appears in: IEEE International Conference on Multimedia Computing and Systems, June 7-11 1999, Firenze Italy

Virtual $3D$ Interactions between $2D$ Real Multi-Views

Katia Fintzel Jean-Luc Dugelay

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

Abstract

In this paper, we introduce an image processing tool, Video Spatialization, as a new approach to navigate through a "virtualized" scene, only known by a limited set of $2D$ uncalibrated images (i.e without any $3D$ CAD model). We develop this approach in the context of an interactive application: a multipoint teleconferencing system for very low bit rate links (internet, mobile communications), based on the immersion of the $3D$ virtual models of all the participants, in a common virtualized meeting place controlled by video spatialization. This article contains (i) the recall of an efficient "mesh-oriented" algorithm for the reconstruction of real views and the synthesis of virtual ones from a triplet of uncalibrated views; (ii) some extensions of the original approach from one to n triplets of views to simulate $3D$ navigation and (iii) preliminary investigations using video spatialization for background control within the context of our virtual teleconferencing application.

1. Introduction

This paper discusses the problem of reconstructing real points of view of an arbitrary $3D$ scene and synthesizing virtualized ones, in order to simulate $3D$ navigation from a limited set of $2D$ uncalibrated views without resorting to any $3D$ CAD model of the scene: this is referred to as *Video Spatialization*. This technique aims at offering the possibility for an observer to visualize a scene from anywhere and in any direction, just as in real navigation through a $3D$ place. With this respect, we recall in section 2 an efficient "mesh-oriented" approach for real view regeneration from two neighboring ones, and a set of analytical inferences to synthesize the virtual point of view in relation with the user's motion and orientation (as developped in [10]). Image mosaïcking is therefore used as an image processing approximation to increase the overlapping areas between the initial input data ($2D$ uncalibrated images), and improve the visual rendering and realism of the resynthesized point of view. These just mentionned algorithms, called "intra-triplets" processings in this paper, deal only with three views of a real scene. However, applications for immersive media will use in practice more than three views. In section 3, we extend the "intra-triplets" process combining the synthesis method from several triplets of uncalibrated images, with image mosaïcking approximations applied to virtual output views, in order to simulate the larger real motions of an observer in a "virtualized" $3D$ scene. Such extensions are called "inter-triplets" processings. Finally, we describe in section 4 our future investigations in the context of the TRAIVI project, which takes advantage of video spatialization techniques for virtual teleconferencing systems, in which we introduce $3D$ clones inside a virtualized $2D$ meeting room.

2. Image Transfer

2.1. A "Mesh-Oriented" Approach

By extension of the stereovision concepts [8, 7], we proposed in [4, 10] an algorithm for real view reconstruction from uncalibrated $2D$ views of a $3D$ scene. This was based on trilinear tensors, first modeled by Spetsakis and Aloimonos [19] in the calibrated case and by Shashua in the uncalibrated case [16]. An existing view can therefore be reconstructed from two other neighboring views without any explicit calibration stage as follows:

- Analysis: Using seven or more corresponding points in three original uncalibrated views, eighteen parameters of a trilinear form can be estimated (for more details about the definition of trilinear parameters see [16, 17, 18] and [4]).
- Synthesis: The central view is reconstructed using all corresponding points of the external images (i.e left and right) and the estimated parameters from the analysis, as shown in figure 1.

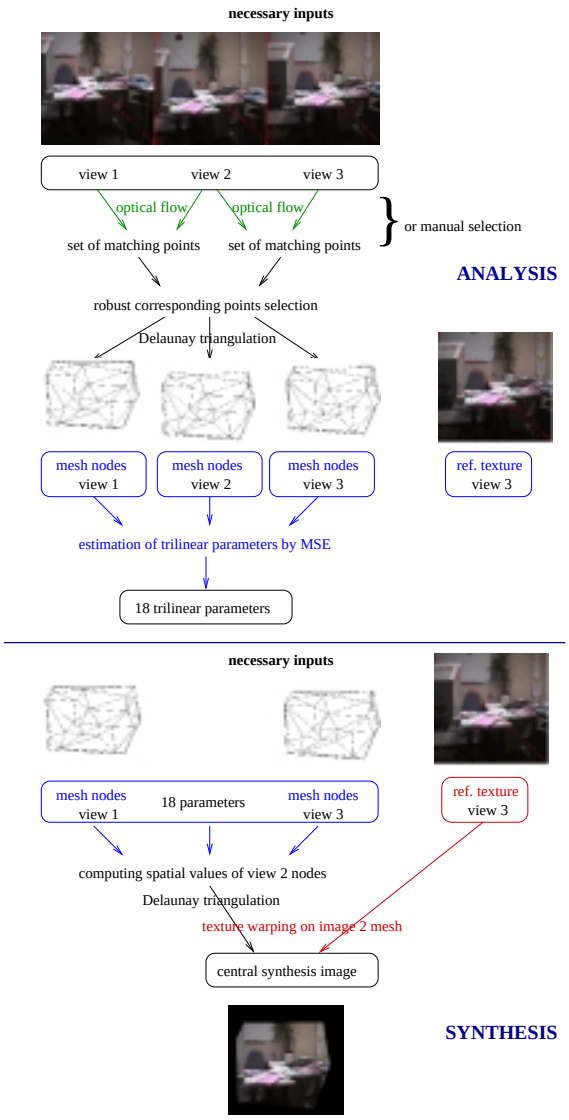
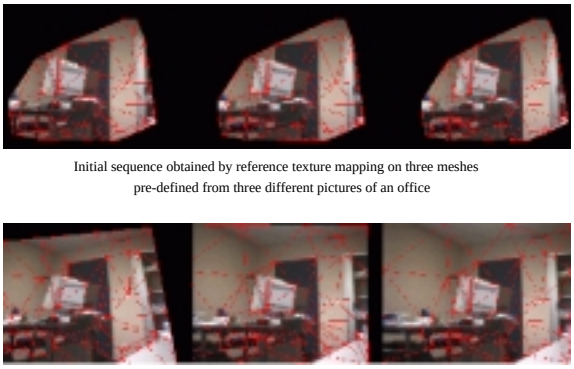


Figure 1. Regeneration process of a real view

Contrary to the image-based rendering methods for view synthesis typically found in the literature [1, 17, 18, 2] and resorting to dense correspondences, one of our contributions is to use a "mesh-oriented" approach. We represent the three original images as a reference texture, mapped on three associated meshes, defined using the Delaunay triangulation [21] on homologous points from the three initial images. This choice of representation is fully justified by the compatibility with real-time constraints of our applications and the increase of the visual comfort rather than the reconstruction accuracy, as explained in [10, 5]. As opposed to classical methods, we obtain plain reconstructions (without any mis/non or over-informed point) visually acceptable for our kind of applications. The coverage of the reconstruction obviously depends on the size of the common area of the three initial meshes, which can be very limited. As a solution to this problem, we introduce in the next subsection a module of image mosaïcking as a pre-processing step of the reconstruction method.

2.2. Mosaïcking on Original Triplets of Pictures

The initial image triplet used for the reconstruction method is represented by three meshes limited to a common covering area and a reference texture corresponding to the third image in our case. Using the theory of homographic transforms [6], we extend the initial meshes to the entire reference texture by approximations (as shown in figure 2).



Initial sequence obtained by reference texture mapping on three meshes pre-defined from three different pictures of an office

Extended initial sequence obtained after image mosaicking process

Figure 2. Initial and extended triplets

We hence obtain larger meshes, which may cover an area of the original image that was not investigated when initializing the meshes. By using image mosaïcking, we combine all the texture information of the initial data in the reconstruction.

2.3. Unknown Views Synthesis

The vector of trilinear parameters is analytically altered to simulate a virtual change of the focal length or a geometrical $3D$ displacement of the camera relative to the reconstructed view and synthesize unknown virtual points of view. Only some synthesis steps are required to simulate human relative motions or relative changes of the current point of view, whereas the analysis step remains unchanged. New views can therefore be virtualized in real time from well-adapted and relevant meshes, which depend on the $3D$ scene complexity and the desired quality for the synthesized images.

All the possible manipulations of the trilinear parameters and the inputs needed to simulate each kind of transformations, due to a change of the intrinsic or extrinsic parameters of the central video camera are summerized in [10]. And the complete developments concerning the modifications of the trilinear parameters, which render virtualized consistent points of view of the scene, are reported in [4]

for the simulations of the focal length changes and in [9] for all kinds of rotations and translations. Rotations and particularly translations are not straightforward without an explicit calibration, but our work aims at keeping the calibration step implicit by restoring the inputs (especially the relative rotations between the initial positions of the video cameras) directly from the trilinear parameters estimated from the triplet of initial images as explained in [9, 15].

To test the validity of the trilinear parameters estimation and manipulation, our method was first applied on several triplets of views extracted from a synthetic scene, before generating virtualized points of view from real scenes as shown in figure 3.

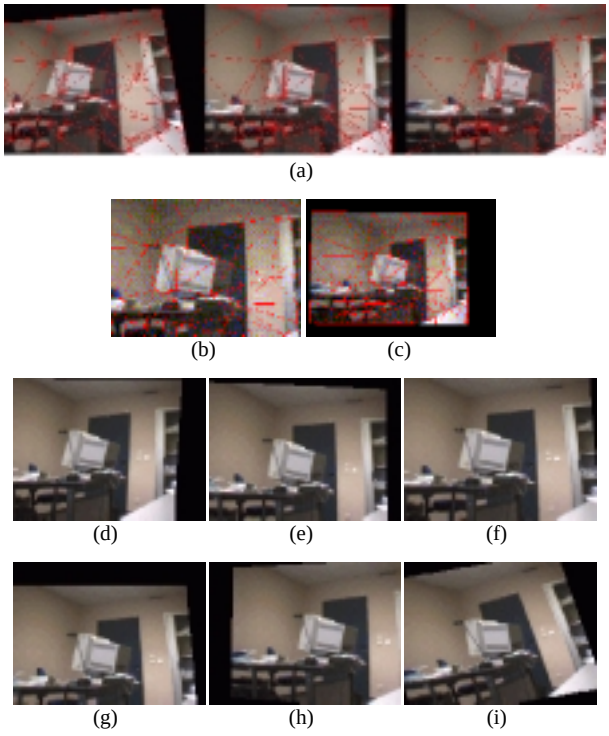


Figure 3. Synthesized points of view from extended meshes:
(a) initial triplet, (b,c) central video camera focal change, (d) central video camera translation along the horizontal axis, (e) along the vertical axis, (f) along the optical axis and (g) rotation around the horizontal axis , (h) around the vertical axis and (i) around the optical axis

3. Video Spatialization from Multi-Triplets

In order to simulate navigation through a virtualized scene, we must generate enough different synthesized images of the environment from several triplets of initial views, and link these resulting syntheses together as if the observer moves freely in the scene. The user's eyes are then considered as a virtual camera, whose positions and motions allow us to synthesize continually his coherent visual feedback of the scene.

Let us consider several triplets of images, represented by the needed reference textures and their associated meshes (a texture per triplet of meshes). We are able to simulate motions around each triplet, as described in section 2, and using consecutive triplets of views (in terms of movement), we can propagate the same type of motion from a triplet to the following one (as presented in figure 4) testing at each time the credibility of the synthesized view. The interested reader can find more details in [9] and examples of Mpeg encoded sequences of video camera simulated motions at *http://www.eurecom.fr/˜image/spatialisation.html*

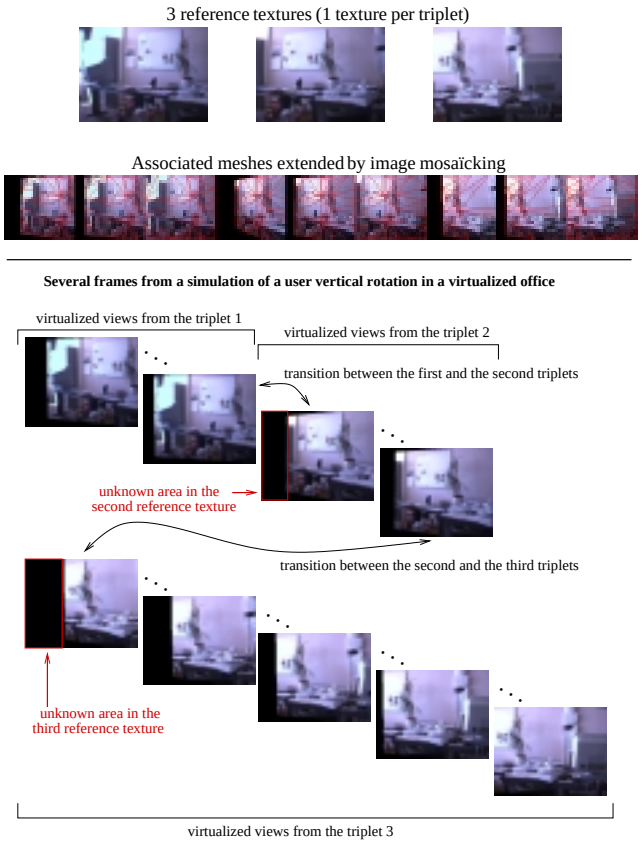


Figure 4. Synthesis of a user large rotation

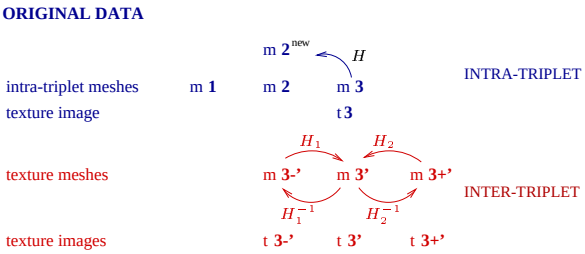
These travelling simulations are fair but visually uncomfortable for the user, because of the transitions between the initial tri-view sequences. In fact, when switching from a triplet to the next or previous one, annoying visual artefacts are introduced in the virtualized picture. The last view generated from a triplet looks different from the first image synthesized from the following triplet: this is referred to as *triplets transition* (highlighted in figure 4). If we work on several triplets, image mosaïcking can be applied between the synthesized output resulting image and the reference

textures of the previous and the next neighboring triplets of images, in order to limit the non-informed area of the resulting image and to make up for the triplets transitions. The main idea here is to surimpose three images to render a realistic point of view, even if there is a switch of triplets:

- The usual intra-triplet synthesized view, really simulating $3D$ as described in section 2, is then displayed in front of two visual approximations, computed as explaned in the next step.

- The two approximations, denoted *first underlayer* and *second underlayer* in figure 5 are obtained by image mosaïcking between the reference texture of the previous and the current triplets for the first view, and the reference texture of the current and the following triplets for the second view. In practice, an underlayer is defined using a combination of two homographic transforms: an inter-triplet homography H_i between two reference textures (including the reference texture of the current triplet) just computed once and for all, and the intra-triplet homography H between the mesh of the reference texture of the triplet and the mesh corresponding to the current synthesized view (previous step), updated at each instant.

This method, entirely sketched in figure 5, is probably sub-optimal because the resulting views are composed by a virtualized image really simulating the user's $3D$ motion, on which attention should be focused, displayed over two approximated images called *underlayers* (more details can be found in [9]). These underlayers are obtained by operating image mosaïcking between the current virtualized view and the reference texture taken from the previous triplet of data for the first under-image and between the same virtualized view and the reference texture taken from the next triplet concerning the second underlayer. The underlayers are only approximations (except in the case of pure rotations) used to increase the visual comfort of the observer, making up for the non-informed areas of his virtualized point of view of the scene. Figure 6 presents visual results of the synthesis of virtualized points of view from altered trilinear parameters, with propagation on several triplets of pictures by image mosaïcking and underlayers method. With respect to figure 4 we note in particular that the mosaïcking steps to create the underlayers allow to increase the fluidity between consecutive triplets of views, smoothing the triplets transition between the last point of view generated from a triplet and the first one obtained from the following triplet. Such Mpeg encoded sequences are also available at *http://www.eurecom.fr/~image/spatialisation.html* for comparison with the previous sequences (i.e without mosaïcking procedure as shown in figure 4).



ALGORITHM FOR THE UNDERLAYERS DEFINITION:

Estimated only once and for all

1. Trilinear parameters (α_i) estimation

   $\Longrightarrow m_2$ resynthesis from m_1 and m_3
2. Definition of $H_1 \|$ $H_1(m3-') = m3'$
3. Definition of $H_1^{-1} \|$ $H_1^{-1}(m3') = m3-'$
4. Definition of $H_2 \|$ $H_2(m3+') = m3'$
5. Definition of $H_2^{-1} \|$ $H_2^{-1}(m3') = m3+'$

For each video camera motion: Estimated at each instant

1. $(\alpha_i) \longrightarrow (\alpha_i')$ altered
2. $m1, m3, (\alpha_i') \longrightarrow$ synthesized mesh $m2^{new}$ view 2 virtualized
3. defintion of $H \|$ $H(m3) = m2^{new}$
4. addition of n nodes to the mesh $m3-' \longrightarrow m3-'^{*}$
5. $H_1^{-1}(m3) = m3-'^{bis}$
6. $m3-'^{*} + m3-'^{bis} \longrightarrow m3-'^{**}$
7. $H(H_1(m3-'^{**})) = m2^{new**t3-}$ first underlayer
8. addition of n nodes to the mesh $m3+' \longrightarrow m3+'^{*}$
9. $H_2^{-1}(m3) = m3+'^{bis}$
10. $m3+'^{*} + m3+'^{bis} \longrightarrow m3+'^{**}$
11. $H(H_2(m3+'^{**})) = m2^{new**t3+}$ second underlayer

Figure 5. Triplets transition make up

m_i = intra-triplet meshes
m_i' = texture meshes, only defined from the view used as reference texture for a triplet (one texture mesh for a triplet) $\neq m_i$ for reasons of limited coverage
$m_i'-$ = previous texture mesh of m_i'
$m_i'+$ = following texture mesh of m_i'
$H \| H(m_i) = m_j$ = homographic transform from the mesh m_i to m_j
m_i^{new} = mesh m_i obtained by the simulation of a video camera motion
m_i^{*} = texture mesh m_i' extended by nodes addition

4. Conclusion

4.1. Future Work

In this paper, we have presented the extension on several triplets of views of our "mesh-oriented" approach for virtual views synthesis combined with an image-mosaïcking-based method, to increase the visual realism of *virtualized immersion*. This was called *video spatialization of a real $3D$ scene from multi-triplets of $2D$ views* introducing the concept of underlayers of a virtualized image, to offer the observer a better visual comfort.

Our future perspectives are focused on the integration of $3D$ objects in the $2D$ synthesized points of view of a scene. This is a necessary stage to offer users more interactivity in applications like virtual teleconferencing systems. But in this case, lots of problems like occlusions or collisions have to be studied (see figure 7).

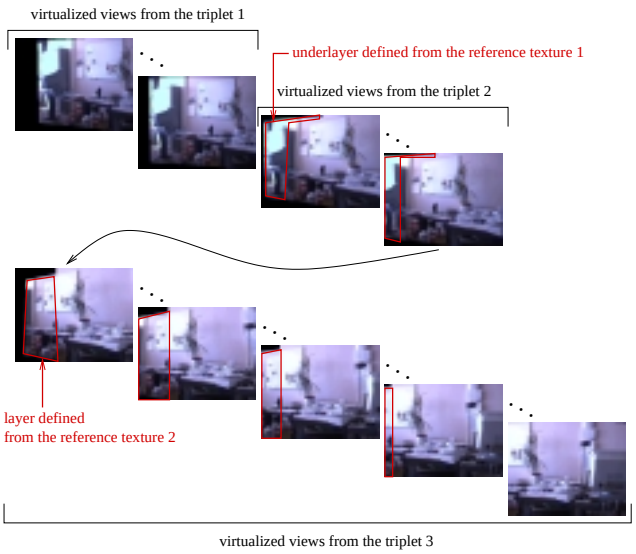


Figure 6. Minimization of the texture transitions using underlayers

The positions and orientations of the $3D$ objects inserted in the scene have to be coherent with the currently rendered $2D$ point of view of the user, unfortunately *a priori* unknown. Collisions between the inserted $3D$ objects and the initial objects present in the scene (only known by $2D$ images) have to be taken into account at each instant. To this extend, we have to define some main depth planes in the scene, to obtain a partial map of relative depth of the static objects of the scene. Our first investigations in this domain allow us to restore a discrete map of relative depths of the $3D$ scene, recovering perspective projection from the estimation of the trilinear parameters.

The complete management of $3D$ objects inserted in $2D$ spatialized environments is our future domain of interest, solving difficulties such as: differences of scale and lighting, and occlusions between objects of varying dimensionnality. Recent standards like MPEG-4 [13] consider this issue: one of the fundamental aims mentionned in the MPEG-4 SNHC call for proposals from the integration group is to "efficiently code interactive $2D$ and $3D$ environments consisting of real-time audio video and synthetic objects" [13]. The integration group experts focus on requirements for $2D/3D$ synthetic and natural data coding, seeking the integration of video coding (based on $2D$ feature analysis and model-based coding) and coding of structured $2D/3D$ graphical synthetic environments (including modeling, communication, run-time efficiency, real-time interaction and rendering of them), given the number of potential applications. Our image processing tool based on video spatialization is independent from the standard MPEG-4, but our work shares some of its major concerns and appears to be applicable in the context of an MPEG-4 encoder/decoder. The SNHC *mesh object* is a representation of a $2D$ deformable geometric shape, from which video objects may be created during a composition process at the decoder, by spatially piece–wise warping of existing video object planes or still texture objects. For this reason, video spatialization is an interesting technique to create a virtual environment without any explicit CAD model, using efficient image–rendering procedures for visualization.

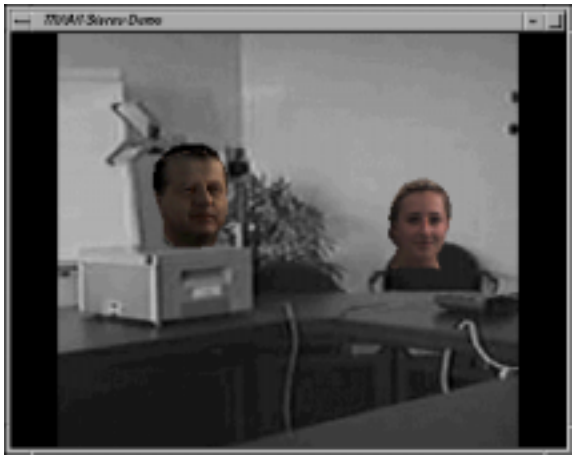


Figure 7. Early insertion of $3D$ clones in a $2D$ environment

4.2. In the Context of a Televirtuality Project

Our work on video spatialization takes place originally in the larger TRAIVI[1] project, whose goal is to create a complete virtual teleconferencing system. In fact, the use of teleconferencing systems between multiple sites has considerably increased [14], because of industrial demands, but generally offers a poor quality of service [12]. The immersion of the participants in the same virtual and realistic environment, with the ability to move and look at the other participants, could make up for the lack of realism of classical systems and offer new ergonomic possibilities [11]. The TRAIVI project proposes an integrated approach that adresses both the participant's representation and the virtual meeting room background by mixing synthetic textured $3D$ face models with spatialized natural images. This virtualized vision of the real world is an alternative to the arbitrary artificial worlds, used in projects like [3] (where videos representing the participants of a videoconference are displayed in the virtual $3D$ model of a meeting area). The stake is then to render the real world in a way that is visually coherent and comfortable for its users.

Video spatialization for background control is one of the video processings we have to master in combination with

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

model-based coding for participants control [20], to achieve a satisfactory level of visual realism in the development of a virtual teleconferencing system. That is why we focus on the synthesis of office or meeting-room images, with an emphasis for real-time visualization and realism of regenerated or unknown synthesized images, as opposed to the reconstruction accuracy. To that extent our "mesh-oriented" approach is a good trade-off in the context of the TRAIVI project, which requires realism and real-time.

The synthesis of virtual views is particularly interesting for the TRAIVI application: we can now imagine a virtual meeting composed of a pre-processing stage before the session. During this stage, information related to the user (his $3D$ model and his initial position) and the choice of the meeting area will be transmitted to a central site, which will pre-compute, from a few real uncalibrated views, the corresponding vectors of trilinear parameters and inter-triplets homography links, uploaded to each remote site. During the session, each site, independently from each others, will be able to create locally, by algebraïc processing applied on the trilinear parameters and intra-triplet homography, new coherent points of view for its user, based on his virtual position, motion parameters and center of interest in the meeting room, without sending any other information [5].

Our perspectives for the TRAIVI project are:

- the implementation of a complete room spatialization system, dealing with the quantity of pre-downloaded textures and the user's permitted motion granularity.
- the coherent integration of $3D$ models of the participants and background $2D$ images, which is still an open problem.

Acknowledgements The authors wish to thank Espri Concept for their support and contribution to this paper.

References

[1] D. Beymer, A. Shashua, and T. Poggio. Example Based Image Analysis and Synthesis. Technical Report 1431, MIT, 1993.

[2] J. Blanc and R. Mohr. Towards Fast and Realistic Image Synthesis from Real Views. In *SCIA'97*, Lappeenranta, Finland, 1997.

[3] C. Breitender, S. Gibbs, and C. Arapis. TELEPORT - An Augmented Reality Teleconferencing Environment. In *3rd Eurographics Workshop on Virtual Environments Coexistence & Collaboration*, Monte Carlo, Monaco, February 1996.

[4] J.-L. Dugelay and K. Fintzel. Image Reconstruction and Interpolation in Trinocular Vision. In *IMAGE'COM 96*, Bordeaux, France, Mai 1996.

[5] J.-L. Dugelay, K. Fintzel, and S. Valente. Synthetic Natural Hybrid Video Processings for Virtual Teleconferencing System. In *PCS*, Portland, Oregon, April 1999.

[6] O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. The MIT PRESS, 1993.

[7] O. Faugeras and B. Mourrain. Algebraic and Geometric Properties of Point Correspondences between N Images. In *ICCV'95*, Boston,MA, June 1995. IEEE Computer Society.

[8] O. Faugeras and L. Robert. What Can Two Images Tell us about a Third One? *The International Journal of Computer Vision*, 1994.

[9] K. Fintzel and J.-L. Dugelay. a: Analytical Manipulations of Trilinear Parameters to Synthesize *a priori* Unknown Views, b: Initial Rotation Parameters Recovery from Trilinear Parameters of a Three Video Cameras System, c: Initial Perspective Projection Matrices Recovery from Trilinear Parameters of a Three Video Cameras System, d: Addition of Underlayers Visually Coherent to a Virtualized Image Obtained by Trilinear Synthesis. Technical report, EURECOM, Sophia Antipolis, France, 1997-98. in french.

[10] K. Fintzel and J.-L. Dugelay. Visual Spatialization of a Meeting Room from $2D$ Uncalibrated Views. In *IEEE IMDSP'98 Workshop*, Alpbach, Austria, July 1998.

[11] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, H. Farid, and T. Kanade. Virtual Space Teleconferencing using a Sea of Cameras. In *First International Symposium on Medical Robotics and Computer-Assisted Surgery 2*, pages 161–167, Pittsburgh, Pa, September 1994.

[12] K. Jeffay, D.-L. Stone, T. Talley, and F.-D. Smith. Adaptive, Best effort Delivery of Audio and Video Across Packet-Switched Networks. In 3^{rnd} *Intl. Workshop on Network and OS Support for Digital Audio and Video*, SanDiego, CA, November 1992.

[13] MPEG–4 Synthetic/Natural Hybrid Coding. URL http://www.es.com/mpeg4-snhc/ .

[14] K. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita. Multiparty Videoconferencing at Virtual Social Distance: Majic Design. In *ACM'94*, pages 385–393, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223 Japan, October 1994.

[15] B. Rousso, S. Avidan, A. Shashua, and S. Peleg. Robust Recovery of Camera Rotation from Three Frames. In *CVPR'96*, June 1996.

[16] A. Shashua. On Geometric and Algebraïc Aspect of $3D$ Affine and Projective Structures from Perspective $2D$ Views. In J.-L. Mundy, A. Zisserman, and D. Forsyth, editors, *Applications of Invariance in Computer Vision*. Second European Workshop Invariants, Ponta Delagada, Azores, October 1993.

[17] 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.

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

[19] M.-E. Spetsakis and J. Aloimonos. A Unified Theory of Structure from Motion. In *DARPA IU Workshop*, pages 271–283, 1990.

[20] S. Valente and J.-L. Dugelay. Face Tracking and Realistic Animations for Telecommunicant Clones. In *ICMCS'99*, Firenze, Italy, June 1999.

[21] D. Watson. *A Guide to the Analysis and Display of Spatial Data*. Pergamon PRESS, 1992.