The ORIGAMI project: advanced tools for creating and mixing real and virtual content in film and TV production

J.F. Evers-Senne(CAU[1]), O. Grau (BBC[2]), R. Koch (CAU), F. Lavagetto (DIST[3]), M. Milne (FmS[4]), M. Price (BBC), O. Razzoli (Quest[5]), A. Sarti (PdM[6]), S. Tubaro (PdM)

Abstract

ORIGAMI is a EU-funded IST project with the goal of developing advanced tools and new production techniques for high-quality mixing of real and virtual content for film and TV productions. In particular, the project focuses on pre-production tools for set extension through image-based 3D modelling of environments and object. One key goal of the project is to achieve real-time in-studio previsualisation of the virtual elements (objects and actors) of the extended set. In this paper we illustrate the studio pre-visualisation system that has been developed for the project, and we describe its usage within a high-end film and TV production environment. We also give an overview of the developed solutions for an automatic generation of 3D models of environments and objects.

1. Introduction

The use of high-end animated computer models in film or even TV production is becoming more and more common. However, computer animation films where real and virtual characters act together in a seamless fashion are still too expensive to be common practice. In fact, aside from the cost related to 3D model creation, animation and rendering, there is the actual cost of shooting the "takes" where the real actors give the impression of seeing the virtual elements of the set and the interacting virtual actors. The interaction with virtual objects and/or virtual characters tends to be a very challenging task for actors, as they have little, if any at all, feedback to indicate the relative position or activity of the virtual object that they are to interact with. We must keep in mind, in fact, that humans are very good at detecting what a person is looking at, therefore, if an actor were looking at a video monitor instead of an specific expected point, we would not miss it. This lack of a visual feedback of the key virtual elements of the scene make actors work in rather difficult conditions, as positional and synchronisation cues are usually given in terms of rough marks on the floor and timed gestures made by the crew.

The situation is not easier for camera crew and director. Those film directors who choose to make extensive use virtual set extension and virtual actors, in fact, must accept dramatic changes in their way of working, as they cannot easily have a preview of what the result will look like at the end of the postproduction work. The lack of a WYSIWYG (What-You-See-Is-What-You-Get) approach to the filming of live action in a chroma-key studio (a controlled environment where actors perform in front of a blue screen for chroma-keying purposes), for example, makes it difficult to plan camera trajectories and lighting. As the film director and the director of photography can do little but guessing the final outcome, they often end up either keeping what they blindly obtain (and settling for less than optimal results), or having to pay for a number of expensive screen tests (prototypes of digital effects).

The Origami project is aimed at overcoming many of such difficulties, as it is developing a set of new tools for planning and pre-visualisation in film and TV production, from image-based modelling to real-time visual feedback techniques. We will provide an overview of the studio pre-visualisation system that has been developed for the project, and we will describe its usage within a high-end film and TV production environment. We will also illustrate our solutions for an automatic generation of 3D models of environments and objects.

2. Overview

The goal of the Origami project is to develop an integrated environment for the mixing of real and virtual content in TV/film production. The project focuses on two

[1] Multimedia Information Processing, Christian-Albrechts-Universitaet zu Kiel, Institut fuer Informatik und Praktische Mathematik. Herman-Rodewaldstr. 3, D-24098 Kiel, Germany
[2] BBC Research and Development, Kingswood Warren, Tadworth, Surrey, KT20 6NP
[3] Dipartimento di Informatica, Sistemistica e Telematica, Via all'Opera Pia 13 - 16145 Genova, Italy
[4] FrameStore. 9 Noel Street, London W1V 8GH, England
[5] Quest, SRL, Via Pontina Km 23,270 – Roma, Italy.
[6] Dipartimento di Elettronica e Informazione - Politecnico di Milano. Piazza Leonardo Da Vinci 32, 20133 Milano, Italy

application scenarios that privilege different aspects of the software environment: off-line productions and television productions with dynamic content. The former case mostly deals with off-line virtualisation and authoring of 3D environments and objects; while the latter case is more focused on bi-directional Real-Virtual (R-V) feedback and require a real-time previsualisation of the composited scene.

The specific area of application that the ORIGAMI project is dealing with and the related needs require us to deal with numerous challenging problems. As far as the 3D modelling aspects are concerned, we have strict requirements of high quality results in a limited acquisition/processing time. These requirements are to be achieved with flexible and low-cost procedures in order to be able to accommodate a wide range of modelling situations with modest costs. For these reasons, the Consortium focused on the development of high-end free-form camera motion estimation methods as well as high-quality geometric and radiometric modelling solutions. As for problems of mutual feedback between real and virtual actors and scenes, there are severe requirements of temporal/positional cueing, gazing direction (eye contact), and, most of all, real-time visual feedback for actors, director and crew. In order to overcome these difficulties, the Origami Project developed a studio that enables a real-time bi-directional feedback between real and virtual.

As we can see in Fig. 1, the Origami's vision of film production consists of an off-line phase in which objects and environments are acquired and modeled using advanced image-based methods. Live action takes place in the studio, where the real dynamic content is acquired and modeled on the fly. The 3D model of actors makes virtual actors aware of their location, while real actors are able to see the virtual elements thanks to a novel technology that enables visual feedback without affecting the studio's ability to perform chroma keying. Real-time previsualisation, of course, is also available for director and crew.

3. Generation of 3D content

Content virtualisation is based on different solutions,

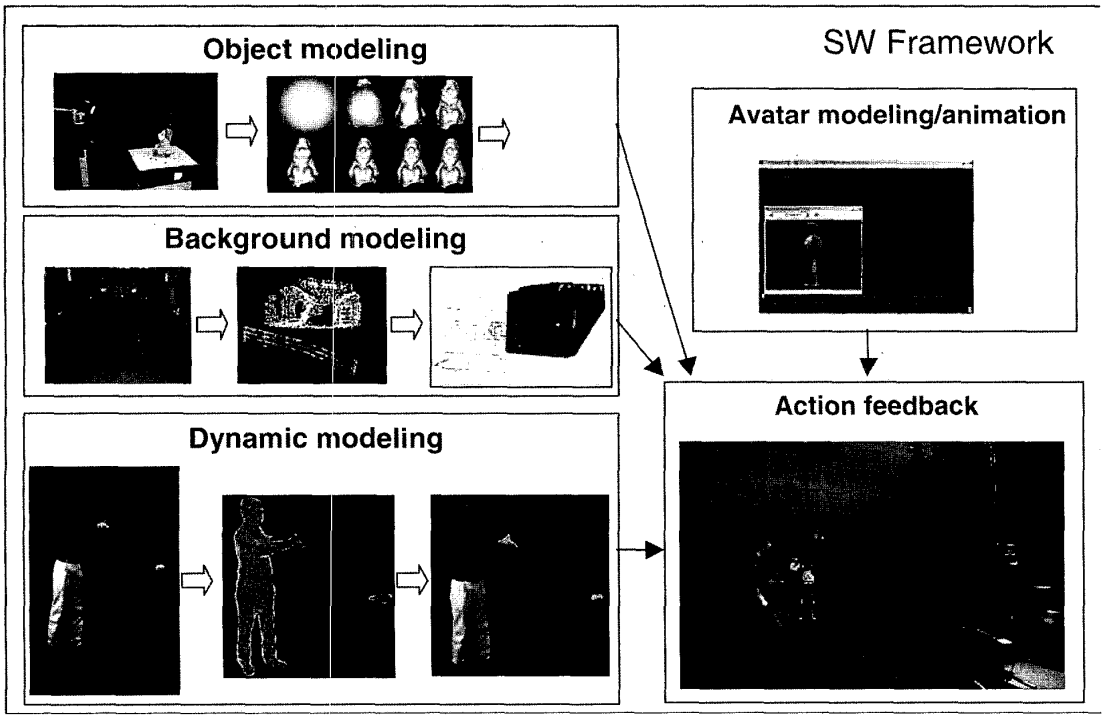


Fig. 1 – ORIGAMI project workflow: objects, background and avatars are modelled off-line. The actor's model is constructed as the action takes place in the studio, which is able to provide a visual feedback of the virtual scene elements to the actor.

depending on the origin. When dealing with full-3D objects, it is important to guarantee the actor the possibility to interact with their geometry (to be able to "handle" them during the live action). This means that the geometry will be modelled in an accurate and viewpoint-consistent fashion. When dealing with the environment, what matters is the appearance from the desired position. In this case we adopt image-based modelling strategies. When dealing with real-time modelling, we need to have geometric models obtained at modest computational cost, therefore we adopt volumetric solutions based on the geometry of visual hulls (volumetric intersection).

3.1. Object modelling

The method adopted by the ORIGAMI project for building complete 3D models of objects from video sequences is based on the temporal evolution of the level-set zero of a volumetric function [2]. This function is usually defined as the signed distance from the evolving surface and its evolution is such that the front will "sweep" the whole volume of interest until it takes on the desired shape under the influence of some properly defined "external action". In our solutions, the level-set evolution is defined by a Hamilton-Jacobi partial differential equation, which is discretised into a front-evolution update equation, controlled by a properly defined velocity function. The front velocity can be quite arbitrarily defined in order to steer the front propagation toward a desired shape. Terms that may appear into its expression are: local curvature, which promotes a maximally smooth implosion of the surface; distance from 3D data, which promotes data fitting; inertia, which promotes topological changes (object splitting or generation of holes); texture agreement – which maximizes the similarity between the appearance of the modelled surface and its available views. Besides such terms, we are free to define new velocity terms that attribute the surface evolution some desired behaviour. One key characteristic of the solutions that we developed for image-based modelling is in the fact that they work in a multi-resolution fashion, which speeds up the computation to a manageable level [2].

We also developed a novel approach to the modelling of surfaces from sets of unorganised sample points (e.g. data coming from range cameras or 3D laser scanners), again based on the temporal evolution of a volumetric function's level-set. The evolving front, in this case, can be thought of as the surface that separates two different fluids that obey specific laws of fluid dynamics. One remarkable feature of this approach is its ability to model complex topologies thanks to a novel strategy that allows us to steer the front evolution using Voronoi surfaces in 3D space [3]. Another remarkable feature of this algorithm is its computational efficiency, which proved to be between one and two orders of magnitude better than traditional level-set approaches.

3.2. Modelling the environment

In order to construct the environment that surrounds actors and virtual objects, we developed solutions for fully automatic generation and rendering of 3D models of natural environments from uncalibrated image sequences.

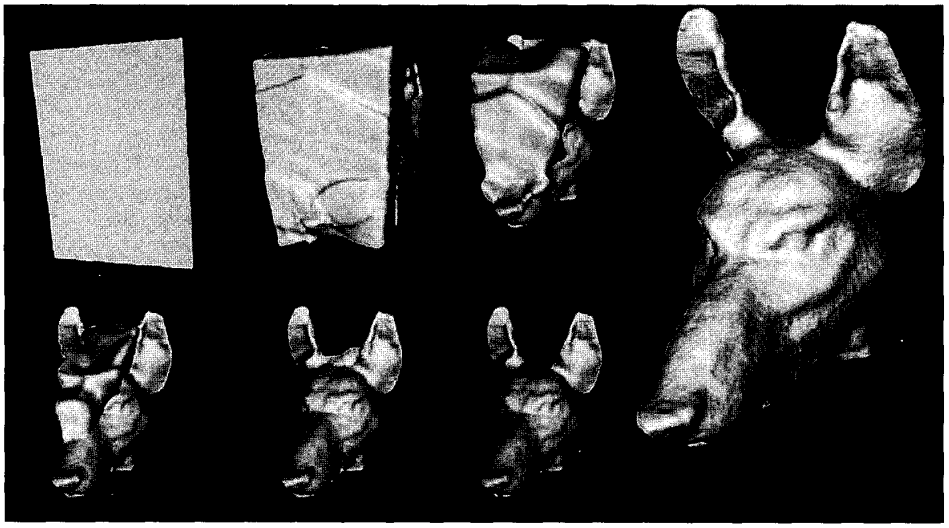


Figure 2 – an example of image-driven level-set evolution applied to 3D object modelling.

The acquisition is made with a hand-held camera or a multi-camera rig, moving with no restrictions on its motion. Since we want to avoid having to place markers in the scene, the camera pose and calibration is estimated directly from the image sequence.

Figure 3 – image data acquistion with multi camera rig

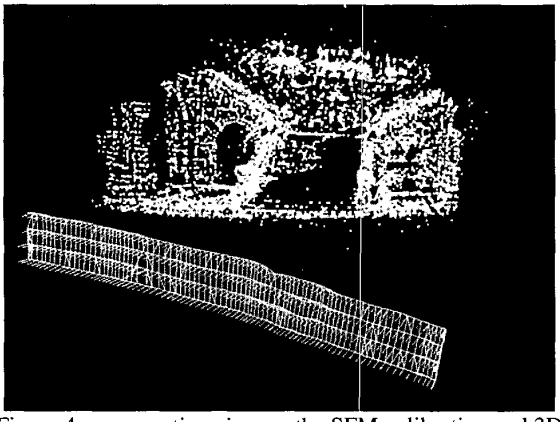


Figure 4 – perspective view on the SFM-calibration and 3D feature points

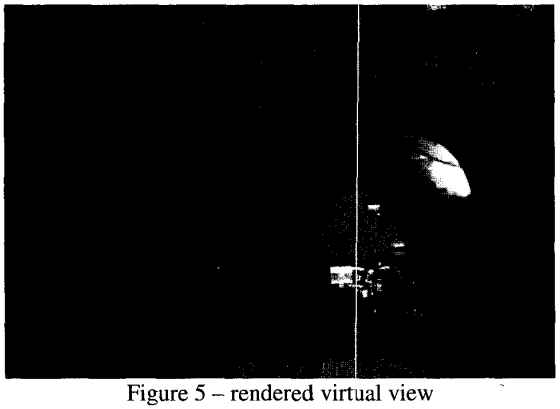


Figure 5 – rendered virtual view

Tracking and calibration are based on an extension of the structure from motion (SFM) approach of Pollefeys and Koch [4,5]. Our SFM approach is based on the tracking of 3D features (corners) throughout the image sequences, and computes the intrinsic and extrinsic camera calibration parameters in a fully automatic fashion. With the calibrated image sequence at hand, one can obtain dense depth maps with multi-stereo image disparity estimation. We extended the method of Koch et al. [6] to a multi camera configuration. A 3D surface representation of the background scene can then be computed from the calibrated images and depth maps.

Scanning the scene sequentially with one camera at different heights or simultaneously with a multi-camera rig improves the quality of depth estimation. One can exploit the connectivity of adjacent viewpoints between the cameras. Fusion of pixel correspondences between neighboring views allows us to fill occluded regions and to significantly improve the precision of the depth maps. The density and the resolution of the resulting depth maps are very high.

The rendering of virtual views from background models is based on two approaches. Starting from the acquired calibrated views and associated depth maps, we can generate geometrical surface mesh models. These models are useful for view planning and mixed reality applications. However, due to incorrect camera calibration and non-static scenes it is often not possible to automatically generate one globally consistent 3D-model from long image sequences. We follow another novel approach [7] for interactive rendering of virtual views from real image sequences. Concepts of depth-compensated image warping and view-dependent texture mapping are here combined to generate virtual views directly from textured depth maps. This image-based modeling approach can handle large and complex scenes while avoiding the need of reconstructing a globally consistent surface geometry. For rendering, a view-dependent warping surface is constructed on the fly, and depth-compensated image interpolation is applied with view-dependent texture mapping. The rendering quality is scalable in order to allow fast pre-visualisation and high-end quality with the same approach. The system can handle large and geometrically complex scenes with many hundreds of real images at interactive frame rates.

3.3. Dynamic modelling

Dynamic content is captured with the Origami studio system. The system is based on a multi-camera approach in a studio equipped with a chroma-keying facility. Currently up to 12 cameras are used for this purpose. For the on-set visualisation three different real-time methods for the computation of the visual hull from silhouette images were implemented and compared. Two methods

use a volumetric representation (3D-array and octree representation). The third method uses a new line representation and has been shown to deliver more accurate surface approximations of objects [8,9].
The dynamic modelling component delivers a 3D triangular mesh in real-time that is used for the on-set visualisation. A directional texture mapping is implemented in the visualisation module as described in the next section. The real-time modelling further determines the 3D head position of the actor, which is used for a view-dependent projection.

4. On-Set Visualisation

The on-set visualisation provides the actor with a view-dependent feedback and the director and operators with a pre-visualisation of the composited programme. It is implemented as a flexible, distributed system and can be configured to the production needs. The main components are: A virtual scene server, the visualisation servers that drive the data projectors and instances of the 3D preview as part of the control panel and an avatar animation module. The virtual scene server keeps the 'master' copy of the entire scene, including static virtual background, actor model and the animated avatar. Further it is taking care of a synchronized update of the (slave) visualisation clients.
The distributed system implements several methods and services: The mask generation to prevent actors being lit by the projectors, a real-time texturing module, the pre-visualisation modules for the director and the view-dependant rendering. These modules make use of data provided by the 3D shape reconstruction, the head tracker and the live user input from the control panel.

4.1. Mask generation

A mask is needed to prevent the projector's light from falling onto the actor. This is generated from the 3D surface model. The most recently computed surface model of the actor is therefore placed into the virtual scene and rendered completely in black with the z-buffer of the rendering system disabled. This guarantees that from the viewpoint of the projector all light rays that could fall onto the actor surface are masked. Due to the latency of the system, light may still fall on the edge of a moving actor, particularly during fast movement. Therefore the mask is enlarged by a security factor that can be adapted to the latency and the fastest motion of the actor that is expected.

4.2. View-dependant rendering

The rendering engines in the visualisation server for the projectors request the latest head position from the head tracker and use this to calculate and render the projected image from the point of view of the actor. The view-dependant rendering allows the actor, if required, to keep looking at the face of the virtual character as he walks around him. That means the virtual scene components appear in space and the actor is immersed.
The renderer also receives any updates in the scene from the virtual scene server. If the virtual character were to move then the actor would see these movements. The combination of the scene updates and the viewpoint-dependant rendering thus allows complex interaction between the virtual and the real scene elements.

4.3. Texturing

For the director or camera operator a renderer is provided that gives a pre-visualisation of the final composited scene, i.e. the virtual and real scene elements. This renderer receives updates from the virtual scene server and grabs the latest 3D shape model of the actors. It can then generate a 3D representation of the scene and allows the director to view the scene from any position. This position can be dynamically updated to allow simulation of shots where the camera is moving.
In order to give a more realistic image the 3D shape model of the actor is textured with a view from one of the cameras. Therefore the renderer determines the studio camera that has the smallest angle to the virtual camera. The 3D shape model is stamped with the time-code of the alpha masks used to generate it, so the renderer requests the image from that time-code from the relevant capturing server, and uses it to texture the 3D shape model.

4.4. Avatar animation

Two avatar animation engines have been developed by DIST for the Origami project, which are able to animate 3D polygon meshes: one for the modelling of a human

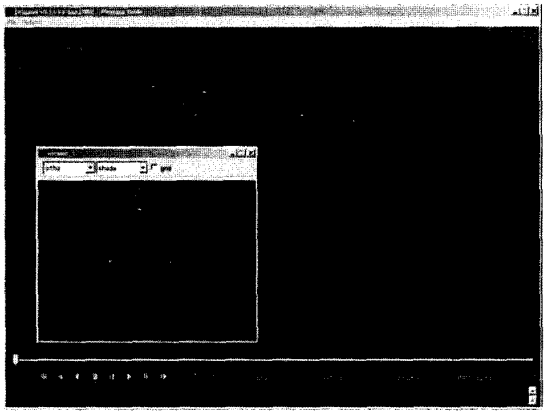


Figure 6 – A view of the avatar animation solver integrated in the FW

face and one for the modelling of full-body humanoids.
The human face and the avatar are integrated in the framework and are rendered in the studio in order give the actor an immediate feedback on the position, the size and even the mood of the virtual characters that will be added in a post-production phase. The rendering will constitute a modest-quality preview of the final scene for the director. Moreover, the director has the possibility to control the reactions and the behaviour of the avatar in real time, during the shooting of the take.
The job of the Face Animation Engine and the Body Animation Engine is to warp the 3D meshes according to the MPEG4-compliant animation data input stream (Facial Animation Parameters or Body Animation Parameters). The Body Animation Engine can also be driven by AMC/ASF (Acclaim Motion Capture) Animation Data, which constitute the de-facto standard for most types of Motion Capture equipment. The animation engines return a reshaped mesh that can be saved as a VRML frame or can be added to an existing Open Inventor tree (i. e. the one provided by the framework).
Both animations are parameter-based, which means that with the same small set of parameters the software is able to animate models of different shape and complexity and both engines are able to perform the necessary computation in real time and guarantee the possibility to interface with real-time motion capturing equipment. The Face Animation Engine and Body Animation Engine have been integrated as a plug-in of the off-line Origami Framework provided by Quest. These plug-ins accept a stream of Animation Parameters and provide as output a 3D mesh that can be used as input for any other plug-in of the framework (i.e. can be viewed connecting its output with the input of the plug-in viewer provided by Quest). The stream of Animation Parameters can be read from a pre-written file BAP/FAP file or can be generated from a script allowing the director to modify in real time the time line of the animation. The main idea is to mix, with the help of a graphical time line composed of different tracks, sequences of simple actions (i. e.: walk and greet) to obtain more complex movements. In the next weeks we expect to be able to acquire a first database of basic movements that will grow in the near future.
The naturalness of the animation has been achieved by taking into account visual prosody cues associated with the expressive face and body gestures typically used by humans in interpersonal verbal and non-verbal communication. Face animation also guarantees lip synchronization with speech.

4.5. Software framework

The ORIGAMI project integrates all software modules within a common flexible environment that provides the user with a friendly and intuitive graphical interface. The SW architecture is based on common object sharing, which constitutes a generalisation of the Object Linking and Embedding (OLE) structure. A powerful and flexible Application Programming Interface (API), has been developed in order to facilitate SW integration. The user is also given a programmable non-modal Graphical User Interface (GUI) based on connection trees, for a WYSIWYG approach to application development.
The basic software framework components are the Solvers, whose aim is to perform specific tasks. Such solvers communicate through ports (sockets) according to a "connection tree", which describes how solvers are logically interconnected. The connection tree can be freely defined by the user in order to construct novel algorithms starting from existing solvers.
Another interesting characteristic of the FW is that solvers can be controlled through parameters that are connected via TCP sockets to remote workstations. In addition, the framework can distribute (and update) the current scene graph to remote viewers. This enables, for example, a camera viewpoint change on a remote PC that controls a projector.

5. Results

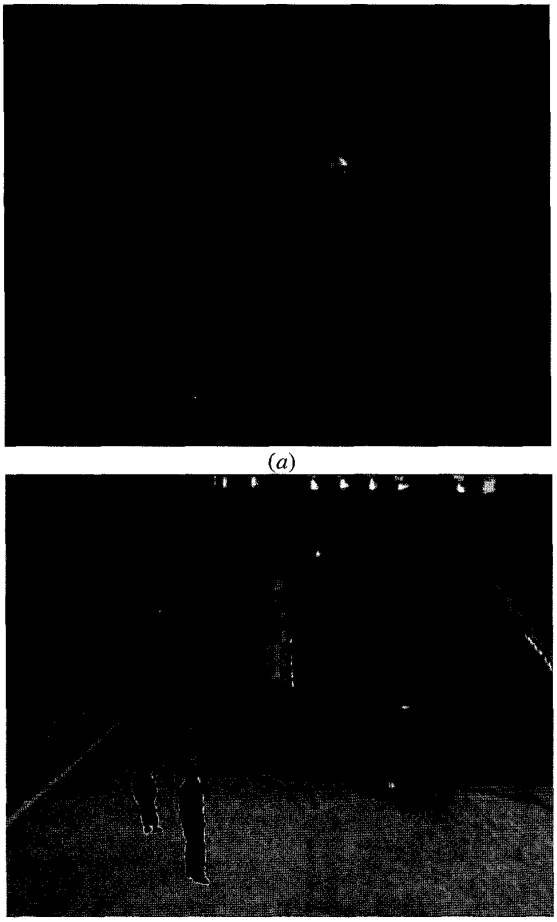
The system was recently used in a demo production. The inside of the entrance hall of a museum was digitised and was then used as a virtual set. This scenario was the basis for pre-visualisation of the entire production, including actors modelled by our system and additional virtual objects.

Figure 7*a* shows a scene in the studio during the production. Figure 7*b* shows the director's view, i.e. a preview of the composited scene. Based on this preview the director of the production is able to decide on changes of positions of actors, props or the camera perspective.

Figure 8 shows the view-dependent projection system. The head position of the actor is tracked and an image according to his position is projected onto the wall. The actor can interact with the scene, which means that as he walks he is immersed into the virtual scene and he can avoid "walking into" virtual objects or he can keep eye contact to objects, as with the 'floating skull' in the image.

6. Conclusions

In this paper we showed the current status of advancement of the Origami project and described the developed technologies for image-based modeling/rendering and for real-time in-studio previsualisation. A public demonstration of this system can be seen at the exhibit of CVMP 2004.



(*a*)

(*b*)

Figure 7 – Original camera view (*a*) and corresponding director's view (*b*)

The ORIGAMI project exhibits a mix of scientifically challenging goals and very pragmatic application scenarios, with measurable performance in terms of cost-effectiveness. In fact, one peculiarity of this project is in the fact that the quality of the results is not a goal but a constraint.

References

[1] IST-2000-28436 "ORIGAMI: A new paradigm for high-quality mixing of real and virtual". INFORMATION SOCIETY TECHNOLOGIES (IST) PROGRAMME, fifth Framework Programme. http://www-dsp.elet.polimi.it/origami/

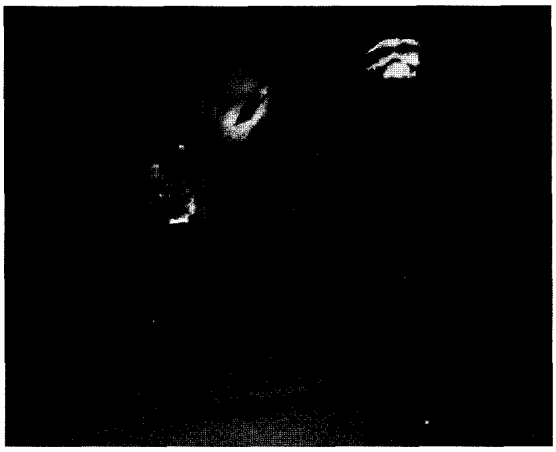


Figure 8 – Actor's view

[2] A. Sarti, S. Tubaro: "Image-Based Multiresolution Implicit Object Modeling". J. on Applied Signal Processing, Vol. 2002, No. 10, Oct. 2002, pp. 1053-1066.

[3] M. Marcon, L. Piccarreta, A. Sarti, S. Tubaro, "A Fast Level-set Approach to 2D and 3D Reconstruction from Unorganized Sample Points". 3rd Intl. Symp. on Image and Signal Proc. and Analysis, ISPA 2003, September 18-20, 2003, Rome, Italy.

[4] R. Koch, M. Pollefeys, B. Heigl, L. Van Gool, and H. Niemann. "Calibration of handheld camera sequences for plenoptic modeling". Proc. ICCV 99, Korfu, Greece, 1999.

[5] Marc Pollefeys, Reinhard Koch, and Luc J. Van Gool. "Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters", International Journal of Computer Vision, 32(1):7-25, 1999

[6] R. Koch, Pollefeys M., and L. Van Gool. "Multi viewpoint stereo from uncalibrated video sequences". In Proc. ECCV'98, number 1406 in LNCS, Springer, 1998

[7] J.F. Evers-Senne, R. Koch. "Image Based Interactive Rendering with View Dependent Geometry", Eurographics 2003, Granada, Spain, September 2003, to appear.

[8] O. Grau, "A Studio Production System for Dynamic 3D Content", accepted paper for Visual Communications and Image Processing 2003, Lugano, Switzerland, 8-11 July 2003.

[9] O. Grau and A. Dearden, "A fast and accurate method for 3d surface reconstruction from image silhouettes," in Proc. of 4th European Workshop on Image Analysis

for Multimedia Interactive Services (WIAMIS), London, UK, April 2003.

[10]O. Grau, M. Price, G.A. Thomas, "Use of 3D Techniques for Virtual Production", in Proc. of SPIE, Conf. Proc. of Videometrics and Optical Methods for 3D Shape Measurement, Vol. 4309, Jan. 2001, San Jose, USA.

[11]F.Lavagetto, R. Pockaj, "The Face Animation Engine: towards a high-level interface for the design of MPEG-4 compliant animated faces", IEEE Tr. Circ. and Sys. for Video Techn., Vol. 9, N.2, March 1999.

[12]G.A. Thomas, O. Grau, "3D Image Sequence Acquisition for TV & Film Production", Proc. of 1st Int. Sym. on 3D Data Processing Vis. and Transm. (3DPVT 2002), Padova, Italy, Jun 19-21, 2002.