ACQUISITION ET ANIMATION DE CLONES REALISTES POUR LES TELECOMMUNICATIONS

A. C. Andrés del Valle, J.-L. Dugelay, E. Garcia et S. Valente[♦]

Institut Eurécom
2229, route des Crêtes
06904 Sophia Antipolis
http://www.eurecom.fr/~image

1. INTRODUCTION

Les clones de visages sont de plus en plus utilisés dans de nombreuses applications multimédia, et figurent d'ores et déjà dans des standards émergeants comme MPEG-4. On peut distinguer deux familles de visages synthétiques : avatars et clones. Les avatars sont apparus les premiers et ne sont généralement qu'une représentation approximative ou symbolique d'une personne. Les clones sont quant à eux plus réalistes et doivent ressembler à un visage réel. Pour des nombreuses applications, il est souhaitable d'utiliser des clones réalistes (i.e. monolocuteur). Dans ce cas, plusieurs problèmes se posent en termes d'acquisition et d'animation. En effet, contrairement aux avatars qui sont générés à partir de modèles mathématiques simples, les modèles réalistes sont difficiles à acquérir sans utiliser un équipement spécialisé et coûteux (e.g. Cyberware) mais également difficiles à manipuler, surtout en temps réel, de par leur complexité. Dans cet article nous présentons nos travaux liés à l'acquisition (section 2) puis l'animation (section 3) de visages réalistes à partir de séquences vidéo réelles.

2. ACQUISITION DE MODELES 3D DE VISAGES

Nous avons développé un procédé d'acquisition de modèles 3D réalistes de visages nécessitant peu de moyens matériels : un rétroprojecteur, une caméra et un objet de calibration plan. Le principe [1] est de projeter une grille lumineuse sur le visage à reconstruire et d'analyser les déformations de cette grille dans l'image d'une caméra pour en déduire la géométrie 3D du visage (Fig. 2). La texture est obtenue à partir d'images du visage en l'absence de la grille lumineuse. Le système est détaillé dans [2].

Les étapes de notre algorithme sont présentées sur la Figure 1. Elles sont regroupés en 3 parties, comme détaillé ci-dessous:

- L'extraction des traits du visage et la calibration des images à l'aide des techniques de traitement d'image, (1) et (2) ;

[♦] Nouvelle adresse: Video and Communication Group. Laboratoires d'Electronique Philips (LEP), France.

- la calibration de la caméra et du rétroprojecteur basée sur la géométrie projective, (3) ;
- la reconstruction, l'enregistrement/fusion des parties du visage, et le texturage, (4), (5) et (6).

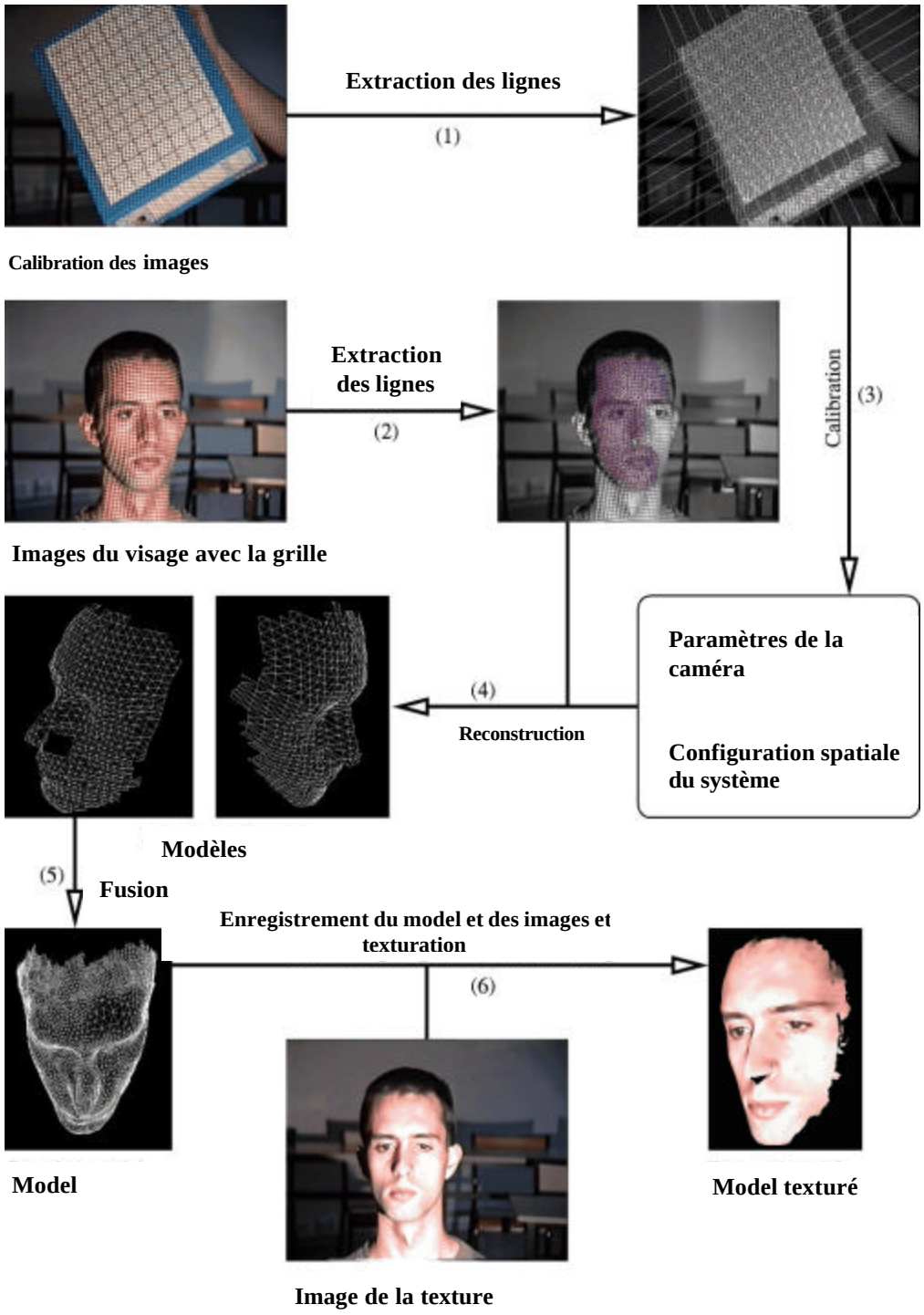

Figure 1 Diagramme du procédé complet d'acquisition de modèles de visage.

2.1 EXTRACTION DES TRAITS DU VISAGE ET CALIBRATION DES IMAGES

Avant d'effectuer les calculs liés à la calibration de la caméra et du rétroprojecteur, il faut tout d'abord estimer les coordonnées des nœuds des quadrillages observés dans l'image. Pour cela, nous utilisons différentes techniques de traitement d'image.

2.1.1. Analyse des images

Pour analyser les images de calibration, nous tirons parti du fait qu'elles comportent 4 faisceaux de droites presque parallèles (en adoptant adopte un modèle de projection perspective sans distorsion pour la caméra). Nous cherchons donc à déterminer la position de toutes ces droites et à caractériser les quadrillages observés par ces droites (ou les nœuds qui se trouvent à leurs intersections).

La détection des nombreuses droites est effectuée en utilisant la transformée de Radon (projection de l'image dans toutes les directions, similaire à la transformée de Hough). Les droites claires ou foncées de l'image correspondent à des points d'intensité maximum ou minimum dans la transformée de Radon. De plus, les droites appartenant à un même faisceau de droites presque parallèles dans l'image correspondent à un ensemble de points presque parfaitement alignés et équidistants dans l'espace de Radon.

La recherche de l'ensemble des droites se réduit alors à la recherche de 4 ensembles de points alignés. Nous appliquons ce principe en cherchant d'abord les 4 droites support des points alignés dans l'espace de Radon puis nous déterminons la position de chacun de ces points individuels, ce qui nous donne la position des droites des quadrillages dans l'image de départ.

Les images des quadrillages dans la caméra ou dans le rétroprojecteur, ainsi que leur géométrie dans l'espace peuvent se décrire via un ensemble d'homographies. Plus précisément, nous exprimons le problème de la calibration en termes de 7 homographies (Fig. 3). Le problème de la calibration revient à estimer ces homographies. Elles sont estimées directement à partir des coordonnées des nœuds des quadrillages dans l'image de la caméra, dont nous avons expliqué la détection plus haut.

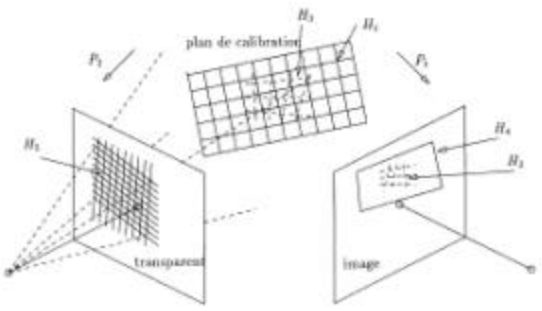

Figure 3 Diagramme des homographies.

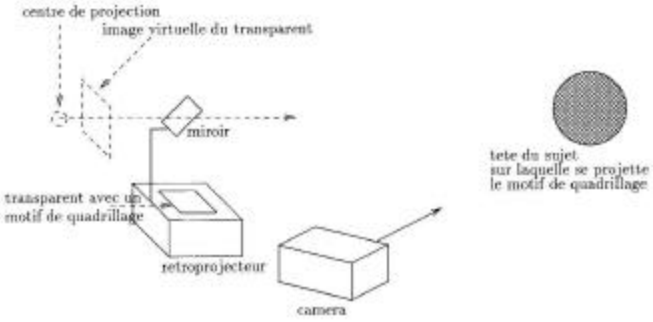

Figure 2 Situation des éléments utilisés pour l'acquisition.

2.2.1 Extraction du quadrillage déformé sur le visage

La détection de la position des nœuds du quadrillage lumineux projeté sur le visage s'effectue par traitement d'image. Nous détectons séparément les lignes horizontales et verticales par filtrage et morphologie mathématique. Nous définissons ensuite les nœuds du

quadrillage comme intersection de ces lignes. Enfin nous ordonnons les nœuds obtenus selon leur position dans la grille, exprimée par deux coordonnées entières.

2.2 CALIBRATION DE LA CAMERA ET DU RETROPROJECTEUR

La calibration de la caméra consiste à trouver ses paramètres intrinsèques qui sont représentables par 5 paramètres indépendants. Nous les calculons en posant que le quadrillage dessiné sur le plan de calibration est carré, ce qui donne deux contraintes (quadrillage orthogonal et de même pas dans les deux directions) liant les paramètres intrinsèques de la caméra et l'image du quadrillage observée (modélisée par une homographie). Avec 3 images, il nous est donc possible de calibrer la caméra.

La calibration du rétroprojecteur consiste à déterminer sa position dans l'espace par rapport à la caméra ainsi que la géométrie du motif de quadrillage (carré) dessiné sur le transparent projeté. Pour cela, nous calculons, pour chaque vue du plan de calibration considérée, la position du plan par rapport à la caméra dans l'espace (en utilisant la connaissance de l'homographie décrivant la projection dans l'image du quadrillage dessiné sur le plan ainsi que les paramètres intrinsèques de la caméra).

Ensuite, nous calculons l'homographie qui définit dans l'espace la projection du motif lumineux sur le plan de calibration (en utilisant la connaissance de la position du plan, des paramètres intrinsèques de la caméra, et de l'homographie qui représente l'image dans la caméra de la projection du motif lumineux sur le plan de calibration).

Cette grille déformée est le résultat de la projection sur un plan d'un quadrillage carré. Ceci nous donne une contrainte sur la position du rétroprojecteur par rapport au plan considéré; nous savons en effet que le centre de projection du rétroprojecteur se trouve obligatoirement sur un cercle associé à l'homographie qui définit la grille déformée sur le plan de calibration. Ainsi, deux configurations différentes du plan de calibration définissent deux cercles où se trouve le centre de projection du rétroprojecteur, ce qui suffit à le définir.

Connaissant la position du rétroprojecteur et de la grille projetée sur le plan de calibration, il suffit de projeter cette grille par rapport au centre de projection du rétroprojecteur pour déterminer le motif du quadrillage carré projeté.

2.3 RECONSTRUCTION 3D

La reconstruction 3D de la surface du visage est effectuée comme par stéréoscopie, par triangulation. Nous avons d'une part détecté la localisation des nœuds du visage dans l'image et d'autre part, pendant la calibration, nous avons estimé leur position sur le transparent du rétroprojecteur

Nous construisons donc plusieurs modèles 3D partiels à partir de photos où le sujet tourne la tête dans différentes directions. Ensuite nous les recalons les uns par rapport aux autres en minimisant une fonction de potentiel qui traduit le fait que deux surfaces sont collées l'une à l'autre. Cette étape a besoin d'une intervention manuelle : nous marquons quelques nœuds communs sur les différents modèles partiaux afin d'obtenir un premier décalage. Pour se faire, nous sélectionnons des nœuds FDP déjà compatibles avec le standard MPEG-4. Enfin, nous construisons un maillage triangulaire global du modèle (voir sections suivantes).

2.4 SYNTHESE ET ANIMATION DES MODELES RESULTANTS

Evaluer la qualité des modèles obtenus à partir du processus précédent est une tâche assez subjective. Le but principal de notre système d'acquisition est d'obtenir un modèle très réaliste et par conséquent le degré de similarité entre la personne et son clone dévient notre mesure de qualité. Même si dans nos premières expériences, nous avons utilisé des modèles frontaux (limités par une seule texture), les résultats sont néanmoins positifs et prometteurs. Cependant il ne suffit pas d'obtenir des modèles visuellement réalistes, mais il faut également pouvoir les utiliser dans toutes nos applications, à la place des clones produits habituellement par des scanners Cyberware^{TM}. Nous étudions déjà comment obtenir une texture adéquate en fusionnant plusieurs textures extraites de diverses photographies.

Figure 4 Nous avons réussi à utiliser des clones obtenus par le système d'acquisition par projection d'une grille dans notre application de suivi de visages.

Pour évaluer nos résultats d'une façon plus objective, nous avons donc utilisé un modèle dans une application de suivi des mouvements du visage dans des séquences vidéo dont la base algorithmique s'appuie très fortement sur le réalisme du clone. Dans les sections suivantes Nous expliquerons la base théorique de notre approche analyse/synthèse pour le suivi des mouvements du visage. Pour nos expériences, nous avions toujours utilisé des clones Cyberware^{TM} qui nous garantissent des résultats optimaux. Notre système de suivi fonctionne avec nos nouveaux modèles malgré leur limitation en rotation (lié au manque de texture) . Nous envisageons à l'avenir une totale substitution des modèles Cyberware^{TM} par ces autres modèles, beaucoup moins chers.

3. ANALYSE, SYNTHESE ET ANIMATION DE CLONES

Etre capable de reproduire exactement sur un clone les mouvements et les expressions d'un visage vus dans une séquence vidéo a plusieurs avantages : les visages deviennent des éléments actifs qu'on peut introduire facilement dans autres environnements, pour la transmission vidéo des éléments, les taux de transmission sont supérieurs à ceux des systèmes traditionnels puisqu'on substitue les trames par des paramètres descriptifs d'animation (les modèles des participants ayant été téléchargés préalablement).

Le standard multimédia MPEG-4 introduit les notions d'objet et de scène. Les objets naturels ou synthétiques sont les différents éléments qui composent une scène. En général, le maillage d'un objet synthétique suffit pour bien le définir et on anime cet objet en envoyant des actions codées. Compte tenu de l'importance qu'une bonne synthèse de visages peut avoir pour certaines applications (acteurs virtuels, visiotéléphones, téléconférence virtuelle, …), on trouve dans le standard une définition plus concrète de l'*objet visage* ("face object"). Le standard définit des règles plus complètes pour l'animation de cet objet. En effet, les paramètres d'animation faciale (FAP) qui permettent d'utiliser

facilement les modèles de visage et leur animation en plusieurs décodeurs ont été spécifiés. Pour profiter d'un tel avantage il est nécessaire que le modèle 3D du visage ait quelques nœuds prédéterminés (FDP) [3]. On peut obtenir des FAPS par analyse des images vidéo à partir de l'audio, du texte, ou bien encore, on peut les définir manuellement et les stocker comme des commandes.

Notre recherche est centrée sur l'analyse de séquences vidéo, pour en extraire des paramètres d'animation faciale, soit déjà compatibles avec le standard MPEG-4, soit plus complexes mais toujours décomposables en plusieurs FAPs. Notre système est formé par deux parties principales. D'abord la localisation globale du visage dans la scène est estimée, ensuite, connaissant la position exacte du visage, on analyse ses expressions.

Le processus d'extraction des mouvements globaux est développé dans la section 3.1, et la base de la recherche sur l'analyse des expressions faciales ainsi que nos premières expériences et résultats dans la section 3.2.

3.1. EXTRACTION DES MOUVEMENTS GLOBAUX

Le système de suivi du visage qui détermine la pose globale fonctionne de la manière suivante:

Initialisation :

i. Le modèle et l'utilisateur sont alignés. L'algorithme a besoin de la position initiale de la personne pour se synchroniser;

ii. ensuite on applique un algorithme de compensation d'illumination 3D sur le clone qui estime les paramètres d'éclairage et minimise les différences photométriques entre le visage synthétique et le visage naturel.

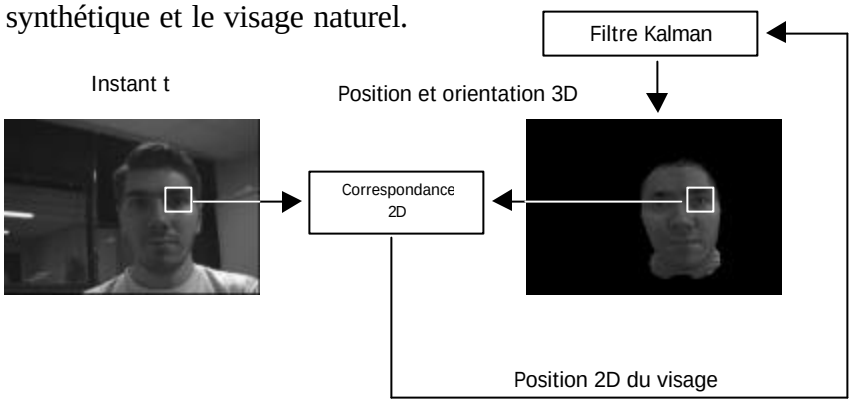

Figure 5 Diagramme de fonctionnement de notre système de suivi de visages.

Boucle principale (Fig. 5):

i. Un filtre de Kalman prédit la position et l'orientation 3D à l'instant *t* ;

ii. le modèle synthétique bouge pour s'adapter à la situation du visage sur la trame à l'instant *t*; cette approximation comprend des distorsions géométriques, d'échelle et de changement d'éclairage ;

iii. on extrait quelques imagettes représentatives de certains traits faciaux (des yeux, des sourcils, des coins de la bouche, etc.) ;

iv. on cherche la correspondance entre les imagettes et les traits de l'utilisateur dans la trame vidéo en utilisant un algorithme différentiel d'appariement de blocs ;

v. on passe au filtre de Kalman les coordonnées 2D des positions trouvées, qui estime la position et l'orientation 3D à cet instant.

L'intérêt de la boucle visuelle de retour dans cette architecture est qu'elle prend implicitement en compte les changements d'échelle, d'éclairage, de fond derrière le locuteur et les déformations géométriques du visage au cours du mouvement (mais pas encore celles liées aux expressions faciales). Cette coopération analyse/synthèse améliorée résulte en un suivi plus stable sans avoir besoin de recourir à des marqueurs faciaux ou du maquillage, tout en autorisant de grandes rotations en dehors du plan image, et ce même sous un faible éclairage. Le système est décrit avec plus de détails dans [4].

L'algorithme nous permet d'obtenir six paramètres d'animation : trois qui représentent les mouvements de translation de la tête et trois qui représentent sa rotation. Ces paramètres peuvent être décrits en suivant la norme MPEG-4 [5]. Le "face object" contient 3 FAPs qui définissent la rotation du visage. En ce qui concerne les paramètres de translation, on ne trouve pas de FAPs équivalents et on doit les traiter comme des transformations générales qui agissent sur tout le "face object".

3.2. ESTIMATION DES EXPRESSIONS FACIALES

La définition d'algorithmes robustes pour un système d'analyse d'expressions faciales est la principale préoccupation de notre recherche actuelle. On développe ces algorithmes en considérant les hypothèses et restrictions suivantes :

- La position globale du visage sur l'écran est connue grâce au processus précédent ;
- suivant la méthode de travail sur la coopération analyse/synthèse, on maximisera l'utilisation des clones, afin d'éviter au locuteur des phases lourdes d'entraînement ;
- les algorithmes devront être robustes aux changements éventuels d'éclairage ;
- on analysera le visage, en appliquant des techniques de traitement d'image sur la vidéo dans le domaine 2D, par opposition à d'autres approches nécessitant des informations 3D pendant l'analyse des expressions ;
- l'analyse devra obtenir directement des paramètres faciaux.

On a élaboré un premier système [6] basé sur l'analyse des images des traits faciaux principaux du visage (yeux, sourcils et bouche). A l'aide du clone, une base de données composée par des imagettes des traits faciaux est créé, avec diverses expressions et sous différents éclairages. En utilisant une analyse en composantes principales (PCA), on réduit la taille de cette base de données. Pour analyser une expression de l'utilisateur, on extrait d'abord, des imagettes de la trame vidéo et on les décompose suivant la base de données, pour obtenir des coefficients de corrélation (λ). Ces coefficients nous indiquent le degré de similarité avec les imagettes mais n'indiquent pas le rapport avec les actions faciales (μ) qui les ont générées. Il faudra un entraînement préalable à l'analyse pour être capable d'estimer μ à partir de λ. Cet entraînement nous assurera qu'une fois décomposée, on saura l'équivalence de l'imagette en termes de FAPs. Pour entraîner le système on utilise le clone plutôt que le locuteur réel.

Les premiers tests sur le système nous ont montré que la construction d'une base de données avec des imagettes extraites du clone sous différents éclairages a quelques

limitations. D'abord, on doit générer une grande base de données parce que chaque expression est analysée sous plusieurs éclairages possibles. Cela nous oblige a construire une base de données d'une taille (nombre de poses x variations d'éclairage x nombre d'expressions). En plus, l'échantillonnage de l'espace de variation d'illumination ne suffit pas à obtenir une indépendance totale par rapport à l'éclairage. Dans le but d'acquérir plus de robustesse par rapport à la lumière, nous avons essayé quelques pré-traitements (normalisation, gradient et flot optique) sur les imagettes à analyser, ainsi que sur celles de la base de données. Nous avons aussi testé divers estimateurs (linéaire et basé sur RBF) afin de trouver la meilleure définition du rapport λ-μ monolocuteur. Ce rapport est dépendant de l'utilisateur et est défini pendant l'entraînement.

Les résultats finaux nous montrent qu'une analyse d'image des traits faciaux principaux est un moyen prometteur d'extraire les paramètres faciaux d'une séquence vidéo. Cependant il est nécessaire d'améliorer l'étape d'extraction de l'information visuelle des imagettes afin de le rendre plus indépendant de l'éclairage et d'optimiser l'entraînement. En effet, à cet instant, il est limité par les actions du clone et ne peut pas prévoir des expressions qui ne sont pas incluses pendant la phase l'entraînement.

Nous explorons actuellement de nouvelles voies, toujours en suivant les contraintes applicatives indiquées en début de sous section, afin de développer le système d'analyse d'expressions sur vidéo. La base structurelle restera identique. Néanmoins on cherchera des algorithmes plus robustes aux variations d'illumination et à l'association imagette-FAP. On envisage en particulier, la possibilité d'analyser les imagettes d'un point de vue morphologique en appliquant des techniques de segmentation.

4. CONCLUSIONS

Nous avons présenté dans cet article une technique de modélisation de visages qui permet d'obtenir des modèles réalistes (clones) sans équipement(s) spécialisé(s). Nous avons montré aussi comment, à partir d'images réelles, nous obtenons des paramètres d'animation facial (FAP) qui peuvent animer des clones compatibles avec la norme MPEG-4. Nos développements exploitent le fait d'utiliser des clones réalistes, en particulier, pour les étapes d'apprentissage et pour une coopération analyse/synthèse renforcée.

5. REFERENCES

[1] Marc Proesmas, Luc Van Gool, *One-shot 3D-shape and Texture Acquisition of Facial Data*, Proceedings First International Conference on Audio- and Video-based Biometric Person Authetication, AVBPA'97, lecture notes in computer science, vol. 1206, pp. 411-418, 1997.

[2] E. Garcia. *Reconstruction 3D de visages.* Rapport de thèse professionnelle, Département des Communication MM, Institut Eurécom Septembre 2000

[3] Information technology - coding of audio-visual objects: Visual - ISO/IEC 14496-2. Maoui, December 1999.

[4] S. Valente. *Analyse, Synthèse et Animatin de Clones dans un Contexte de Téléréunion Virtuelle*. Thèse de doctorat, Ecole Polytechnique Fédérale de Lausanne, Institut Eurécom, France, 1999.

[5] Information technology - coding of audio-visual objects: Systems - ISO/IEC 14496-1. Atlantic City, November 1998.

[6] S. Valente, A.C. Andrés, J.-L. Dugelay. *Analysis and Reproduction of Facial Expressions for Realistic Communicating Clones*. The Journal of VLSI and Signal Processing, à paraître en automne 2000.