Sixth Hungarian Conference on Computer Graphics and Geometry, Budapest, 2012

4D Reconstruction Studio: Creating dynamic 3D models of moving actors

Zsolt Jankó[1], Dmitry Chetverikov[1,2] and József Hapák[1,2]

[1] Computer and Automation Research Institute, Budapest
[2] Eötvös Loránd University, Faculty of Informatics, Budapest

Abstract

Recently, a 4D reconstruction studio has been built at the Computer and Automation Research Institute of the Hungarian Academy of Sciences (MTA SZTAKI). The studio features 13 synchronised, calibrated high-resolution video cameras whose output is used to create dynamic 3D models of moving actors. This is a pioneering project in Central and Eastern Europe, and the software is still under development. We discuss the application areas of 4D reconstruction studios, then give a brief overview of advanced studios operating around the world. Finally, the main hardware and software components of our studio are presented. Both hardware and software contain novel, innovative elements which are discussed in more detail.

1. Introduction

A typical 4D reconstruction studio is a large room with uniform background (e.g., green or blue) and appropriate illumination, equipped with multiple $(6-15)$ calibrated, synchronised video cameras. Its main objectives are capturing videos of a scene from multiple viewpoints and creating dynamic 3D models of articulated objects moving in the scene. Figure 1 provides a panoramic view of the interior of the 4D studio developed at SZTAKI in Budapest, Hungary. A sketch of the studio is given in Figure 2. There is a number of 4D studios in Western Europe and the USA; most of them have similar configurations.

The term 4D studio refers to the spatio-temporal domain where the dynamic models are built: 3D plus time. Acquisition of dynamic, non-rigid objects differs from that of the static ones in several critical aspects. For *static*, rigid objects, one or two cameras can be sufficient, and the camera(s), or a scanner, can move around the object to be reconstructed. (Alternatively, the object can be rotated with respect to static sensor(s).) Large number of different views can be acquired at high resolution and quality, resulting in higher accuracy of reconstruction in both geometry and surface texture.

Figure 1: *A panoramic view of the 4D Studio at SZTAKI.*

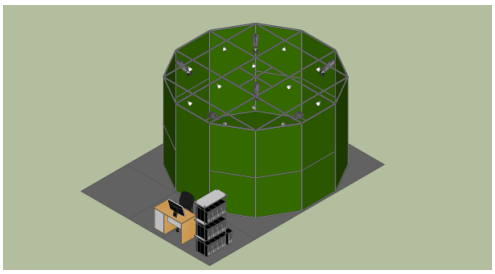

Figure 2: *A sketch of the 4D Studio at SZTAKI.*

3D reconstruction of *dynamic*, non-rigid objects requires a temporal sequence of simultaneous images taken from multiple viewpoints, which needs a set of multiple, fixed video cameras. The number of the views is limited by the number of the cameras, which is, in turn, limited by the data process-

ing capacity of the system. Also, the resolution of videos is usually lower than that of single snapshots. These factors tend to decrease the reconstruction accuracy compared to static objects. To a certain extent, this can be compensated by utilising the high redundancy of the videos.

Main *application domains* of 4D studios are computer games, interactive media, film production, and motion analysis in different areas. TV production of sports events also uses similar techniques.

Computer games and interactive media may need dynamic 3D models of real objects and actors; motion transfer from actor to model, including character animation; motion transfer from model to actor, including motion learning in sports, dancing, etc.; human-computer interaction, such as gesture and activity recognition.

Film production may involve content creation for movies; virtualised reality, including virtual characters in real scenes and real actors in virtual scenes; character multiplication and animation, for example, for crowd or battlefield scenes. Human motion analysis and understanding may aim at treatment of motion disorders, human motion recognition, and person identification by motion, e.g, gait. Animal motion analysis aims at studying the motion of animals for scientific and engineering purposes.

The rest of this paper is structured as follows. In Section 2, we give a brief overview of advanced 4D studios in Europe and the USA. The main contribution of this paper is the description of the hardware and software components of our studio given in Section 3. Our system contains a number of novel solutions which are discussed in more detail. The experience we gained to date, including the current problems and their possible remedies, is discussed in Section 4, where our future plans are also presented.

2. Previous Work

Building a 4D studio is a relatively expensive high-tech project. A number of studios exist in some developed countries of Western Europe and in the United States. The studios share main operation principles but differ in technical solutions. In this section, we briefly discuss a few advanced 4D studios focusing on the common principles and the specifics. Our goal is to place the development of our studio in the context of the recent trends in multiview dynamic scene reconstruction from videos. To save space, in this section we will refer the reader to the web pages of the projects discussed, where relevant publications, demos and applications can be found. Additional information on novel methods and applications can be found in the proceedings of the recent ICCV workshop [7].

To help the reader understand the operation of a typical 4D studio, let us first briefly discuss some standard steps of multiview reconstruction from videos. Most systems mentioned in this paper use the *visual hull* [11] as the initial volumetric 3D model. Images are first segmented into object and background. This critical operation is facilitated by ensuring good imaging conditions and using a specific, uniform background colour in the studio. In Section 3, we will give examples of segmentation.

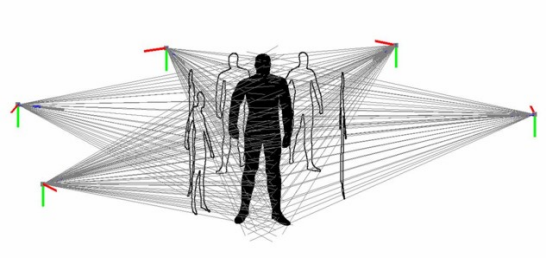

Figure 3: *Extracting the visual hull from silhouettes.*

Object silhouettes obtained by the cameras are then back-projected to 3D space as the generalised cones whose intersection gives the visual hull, a bounding geometry of the actual 3D object. Using more cameras results in a finer volumetric model, but some concave details may be lost anyway. The process of obtaining the visual hull from silhouettes is illustrated in Figure 3.

The volumetric visual hull is usually transformed into a surface mesh which is textured by selecting the most appropriate view for each unit of the mesh based on visibility, or by combining several views. The mesh is typically obtained from the hull using the standard Marching Cubes algorithm [12], while texturing techniques show greater variety. As the local geometry of the visual hull may significantly differ from the true one, various methods are used to enhance the shape. A critical issue is handling the possible topological changes of the non-rigid shape, such as the hands touching the body.

Our project is related to similar projects at INRIA [8, 14], MIT [5], MPI Informatik [15] and University of Surrey [2]. An essential difference between our approach and INRIA, MIT and MPII is that they go beyond independent frame-by-frame modelling we currently use. They take advantage of the continuity of motion and exploit the high redundancy of video sequences. Working in the *spatio-temporal* domain results in better geometry and texturing, but needs much more computing power.

Figure 4 demonstrates sample input images and a textureless 3D model reconstructed from video streams in the 4D studio created by the Computer Graphics Group at the Massachusetts University of Technology, USA. Figure 5 shows in more detail another high-quality result obtained by the Group. The dynamic reconstruction process is initialised by a high-quality static 3D model obtained by a laser scanner. Articulated mesh animation from multiview silhouettes [5] is

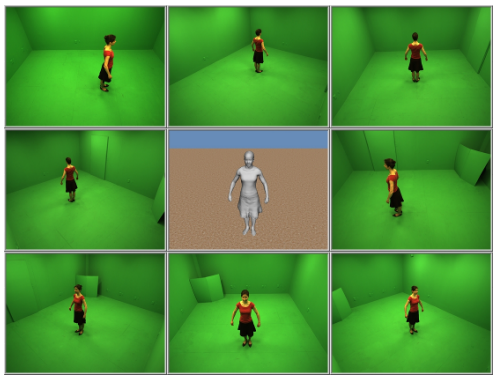

Figure 4: *MIT: Sample input images and reconstructed textureless 3D model.*

Figure 5: *Example of reconstruction at MIT.*

achieved using a simplified *skeleton model* of human body. Manual correction is applied to improve the mesh. The skeleton model facilitates motion transfer from human to model and animation of 3D human models.

Better capture of the local geometry can be achieved by a different 3D reconstruction approach called *photometric stereo* [6] which yields smoother surfaces and finer details. The Computer Graphics Group created another studio working on this principle. This studio is a large semi-sphere with numerous programmable lights. It can be used for shape and motion capture of people and their clothes. However, photometric stereo is less robust than multiview reconstruction techniques, and it is prone to ambiguities, which may lead to global errors in geometry. A good solution could be combining the two approaches [18].

Similarly to the Computer Graphics Group of MIT, the Graphics, Vision and Video Group [15] of the Max Planck Institut Informatik, Germany, also uses a laser-scanned shape to initialise the dynamic reconstruction process. However, the approach [15] does not apply skeleton model of human body. Instead, feature points detected in surface texture are

Figure 6: *MPI Informatik: Example of reconstructed 3D model in motion.*

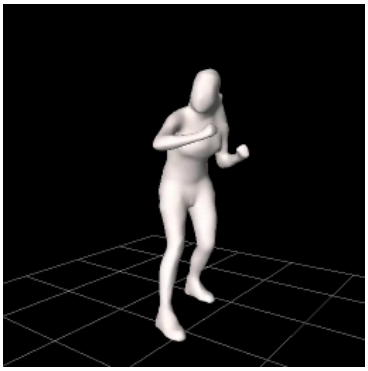

Figure 7: *Example of reconstruction at University of Surrey.*

used to support handling the shape deformations. Similar to INRIA, *photo-consistency* [10] is used for fine tuning of the result. Figure 6 gives an example of 3D model of a man in motion reconstructed at the MPI Informatik.

The Centre for Vision, Speech and Signal Processing at the University of Surrey, UK, has developed an advanced 4D studio [2] that combines multiview silhouettes with *shape-from-shading* [22]. The initial 3D model obtained from the silhouettes is enhanced using shape-from-shading. A skeleton model of human body is applied. Figure 7 shows a textureless 3D model reconstructed from video streams. The software developed at the Centre provides *free-viewpoint video* [1], that is, allows one to view the recorded event from any viewpoint during video capture. The main application areas are visual content production, computer games and interactive media, and sports TV production.

The Institute of Computer Science (FORTH, Crete, Greece) has created a smaller studio [4] for smaller articulated objects such as human hands. The cameras and lights are set around a table. Otherwise, the algorithmic principles are similar to those adopted by SZTAKI. All processing steps are implemented on a GPU, which provides real-time operation for relatively slow hand motion. The project primarily aims at markerless hand pose recovery in 3D for applications such as human-computer interaction and virtualised reality. For precise solution, a 26-DOF model of human hand is used, including the five fingers.

3. 4D Reconstruction Studio at MTA SZTAKI

A 4D studio is an advanced, intelligent sensory environment operated by sophisticated programming tools. This environment can be used for computer vision research as well as for technological development in a variety of applications, as discussed in the previous Section. To our best knowledge, the 4D Reconstruction Studio at MTA SZTAKI is a pioneering project in Central and Eastern Europe. The main motivation for building the Studio was the desire to bring advanced knowledge and technology to this region in order to facilitate testing new ideas and developing new methods, tools and applications.

In this section, we discuss the main hardware and software elements of the 4D Reconstruction Studio [17] being developed at MTA SZTAKI by the authors of this paper. It should be mentioned that two former collaborators, Bálint Fodor and Attila Egri, have also contributed to the project.

Most of the components of our studio are existing solutions which will be presented below very briefly; more attention will be paid to a few novel solutions we designed and applied in the Studio.

3.1. Hardware

The 4D Reconstruction Studio is a 'green box': green curtains and carpet provide homogeneous background. The massive, firm steel frame is a cylinder with dodecagon base. The size of the frame is limited by the size of the room. The diameter is around five meters; originally, a seven-meter studio was planned. The frame carries 12 video cameras placed uniformly around the scene and one additional camera on the top in the middle. (See Figures 1 and 2.)

The cameras are equipped with wide-angle lenses to cope with relatively close views; this necessitates careful calibration against radial distortion. The resolution of the cameras is 1624×1236 pixels; they operate at 25 fps and use GigE (Gigabit Ethernet).

Special, innovative lighting has been designed for the Studio to achieve better illumination. Apart from the standard diffuse light sources, we use light-emitting diodes (LEDs) placed around each camera, as illustrated in Figure 8. The LEDs can be turned on and off with high frequency. A micro-controller synchronises the cameras and the LEDs: when a camera takes a picture, the LEDS opposite to the camera are turned off. This solution improves illumination and allows for more flexible configuration of the cameras. The Studio uses seven conventional PCs; each of them but one handles two cameras.

3.2. Software

The Studio has two main software blocks: the image acquisition software for video recording and the 3D reconstruction software for creation of dynamic 3D models. The software system includes elements from the OpenCV [19]; otherwise, the entire system has been developed at SZTAKI.

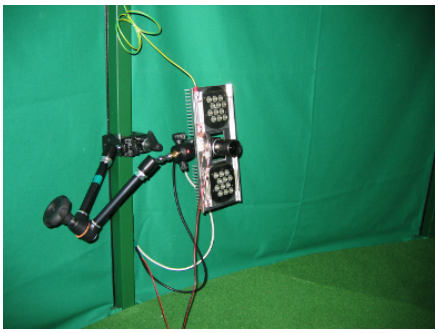

Figure 8: *Adjustable platform with a video camera and LEDs mounted on the frame.*

The image acquisition software configures and calibrates the cameras and selects a subset of the cameras for video recording. The easy-to-use, robust and efficient Z. Zhang's method [23] implemented based on OpenCV routines is used for intrinsic and extrinsic camera calibration and calculation of the parameters of radial distortion. During calibration, the operator repeatedly shows a flat chessboard pattern to the cameras. The complete procedure takes a few minutes; it is normally applied prior to every new acquisition.

The main steps of the 3D reconstruction process are as follows:

1. Extract *colour images* from the raw data captured.
2. Segment each colour image to *foreground and background*.
3. Create *volumetric model* using the Visual Hull algorithm.
4. Create *triangulated mesh* from the volumetric model using the Marching Cube algorithm.
5. Add *texture* to the triangulated mesh.

Similarly to the other studios mentioned in this paper, we use a shape-from-silhouettes technique to obtain a volumetric model of the dynamic shape. Currently, video frames are processed separately, i.e., the dynamic model obtained is a sequence of separate, instantaneous shapes. Since the visual hull is sensitive to errors in the silhouettes, segmenting input images into foreground and background is a critical step. Figure 9 shows sample input images acquired in our Studio. The binary segmented images are demonstrated in Figure 10.

Our image segmentation procedure is a novel method developed at SZTAKI for this project. The method assumes that the background is larger than the object, which is normally the case since the object needs room to move in the scene. The principles of segmentation are listed below.

- Acquire a *reference background image* in the absence of any object.

Figure 9: *Sample input images of the Studio.*

Figure 10: *Segmentation of the images shown in Figure 9.*

- Convert the input RGB image to the *spherical colour representation*.
- Calculate the absolute *difference* between the input image and the reference background image.
- In the difference image, select object pixels as outliers using robust *outlier detection*.
- Clean the resulting object image using *morphologic operations* such as erosion and dilation by disc.

Our recent study on optical flow estimation [13] has demonstrated that spherical colour representation improves robustness to illumination changes. Given the R,G,B values of a pixel, the spherical colour coordinates are defined as

$$\begin{aligned} \rho &= \sqrt{R^2+G^2+B^2} \\ \theta &= \arctan\left(\frac{G}{R}\right) \\ \phi &= \arcsin\left(\frac{\sqrt{R^2+G^2}}{\sqrt{R^2+G^2+B^2}}\right) \end{aligned} \tag{1}$$

The angles θ,ϕ are photometric invariants of the dichromatic reflection model [21]. They are less sensitive to illumination changes, shadow and shading. Although ρ is not an invariant, we still use ρ to account for meaningful differences in brightness. We normalise the three spherical coordinates and calculate the difference image using a smaller weight for ρ than for the two angles.

The difference image $I_D(x,y)$ is thresholded to detect object pixels as large-value outliers. To set the threshold, we first calculate the median of the difference image, μ_D. The median of the positive deviations from μ_D, that is, the median of $I_D(x,y)-\mu_D$ for $I_D(x,y)>\mu_D$ evaluates the inlier variation in the difference image due to noise. This latter median serves as the basis of robust outlier detection as described in the standard textbook [20].

The algorithm for texturing the triangulated surface [9] calculates a measure of visibility for each triangle and each camera. The triangle should be visible from the camera and its normal vector should point towards camera. Then, a cost function is formed with visibility and regularisation terms to balance between visibility of triangle and smoothness of texture. The regularisation term reduces sharp texture edges between adjacent triangles. The cost function is minimised using graph cuts.

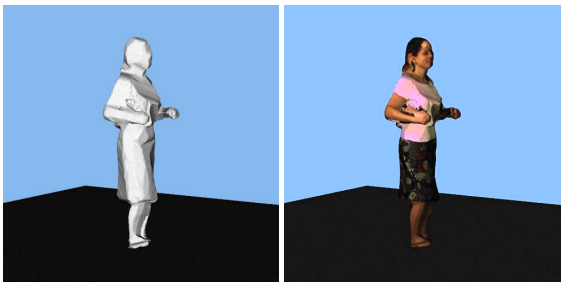

Figure 11: *Examples of textureless and textured models.*

Figure 12: *Another example of textured model.*

Figure 11 shows examples of textureless and textured models. Another example of textured model is demonstrated in Figure 12. Figure 13 illustrates our system's capability to create mixed reality. Three different models were multiplied at varying phases of motion and placed into a virtual environment.

4. Discussion and conclusions

We have presented the main current features of the 4D Reconstruction Studio being developed at the MTA SZTAKI.

Figure 13: *Mixed reality: Three different dynamic models multiplied and placed into a virtual environment.*

As far as the Studio's hardware is concerned, we are basically satisfied with its operation and performance. Recently, a modern graphics card has been added to the system to achieve real-time performance. Otherwise, we believe that the current configuration is appropriate.

A few remarks are still to be made in relation to the configuration. First of all, the size of the Studio is not sufficient for a tall person or a group of persons to move freely in the scene. This is not a principal limitation, but it is inconvenient from practical point of view. Second, our lighting solution with programmable LEDs is quite efficient, but it may be unpleasant for human eyes because of the vibrating illumination.

Concerning the reconstruction software, the quality of texturing depends on the precision of surface geometry which is not perfect. The Visual Hull and the Marching Cube algorithms may yield imprecise surface normals, which may in turn deteriorate the calculation of visibility and lead to incorrect texture mapping. Along with some concave shape details, texture details may be lost or distorted. In addition, the frame-by-frame processing may lead to quick small-size temporal variations in texture called texture flickering, which are minor but still perceived by human eye.

We are now working on improving the quality of the model. This includes better segmentation as well as better shape and texturing, in particular, by utilising the spatio-temporal coherence. As a part of this plan, we are developing a program for interactive correction of the triangulated mesh, which will result in better shape and consistent handling of topological changes. Such programs are used by other studios as well, e.g., at the MIT.

We have already implemented all phases of the reconstruction process on a modern GPU and achieved real-time performance and free-viewpoint video. The dynamic model can be built and viewed from any viewpoint in real time, without transmitting and storing the video data. (The offline version of the software needs about 30 minutes to process a ten-second video.) For efficient GPU implementation, some steps of processing, including segmentation and texturing, had to be simplified. Fortunately, the quality of the model is still acceptable. Work in this direction will be continued, and the quality will be improved. The GPU implementation of the system will be presented in a forthcoming paper.

In future, we plan to address applications beyond human motion. In particular, it will be interesting to help physicists in spatio-temporal modelling of natural processes, such as fire, water, gases, or vegetation in the wind. This would need segmentation of *dynamic texture* [3], a topic we have recently worked on and gained significant experience in.

It is also planned to connect our Studio to the Virtual Collaboration Arena (VirCA) [16] developed by another unit of MTA SZTAKI led by Péter Baranyi. VirCA is situated in a neighbouring room. It is a 3-wall real-time virtual environment which allows one to act in a virtual world and add real-world models to a virtual world to create mixed reality. We plan to transmit models from 4D studio to VirCA and build them into virtual worlds. This will allow, for example, a dancer to move around his/her own 3D model in motion and watch it. Finally, we were invited to participate in the planned consortium of European 4D studios that includes leading West-European research centres and media companies. We hope that, due to the Studio, our part of Europe will also be represented in the consortium.

Acknowledgments

This work was supported by the NKTH-OTKA grant CK 78409, by the European Union and the European Social Fund under the grant agreement TÁMOP 4.2.1./B-09/KMR-2010-0003, and by the HUNOROB project (HU0045, 0045/NA/2006-2/ÖP-9), a grant from Iceland, Liechtenstein and Norway through the EEA Financial Mechanism and the Hungarian National Development Agency. The authors acknowledge the valuable contribution of Bálint Fodor to the development of the image acquisition software.

References

1. J. Carranza, C. Theobalt, M.A. Magnor, and H.P. Seidel. Free-viewpoint video of human actors. *ACM Transactions on Graphics*, 22:569–577, 2003.

2. Centre for Vision, Speech and Signal Processing, University of Surrey. SurfCap: Surface Motion Capture. `http://kahlan.eps.surrey.ac.uk/Personal/AdrianHilton/Research.html`, 2008.

3. D. Chetverikov and R. Péteri. A brief survey of dynamic texture description and recognition. In *Proc. International Conference on Computer Recognition Systems*, pages 17–26. Springer Advances in Soft Computing, 2005.
4. FORTH Institute of Computer Science. From multiple views to textured 3D meshes: a GPU-powered approach. `www.ics.forth.gr/~argyros/research/gpu3Drec.htm`, 2010.
5. MIT Computer Graphics Group. Dynamic Shape Capture and Articulated Shape Animation. `people.csail.mit.edu/drdaniel/`, 2011.
6. S. Herbort and C. Wöhler. An introduction to image-based 3D surface reconstruction and a survey of photometric stereo methods. *3D Research*, 2:1–18, 2011.
7. S. Ilic, E. Boyer, and A. Hilton, editors. *ICCV 2011 Workshop on Dynamic Shape Capture and Analysis*. IEEE, 2011.
8. INRIA Rhône-Alpes. The Grid and Image Initiative. `grimage.inrialpes.fr/`, 2012.
9. Z. Janko and J.-P. Pons. Spatio-temporal image-based texture atlases for dynamic 3-D models. In *Proc. ICCV Workshop 3DIM'09*, pages 1646–1653, 2009.
10. K.N. Kutulakos and S.M. Seitz. A theory of shape by space carving. In *Proc. International Conference on Computer Vision*, volume 1, pages 307–314, 1999.
11. A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16:150–162, 1994.
12. W.E. Lorensen and H.E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *Proc. ACM SIGGRAPH*, volume 21, pages 163–169, 1987.
13. J. Molnár, D. Chetverikov, and S. Fazekas. Illumination-robust variational optical flow using cross-correlation. *Computer Vision and Image Understanding*, 114:1104–1114, 2010.
14. Morpheo Team. Capture and Analysis of Shapes in Motion. `morpheo.inrialpes.fr/`, 2012.
15. MPI Informatik, Graphics, Vision and Video Group. Dynamic Scene Reconstruction. `www.mpi-inf.mpg.de/~theobalt/`, 2012.
16. MTA SZTAKI Cognitive Informatics Research Group. The Virtual Collaboration Arena. `www.virca.hu/`, 2012.
17. MTA SZTAKI Geometric Modelling and Computer Vision Research Lab. The 4D Reconstruction Studio. `vision.sztaki.hu/4Dstudio/index.php`, 2011.
18. D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics*, 24:536–543, 2005.
19. OpenCV. Open Computer Vision Library. `sourceforge.net/projects/opencvlibrary/`, 2012.
20. P.J. Rousseeuw and A.M. Leroy. *Robust regression and outlier detection*. Wiley, 1987.
21. S.A. Shafer. Using color to separate reflection components. *Color Research and Applications*, 10:210–218, 1985.
22. R. Zhang, P.S. Tsai, J.E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21:690–706, 1999.
23. Z. Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1330–1334, 2000.