3D-TV – THE FUTURE OF VISUAL ENTERTAINMENT

M. MAGNOR
MPI Informatik
Stuhlsatzenhausweg 85
Saarbrücken, Germany
e-mail: magnor@mpi-sb.mpg.de

Television is the most favorite pastime activity of the world. Remarkably, it has so far ignored the digital revolution; the way we watch television hasn't changed since its invention 75 years ago. But the time of passive TV consumption may be over soon: Advances in video acquisition technology, novel image analysis algorithms, and the pace of progress in computer graphics hardware together drive the development of a new type of visual entertainment medium. The scientific and technological obstacles towards realizing 3D-TV, the experience of interactively watching real-world dynamic scenes from arbitrary perspective, are currently being put out of the way by researchers all over the world.

1. Introduction

According to a recent study[1], the average US American citizen watches television 4 hours and 20 minutes every day. While watching TV is the most favorite leisure activity in the world, it is interesting to note that TV technology has shown remarkable resistance against any change. Television sets today still have the computational capacity of a light switch, while modern PC and graphics cards work away at giga-flop rates to entertain the youth with the latest computer games.

The idea of making more out of television is not new. Fifty years ago, Ralph Baer began to think about how to add interactivity to television. He invented the video game and developed the first game console, thus becoming the founding father of the electronic entertainment industry, a business segment whose worldwide economic impact has by now even surpassed the movie industry[2]. Driven by technological progress, economic competition and an ever-growing number of users, computer games have become more and more realistic over the years, while TV has remained the passive medium of its first days.

In recent times, however, scientists from different fields have joined

forces to tear down the wall between interactive virtual worlds and the real world[3,4,5]. Their common goal is to give the consumer the freedom to watch natural, time-varying scenes from any arbitrary perspective: A soccer match can be watched from the referee's, goal keeper's or even the ball's point of view, a crime story might be experienced from the villain's or the victim's perspective, and while watching a movie, the viewer is seated in the director's chair. This paper intends to give an overview of current research in 3D-TV acquisition, coding, and display.

2. 3D-TV Content Creation

To display an object from arbitrary perspective, its three-dimensional shape must be known. For static objects, 3D geometry can be acquired, e.g., by using commercial laser scanners. But how can the constantly changing shape of dynamic events be captured ?

Currently, optically recording the scene from different viewpoints is the only financially and logistically feasible way of acquiring dynamic 3D geometry information, albeit implicitly[6]. The scene is captured with a handful of synchronized video cameras. To recover time-varying scene geometry from such multi-video footage, relative camera recording positions must be known with high accuracy[7].

The visual hull has been frequently used as geometry proxy for time-critical reconstruction applications. It can be computed efficiently and represents an approximate, conservative model of object geometry[8]. An object's visual hull is reconstructed by segmenting object outlines in different views and re-projecting these silhouettes as 3D cones back into the scene. The intersection volume of all silhouette cones encompasses the true geometry of the object. Today, the complete processing pipeline for on-line 3D-TV broadcast applications can be implemented based on the visual hull approach[9].

Unfortunately, attainable image quality is limited due to the visual hull's approximate nature. Allowing for off-line processing during geometry reconstruction, refined shape descriptions can be obtained by taking local photo-consistency into account. Space carving[10] and voxel coloring[11] methods divide the volume of the scene into small elements (voxels). Consequently, each voxel is tested whether its projection into all unoccluded camera views corresponds to roughly the same color. If a voxel's color is different when viewed from different cameras, it is deleted. Iteratively, a photo-consistent hull of the object surface is "carved" out of the scene

volume.

2.1. *Spacetime Isosurfaces*

Both the visual hull approach as well as the space carving/voxel coloring methods have been developed with static scenes in mind. When applied to multi-video footage, these techniques reconstruct object geometry one time step after the other, making no use of the inherently continuous temporal evolution of any natural event. Viewed as an animated sequence, the resulting scene geometry potentially exhibits discontinuous jumps and jerky motion. A completely new class of reconstruction algorithms is needed to exploit temporal coherence in order to attain robust reconstruction results at excellent quality.

When regarded in 4D spacetime, dynamic object surfaces represent smooth 3D hyper-surfaces. Given multi-video data, the goal is to find a smooth 3D hyper-surface that is photo-consistent with all images recorded from all cameras over the entire time span of the sequence. This approach can be elegantly formulated as a minimization problem whose solution is a minimal 3D hyper-surface in 4D spacetime. The weight function incorporates a measure of photo-consistency, while temporal smoothness is ensured because the sought-after minimal hyper-surface minimizes the integral of the weight function. The intersection of the minimal 3D hyper-surface with a 3D hyper-plane perpendicular to the temporal dimension then corresponds to the 2D object surface at a fixed point in time.

The algorithmic problem remains how to actually find the minimal hyper-surface. Fortunately, it can be shown[12] that a k-dimensional surface which minimizes a rather general type of functional is the solution of an Euler-Lagrange equation. In this form, the problem becomes amenable to numerical solution. A surface evolution approach, implemented based on level sets, allows one to find the minimal hyper-surface[13]. In comparison to conventional photo-consistency methods that do not take temporal coherence into account, this spacetime-isosurface reconstruction technique yields considerably better geometry results. In addition, scene regions which are temporarily not visible from any camera are automatically interpolated from previous and future time steps.

2.2. *Model-based Scene Analysis*

A different approach to dynamic geometry recovery can be pursued by exploiting a-priori knowledge about the scene's content[14]. Given a parame-

terized geometry model of the object in an scene, the model can be matched to the video images. For automatic and robust fitting of the model to the images, object silhouette information is used. As matching criterion, the overlapping area of the rendered model and the segmented object silhouettes is employed[15]. The overlap is efficiently computed by rendering the model for all camera viewpoints and performing an exclusive-or (XOR) operation between the rendered model and the segmented images. The task of finding the best model parameter values thus becomes an optimization problem that can be tackled, e.g., by Powell's optimization scheme. In an analysis-by-synthesis loop[16], all model parameters are varied until the rendered model optimally matches the recorded object silhouettes.

Making use of image silhouettes to compare model pose to object appearance has numerous advantages:

- Silhouettes can be easily and robustly extracted,
- they provide a large number of pixels, effectively over-determining the model parameter search,
- silhouettes of the geometry model can be rendered very efficiently on modern graphics hardware, and
- also the XOR operation can be performed on graphics hardware.

Model-based analysis can additionally be parallelized to accelerate convergence[17]. In addition to silhouettes, texture information can be exploited to also capture small movements[18]. One major advantage of model-based analysis is the comparatively low dimensionality of the parameter search space: only a few dozen degrees of freedom need to be optimized. In addition, constraints are easily enforced by making sure that during optimization, all parameter values stay within their physically plausible range. Finally, temporal coherence is maintained by allowing only a maximal change in magnitude for each parameter from one time step to the next.

3. Compression

Multi-video recordings constitute a huge amount of raw image data. By applying standard video compression techniques to each stream individually, the high degree of redundancy among the streams is not exploited. However, the 3D geometry model can be used to relate video images recorded from different viewpoints, offering the opportunity to exploit inter-video correlation for compression purposes. Model-based video coding schemes

have been investigated for single-stream video data, and MPEG4[19] provides suitable techniques to encode the animated geometry, e.g. by updating model parameter values using differential coding. For 3D-TV, however, multiple synchronized video streams depicting the same scene from different viewpoints must be encoded, calling for new coding algorithms to compress multi-video content.

To efficiently encode the multi-video data using object geometry, the images may be regarded as object textures. In the texture domain, a point on the object surface has fixed coordinates, and its color (texture) varies only due to illumination changes and/or non-Lambertian reflectance characteristics. For model-based coding, a texture parameterization is first constructed for the geometry model[20]. Having transformed all multi-video frames to textures, the multi-view textures are then processed to de-correlate them with respect to temporal evolution as well as viewing direction[21]. Shape-adaptive[22] as well as multi-dimensional wavelet coding schemes[20]lend themselves to efficient, progressive compression of texture information. Temporarily invisible texture regions can be interpolated from previous and/or future textures, and generic texture information can be used to fill in regions that have not been recorded at all. This way, any object region can later be displayed without holes in the texture due to missing input image data.

For spacetime-isosurface reconstruction, deriving one common texture parameterization for all time instants is not trivial since the reconstruction algorithm does not provide surface correspondences over time. Encoding the time-varying geometry is also more complex than in the case of model-based analysis. Current research therefore focuses on additionally retrieving correspondence information during isosurface reconstruction.

4. Interactive Display

The third component of any 3D-TV system consists of the viewing hard- and software. In a 3D-TV set, one key role will be played by the graphics board: It enables displaying complex geometry objects made up of hundreds of thousands of polygons from arbitrary perspective at interactive frame rates.

The bottleneck of current, PC-based prototypes constitutes the limited bandwidth between storage drive and main memory, and, to a lesser extent, between main memory and the graphics card. Object geometry as well as texture must be updated on the graphics board at 25 frames per

second. While geometry animation data is negligible, the throughput capacity of today's PC bus systems is sufficient only for transferring low- to moderate-resolution texture information. Since any user naturally wants to exploit the freedom of 3D-TV to zoom into the scene and to observe object details from close-up, object texture must be updated continuously at high resolution to offer the ultimate viewing experience.

To overcome the bandwidth bottleneck, object texture must be stored in some form that is at the same time efficient to transfer as well as fast to render on the graphics board. In addition, rendered image quality shall not be degraded, preserving the realistic, natural impression of the original multi-video images. These requirements can be met by decomposing object texture into its constituents: local diffuse color and reflectance, and shadow effects. Since typically neither object color nor reflectance change over time (only exceptions: chameleons and sepiae), diffuse color texture and reflectance characteristics need to be transferred only once. Given this static texture description, the graphics board is capable of computing very efficiently object appearance for any illumination and viewing perspective. The remaining difference between rendered and recorded object appearance is due to small-scale, un-modeled geometry variations, e.g. clothing creases. Only these time-dependent texture variations need to be updated per frame, either as image information, or, more elegantly, as dynamic geometry displacement maps on the object's surface. One beneficial side-effect of representing object texture in this form is the ability to vary object illumination: The object can be placed into arbitrary environments while retaining its natural appearance.

To represent object texture in the above-described way, new analysis algorithms need to be developed. These must be capable of recovering reflectance characteristics as well as surface normal orientation from multi-video footage. While research along these lines has only just begun, first results are encouraging, and multi-video textures have already been robustly decomposed into diffuse and specular texture components.

5. Outlook

So will 3D-TV supersede conventional TV anytime soon ? The honest answer is: probably not this year. The TV market exhibits enormous momentum and has already defied a number of previous attempts at technological advances, e.g. HDTV and digital broadcast.

The new possibilities interactive 3D-TV offers to the user, however, are

too attractive to be ignored for long. In a few years, 3D-TV will start as a new application for conventional PCs (much like the RealPlayer© some years ago), probably with later adaptation to game consoles, which are then already hooked up to the TV set situated in the living room. The pace of progress will depend on the effort required to create attractive content. Nevertheless, from today's state-of-the-art one can be optimistic that the scientific and technological challenges of 3D-TV will be surmountable: A brave, new, and interactive visual entertainment world lies ahead.

References

1. P. Lyman and H. Varian. How much information, 2003. University of California Berkeley, http://www.sims.berkeley.edu/research/projects/how-much-info-2003/.
2. J. Gaudiosi. Games, movies tie the knot, December 2003. http://www.wired.com/news/games/0,2101,61358,00.html.
3. M. Op de Beeck and A. Redert. Three dimensional video for the home. *Proc. EUROIMAGE International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging (ICAV3D'01), Mykonos, Greece*, pages 188–191, May 2001.
4. C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, E. Ofek L. Van Gool, and Sexton I. An evolutionary and optimised approach on 3D-TV. *Proc. International Broadcast Conference (IBC'02)*, pages 357–365, September 2002.
5. K. Klein, W. Cornelius, T. Wiebesiek, and J. Wingbermühle. Creating a "personalised, immersive sports tv experience" via 3D reconstruction of moving athletes. In Abramowitz and Witold, editors, *Proc. Business Information Systems*, 2002.
6. L. Ahrenberg, I. Ihrke, and M. Magnor. A mobile system for multi-video recording. In *1st European Conference on Visual Media Production (CVMP)*, pages 127–132. IEE, 2004.
7. I. Ihrke, L. Ahrenberg, and M. Magnor. External camera calibration for synchronized multi-video systems. *Journal of WSCG*, 12(1-3):537–544, January 2004.
8. A. Laurentini. The visual hull concept for sihouette-based image understanding. *IEEE Trans. Pattern Analysis and Machine Vision*, 16(2):150–162, February 1994.
9. M. Magnor and H.-P. Seidel. Capturing the shape of a dynamic world - fast ! *Proc. International Conference on Shape Modelling and Applications (SMI'03), Seoul, South Korea*, pages 3–9, May 2003.
10. K.N. Kutulakos and S.M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.
11. S.M. Seitz and C.R. Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999.
12. B. Goldluecke and M. Magnor. Weighted minimal hypersurfaces and their

applications in computer vision. *Proc. European Conference on Computer Vision (ECCV'04), Prague, Czech Republic*, 2:366–378, May 2004.
13. B. Goldluecke and M. Magnor. Space-time isosurface evolution for temporally coherent 3D reconstruction. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04), Washington, USA*, June 2004. to appear.
14. J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *ACM Trans. Computer Graphics (Siggraph'03)*, 22(3):569–577, July 2003.
15. M. Magnor and C. Theobalt. Model-based analysis of multi-video data. *Proc. IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI-2004), Lake Tahoe, USA*, pages 41–45, March 2004.
16. R. Koch. Dynamic 3D scene analysis through synthesis feedback control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):346–351, July 1993.
17. C. Theobalt, J. Carranza, M. Magnor, and H.-P. Seidel. A parallel framework for silhouette-based human motion capture. *Proc. Vision, Modeling, and Visualization (VMV-2003), Munich, Germany*, pages 207–214, November 2003.
18. C. Theobalt, J. Carranza, M. Magnor, J. Lang, and H.-P. Seidel. Enhancing silhouette-based human motion capture with 3D motion fields. *Proc. IEEE Pacific Graphics 2003, Canmore, Canada*, pages 185–193, October 2003.
19. Motion Picture Experts Group (MPEG). N1666: SNHC systems verification model 4.0, April 1997.
20. M. Magnor, P. Ramanathan, and B. Girod. Multi-view coding for image-based rendering using 3-D scene geometry. *IEEE Trans. Circuits and Systems for Video Technology*, 13(11):1092–1106, November 2003.
21. G. Ziegler, H. Lensch, N. Ahmed, M. Magnor, and H.-P. Seidel. Multi-video compression in texture space. *Proc. IEEE International Conference on Image Processing (ICIP'04), Singapore*, September 2004. accepted.
22. H. Danyali and A. Mertins. Fully scalable texture coding of arbitrarily shaped video objects. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, pages 393–396, April 2003.