.6 Mises à l'échelle et copie des yeux

Comme les yeux ne sont pas mobiles et ne changent pas de taille, le facteur d'échelle horizontal pour la copie de l'œil est parfaitement spécifié, ainsi que sa destination : **par la copie, les segments inférieurs doivent être mis en correspondance**, en position et en largeur, comme sur la figure 11.8.

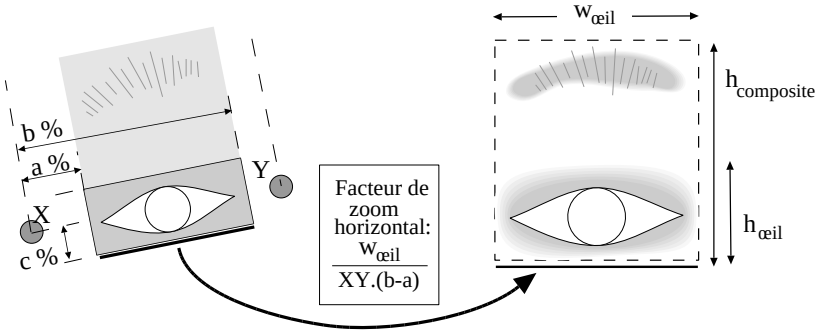


FIG. 11.8 – *Réalisation de la copie de l'œil*

L'échelle verticale est liée à l'échelle horizontale, selon le ratio précédemment explicité, qui corrige l'aspect des pixels dans la texture cylindrique. Comme on l'avait déjà laissé entrevoir en introduisant le repère de définition de la zone des yeux, c'est donc la taille de la fenêtre destination qui détermine le rapport entre XY et la hauteur du segment vidéo qui sera copié.

11.5.7 Mises à l'échelle et copie de la bouche

La bouche change de forme, et ses marqueurs ne nous renseignent donc pas sur sa taille dans la texture, seulement sur sa taille apparente dans la vidéo. C'est le rapport d'échelle des yeux qui sert à estimer l'échelle qui sera appliquée à la bouche. La taille effective de la bouche une fois copiée dans la texture est donc variable. En général, et comme c'est le cas sur la figure 11.9, la bouche n'occupe donc pas tout l'espace qui, lors de la «neutralisation» du clone, a été défini comme légal. Le masque elliptique pour la bouche est bien sûr utilisé suivant la taille finale de l'incrustation.

Toujours par une approche de proportions, la position horizontale de la bouche peut être mesurée sur la vidéo et reproduite dans la texture. En l'absence de plus d'informations, la position verticale de la bouche ne peut pas être reproduite : la ligne des deux marqueurs est envoyée sur la ligne qui a été définie sur la texture cylindrique avant neutralisation.

Comme le menton n'est pas mobile, l'espace où la bouche peut être incrustée ne s'agrandira pas. Pour ne pas sortir de cette limite quand la bouche est grande ouverte, on s'autorise à diminuer l'échelle (autant horizontalement que verticalement) de la copie. En pratique, lorsque l'échelle atteint un maximum, elle n'est pas autorisée à le dépasser. Cela ne provoque donc pas de discontinuité visuelle à la reconstruction, aucun détail de la texture statique ne sera écrasé et les proportions de la bouche originale ne sont pas changées.

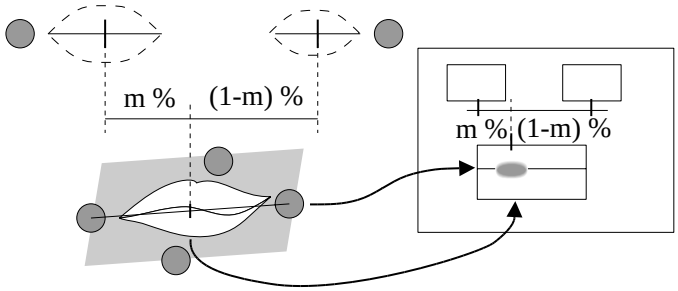
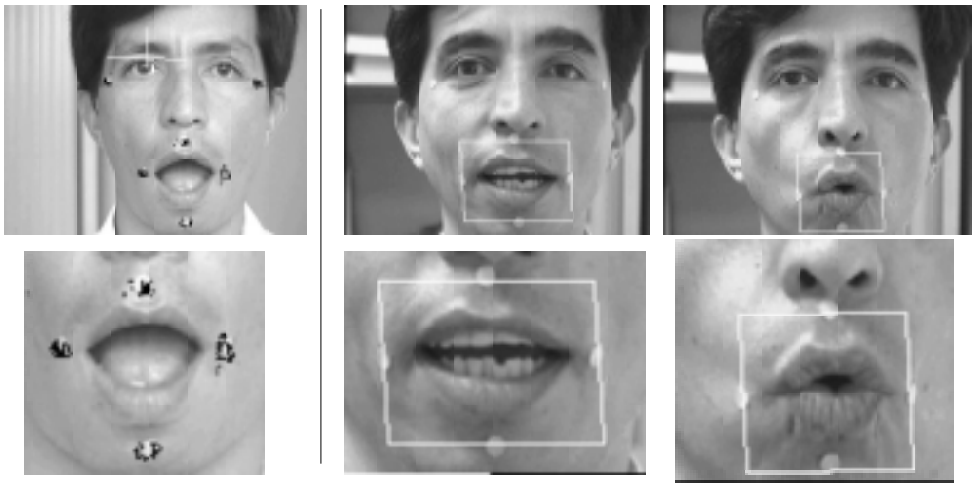


FIG. 11.9 – *Position et copie de la zone de la bouche*

Contrairement au cas de l'œil, les marqueurs sont partiellement visibles dans la source vidéo sélectionnée pour la copie. Au moment de la copie, chaque pixel est testé selon un modèle des pixels marqueurs (plus lâche que celui utilisé pour le suivi, pour ne pas laisser passer les pixels qui seraient partiellement verts). S'il est reconnu marqueur, une transparence nulle est utilisée à la place de celle lue dans le masque elliptique.

11.5.8 Résultats du suivi

Par ce suivi de quelques marqueurs sur la peau, on peut donc matérialiser une réalisation de l'hypothèse initiale : disposer d'un flux vidéo des caractéristiques faciales que l'on veut incruster dans la texture du clone. La figure 11.10 montre le résultat du suivi des marqueurs (et leur difficile classification), ainsi que la zone d'intérêt de la bouche.



Sur la colonne de gauche, les points noirs sont ceux que la classification a reconnus comme « marqueurs » sans incertitude. Les images de droite montrent la zone source de la vidéo que définissent les capteurs de la bouche.

FIG. 11.10 – *Le suivi des marqueurs en action*

Du fait de sa simplicité, cette méthode de suivi direct (sans synthèse pour analyse par exemple) ne consomme quasiment pas de temps CPU, de sorte que l'acquisition vidéo et le rendu[7] restent à priori les facteurs limitants.

11.6 Résultats des clones animés

On dispose alors d'une boucle analyse-synthèse très rapide, où l'on peut vraiment juger de l'impact des expressions, et de la corrélation son/image : **le clone semble réellement vivant, ce qui favorise les réactions empathiques du spectateur** (il n'y a pas besoin d'un sourire appuyé pour qu'il soit visible, et le moindre mouvement de sourcil est laissé à l'appréciation du correspondant). **En ce sens, le clone est bien vidéo-réaliste**.

FIG. 11.11 – *Les résultats de l'animation temps-réel : source vidéo et modèle 3D incrusté*

7. en fait, sur le NC où la carte vidéo ne propose pas de DMA (accès direct de la mémoire centrale par le périphérique, pour des transferts rapides), c'est le processeur qui doit prendre le temps (les cycles) d'aller récupérer ces données en opérations d'Entrées/Sortie explicites, ce qui limite la résolution de la vidéo à 256x200 pour 20 images par seconde. Le rendu et le suivi ne font pas baisser ce taux de rafraîchissement.

La figure 11.11 montre, sans en capturer bien sûr la dynamique, quelques images de la vidéo qui est disponible en MPEG, sans le son, à l'adresse http://www-vis.imag.fr/Elisei/ClonesVideo/.

Ces arrêts sur image montrent que les dents, les yeux et les sourcils sont bien ceux du modèle réel et que le clone est complètement texturé : contrairement à certaines approches, on ne remarque pas que l'œil ou les dents sont artificiels, et il n'y a de coupure franche sous forme de limite polygonale aux ouvertures du modèle. Le plus grossier défaut de ressemblance est visiblement la forme (statique) du clone que l'on a utilisé, qui n'existerait pas si l'on avait par exemple eu accès à un scanner 3D ou un *toolkit* de création. À défaut, la forme de ce clone a été construite très approximativement à partir du contour de plusieurs vues du modèle original [Ben97, Eli97]. Si l'hypothèse d'une faible perspective est compatible avec la construction de la texture cylindrique, elle a fortement perturbé celle de sa forme.

11.6.1 Analyse des résultats visuels

Sur l'ensemble des expériences réalisées (toujours sur la même personne), et si l'on excuse la forme du clone statique, il faut détailler les réussites et artefacts suivants :

- la zone intérieure de la bouche est généralement bien restituée, et on retrouve toujours les dents ou la langue originales, si difficiles à restituer fidèlement dans les approches 3D. L'extérieur par contre est parfois flou, particulièrement pour les sourires les plus larges : le masque elliptique gomme en effet ces extrémités. En adaptant le masque selon la largeur de la bouche (modèle à priori, donc avec un risque quant à sa fidélité dans tous les cas de figure) ou en détectant les commissures (ce sont deux des rares points du visage que l'on arrive à suivre de façon robuste) pour ajuster dynamiquement le masque de la bouche comme cela est fait pour les yeux, ce problème devrait disparaître. Plutôt que les approches classiques de *snakes* utilisables lorsqu'une caméra fixe les lèvres en gros plan (pour l'aide à la reconnaissance de la parole, ou la lecture directe), une approche statistique [BCS97] semble pouvoir être assez robuste pour de la vidéo.

- la restitution des yeux est très satisfaisante, sans problème visible de transition avec la peau statique, puisque les pixels diffèrent beaucoup, comme prévu.

- les expressions peuvent être visualisées sous des angles de vue assez éloignés de ceux de l'acquisition, comme sur la figure 11.12. Sauf quand la langue est tirée, le résultat est très convaincant.

- les sourcils ont bien les formes et la vivacité de leurs homologues réels, cependant les transitions aux frontières ne se font pas linéairement mais sont accentuées au profit de la peau. Couvrant la même surface qu'à l'origine, ils sont néanmoins d'autant plus délavés qu'ils étaient clairsemés (visible uniquement sur la figure 11.11 et la planche couleur page 120).

- le blanc des dents est un révélateur de la correction de la couleur qui est apportée à la vidéo. Lorsque les conditions d'éclairage sont très différentes, un modèle de correction de couleur additif montre ses limites : le blanc se teinte, voire détonne. Avec un modèle spectral (pour

établir la table de correction à l'initialisation, donc sans charge supplémentaire pendant l'utilisation), ce problème n'existerait pas non plus.

– les éléments expressifs autres que ceux sélectionnés sont masqués : on perd les contractions musculaires des joues et le mouvement des rides (mais on pourrait envisager, si l'on sait les détecter, de les synthétiser à la surface de la texture par une image de différence, ou une courbe paramétrée). La figure 11.13 montre de tels cas.

– le menton n'est pas mobile, et la bouche restituée est parfois rapetissée (en gardant ses proportions) pour se contenir dans l'espace disponible (le rectangle où les mises-à-jour de la texture sont autorisées). Dans une approche avec un modèle 3D générique, la transmission et l'utilisation d'un paramètre supplémentaire pour la rotation de la mâchoire est tout-à-fait réalisable (sa capture fait partie de nombreuses approches compatibles MPEG-4). On peut aussi envisager une approximation dans le cas du rendu rapide, qui consisterait à rééchantillonner la partie inférieure du modèle (plus de tranches horizontales, donc plus de lignes d'écran) pour accroître l'espace (aux dépends de la forme exacte de l'arrière du visage) en fonction de la hauteur de la zone vidéo à copier pour la bouche.

Sur ces images d'une ancienne version (qui présentait le défaut de couper les sourcils et exigeait qu'ils soient accentués), on peut apprécier diverses expressions sur des angles de vue assez éloignés de ceux de la caméra.

FIG. 11.12 – *Les degrés de libertés du rendu*

11.6.2 Utilisation pour la communication

L'utilisation de ce principe d'animation par mises-à-jour de la texture amène plusieurs questions et remarques, si l'on veut l'utiliser pour la communication, par exemple avec le prototype du chapitre précédent :

– comment encoder efficacement ces mises à jour, pour les envoyer sur le réseau ? Si l'on envoie les fenêtres maximales, on peut considérer que l'on a trois vidéos indépendantes, et utiliser n'importe quelle méthode de compression de vidéo. Cette approche est la plus simple, mais n'est probablement pas celle qui générera la meilleure qualité ni le plus bas

La source vidéo présente des expressions où des rides importantes apparaissaient sur le front et sur le nez. Le clone vidéo ne les reproduit pas (sauf à l'extrémité des sourcils, qui pour ces images préliminaires étaient accentués). L'expression transmise est donc moins fidèle.

FIG. 11.13 – *Les limites d'une incrustation par morceaux*

débit. En effet, avec un codeur qui estime le mouvement par blocs (comme H261), la bouche va souvent apparaître dégradée (comme dans la partie texture de la vidéo MPEG précédemment citée). Si l'on choisit par contre de ne transmettre que les blocs destination de la copie, il faudra tenir compte de leur taille variable, comme de leur position dans la destination, ce qui nécessite un codage adapté ou un canal supplémentaire.

– comment baisser encore plus le débit de la vidéo? Au détriment de la qualité (fidélité) bien sûr, on peut adopter un codage du type *Eigenfeatures*. On suppose que les blocs de mise à jour sont un sous-espace linéaire de petite dimension. L'analyse (projection par produit scalaire) fournit un petit nombre de coefficients, utilisés à la synthèse comme pondérations des éléments de la sous-base. Celle-ci peut être statique (établie automatiquement en «grimaçant» quelques secondes devant la caméra avant la communication) ou dynamique (se transmettre et se remettre à jour pendant l'utilisation).

– l'utilisateur doit-il se voir dans une fenêtre de contrôle? On a déjà noté qu'il ne fallait pas envoyer d'une personne une image qui pourrait travestir ses expressions réelles, suite à une défaillance ou à l'imprécision de la méthode de codage. Une fenêtre dédiée au retour «vidéo» n'est, dans cette application, pas nécessaire si tous les participants reconstruisent le même débat (ce qui permet de ne diffuser les vidéos incrémentales que quand elles seront utilisées). Dans l'autre hypothèse, les flux ne peuvent pas être économisés aussi facilement, et une personne ne sait jamais si son clone est regardé (contrairement à la réalité ou à certains *Mediaspaces*), alors qu'elle relâche son attention.

Le déploiement du prototype sur plus de machines nous permettra probablement de tester d'autres scénarios applicatifs dans un proche avenir. Dans le cadre d'une CTI avec le CNET,

qui a positivement influencé et stimulé ce travail en permettant de nombreux échanges pluri-disciplinaires, tous les scénarios qui ont pu être testés grâce au travail présenté ici et qui ont été retenus sont ou seront réutilisés pour un projet de plus grande envergure, correspondant à plusieurs thèses. Dans un environnement 3D complet (clones, mais aussi éléments physiques de salle) en Java 3D, et avec le spatialisateur de l'IRCAM. Bien sûr, pour ne pas handicaper les performances, ce sont des stations graphiques et non plus des *SetTopBox* qui sont utilisées.

11.7 Conclusions et perspectives

Il est donc possible de communiquer des expressions par un «vidéo-clone», un clone statique où seule la texture est remise à jour, même s'il s'agit seulement de morceaux choisis. L'approche développée apporte à l'étape de synthèse des solutions qui approchent le vidéo-réalisme complet : la plupart des expressions, l'articulation et sa synchronisation aux sons sont bien plus fidèles que dans une approche tout 3D. La ressemblance statique n'est pas non plus compromise puisque l'on n'est pas obligé d'adapter un modèle générique «instrumenté», et les clones les plus précis peuvent (doivent?) être utilisés comme support.

L'approche d'analyse est elle aussi radicalement transformée: au lieu d'estimer de très nombreux paramètres[8], avec des points de contrôle parfois très proches ou invisibles (dans la bouche) et correspondant à des entités sémantiques assez fortes (trouver «le milieu supérieur de la lèvre» lorsque la bouche est en cœur demande de connaître la direction vers le haut du modèle global), on descend à un niveau plus bas de vision, où moins d'abstractions sont nécessaires (même sans les marqueurs). Il faut noter que l'on n'a pas besoin d'un modèle (ni exact, ni générique) de la personne dans cette phase d'analyse, ce qui est une contrainte de moins, si l'on accepte un placement des éléments de vidéo sur un clone générique neutre.

En supposant un module de suivi rapide (joué par l'adjonction des marqueurs), on n'est pas tellement loin de ce que ce domaine de recherche proposera sûrement bientôt ou pourrait proposer sur des machines plus puissantes.

Cependant, il serait illusoire de croire que l'outil final de représentation est là, ne dépendant que de l'existence de cette brique de vision en temps réel :

- l'ensemble du clone est complètement texturé, mais les zones animées sont restreintes (pas de mouvement du menton, des tissus de la joue ou de rides d'expression),

- il subsiste trop d'étapes d'initialisation non-automatisées: pointer des exemples pour la peau et les sourcils, définir les paramètres de positionnement relatif des yeux ou neutraliser le clone,

- la communication ne se limite pas aux seules attitudes faciales. Les gestes de la main comme les postures du corps constituent un canal riche, qui facilite une communication naturelle.

8. les paramètres sont en fait implicites : ce sont les valeurs des pixels qui seront incrustés, mais ils sont estimés d'un seul coup.

11.7 Conclusions et perspectives

Il est probable qu'une approche combinée puisse être envisagée, qui offrirait «le meilleur des deux mondes» : des fragments de vidéo, incrustés sur un modèle 3D où seuls les joues et le menton bougent selon une simulation physique des muscles, pilotée d'après une estimation visuelle. Resterait alors à reproduire les rides d'expressions ou le froncement du nez. Des solutions de synthèse existent déjà [VY92, WKMT96], mais on leur préférerait une solution basée texture, et pour laquelle trouver une solution d'analyse semble plus raisonnable.

Concernant les gestes, l'intégration dans MPEG-4 qui spécifie un décodeur ad-hoc (MPEG-4 *Body*) permettrait de profiter des prochains progrès en analyse puisque l'on ne souhaite pas ajouter de nouveaux capteurs aux participants.

Dans tous les cas, avec les spécifications finales et publiques de MPEG-4, il serait souhaitable de valoriser nos résultats en générant des flux compatibles.

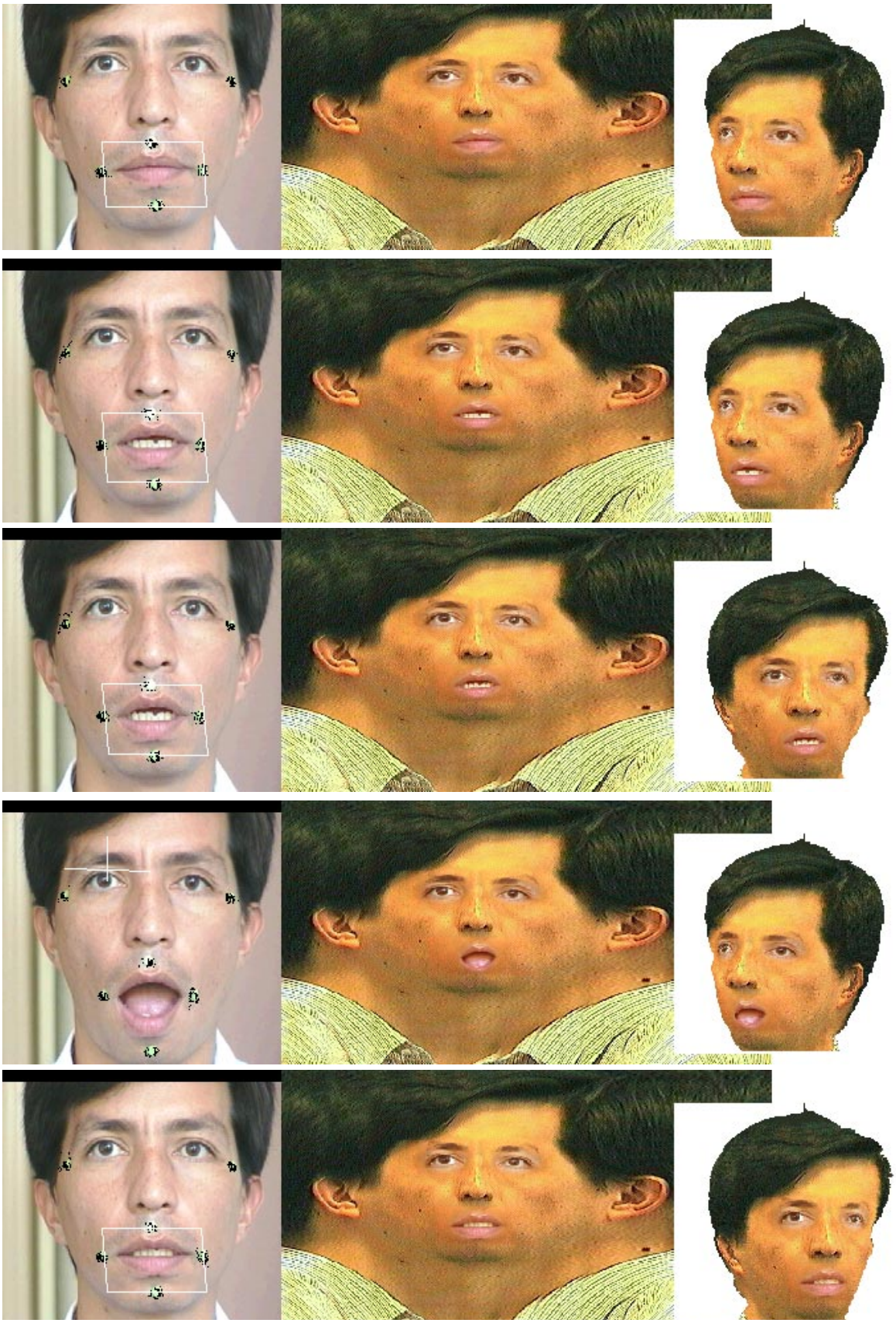


FIG. 11.14 – *Planche couleur : Vidéo, texture et vidéo-clone*

Conclusions et perspectives

Une première motivation, pratique, pour ce travail a germé avec l'expérience de quelques groupes de travail distants. Entre les réunions en un même lieu physique, le constat de l'absence d'un outil simple et informel d'échanges s'imposait. D'un point de vue plus scientifique, le projet a vraiment débuté par le rendu de quelques clones 3D, et la volonté de les utiliser, sinon dans un but de communication assistée par ordinateur, au moins pour juger de la pertinence d'un tel outil pour cette approche.

Lors de l'analyse, il a évidemment fallu élargir l'approche et s'ouvrir à des domaines très variés, en partie représentés dans l'état de l'art de ce document. Intégrant des approches de plusieurs horizons avec des solutions personnelles, on a pu construire différentes réponses théoriques et pratiques aux problèmes initialement définis.

Bilan des contributions

On a tout d'abord proposé **un algorithme de rendu spécifique**, pour des objets texturés à géométrie quasi-cylindrique, qu'on utilise pour des clones 3D statiques et photo-réalistes. Parce qu'il restreint les angles de vue autorisés, il permet un rendu simple et rapide qui profite aux machines très modestes programmées avec du code natif, comme aux machines plus puissantes lorsqu'il faut interpréter du code portable, du type *bytecode* sur le Web. La structure de données retenue est cependant compatible avec une représentation par polygone, de sorte que le client peut tout-à-fait décider d'utiliser un moteur de rendu classique, avec tous les degrés de libertés.

On a ensuite proposé une **architecture** qui permet de gérer l'image de ces clones, leur position à l'écran. Pour permettre un plus grand nombre de participants simultanés sans compromettre la qualité de l'image proposée, on adopte le principe **d'une régie automatique**, qui alterne les vues **pour filmer la conférence comme s'il s'agissait d'un débat dans un lieu unique**. En choisissant une approche automatique, on libère l'utilisateur du système, qui peut se concentrer sur le débat, à la fois comme spectateur et comme participant. Le cadre formel proposé a été validé avec différents types de caméras virtuelles, qui fournissent de tous les participants une image reconstruite à base de clones. À cette occasion, l'utilisation d'un environnement sonore, enrichi pour mieux restituer à l'auditeur les interventions distantes, a été discutée, en particulier à l'aide **d'expérimentations pour le débat d'associations entre l'image et le son spatialisé**.

Finalement, on a défriché une piste plutôt neuve pour une **télé-représentation vidéo-réaliste pilotée par caméra**, qui se démarque des solutions de synthèse pure utilisée pour l'animation des clones. En particulier, les besoins d'analyse sont bien plus faibles, puisque l'on se contente **d'incruster l'image des yeux, des sourcils et de la bouche à la surface d'un clone statique**. On laisse ainsi au spectateur la responsabilité **d'interpréter les expressions**

visibles, dans toute leur dimension vidéo. À l'aide d'une brique sommaire de vision, les méthodes de composition proposées ont pu être validées en temps-réel.

Quoique ces contributions soient parfaitement indépendantes, ce qui à nos yeux fait l'unité de cette thèse est d'avoir rendu possible un **prototype efficace**, correspondant à un scénario applicatif assez ambitieux, et dont les performances en terme de vitesse permettent vraiment de se rendre compte de ce que serait un outil pleinement fonctionnel **favorisant communication sonore et empathie visuelle**.

Clairement, ce prototype n'est pas et ne se veut pas l'avant-projet d'un produit commercial, mais avec les contraintes qu'on a tolérées et à petite échelle, il a déjà permis d'expérimenter plusieurs scénarios réalistes.

Perspectives des contributions

Bien sûr, dans comme dans toute démarche scientifique, il faut remettre en cause chaque hypothèse pour s'assurer de sa validité, se demander si le résultat ne pourrait pas être obtenu autrement, et s'il ne pourrait pas être appliqué à d'autres domaines ou généralisé.

Clairement, l'approche employée pour la vision est le point faible dans le cadre applicatif envisagé. Avec l'évolution des résultats [YCH89, BY95, Rei95, HB96, OPB97], il ne fait aucun doute qu'elle pourra un jour être remplacée.

Avec le développement du matériel, il est probable que le champ d'application de l'algorithme de rendu se restreigne (même si les contraintes d'angles peuvent permettre de gagner sur la complexité du modèle et donc rester imposées avec un moteur de rendu générique). Il pourrait cependant être intéressant d'envisager son utilisation dans le cadre d'un rendu de foules (en déplacement ou sur des gradins). La restriction sur l'angle de vue de la caméra n'empêche pas que les clones modélisés puissent présenter une tête inclinée, lors de la capture.

La perspective la plus prometteuse est bien sûr de générer un codeur compatible MPEG-4. Un support 3D, avec une texture mise-à-jour est en effet tout à fait réalisable dans la norme existante. En embrassant ce standard, on bénéficie de tous les outils puissants et efficaces qu'il intègre et propose, pour compresser et transmettre efficacement les flux audios et vidéos que l'on a générés. Toujours dans cette norme, l'intégration de décors 3D, des corps et de leur gestuelle ou d'autres services pendant la conférence est grandement facilitée.

Perspectives du domaine

Les nombreuses contributions et projets qui tendent au développement de nouveaux outils de communication assistée par ordinateur ne laissent aucun doute sur les possibles succès scientifiques, puis technologiques qui sont en train de se jouer : un jour, le vidéophone de haute qualité sur le réseau téléphonique ne sera plus une vision, mais une réalité courante. Mais il se pourrait bien que ce soit en 3D. Et que de nombreux autres services à base de codages hybrides, synthétiques et naturels, l'aient précédé sur d'autres médias.

Annexes

Annexe A

Modèle de couleur

Il est nécessaire de pouvoir faire la différence entre les marqueurs (bleus ou verts), la peau et par exemple les sourcils. L'éclairage n'étant pas uniforme sur le visage, voire changeant, on ne pourra pas obtenir de résultat robuste si l'on ne fait pas abstraction de la luminosité. Il faudra se restreindre à des classifications et comparaisons sur la teinte.

A.1 Le codage des informations visuelles

A.1.1 L'équivalence des modèles de couleurs pour la vision humaine

Que ce soit sur un tube cathodique ou à l'impression papier, on cherche à restituer pour la vision humaine sa sensation de couleur. Physiquement, l'oeil est sensible à tout un intervalle de longueurs d'ondes, de $360nm$ à $830nm$, l'intervalle du visible. Mais les capteurs rétiniens ne font pas la différence entre les longueurs d'ondes qui composent une information lumineuse, mais seulement en fonction de leur impact physiologique sur notre système récepteur, nous classant parmi les tri-chromate. Il y a donc une infinité de façons de coder une même sensation de couleur. Par exemple, on ne peut pas savoir si la même impression de couleur orange est produite par une seule longueur d'onde très pure ou par différents mélanges de rouges et de jaunes. L'analogie auditive serait qu'on fabriquerait un MI parfait en mélangeant un DO et un LA.

Une conséquence utile de la perception humaine par couleurs, c'est que de nombreux modèles de couleurs pourront servir de façon équivalente à encoder une information visuelle.

A.1.2 Le modèle YCbCr

Il correspond aux signaux qu'on mesure depuis certaines caméras :

- **Y** désigne la luminance, qui permet d'obtenir un signal monochrome («noir et blanc», en fait une gamme continue de luminosités) de la scène,

- **Cb, Cr** sont des signaux de chrominance, qui complètent le signal précédent en exprimant les variations de teinte.

A.1.3 Le modèle RGB

Il correspond aux signaux qu'affichent les tubes cathodiques des téléviseurs ou moniteurs, et est parfois le seul format couleur disponible en digitalisation. Il n'y a pas de séparation nette entre le codage de la couleur et celle de la luminosité, mais cela n'est pas un problème puisqu'il y a équivalence, par exemple avec le modèle précédent, à l'aide de transformations linéaires :

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} .M \text{ ou bien } \begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} R \\ G \\ B \end{bmatrix} .M^{-1}$$

A.2 Un modèle d'éclairage sommaire

On cherche à prévoir quelles seront les observations possibles d'une série de marqueurs qui sont initialement identiques (même couleur, reflectivité etc) mais seront perçus différemment, à cause de leurs positions par rapport aux sources d'éclairages.

Un modèle théorique simple consiste à dire que la teinte perçue n'est pas modifiée, seulement la luminosité. Cela présuppose entre autres que les seules sources d'éclairages sont en lumière blanche.

A.3 Les problèmes pratiques

Trop lumineux, on a un risque de saturation de certaines composantes du signal vidéo, ce qui provoque une dérive de la teinte mesurée. En pratique, on ne doit pas faire confiance à un pixel dont une des composantes est saturée. Ainsi, un reflet «trop blanc» ou présentant une dérive de couleur peut apparaître sur l'image d'un marqueur vert ou très près d'un marqueur.

Trop noir, la confiance sur la teinte est faible : les caméras à base de CCD captureront en effet un signal bruité et incertain. C'est encore plus vrai dans un environnement où des tubes néons qui papillotent participent à l'éclairage. De même, l'écran d'un moniteur émet une pollution lumineuse/chromatique qui n'est généralement pas en phase avec le dispositif de capture (par exemple 72 Hz pour un écran, et 50 Hz ou 60 Hz pour la caméra).

En conséquence, en présence d'une mesure où la luminosité est trop faible, ou qui présente au moins une composante saturée, on ne peut pas décider si le point associé appartient ou n'appartient pas à la classe recherchée. Sans critères supplémentaires, par exemple de connexité ou de proximité, il faut donc trois classes pour étiqueter les résultats.

A.4 Classification des pixels

Étant donné un modèle de référence (Y_0, Cb_0, Cr_0), une tolérance t et une distance dans l'espace des chrominances $d(,)$, on classifiera un pixel de couleur $(Y, Cb, Cr) \equiv (R, G, B)$ comme :

- **'Indéterminé'**, si $Y < 10$ ou $R > 254$ ou $G > 254$ ou $B > 254$
- **'Marqueur'**, s'il n'est pas indéterminé et que $d((Cb, Cr), (Cb_0, Cr_0)) \leq t$

– **'Non marqueur'**, sinon.

Plutôt qu'une distance euclidienne, qui n'introduit aucune direction privilégiée ni aucun couplage entre les composantes, on peut par exemple utiliser la distance de Mahlanobis et une analyse statistique d'échantillons de référence de peau ou des marqueurs verts pour trouver des valeurs représentatives de (Y_0, Cb_0, Cr_0). L'ensemble des points de l'espace de couleur à une distance donnée de la valeur moyenne est alors un ellipsoïde, dont les axes et les rayons ont été fixés par la distribution des échantillons de référence.

Annexe B

Formats d'images du Web

Cette annexe rappelle les principes et différences entre les trois formats de compression d'image les plus utilisés sur le Web : GIF, PNG et JPEG.

B.1 Compression sans perte

Le principe consiste à chercher et factoriser une information qui est redondante dans l'image. Par exemple lorsque des groupes (en ligne ou en blocs) de pixels voisins se ressemblent (couleur constante) ou apparaissent en divers endroits (des trames par exemple), ou à diverses échelles.

Le format GIF, qui connaît une grande utilisation sur le Web – malgré des restrictions de droits et de brevets sur une partie de l'algorithme – consiste à étiqueter, avec l'espoir de les réutiliser, des séquences de plus en plus longues de pixels (en ligne seulement). Lorsqu'une séquence étiquetée réapparaît plus loin dans l'image, on pourra la référencer, et compresser par exemple les à-plats ou trames de l'image. Cette technique (et de nombreuses variantes, par exemple pour certains FAX) est donc efficace pour les schémas, ou les images avec peu de couleurs (avec seulement 16 couleurs, le nombre de combinaisons possibles des successions de pixels est relativement réduit).

Plus récent, le format PNG optimise le choix des séquences de pixels en s'intéressant à celles qui sont réutilisées, et choisit la taille et le codage de chacune des étiquettes selon le nombre effectif de ses occurrences (par un codage d'Huffman, libre de droits). Parce qu'il compresse de façon généralement plus compacte que le GIF, sans handicaper notablement le temps de décompression et que son utilisation n'est contrainte par aucun brevet, ce format devrait remplacer GIF, notamment sur le Web, comme le recommande W3C, le consortium chargé de son évolution.

Limites de ces approches

Avec les deux formats précédents, l'algorithme est en fait basé sur la compression d'un signal 1D, succession des pixels de gauche à droite. On ne tire pas directement parti de la ressemblance fortement probable entre les pixels d'une ligne et ceux des lignes voisines. Ce n'est pas le cas de

codages par régions, par exemple avec des maillages réguliers comme les quadtrees ou adaptatifs. Hélas, ceux-ci ne sont pas (pas encore) standardisés et sont plus adaptés aux compressions avec pertes.

En effet, avec la contrainte d'une fidélité parfaite, les taux de compression ne sont jamais très élevés sur des images naturelles, comme celles d'un extérieur ou d'un visage. Dans le cadre d'une communication vidéo, ce ne sont donc pas de telles techniques qui sont employées.

B.2 Compression avec pertes

Si l'on tolère que l'image soit altérée, les opportunités de compression sont forcément égales ou meilleures. Si le document ne devait pas servir à effectuer un diagnostique, ou que l'on ne dispose pas du document original pour effectuer la comparaison, des taux de compression plus intéressants sont possibles avec des techniques moins fidèles. D'autant que les erreurs (ou les tatouages de protection des droits) introduites ne sont pas forcément décelables si elles restent probables.

Idéalement, on pourrait tirer parti, pour modifier l'image avec le minimum d'impact visible, des modèles sur les imperfections de la perception humaine, comme c'est le cas pour le son.

En pratique, le format JPEG n'applique ce précepte que pour décimer les données de chrominance (par exemple, en diminuant leur résolution horizontalement et/ou verticalement par rapport au plan de luminance). Chacun des blocs 8 x 8 de l'image est ensuite transformé sous une forme fréquentielle par une DCT (transformée en cosinus discrets). Ce sont les coefficients obtenus qui sont modifiés (arrondis puis quantifiés) pour diminuer le nombre de valeurs observées et non-nulles. En élaguant préférentiellement c'est à dire plus fortement les coefficients associés aux fréquences plus hautes, on altère surtout des détails ou du bruit. La matrice de coefficients quantifiés (qui inclue la valeur moyenne, généralement proche de la moyenne du bloc précédent) se prête alors bien à un encodage : seul un sous-ensemble de certaines valeurs reste observable, séparé par un nombre espéré important de coefficients nuls. Ces deux informations sont conjointement transcrites avec un code d'Huffman pour former le flux de bit qui codera l'image finale, avec des marqueurs spéciaux pour la détection (et la resynchronisation) en cas d'erreurs lors de la transmission.

Ces opérations lui permettent de compresser et coder efficacement des images plus ou moins dégradées dès lors qu'on ne zoome pas sur les détails, qui font apparaître les discontinuités entre les blocs 8 x 8 de l'image.

Conclusion

Bien sûr, il existe de nombreux autres formats de compression d'images, mais leur manque de standardisation et la généralisation de brevets grèvent leur utilisation, dans le cadre d'applications gratuites pour et sur le Web notamment. C'est pourquoi les formats PNG (GIF dans les faits) et JPEG restent si utiles et tellement usités.

Annexe C

Son localisé

Par la perception de réverbérations ou la composition spectrale d'un son connu, un auditeur déduit naturellement de nombreuses informations de son environnement. On peut donc penser à simuler pour un auditeur l'impression qu'une source sonore est à une position donnée, dans un contexte donné (petite salle vide ou extérieur par exemple). Bien sûr, les calculs à mettre en œuvre peuvent être plus ou moins complexes, et dépendent en particulier des moyens de restitution qui seront employés : casque, paire ou réseau d'enceintes.

Très modestement, il est possible de simuler avec une paire d'enceintes l'impression qu'un son provient d'une direction donnée, en jouant seulement sur les volumes et retards relatifs des deux canaux, deux indices de la localisation sonore qui étaient déjà recensés par Lord Rayleigh au début du siècle.

C.1 Modèle de propagation

Dans un espace libre, les ondes sonores sont sphériques. À longue distance, par rapport à la taille de la source et aux longueurs d'ondes mises en jeu, on néglige les phénomènes d'interférences, et on peut considérer les ondes comme localement planes.

FIG. C.1 – *Approximation d'une onde sonore par une onde plane*

On suppose aussi qu'il n'y a pas de vent et que la température et l'hygrométrie sont constantes. Dans ces conditions, la vitesse de propagation est constante, et seuls les paramètres liés au son, notamment sa décomposition fréquentielle, ne sont pas fixés.

C.2 Modélisation du retard de perception

On approxime la tête de l'auditeur par une sphère parfaite de rayon R, avec les oreilles situées aux extrémités d'un diamètre. On cherche quel retard de perception entre les deux oreilles résulterait de la différence de temps de propagation, selon la direction d'où provient le son.

Dans le cas particulier d'une onde provenant exactement d'en face, les deux oreilles sont à égale distance et aucun retard relatif n'est perçu, comme l'illustre la partie gauche de la figure C.2.

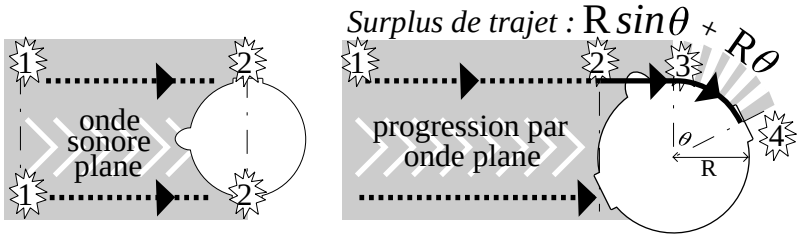


FIG. C.2 – *Trajet et perception d'une onde sonore*

Dans le cas général, l'onde quasi-plane n'arrivait pas de face et a atteint l'une des deux oreilles en premier. Après cette date, le front de l'onde a continué à se propager de façon rectiligne pour dépasser la tête, mais devra contourner le visage pour atteindre l'autre oreille en décrivant un arc de cercle[1]. Comme le résume la figure C.2, la distance supplémentaire parcourue par l'onde depuis la première oreille est donc $\Delta D = R\sin\theta + R\theta$, d'où un retard temporel perçu de :

$$\Delta t = \frac{R}{C}(\sin\theta + \theta) \tag{C.1}$$

où C figure la vitesse de propagation (moyenne) du son dans l'air. On peut noter que cette équation, tirée du modèle de Woodsworth présenté dans [Kit94], s'applique aussi pour le cas de symétrie où $\theta = 0$, comme pour des valeurs de θ étendues au cas signé, puisqu'on s'intéresse au retard ou à l'avance relatives.

C.3 Modélisation de l'atténuation en volume

Plus que l'atténuation depuis la source jusqu'aux deux oreilles, ce qui nous intéresse est un rapport des volumes perçus. En effet, dans le cadre d'une discussion, on souhaite que le message sonore soit toujours audible, avec un volume moyen qui semble constant. On ne cherche pas à

1. il s'agit d'une onde sonore, pas de photons ou d'un faisceau de particules lancées en ligne droite. D'après le théorème des sources secondaires, chaque point de l'espace peut être considéré comme une source qui rayonnerait une part de l'énergie reçue. La propagation se fait de proche en proche, et la propagation rectiligne des ondes sonores en espace libre peut n'être vue que comme la résultante de ces contributions, au même titre que la propagation autour des obstacles à ces longueurs d'ondes.

rendre la notion de distance de la source, seulement la direction et le déplacement de la source, qui se manifestent par la dissymétrie de la perception et son évolution.

Dans ces conditions, la différence de distance parcourue par les deux parties de l'onde est faible, au maximum le diamètre de la tête. On négligera alors la plupart des causes d'atténuation, notamment l'altération de la composition fréquentielle dues aux frictions et relaxations des molécules de l'air, pour ne garder que l'atténuation géométrique : la puissance émise se répartit sur des sphères dont la surface est de plus en plus grande, et décroît donc suivant une loi en $\frac{1}{r^2}$.

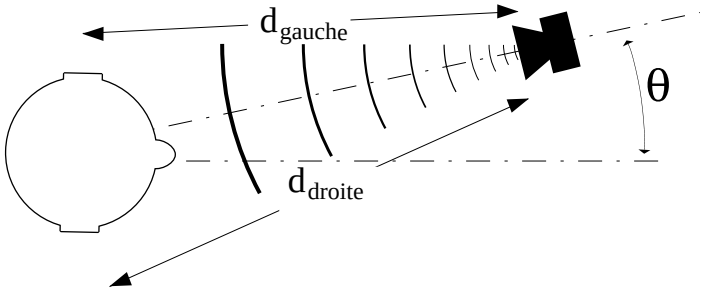


FIG. C.3 – *Positions relatives de l'auditeur et de la source*

Si l'on utilise directement cette formule, en appliquant à chaque oreille une atténuation suivant l'inverse du carré de sa distance à la source (d_{gauche} et d_{droite} respectivement, sur la figure C.3), on reproduit bien un effet de distance et de position. Cependant, on ne veut pas pénaliser certains locuteurs par rapport à d'autres, par exemple en les rendant moins audibles. Pour conserver l'intelligibilité de tous, on décide de ne pas modifier le volume global perçu. Comme si un ingénieur du son compensait la distance en amplifiant plus ou moins le son capté, on va réaliser une commande automatique du gain entre les deux canaux.

Si l'on veut préserver l'intensité globale perçue par l'oreille, il faut conserver la quantité à laquelle elle est sensible, qui n'est pas l'amplitude mais plutôt l'intensité [Roa96] :

$$I_{tot} = \sqrt{\left(\frac{Amp}{d^2_{gauche}}\right)^2 + \left(\frac{Amp}{d^2_{droite}}\right)^2}$$

Ainsi renormalisée pour ne pas changer notablement son volume global, la source sonore sera répartie entre les canaux gauches et droits, par des atténuations respectives de $1/d^2_{gauche}$ et $1/d^2_{droite}$, pour simuler une direction de perception d'angle θ.

C.4 Conclusion : utilisation pratique

La méthode des déphasages produit une impression plus convaincante, surtout sur les sons graves et de fréquence assez stable. Mais l'oreille perçoit peu les déphasages sur les sons aigus, et le cerveau se base alors sur la différence de volume apparent pour estimer la direction du son.

Il y a donc nécessité d'utiliser les deux effets combinés lors de la restitution du son : modification des volumes et phases, relativement aux deux canaux.

Bibliographie

[ACK94] S. Aoki, M. Cohen, and N. Koizumi. Design and control of shared conferencing environments for audio telecommunication using individually measured HRTFs. *Presence*, 3(1):60–72, 1994.

[AHMS97] M. Ackerman, D. Hindus, S. Mainwaring, and B. Starr. Hanging on the 'wire : A field study of an audio-only media-space. *ACM transactions on Computer-Human Interaction*, 4(1):39–66, March 1997.

[ASW93] T. Akimoto, Y. Suenega, and R. S. Wallace. Automatic creation of 3D facial models. *IEEE Computer Graphics and Applications*, 13(5):16–22, 1993.

[BA99] C. Beumier and M. Archeroy. 3D facial surface acquisition by structured light. *International Workshop on SNHC and 3D Imaging*, 1999.

[Bas79] J. N. Bassili. Emotion recognition : the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 37, 1979.

[BCS97] C. Bregler, M. Covell, and M. Slaney. Video rewrite : Driving visual speech with audio. *Proceedings of SIGGRAPH '97*, pages 353–360, August 1997.

[Ben97] Wm. Cullen Bengston. Digitizing on a shoestring. *Computer Graphics World*, pages 75–76, April 1997.

[BHI93] S. A. Bly, S. R. Harrisson, and S. Irwin. Media spaces : Bringing people together in a video, audio and computing environment. *Communications of the ACM*, 36(1):28–47, January 1993.

[Bir97] Stan Birchfield. An elliptical head tracker. *31st Asilomar Conference on Signals, Systems, and Computers*, November 1997.

[BL94] Patrice Bourcet and Pierre Liénard. *Le livre des techniques du son*, volume 1, chapter Accoustique fondamentale, pages 13–44. Eyrolles, fréquences edition, 1994.

[BLM92] C. Benoît, M. T. Lallouache, and T. Mohamadi. A set of french visemes for visual speech synthesis. *Talking machines : theories, models and designs*, pages 485–504, 1992.

[BN92] T. Beier and S. Neely. Feature-based image metamorphosis. *Proceedings of SIGGRAPH'92*, 26(2):35–42, July 1992.

[BSHP97] M. Bourges-Sévenir, P. Horain, and F. Prêteux. Recalage d'un modèle 3D générique sur une sequence d'images 2D. *CORESA*, 1997.

[BSP93] D. Beymer, A. Shashua, and T. Poggio. Example based image analysis and synthesis. *AI Memo No. 1431, MIT*, November 1993.

[BT] British Telecom. the Synthetic Personna project. http://www.labs.bt.com/people/welshwj/.

[BT97] P. J. L. Van Beek and A. M. Tekalp. Object-based video coding using forward tracking 2D mesh layers. *Proc. of VCIP'97*, February 1997.

[Bux92] W. Buxton. Telepresence : integrating shared task and person spaces. *Proceedings of Graphics Interface*, pages 123–129, 1992.

[BV99] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *Computer Graphics Proceedings*, 1999.

[BY95] M. J. Black and Y. Yacoob. Tracking and recognizing facial expressions in image sequences using local parametrized models of image motion. *International Conference on Computer Vision*, pages 374–381, 1995.

[C3D] Turing Institute. C3D 2020 3D portrait camera. http://www.turing.gla.ac.uk/products/2020.htm.

[CBCC98] J. Coutaz, F. Bérard, E. Carraux, and J. Crowley. Early experience with the mediaspace CoMedi. *EHCI'98*, 1998.

[CLH97] A. Colmenarez, R. Lopez, and T. S. Huang. 3D model-based head tracking. *Visual Communications and Image Processing 97*, 1997.

[CPB+94] J. Cassel, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated converstation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversationnal agents. *SIGGRAPH Proceedings*, 1994.

[CPN+97] T. K. Capin, I. S. Pandzic, H. Noser, N. Magnenat-Thalmann, and D. Thalmann. Virtual human representation and communication in VLNet. *IEEE Computer Graphics and Applications, Special Issue on Multimedia Highways*, pages 42–53, March-April 1997.

[CR98] T. Chan and R. R. Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–852, May 1998.

[CW93] S. E. Chen and L. Williaws. View interpolation for image synthesis. *SIGGRAPH '93 Proceedings*, pages 279–288, 1993.

[Cyb] Cyberware laboratory. http://www.cyberware.com/.

[DB92] P. Dourish and S. Bly. Portholes : supporting awareness in a distributed work group. *CHI'92*, pages 541–547, 1992.

[DMS98] D. DeCarlo, D. Metaxas, and M. Stone. An anthropometric face model using variational techniques. *Computer Graphics Proceedings*, 1998.

[DTM96] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs : A hybrid geometry-and image-based approach. *SIGGRAPH Proceedings*, pages 11–20, 1996.

[ECG97] Peter Eisert, Subhasis Chaudhuri, and Bernd Girod. Speech driven synthesis of talking head sequences. *3D Image Analysis and Synthesis*, pages 51–56, November 1997.

[EF] P. Ekman and W. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Palo Alto, CA.

[EG97a] P. Eisert and B. Girod. Facial expression analysis for model-based coding of video sequences. *Proc. Intern. Picture Coding Symposium*, pages 33–38, September 1997.

[EG97b] M. Emerit and A. Gilloire. Application des techniques de spatialisation sonore à la télécommunication de groupe. *CORESA*, 1997.

[EG98] Peter Eisert and Bernd Girod. Model-based coding of facial image sequences at varying illumination conditions. *IMDSP Workshop '98*, July 1998.

[EI99] F. Elisei and P. Inostroza. Video-driven real-time update of eyes and mouth regions on the texture of a 3D head model. *International Workshop on SNHC and 3D Imaging*, 1999. http://www-vis.imag.fr/Elisei/Pubs/e_snhc99.ps.gz.

[Eli97] F. Elisei. Visages pour vidéo-acteurs 3D. *CORESA*, 1997. http://www-vis.imag.fr/Elisei/Pubs/f_coresa97.ps.gz.

[Eli98] Frédéric Elisei. Clones et vidéo. *CORESA*, 1998. http://www-vis.imag.fr/Elisei/Pubs/f_coresa98.ps.gz.

[EMT97] M. Escher and N. Magnenat-Thalmann. Automatic 3D cloning and real-time animation of a human face. *Computer Animation proceedings*, pages 58–66, 1997.

[EP96] T. Ezzat and T. Poggio. Facial analysis and synthesis using image-based models. *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, October 1996.

[EP98] T. Ezzat and T. Poggio. Miketalk : A talking facial display based on morphing visemes. *Proceedings of the Computer Animation Conference*, June 1998.

[Ess95] Irfan A. Essa. *Analysis, Interpretation and Synthesis of Facial Expressions*. PhD thesis, Media Arts and Science, MIT, 1995.

[FANa] Facial analysis links. http://mambo.ucs.edu/psl/fanl.html.

[FANb] Facial animation links. http://mambo.ucs.edu/psl/fan.html.

[FD] D. Forsey and J.-L. Duprat. Facemaker: The human face with MPEG facial action parameters. http://zeppo.cs.ubc.ca:5656/faceMPG.html.

[For97] R. Forchheimer. CANDIDE demo. http://www.bk.isy.liu.se/candide/candemo.html, 1997.

[FT97] P. Fieguth and D. Terzopoulos. Color-based tracking of heads and other mobile objects at video frame rates. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–27, 1997.

[Fua98] Pascal Fua. Face models from uncalibrated video sequences. *Modelling and Motion Captures Techniques for Virtual Environments*, pages 214–227, International Workshop, CAPTECH'98. Lecture Notes in Artifical Intelligence 1537.

[FvDFH95] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics, principles and practice (2nd ed. in C)*, chapter Visible-surface determination of two variables. Addison-Wesley, 1995.

[GGW+98] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin. Making faces. *Computer Graphics Proceedings*, 1998.

[GKB98] A. Georghiades, D. Kriegman, and P. Belhumeur. Illumination cones for recognition under variable lighting : Faces. *Proc. IEEE Conf. on Comp. Vis. and Patt. Recog.*, pages 52–58, 1998.

[Hai84] N. Haig. The effect of feature displacement on face recognition. *Perception*, 13:505–512, 1984.

[Hal95] P. Hallinan. A low-dimensionnal lighting representation of human faces under arbitrary lighting conditions. *Proc. IEEE Conf. on Comp. Vis. and Patt. Recog.*, pages 995–999, 1995.

[HB96] G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 1996.

[HCS96] L.-W. He, M. F. Cohen, and D. H. Salesin. The virtual cinematographer: A paradigm for automatic real-time camera control and directing. *Computer Graphics Proceedings*, 1996.

[HK93] Pat Hanrahan and Wolfgang Krueger. Reflection from layered surfaces due to subsurface scattering. *SIGGRAPH 93 Proceedings*, 1993.

[Hog92] Burne Hogarth. *Le dessin anatomique facile*. Taschen, 1992.

[HPSW97] B. Hofer, F. Parke, D. Sweetland, and K. Waters. Panel on facial animation : Past, present and future. In *SIGGRAPH 97 Proceedings*, pages 434–436, 1997.

[IE] P. Inostroza and F. Elisei. 3D-finger java applet on-line. http://www-vis.imag.fr/3Dfinger/.

[INA] INA. Televirtuality project : Cloning and real-time animation system. http://www.ina.fr/INA/Recherche/TV/.

[INS] Inspeck 3D capturor. http://www.cyberware.com/.

[Kin] Scott King. Facial animation overview. http://www.cis.ohio-state.edu/ sking/FacialAnimation.html.

[Kit94] Mpaya Kitantou. *Le livre des techniques du son*, volume 1, chapter La perception auditive, pages 155–182. Eyrolles, fréquences edition, 1994.

[KK89] J. T. Kajiya and T. Kay. Rendering fur with three dimensional textures. *Computer Graphics Proceedings*, 1989.

[KMMTT91] P. Kalra, A. Mangili, N. Magnenat-Thalmann, and D. Thalmann. SMILE : a multilayered facial animation system. *Proc. IFIP Conference on Modelling in Computer Graphics*, 1991.

[Koc93] Reinhard Koch. Dynamic 3D scene analysis through synthesis feedback control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):556–568, June 1993.

[KTH] Interaction with an animated agent in a spoken dialogue system. http://www.speech.kth.se/August/.

[Lav99] Fabio Lavagetto. VIDAS : analysis/synthesis tools for natural-to-virtual face representation. *Proceedings of ECMAST'99*, pages 348–363, 1999. Lecture Notes in Comp. Science 1629.

[LCH97] R. Lopez, A. Colmenarez, and T. S. Huang. Head and feature tracking for model-based video coding. *International Workshop on SNHC and 3D Imaging*, 1997.

[Lee93] Yuencheng Lee. The construction and animation of functional facial models from cylindrical range reflectance data. Master's thesis, University of Toronto, 1993.

[LLS98] P. Lechat, N. Laurent, and H. Sanson. Représentation d'images et estimation de mouvement basées maillage, application à un codeur tout maillage. *CORESA 98*, Juin 1998.

[LM97] R. Lenz and P. Meer. Illumination independent color image representation using log-eigenspectra. http://www.isy.liu.se/~reiner/coco/workshop/workshop.html, 1997.

[LMT98] W.-S. Lee and N. Magnenat-Thalmann. From real faces to virtual faces : problems and solutions. *Proc. 3IA'98, Limoges*, pages 5–19, 1998.

[LRF93] H. Li, P. Roivainen, and R. Forchheimer. 3D motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555, June 1993.

[LTW93] Y. Lee, D. Terzopoulos, and K. Waters. Constructing physics-based facial models of individuals. *Graphics Interface '93*, pages 1–8, 1993.

[LTW95] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. *Proceedings of SIGGRAPH'95*, pages 55–61, 1995.

[LW94a] A. R. Lebeck and D. A. Wood. Cache profiling and the SPEC benchmarks : A case study. *IEEE Computers*, 27(10):15–26, October 1994.

[LW94b] P. Litwinowicz and L. Williams. Animating images with drawings. *Proceedings of SIGGRAPH'94*, pages 409–412, 1994.

[MFL98] G. Marquant, H. Le Floch, and C. Labit. Génération et suivi de maillages adaptatifs : un état de l'art et quelques résultats. *CORESA 98*, Juin 1998.

[Mir] MIRALab's movies. http://www.miralab.unige.ch/Films/index.html.

[MK92] M. Miyoshi and N. Koizumi. NTT's research on acoustics for future telecomunication services. *Applied Acoustics*, 36:307–326, 1992.

[MPB99] M. Malciu, F. Prêteux, and V. Buzuloiu. 3D global head pose estimation : a robust approach. *IWSNHC3DI*, 1999.

[MPE99] MPEG-4 overview. http://www.cselt.it/ufv/leonardo/mpeg/standards/mpeg-4/mpeg-4.htm, March 1999.

[MTKE98] N. Magnenat-Thalmann, P. Kalra, and M. Escher. Face to virtual face. *Proceedings of the IEEE*, 86(5):870–883, 1998.

[MTKP95] N. Magnenat-Thalmann, P. Kalra, and I. S. Pandzic. Direct face-to-face communications between real and virtual humans. *International Journal of Information Technology*, 1(2):145–157, 1995.

[MTPT88] N. Magnenat-Thalmann, E. Primeau, and D. Thalmann. Abstract muscle action procedures for human face animation. *Visual Computer*, 3:290–297, 1988.

[NEG98] R. Nicol, M. Emerit, and A. Gilloire. Mur de téléprésence pour la visioconférence : une approche holophonique. *CORESA*, 1998.

[OPB97] N. Oliver, A. P. Pentland, and F. Bérard. LAFTER : Lips and face real time tracker. *IEEE CVPR*, 1997.

[OTO+87] M. Oka, K. Tsutsui, A. Ohba, Y. Kurauchi, and T. Tago. Real-time manipulation of texture-mapped surfaces. *Proceedings of SIGGRAPH'87*, 21(4):181–188, 1987.

[OTP] Ontario telepresence project. http://www.dgp.utoronto.ca/tp/tphp.html.

[Pan] ACTS-AC092 PANORAMA (25-mar-1999). http://www.tnt.uni-hannover.de/project/eu/panorama/.

[Par82] F. I. Parke. Parameterized models for facial animation. *IEEE Computer Graphics*, 2(9):61–68, 1982.

[PB81] S. M. Platt and N. I. Badler. Animating facial expression. *ACM SIGGRAPH Conference Proceedings*, 15(3):245–252, 1981.

[PHSS98] F. Pighin, J. Hecker, R. Szeliski, and D. Salesin. Synthesizing realistic facial expressions from photographs. *Computer Graphics Proceedings*, 1998.

[PSS99] F. Pighin, R. Szeliski, and D. Salesin. Resynthesizing facial animation through 3D model-based tracking. *Proceedings of ICCV*, 1999.

[Rei95] M. Reinders. *Model adaptation for image coding*. PhD thesis, Delft University of Technology, Information Theory Group, 1995.

[Roa96] Curtis Roads. *The Computer Music Tutorial*, chapter Localization cues and Simulating the azimuth cue. MIT Press, 1996. pp. 457-470.

[Rom] Sam Romdhani. Face recognition using principal components analysis. http://www.elec.gla.ac.uk/~romdhani/pca.htm.

[Rou97] N. Roussel. Au-delà du mediaspace : un modèle pour la collaboration médiatisée. *Actes Neuvièmes Journées Francophones sur l'Interaction Homme Machine*, Septembre 1997.

[Ryd87] M. Rydfalk. *CANDIDE: A parametrized face*. PhD thesis, Linköping University, Departement of Electrical Engineering, October 1987.

[SGHS98] J. Shade, S. Gortler, L.-W. He, and R. Szeliski. Layered depth images. *Computer Graphics Proceedings*, 1998.

[SHSW99] A. Singer, D. Hindus, L. Stifelman, and S. White. Tangible progress : less is more in somewire audio spaces. *CHI'99 conference Proceeding*, pages 104–111, May 1999.

[SNH] MPEG-4 SNHC Web Home. http://www.es.com/mpeg4-snhc/.

[SNH97] MPEG-4 SNHC FAQ. http://www.cselt.it/ufv/leonardo/mpeg/faq/faq-snhc.htm, April 1997.

[Sol98] O. Soligon. *Modélisation et animation du buste humain pour la compression de séquences d'images visiophoniques*. PhD thesis, Université de Rennes 1, Mai 1998.

[SPA] Le spatialisateur de l'IRCAM. http://www.ircam.fr/equipes/salles/spat/index.html.

[ST93] J. Shi and C. Tomasi. Good features to track. *Technical Report TR-93-1399, Cornell University*, November 1993.

[STS99] N. Sarris, G. Tzanetos, and M. G. Strintzis. Three dimensionnal model adaptation and tracking of a human face. *Proceedings of ECMAST'99*, pages 392–405, 1999. Lecture Notes in Comp. Science 1629.

[SVG95] A. Saulnier, M.-L. Viaud, and D. Geldreich. Real-time facial analysis and synthesis chain. *International Workshop on Automatic Face- and Gesture-Recognition*, pages 86–91, 1995.

[TABG96] N. Tsingos, A. Adjoudani, C. Benoît, and M.-P. Gascuel. 3D models of the lips for realistic speech animation. *Computer Animation'96*, 1996.

[TEGK97] C.-J. Tsai, P. Eisert, B. Girod, and A. K. Katsaggelos. Model-based synthetic view generation from a monocular video sequence. *International Conference on Image Processing*, 1:444–447, October 1997.

[TH98] H. Tao and T. S. Huang. Bézier volume deformation model for facial animation and video tracking. *Modelling and Motion Captures Techniques for Virtual Environments*, pages 242–253, International Workshop, CAPTECH'98. Lecture Notes in Artificial Intelligence 1537.

[TL99] J.-M. Trivi and J. Lemordant. Use of a 3D positionnal interface for the implementation of a versatile graphical mixing console. *107th AES convention*, 5054(P-4), September 1999.

[Tov95] Martin James Tovee. Les gènes de la vision des couleurs. *La Recherche*, 26(272):26–33, Janvier 1995.

[TP94] M. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 1994.

[TW93] 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):569–579, June 1993.

[USA] Ultimate spatial audio index. http://www.dform.com/inquiry/spataudio.html.

[VAS] Simgraphics VActor. http://www.simg.com.

[VC] The multimedia conferencing applications archive. http://k2.avc.ucl.ac.uk/mice/.

[VD99a] S. Valente and J.-L. Dugelay. Analysis and reproduction of facial expressions for communicating clones. *IEEE MMSP'99*, September 1999.

[VD99b] S. Valente and J.-L. Dugelay. Face tracking and realistic animations for telecommunicant clones. *IEEE International Conference on Multimedia Computing and Systems*, June 1999.

[VDD98] Stephane Valente, Jean-Luc Dugelay, and Herve Delingette. An analysis/synthesis cooperation for head tracking and video face cloning. *ECCV Workshop on Perception of Human Action*, June 1998.

[VH] The visible human project. http://www.nlm.nih.gov/research/visible/visible_human.html.

[VL97] S. P. VanderWiel and D. J. Lilja. When caches aren't enough : data prefetching techniques. *IEEE Trans. on Computers*, 30(7):23–30, July 1997.

[VP97] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7), July 1997.

[VY92] M.-L. Viaud and H. Yahia. Facial animation with wrinkles. *3rd Workshop on animation, Eurographics'92*, 1992.

[W3C96] W3C recommandation : PNG specification v 1.0. http://www.w3.org/TR/REC-png, October 1996.

[Wat87] Keith Waters. A muscle model for animating three-dimensional facial expression. *Proceedings of SIGGRAPH'87*, 21(4):17–24, 1987.

[Wil90] Lance Williams. Performance-driven facial animation. *Proceedings of SIGGRAPH'90*, 24(4):235–242, August 1990.

[WKMT96] Y. Wu, P. Kalra, and N. Magnenat-Thalmann. Simulation of static and dynamic wrinkles of skin. *Proc. Computer Animation*, pages 90–97, 1996.

[WL94] K. Waters and T. Levergood. An automatic lip-synchronization algorithm for synthetic faces. *Proceedings of ACM Multimedia*, pages 149–156, 1994.

[WSHS90] W. J. Welsh, A. D. Simons, R. A. Hutchinson, and S. Searby. Synthetic face generation for enhancing a user interface. *Proceeding of Image Com.*, pages 177–182, 1990.

[Yau88] J. F. S. Yau. A texture mapping approach to 3-D facial image synthesis. *Computer Graphics Forum*, 7(2):129–134, 1988.

[YCH89] A. L. Yuille, D. Cohen, and P. Hallinan. Feature extraction from faces using deformable templates. *Proc. Comp. Vis. and Patt. Recog.*, pages 104–109, 1989.

[YS90] T. Yamana and Y. Suenaga. Hair image generation using functionally controlled anisotropic reflection. *Proceeding of Image Com.*, pages 189,193, 1990.

RESUME

Pour la téléconférence, on peut remplacer l'image des correspondants distants par des modèles 3D animés de leurs visages. En plus de taux de compression avantageux, cette approche offre les libertés du virtuel : on peut par exemple composer à l'écran l'impression d'un lieu unique, virtuel, où débattent les représentants 3D.

Cette thèse présente un algorithme de rendu spécifique, applicable à des clones 3D photo-réalistes de visages. En restreignant les angles de vue autorisés, il permet un rendu simple et rapide, même avec des ordinateurs peu puissants ou sur des machines virtuelles.

On propose aussi une architecture de régie automatique, reliée à des caméras virtuelles qui réagissent aux interventions des participants et en proposent une image synthétique. En alternant plusieurs vues (éventuellement partielles) de la scène, on autorise plus de participants simultanés, sans compromettre la qualité de l'image proposée ni l'intelligibilité du débat restitué. Automatique, cette approche libère l'utilisateur du système, qui peut se concentrer sur un débat rendu plus attractif, à la fois comme spectateur et comme participant.

Cette thèse rend aussi compte de la réalisation d'un prototype de communication qui intègre les éléments précédents et permet de juger la qualité de la communication obtenue. À cette occasion, l'utilisation d'un environnement sonore qui intègre les interventions distantes et leur localisation (dans ou hors de l'image) est discutée, avec plusieurs expérimentations sur l'association entre l'image et le son spatialisé.

Enfin, on introduit une solution hybride (3D et vidéo) pour animer les clones des visages. En incrustant à la surface d'un clone statique l'image des yeux, des sourcils et de la bouche vues par une caméra, on laisse aux spectateurs la responsabilité d'interpréter les expressions originales, dans toute leur dimension vidéo (forte résolution spatiale et temporelle). Un second prototype permet de juger de l'empathie visuelle.

TITLE 3D heads for audio and video communication

ABSTRACT

In teleconferencing applications, animated 3D heads can replace the usual video channels. This offers high compression opportunities, as well as the freedom of virtual spaces : one can compose a virtual place on screen, where 3D representations are debating.

This thesis introduces an ad-hoc rendering algorithm, that can be applied to photorealistic 3D heads, at the expense of slightly limited viewing angles. Fast renderings are achieved on simple computers as well as 2D virtual machines.

An automatic control architecture of the virtual cameras and the broadcasted view is also proposed. Cameras produce synthetic views, and react to speaking events. Switching between various (partial) camera views intends to let the debate look more attractive and intelligible. More simultaneous participants can take part, without their image size being lowered too much. The manual-interface-free automatic scheme enables the user to naturally talk and concentrate on the discussions.

The previous parts have been implemented in a prototype. Several sound-scene scenarios have been experimented, playing with the image and sound association in a spatial-sound environment, where events from or outside the viewed area can be simulated.

An hybrid video-based and 3D-based solution to the face-animating problem is defended as well. Partial images of eyes, mouth and eyebrows regions from live performance are inlaid on the clone surface texture. That way, it's up to the spectator to interpret the broadcasted video-like expressions. Another prototype has been built to test the real-time visual empathy.

DISCIPLINE Informatique

MOTS-CLES clones 3D, vidéoconférence, débat virtuel, téléprésence, SNHC, animation faciale.

Laboratoire GRAVIR – INRIA, ZIRST, 655 avenue de l'Europe, 38330 Montbonnot S^t Martin