Telecommunicating via CYBERWARE Interfaces and Virtual Images

S. Valente and J.–L. Dugelay

Institut EURECOM
MultiMedia Communications Department
2229, route des Crêtes, B.P. 193,
06904 Sophia–Antipolis Cedex, France
URL. http://www.eurecom.fr/∼image

E–mail: {valente, dugelay}@eurecom.fr
Tel: +33–93.00.26.{77,41}
Fax: +33–93.00.26.27

Abstract

This paper presents early results in the context of a televirtuality project (named "TRAIVI"). The goal of this project is to create and run virtual meetings via low bit–rate links and virtual reality paradigms. We propose to enable several persons located at different physical sites to meet each other in a common virtual meeting room. In order to preserve a high level of realism, we provide them with $3D$ human–like synthetic interfaces based on animated "CYBERWARE" models.

Keywords: telepresence, televirtuality, video communications, object–oriented compression, human face clonage.

Résumé

Ce papier présente les premiers résultats d'un projet de télévirtualité (appelé "TRAIVI"). Le but de ce projet est de réaliser des espaces virtuels de réunion sur des lignes à bas–débit en utilisant les outils de la réalité virtuelle. Nous proposons de permettre à plusieurs personnes situées dans différents lieux physiques de se rencontrer dans une même salle de réunion virtuelle. Pour préserver un haut degré de réalisme, nous leur offrons des interfaces synthétiques humanoïdes $3D$ basées sur des modèles "CYBERWARE" animés.

Mots clés: téléprésence, télévirtualité, communications video, compression orientée–objet, clonage de visage.

Introduction

In terms of video compression, low bit–rate teleconferencing systems may have two different approaches: they may be signal–oriented (that–is–to–say that they consider video images as $2D$ or $3D$ signals and try to efficiently code the signal redundancy), or they may be object–oriented (they consider an image as a geometrical projection of physical $3D$ objects and try to efficiently code the scene objects and how they are displayed in the image). Most of the currently available commercial teleconferencing systems belong to the first category, whereas virtual imaging techniques belong to the second one.

This introduction will present our project position and philosophy.

Current Signal–Oriented Systems Limitations

Classical low bit–rate teleconferencing systems reach enormous compression rates that all engineers do admire. Nonetheless, the average user does not care about technical figures, and all he pays attention to is the quality of the teleconferencing session. From the point of view of human communications, such systems still suffer from four main limitations:

- these systems are not satisfactory for meetings involving more than two people at the same time, or more generally, more than two sites, because people have to switch between speakers' icons and/or alternate site views in order to see or talk to other persons
- it is not possible to achieve eye contacts with a participant, or to blink only at him by looking either at the display device or the camera. Users have no way to feel directly involved by other users' gaze
- participants have no meeting room common visual references (they can see the room where other participants are speaking from). As a consequence, people have a persistent feeling of distance
- the audio aspects of the communication are often ignored, since most of the efforts in developping a teleconferencing system focus on image compression. Because the point is to allow the speakers to talk in front of a screen, and not into a mike, the audio input has room reverberation effects, and users generally feel that the participants' rooms do not sound like theirs. Besides, when several speakers are talking at the same time, the use of spatial sound techniques can tremendously improve the conversations intelligibility

It seems that human communication would be made much easier if users could be provided with a visual and auditory common environment.

Virtual Teleconferencing *Versus* Signal–Oriented Teleconferencing

The concept of virtual teleconferencing offers elegant solutions to the limitations of classical teleconferencing systems. The key idea is to provide speakers with a common meeting space as if they were all meeting in the same physical room, to synthesize the individual points of view they could naturally experience, and to give them the opportunity of having eye contact with each other via synthetic $3D$ models of the participants (clones). Virtual reality tools are used here to allow people to talk comfortably, forget the distance, and help them in their collaborative work by sharing data [1]. In order to achieve a good level of realism, several audio and video techniques have to be investigated and implemented. Among them are audio spatialization[1] [2] and multiplexing, echo cancellation, video spatialization, speakers face clonage, audio–video synchronization...

In recent years, research has been conducted on topics related to teleconferencing such as human image analysis and reproduction [3], model–assisted video coding and human face clonage [4]. Meanwhile, technology has evolved, and virtual reality tools like $3D$ rendition displays have become more mature and widespread. Therefore, we are able to propose an approach to virtual teleconferencing with the TRAIVI[2] project.

Virtual Realism *versus* Virtual Reality

TRAIVI targets a high level of realism. Our goal is neither to build an imaginary world that has no equivalent in reality, nor to exactly synthesize the real world and the true speakers expressions, but to render the real world in a way that is visually coherent and where the proper (not the real) speakers facial expressions can be seen by every participant.

This project intends to trade virtual realism with virtual reality. Like some authors, such as Kanade [5], we think that *virtualized reality* is quite superior to *virtual reality* since it takes into account the real world fine details and not a simplistic CAD model. To that extent, we propose to use $3D$ textured wire frame models scanned from flesh and blood persons ("CYBERWARE" models) to perform the face clonage, and to mix real and synthetic portions of images during the clone animation process if the level of realism is not high enough[3]. The clones will then be inserted in the (virtual) meeting room. We chose to use *video spatialization* techniques on real room uncalibrated pictures to create the virtualized meeting room, as opposed to building $3D$ room models from scratch (again for realism reasons) [6, 7, 8].

[1] to get $3D$ reverberated sounds closer to natural hearing conditions

[2] TRAIVI stands for "TRAItement des images VIrtuelles" (Virtual Image Processing)

[3] Typically when a speaker's mouth is wide–open, his teeth and tongue might be visible, and should be reproduced with real images on the clone because they are not part of the $3D$ wire frame texture

System Overview

The system that will be implemented looks like Fig. 1. Please keep in mind that our system is geared to virtual realism rather than virtual reality: we must be able to provide the users with a realistic and coherent view of the meeting situation, and this is the reason why we have to analyse the scene in terms of $3D$ parameters which are meaningful under any participant's point of view.

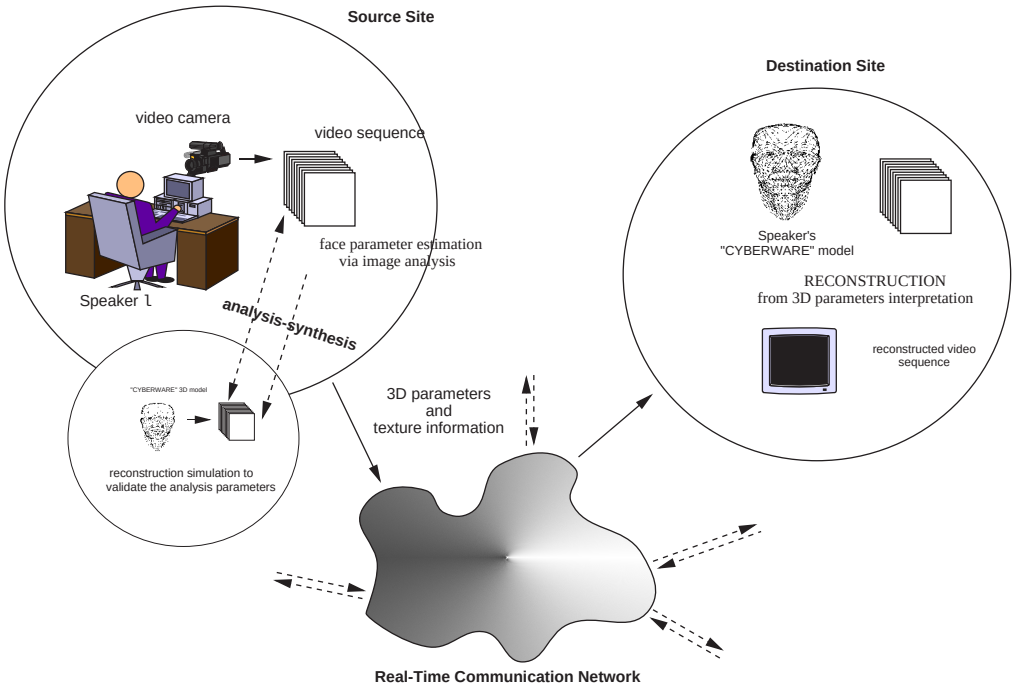


Figure 1: Virtual teleconferencing system configuration

So, in our system, $3D$ parameters are extracted at the source site, sent to destination site(s), and finally interpreted to render the $3D$ scene under another viewpoint (see Fig. 1).

As you can notice, the teleconferencing system is partially illustrated here; another mono–directional connection must be opened from the destination site to the source site to make sure all speakers can see and hear each other. And of course, more than two sites can participate in the meeting.

Furthermore, a central site in the network will have the responsibility to transmit to all teleconferencing sites the common meeting room data, the participants virtual locations within it, and to provide them with the right "CYBERWARE" models (if they are not found in local databases) before the meeting begins.

The protocol and network components that would exchange data between all sites have still to be defined.

We are now going to describe how we analyse the participants' motions, and which modalities were used to render and animate their clones.

1 "CYBERWARE" Models

Several research teams are now working on facial animation and model synthesis. Most of them work on $3D$ face models they implemented on their own, and are now resorting to texture–mapping to augment the level of realism [4, 9, 1, 10]. Their goals differ depending on the research aspects they are studying.

Our approach is original because unlike them, we did not start working on a generic $3D$ model more–or–less adapted to each speaker, but instead, we are using a "CYBERWARE" model.

Such models are produced by three–dimensional cylindrical scanners, and consist in 2 files: the first one contains a set of $3D$ coordinates representing the scanned head geometry (i.e. a wire frame), and the second one holds texture information to be mapped to the geometrical coordinates (i.e. an image, see Fig. 2).

Figure 2: Example of a "CYBERWARE" texture

The pros related to the use of "CYBERWARE" models are:

- such models are highly realistic
- getting one is straightforward as long as you have a "CYBERWARE" scanner
- they output precise information about someone head's geometry
- the geometrical data is precisely tied to the texture information. This can lead to very robust template matching procedures: if we roughly know the speaker's head global position, we can render the clone to dynamically create new $2D$ templates that will take into account the projected nature of the real view. This can make up the fact that some features might disappear (like one eye when the speaker turns his head) and should no longer be searched, or might look different given the speaker's pose (for example the speaker's nose).

However, in the real world, the pros are often balanced by a few cons:

- a "CYBERWARE" model is only valid for a single person, taken at a given time
- the mesh model is unoptimized in terms of complexity. The face $3D$ surface sampling was made along regular cylindrical coordinates, and it turns out that there are as many points to represent smooth surfaces[4] as to represent the sharpest ones[5]
- it has neither anatomical data (like bones and muscles under the skin to model human face deformation possibilities) nor a physical model (to be able to deform it realistically along the time axis) [3].

In the next sections, we will explain how we handle a "CYBERWARE" model both globally (its position in the view) and locally (the way facial features are altered). Since our project intends to improve the human communication aspects of teleconferencing, we wish to make clear that we do not want to set the users ill–at–ease, and using marks taped on user's faces to help tracking facial features is thus out of question.

2 Global Animation

We have currently implemented the following global teletracking scheme in the limited case of 2 sites linked by a mono–directional communication protocol:

- a person (later denoted λ) is located in site 1 (called the *source site*) in front of a video camera;
- the video sequence is digitalized, and images are analyzed in real–time to extract the speaker's head position parameters; a pseudo $2D \rightarrow 3D$ transform is operated on the $2D$ parameters;
- these parameters are then sent to site 2 (called the *destination site*);
- site 2 uses the "CYBERWARE" model of λ, interprets the incoming head position parameters, and consequently renders λ's face.

We will now describe the modalities of the implementation.

2.1 Analysis

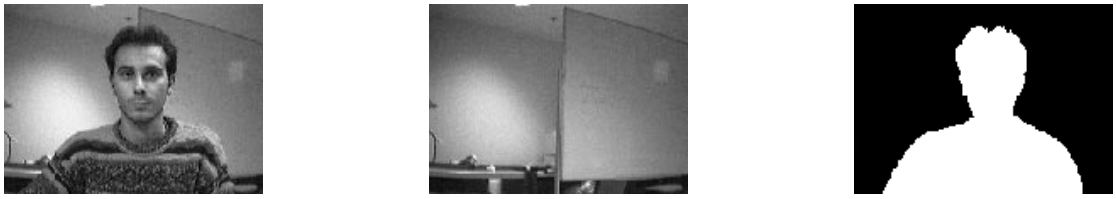
The current analysis method runs in real–time on a Sparc 20 with image processing dedicated hardware.

There is a video camera on the top of the worksation. When there is no user in its field, a reference view of the background scene is digitalized. When the user sits down in front of the workstation, his presence is detected, and the session starts up. Then, during the session, the user's figure is located by substracting its live image with the background reference view and by applying a threshold (see Fig. 3).

Once his figure is located, his face is tracked by a rectangular window W, and within the window, we detect the eyes axis E and the vertical axis V between them (see Fig. 4). From these parameters, we can derive 5 out of the 6 degrees of freedom of the speaker's face global motions:

- **left/right and up/down translations**: given by the window W center coordinates
- **forward/backward translation**: derived from the size of W

[4] like the cheeks
[5] like the nose



speaker's current view | background reference view | speaker's outline

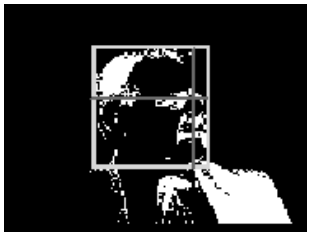
Figure 3: Speaker's outline determination

- **left/right rotation**: given by the position of V within W
- **up/down rotation**: given by the position of E within W

This implementation might seem somewhat simple, and be prone to stability problems, but it corresponds to early results that have no other reason than to show how the system would look like and what the implementation difficulties are. We are currently working on porting the analysis stage on an SGI workstation to ease our future developments by using the SGI's image processing library. We will then have the opportunity to investigate other solutions to globally track a person's face by using more robust parameters.

The speaker is in front of his workstation. His head and eyes positions are located.

The speaker's "CYBERWARE" model is displayed at a distant site.



The speaker turns his head to the left.

The speaker's "CYBERWARE" model also turns to the left.

Figure 4: Speaker's face axis detection

2.2 Transmission

The estimated parameters are sent to a graphics workstation via a TCP/IP socket. In addition, the audio input of the workstation is embedded into the communication stream. Considering the low bandwidth required by the parameters, it would be possible to use a mere telephone line as a digital communication link.

2.3 Synthesis

As it was explained above, the received parameters are used to render the clone. The rendition is performed on a High Impact SGI workstation in real–time. We are now going to discuss the different synthesis choices we made for the restitution of the $3D$ models.

2.3.1 $3D$ Rendition

Since the $3D$ points produced by the "CYBERWARE" scanner are really numerous, we downsample the geometric data by a factor 6 in order to obtain roughly $250,000$ nodes, leading to a complexity compatible

with real–time constraints. This is still an unoptimized wire frame, and further work has to be done to reduce much more the amount of points.

2.3.2 Stereoscopic Rendition

One of the advantage of having an explicit $3D$ model is the possibility to render it in $3D$. We are currently using CrystalEyes glasses with our SGI workstation to render the $3D$ clone. Of course, the glasses hide the user's face, and in this case, the analysis stage cannot achieve fine local expressions tracking. This is not a real problem, because these are only prospective studies to see which synthesis methods are the most appropriate from the user's point of view. Stand–alone $3D$ displays are being developped in research laboratories, and will shortly replace stereoscopic glasses [11].

2.3.3 Clone Immersion

Once we are able to synthesize the participants' clones, we have to make them meet in the same virtual room. We implemented the room metaphor by displaying the room reference view (see section 2.1) behind the textured $3D$ model as a distant plane. In the future, a virtual meeting room will be reconstructed from real images by *Video Spatialization* algorithms based on trilinear constraints [6, 7, 8]. The goal of this technique is to reconstruct existing and virtual points of view from a finite set of real uncalibrated pictures.

3 Local Animation

Now that we have seen how to track and position our $3D$ model globally, we will focus on local animations that will make it be more "human" and show some emotions. Two different strategies can be implemented to realize local animations: the first one consists in applying wire frame deformations, and the second one in altering the texture file.

3.1 Wire Frame Animation

To implement a control scheme over the wire frame, further processings have to be done on the plain "CYBERWARE" model:

- its complexity has to be decreased (in terms of nodes number)
- anatomical and physical knowledge have to be added to the wire frame

Both issues can be addressed by using Delingette's *Simplex Mesh* approach [12]. His mesh model cuts down the number of nodes involved to represent a human face by adapting the mesh complexity to the surface geometry. It also supports realistic active mesh deformations [13] and can generate facial expressions based on the Facial Action Coding System (FACS) [9]. Other studies have demonstrated that simplex meshes could model accurate expression wrinkles (like foreheads wrinkles) [14].

3.2 Texture File Modification

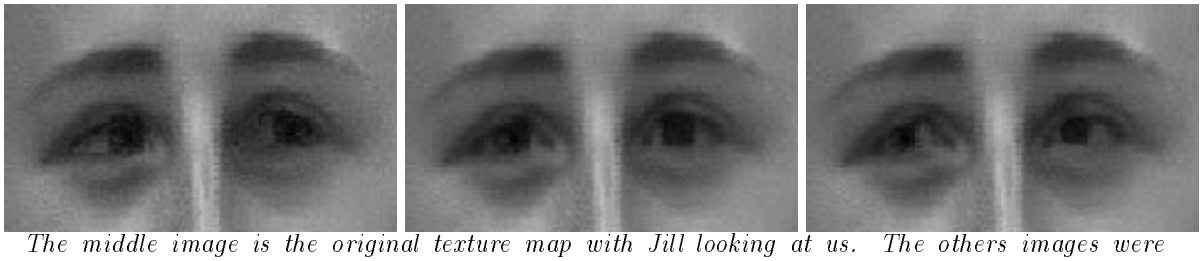
The other alternative to animate the $3D$ model is to alter the texture information attached to the wire frame. In particular, this scheme is quite promissing for facial features like the eyes or the eyebrows.

In our synthesis software, we have implemented routines that dynamically switch between several eye textures representing three different gaze directions, with almost no computational overload. These textures where pre–calculated using commercial Image Editing software. The level of realism of such plug–in methods can be seen on Fig. 5.

Other facial features, like the mouth, will require both texture and mesh modifications to simulate the motion of the speaker's jaws and show the speaker's teeth. Since the teeth do not belong to the original model produced by the scanner[6], we will have to resort to portions of real images (not necessarily live images, but offline digitalized images extracted from typical teleconferencing sessions). This will imply processing real images taken by a video camera to make them match the "CYBERWARE" photometric data scale and its lightning conditions.

Our current implementation validates the fact that modifying the texture information is a powerful and low–cost way to animate our model. We will continue our developments to build a dictionary of local expressions based on real images that will be referenced by the analysis software, and will avoid sending portions of live images to update the $3D$ model texture information.

[6] actually, it depends on the expression the user had at scanning time — we here assume that the most common expression of someone standing in a scanner is the neutral one



The middle image is the original texture map with Jill looking at us. The others images were obtained by plugging–in eyes textures that presented the proper gaze direction to make Jill look at her left or right hand–side.

Figure 5: Gaze control via texture modification

We must emphasize that live video images should be used as less as possible (but not excluded) for two reasons: the first one is that we need parameters that are consistent in the $3D$ space. We cannot directly use live images because they are only valid in the camera point of view. Further works have to be conducted to derive the $2D \rightarrow 3D$ transformations that would make $2D$ live images useable in $3D$. The second reason is that we want to keep the bandwidth requirements of our system low, and we cannot afford sending huge portions of live images to the destination sites. Referencing indexes to alter the texture information is the *simplest* solution for now.

3.3 Chaotic Behavior

We claimed earlier that *virtualized reality* was better than *virtual reality* because of small details that are difficult to model via CAD methods. One of the recurrent reproach made by users is that artificial models are "too perfect"[7]. This will also stand for our animated model if its eyes do not blink. Every one does it unconsciously, and even if we do not explicitely notice other people's blinking, this is a part of the real life that our model lacks.

An unconscious eye blink is very quick, and video shots showed that it generally lasts 3 frames (at 25 frames–per–second). If the system analysis side cannot sustain the full video rate, it will not be possible to detect every single eye blink made by the speaker. As a solution, we are to implement a kind of random process to automatically produce eye blinks via texture modification so that the clone looks more "alive".

3.4 Cooperation between Analysis and Synthesis

Analysing monocular $2D$ images in terms of $3D$ is a quite difficult task, and avoiding taping marks on users' faces cannot make it easier.

However, we can use a feedback loop to make the analysis stage more robust (see Fig. 6). First of all, the global head position is estimated. The idea is to render the synthetic view to approximate the camera point of view and to validate the first estimates. The real and synthetic views are resized, normalized, aligned and compared to determine if their similarity level is satisfying or not. If some dissimilarities are observed, the $2D$ real image areas that were not properly synthesized are extracted, and the system takes them into account to adjust its estimates. This operation can be repeated until the $2D$ view is properly synthesized.

Several similarity measures can be useful, depending on the local features they compare:

- **Image Matching**: synthesized and real image areas are simply correlated. This method can find for instance the best pre–calculated texture for the eyes, or for the user's forehead wrinkles.

- **Eigenfeatures**: the Karhunen-Loeve decomposition has been in use in the Face Recognition Community to characterize human faces (eigenfaces) and their facial features (eigenfeatures) and is a powerful alternative over simple template matching, especially in the context of evaluating the mean square error of a match [15, 16].

- **Contour Snakes**: active contours can successfully evaluate the lips state when the speaker is looking at the video camera [17] by estimating their area, perimeter and/or deformations, but interpreting their shape in $3D$ terms when the user looks at the right hand–side of the camera might not always lead to the right lips shape. Running snakes on both the synthesized and real mouth images could help adjusting the lips $3D$ state.

[7] actually, do straight lines exist in the real world, or are they just a mathematical concept ?

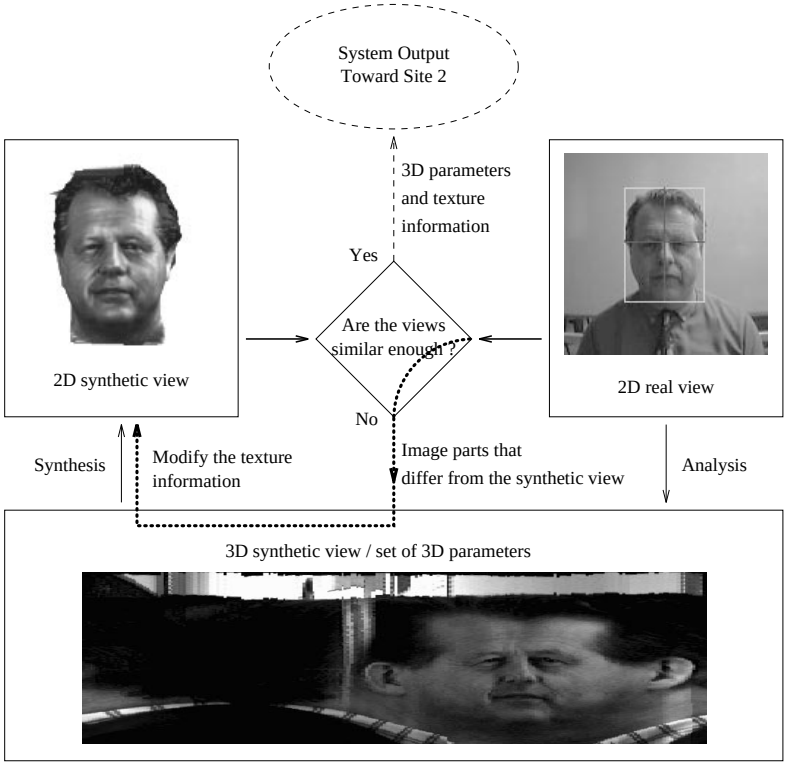


Figure 6: Cooperation between Analysis and Synthesis

3.5 Other Techniques

To animate the clone's mouth, one can foresee that some problems will occure when the speaker's mouth is not visible. In this situation, it will also be quite acceptable to apply a realistic (but not real) deformation on the clone lips based on the audio signal analysis [18].

Conluding Remarks

We presented in this paper early results obtained in a "human–friendly" teleconferencing project using Virtual Realism paradigms.

Video Clonage aims at providing people with $3D$ interfaces to meet each other in a common virtual space under natural dialog conditions. The possibility to achieve eye contacts and to create individual points of view for each participant will be offered. "CYBERWARE" models were used to ensure a high level of realism.

Validation results were discussed, and studies are in progress to animate the $3D$ models' facial features in real–time.

References

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

[2] J.-M. Jot. Synthesizing three–dimensional sound scenes in audio or multimedia production and interactive human–computer interfaces. In *L'interface des mondes réels et virtuels*, Montpellier, France, Mai 1996. To be published.

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

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

[5] T. Kanade, P. J. Narayanan, and P. W. Rander. Virtual*ized* reality: Concepts and early results. In *IEEE Workshop on Representation of Visual Scenes*, Cambridge, Massachusetts, June 1995. In conjonction with ICCV'95.

[6] J.-L. Dugelay and K. Fintzel. Image reconstruction and interpolation in trinocular vision. In *IMAGE'COM*, pages 277–282, Bordeaux, France, May 1996.

[7] P. Bobet, J. Blanc, and R. Mohr. Aspects cachés de la trilinéarité. In *RFIA*, pages 137–146, Rennes, France, 1996.

[8] A. Shashua. Trilinearity in visual recognition by alignment. In *ECCV*, Stockholm, Sweden, May 1994.

[9] B. Girod. Image sequence coding using 3D scene models. In *SPIE Symposium on Visual Communications and Image Processing*, Chicago, September 1994.

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

[11] T. Fujii, R. Ishimi, and J. Hamasaki. Experiments on data compression for an autostereoscopic lenticular 3D TV. In *First International Festival of 3D Images*, Paris, France, September 1991.

[12] H. Delingette. *Modélisation, Déformation et Reconnaissance d'Objets Tridimensionnelles à l'aide de Maillages Simplexes*. PhD thesis, Ecole Centrale de Paris, Sophia-Antipolis, France, 1994. Available via URL http://www.inria.fr/epidaure/personnel/delingette/biblio.html.

[13] H. Delingette, G. Subsol, S. Cotin, and J. Pignon. A craniofacial surgery simulation testbed. Research Report RR–2199, INRIA, 1994.

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

[15] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

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

[17] S. Horbelt and J.-L. Dugelay. Active contours for lipreading: Combining sankes with templates. In *Quinzième coloque GRETSI*, Juan–Les–Pins, France, September 1995.

[18] C. Benoit. On the production and perception of audio-visual speech by man and machines. In *Int. Symp. on Multimedia Communications and Video Coding*, New York City, October 1995. Polytechnic University.