appears in: IEEE Picture Coding Symposium, April 21-23 1999, Portland Oregon USA

Synthetic Natural Hybrid Video Processings for Virtual Teleconferencing Systems

Jean-Luc Dugelay, Katia Fintzel and Stéphane Valente

Institut Eurécom, Multimedia Communications Department
B.P. 193, 06904 Sophia Antipolis Cedex, France
{dugelay, fintzel, valente}@eurecom.fr

Abstract

In this paper, we propose powerful virtual image processing tools, which can be useful for designing new virtual teleconferencing systems with limited bandwidth requirements. We present an integrated approach that mixes real images with computer graphics in order to obtain an efficient and highly realistic model-based coding scheme, while offering new possibilities from the point of view of the ergonomics. In particular, we focus on original or extended existing algorithms to address the speakers' representation problem via face cloning, *and the virtual environment creation issue via* video spatialization.

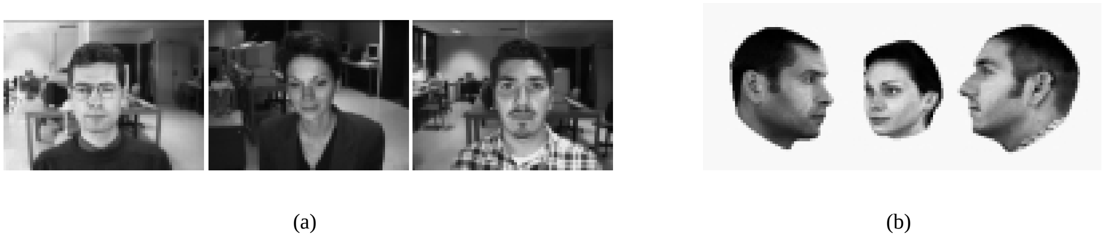
1. Introduction

Low bit-rate video communication systems can be divided into two categories:

signal-processing-oriented, encoding and decoding the real image with as much visual accuracy as possible,
object-oriented, rebuilding more or less realistically the image by encoding objects as opposed to waveforms.

Considering the first systems, human communications become critical, compared to real meeting conditions, with three or more sites. In this case, you have to alternatively display the view of the currently speaking site or, if more bandwidth is available, you have mini-images representing each site (figure 1(a)). In all cases, even if the participants are self-disciplined [9], you cannot feel as being *with* the others. These systems suffer from important limitations, due to the difficulty to achieve convincing eye contact between speakers, and the persistent feeling of distance because of the lack of common meeting room references and poor audio immersion.

The concept of virtual model-based teleconferencing offers elegant solutions to these limitations. The key idea is to provide the participants with a common meeting space as if they were all meeting in the same physical room (figure 1(b)) and to synthesize the individual points of view they would naturally see. In our opinion, model-based coding with *virtual imaging* techniques is an interesting approach from the point of view of ergonomics for teleconferencing systems, that we investigate in the TRAIVI[1] project.



(a)

(b)

Figure 1. (a) What you are likely to see with a *classic* teleconferencing system (b) What is expected from a *virtual* teleconferencing system. In this configuration, you are assumed to be the fourth participant.

[1] TRAIVI stands for "TRAItement des images VIrtuelles" (Processing of Virtual Images)

The TRAIVI project proposes an integrated approach that addresses both the speakers' representation and the meeting room background by mixing synthetic textured face models, built from range data, with spatialized natural images, resulting in a virtualized vision of the real world rather than an arbitrary $3D$ CAD world disconnected from reality. The stake here is not 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 *comfortable* for its users.

This publication highlights the virtual imaging techniques found in the TRAIVI project: (i) the *face cloning* aspect from the face modeling to the head position and orientation estimation algorithms using an enhanced analysis/synthesis feedback loop, (ii) *video spatialization* applied to the construction of virtual environments, based on an original mesh-oriented approach to recreate virtual points of view from a limited set of uncalibrated real room pictures.

2. Face Cloning

Face cloning consists in animating a synthetic face model by analyzing the video sequence of a real speaker [10]. In the context of a virtual teleconference, telecommunication aspects impose specific and challenging constraints on facial cloning, like the face analysis and synthesis frame-rates, the image processing delays and the very low bandwidth networks available to transmit the animation parameters. Face modeling and global motion tracking techniques for such a system should operate without colored marks taped on the speaker's face, deal with unknown lighting conditions and background, allow the users to move freely in front of the camera and yield visually realistic results.

To provide a high level of realism, we propose to represent each speaker within the virtual area by a $3D$ texture-mapped model obtained from cylindrical geometry Cyberware finders [2] and optimized by the theory of deformable simplex meshes [3]. Due to the realism of the speakers' models, a cooperation is made possible between the analysis of the speaker's face and the synthesis of the face model. A global motion tracking software using an original analysis/synthesis feedback loop was developped, as follows:

(i) a Kalman filter predicts the head $3D$ position and orientation for time t;

(ii) the synthetic face model is used to generate an approximation of the way the real face will appear in the video frame at time t; this approximation includes both geometric distortions and shaded lighting due to the speaker's pose, as well as some clues about the location of the background with respect to the face facial features [11];

(iii) patterns are extracted from the synthesized image, representing contrasted facial features (like the eyes, eyebrows, mouth corners, nostrils);

(iv) a differential pattern-correlation algorithm matches these patterns with the user's facial features in the real video frame;

(v) the $2D$ coordinates of the found positions are given to the Kalman filter, which estimates the current head $3D$ position and orientation.

This enhanced analysis/synthesis cooperation makes the face tracking more robust without artificial marks highlighting the facial features, and supports very large rotations out of the image plane, as it can be seen on figure 2, while meeting the requirements detailed in the introduction.

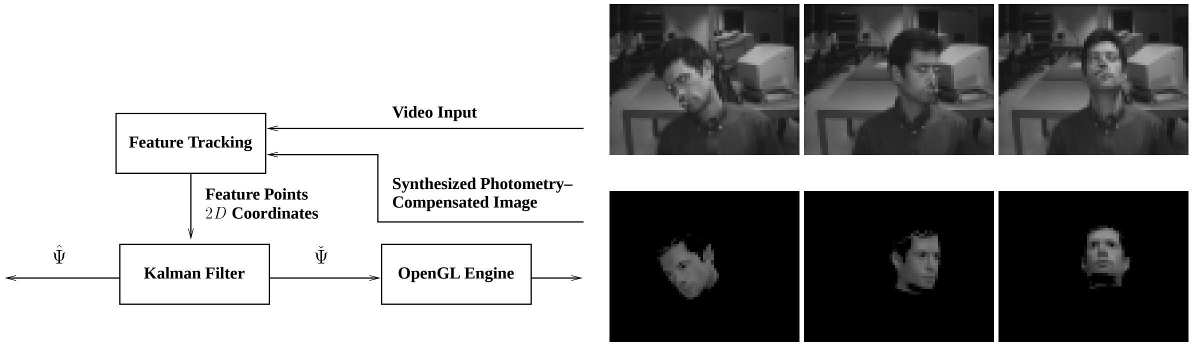


Figure 2. Feedback loop strategy based on a Kalman Filter and Synthetic Images - $\check{\Psi}$ and $\hat{\Psi}$ are the speaker's $3D$ position and orientation predicted and filtered estimates.

The result of the face tracking algorithm is presented in an Mpeg sequence available on the WWW [6] and the current developments are directed towards the tracking of facial expressions for $3D$ models [12].

3. Video Spatialization

The second aspect of video processings we study in order to create a virtual conference area is *video spatialization* for background control, only based on uncalibrated $2D$ views without any $3D$ CAD model. Such a process aims at offering the possibility to visualize the meeting room from anywhere and toward any direction, instead of imposing a unique point of view for each site, like current teleconferencing systems do. In terms of low bit-rate bindings, it is impossible to catch and transmit all the necessary views of the scene. So our work uses the trilinearity theory combined with texture mapping, to compress data for transmission on the one hand and to increase information for the creation of available points of view on the other hand.

We introduce an original "mesh-oriented" algorithm, using a triplet of views, for an existing view regeneration from two other neighboring images, as follows:

(i) an analysis step, using corresponding points (equivalent to the nodes of the meshes) in the three original uncalibrated views, estimates the eighteen parameters of a trilinear form, derived from the trinocular vision theory (see [7] for more details about trilinear parameters definition);

(ii) a synthesis step, using the meshes nodes of the external images and the estimated parameters, reconstructs the mesh of the central image, on which the reference texture of the initial triplet of views is mapped (figure 3 (a)).

The "mesh-oriented" caracterization of this method is really innovative as opposed to "pixel-oriented" methods, widely found in the literature [8, 1]. In the context of virtual teleconferencing systems, we also derived possible extensions of this method in order to create virtual points of view of the $3D$ scene only known by a set of initial $2D$ images. The modification of intrinsic (focal length) or extrinsic ($3D$ position and orientation) parameters of the camera relative to the reconstructed view can be applied directly to the pre-estimated trilinear parameters without an explicit calibration stage: only a synthesis step is required to change the current point of view of a participant (figure 3 (b)). The process speed only depends on the number of mesh nodes, new views can therefore be synthesized in real time from well adapted and pertinent meshes. Readers can find in [4] a complete definition of the parameters modifications due to a change of intrinsic parameters or a virtual motion of the camera, which give the resulting reconstructed view of the virtualized scene (figure 4).

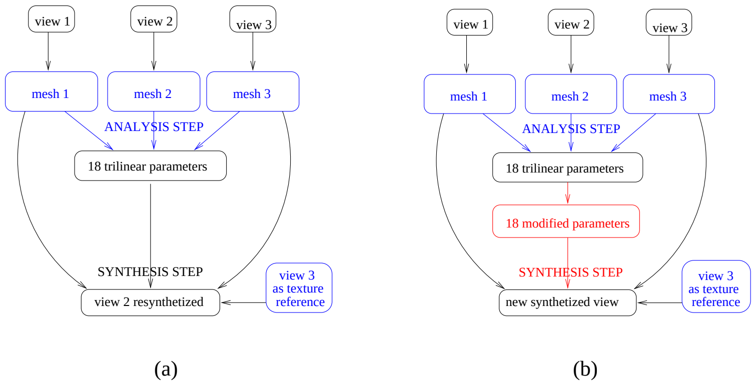


Figure 3. The methods of real and virtual views synthesis: (a) summarizes the real views regeneration method; (b) outlines the procedure to synthesize a new point of view.

In pratice, we introduce an *inter-triplets* approach controlled via an *image mosaïcking* module and an *underlayers* technique [5] in order to increase each virtualized view realism and cover the meeting room. The complete implementation of a room spatialization system, under investigation, is as follows:

(i) (optional) extension of the initial manually-initialized meshes by image mosaïcking (justified in special cases);

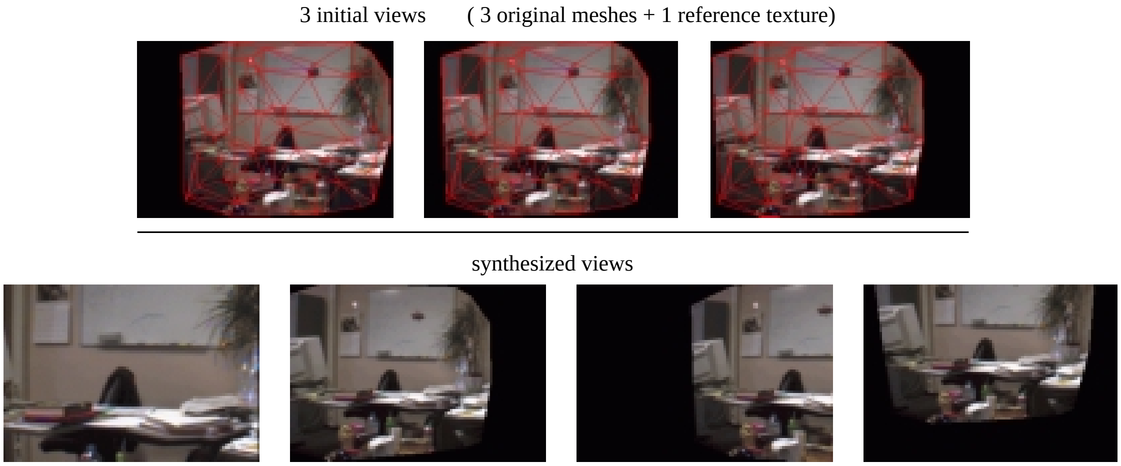
(ii) trilinear parameters estimation from meshes nodes;

(iii) (optional) trilinear parameters alteration;

(iv) synthesis step from the two meshes of the external views and the altered trilinear parameters (or only reconstruction with unaltered parameters);

(v) mosaïcking extension of the synthesized view, if necessary;

(vi) addition of under-layers from the reference texture of the previous and the next triplets of views, if the algorithm is used with more than three views.



focal length change (*1.8) vertical rotation of -5 degrees vertical rotation of 20 degrees horizontal rotation of 5 degrees

Figure 4. Synthesized points of view after a camera focal change or a camera rotation

4. Concluding Remarks

In this paper, we presented an integrated approach using original virtual imaging tools and solutions useful to design new teleconferencing systems. Thanks to the *face cloning* and *video spatialization* modules, such systems could mix synthetic and natural images over very low bit–rate links, and offer some functionalities available in real meetings. We developped an early prototype of a virtual teleconferencing system, combining our preliminary results on *face cloning* and *video spatialization*. In this demonstration, all sites send to all other sites the pose of their local user. At every instant, we have then to define in real time the visual feedback of each participant. Considering one of them, we have to:

- render the $3D$ virtual models of all the other participants, given their individual motions modulated by his own estimated pose;
- control the $2D$ point of view of the background using video spatialization with respect to his personal position and orientation;
- mix the $3D$ updated models and the synthesized background.

Our futur perspectives include now the complete and rigorous insertion of the $3D$ models in the $2D$ spatialized scene (taking into account scale and lighting differences, occlusions and collisions).

References

[1] P. Bobet, J. Blanc, and R. Mohr. Aspects cachés de la trilinéarité. In *Proc. RFIA'96 Conf.*, pages 137–146, Rennes, France, Janvier 1996.

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

[3] H. Delingette. General Object Reconstruction based on Simplex Meshes. Technical Report 3111, INRIA, Sophia Antipolis, France, February 1997.

[4] K. Fintzel and J.-L. Dugelay. Visual Spatialization of a Meeting Room from $2D$ Uncalibrated Views. In *IEEE IMDSP*, Alpbach, Austria, July 1998.

[5] K. Fintzel and J.-L. Dugelay. Virtual $3D$ Interactions on $2D$ Real Multi-Views. In *IEEE ICMCS*, Firenze, Italy, June 1999.

[6] MPEG demo of the face tracking system. URL http://www.eurecom.fr/∼image/TRAIVI/valente-8points.mpg . (1782100 bytes).

[7] A. Shashua. Projective Structure from Uncalibrated Images: Structure from Motion and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):778–790, August 1994.

[8] A. Shashua. Trilinearity in Visual Recognition by Alignment. In *ECCV A*, pages 479–484, 1994.

[9] The Virtual Meeting for Macintosh and Windows, RTZ Software Home Page. URL http://www.rtz.com .

[10] S. Valente and J.-L. Dugelay. A Multi–Site Teleconferencing System using VR Paradigms. In *ECMAST*, Milano, Italy, 1997.

[11] S. Valente and J.-L. Dugelay. $3D$ Face Modeling and Encoding for Virtual Teleconferencing. In *VLBV*, Urbana, IL, October 1998.

[12] S. Valente and J.-L. Dugelay. Face Tracking and Realistic Animations for Telecommunicant Clones. In *IEEE ICMCS*, Firenze, Italy, June 1999.