3D Video – Being Part of the Movie

Christian Theobalt, Joel Carranza, Marcus A. Magnor, Hans-Peter Seidel
MPI Informatik

In recent years, a convergence between the fields of computer graphics and computer vision has been observed [1]. This convergence has been motivated by the idea to create photorealistic visualizations of real-world scenes in a computer not by designing models of shape and appearance, but by reconstructing these models from photographic or video data of the real world.

From the consumer's point of view, up to now video has been a two-dimensional medium. The viewpoint onto a motion picture or a TV broadcast coincides with the viewpoint of the recording image sensor, which cannot be altered by the viewer. Recent technological advancements, such as the advent of novel high-resolution video cameras with high dynamic range, improvements in data storage and transmission technology, as well as the presence of specialized graphics hardware even in consumer-grade electronics, make it seam feasible to lift the medium video onto a novel, an immersive level. The goal of free viewpoint video is to enable the choice of an arbitrary viewpoint onto a dynamic scene, thereby creating a feeling of immersion into the event. Interactive free-viewpoint video and 3D TV will spawn a multitude of novel applications in visual media. To mention only a few, interactive motion pictures will become feasible, and not only to the viewer during a sports broadcast, but also to the coach during his analysis of an athlete's performance. All will profit from the possibility to look at the event from a novel perspective.

Researchers that aim at setting the path for this new technology are facing a sea of algorithmic and technological challenges, the majority of which are still to be solved.

One of the key issues involved is the acquisition, which typically involves recording a multitude of synchronized video streams, making necessary an advanced multi-camera recording environment that can process the large amount of image data. Furthermore, the algorithmic core of 3D video is formed by methods for the reconstruction of shape and appearance models of real-world scenes from video data and algorithms for rendering them in real time. Finally, efficient encoding of the immersive video is a precondition for real-time broadcasting and on-disc delivery.

In the Computer Graphics Group and the Graphics-Optics-Vision Group at the Max-Planck-Institut für Informatik (**http://www.mpi-sb.mpg.de**) in Saarbrücken, Germany, we investigate the algorithmic ingredients of free viewpoint video.

Human actors are the central elements of motion picture scenes and the human visual system is highly sensitive to even slight inaccuracies in a human's motion and look. In consequence, the synthesis of realistic images of humans in motion is a challenging problem. Two aspects of this problem are the creation of natural human motion and the accurate rendering of a person's physical appearance. We have developed a system that acquires both, the large-scale motion of the human body as well as its appearance down to the level of skin details and cloth movement [2]. We employ a marker-free optical motion capture approach from multiple video streams to estimate the parameters of motion of a kinematic body model. Time-varying multi-view surface textures are created from the video frames.

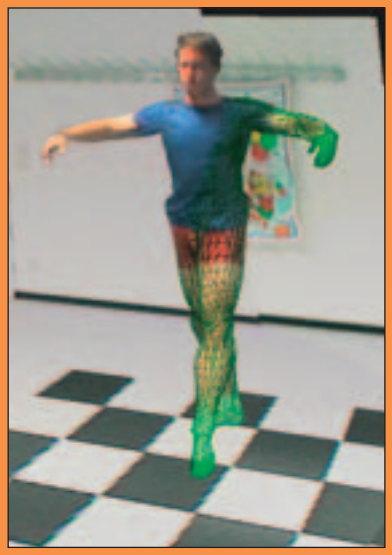

Figure 1: *Blend between the underlying triangle model of the human actor and the textured body in our free viewpoint video renderer.*

During rendering, the surface textures are applied to the moving body model, enabling an interactive choice of an arbitrary viewpoint (Figure 1).

Related Work

In computer vision, the estimation of motion parameters from video has been an intensive field of research for many years. Detailed reviews of work in this field can be found in [3, 7]. Pioneering work in the field of 3D video stems from the field of image-based rendering. Most approaches from that area are based on visual hull reconstruction [4, 5] or stereo reconstruction from a multitude of cameras [8]. At SIGGRAPH 2004, a system for creating virtual novel viewpoints of a scene by means of view interpolation is presented [12]. An approach for recording, broadcasting and stereoscopic display of 3DTV is also shown [6]. For a detailed review of previous work in the field we would like to refer the reader to the appropriate sections in the cited references.

Multi-View Video Recording

The video sequences used as inputs to our system are recorded in our multi-view video studio [11]. IEEE1394 cameras are placed in a convergent setup around the center of the scene. The video sequences used for this article are recorded from eight static viewing positions arranged at approximately equal angles and distances around the center of the room. The cameras are synchronized via an external trigger and all the video data are directly streamed to the hard drives of four control PCs, each of which is connected to two cameras. Video frames are recorded at resolutions of up to 640x480 pixels. The frame rate is fundamentally limited to 15 fps by the external trigger. The cameras' intrinsic and extrinsic parameters are determined, thereby calibrating every camera into a common global coordinate system. The lighting conditions are controlled and all cameras are color-calibrated. In each video frame, the person in the foreground is cut out via background subtraction based on per-pixel color statistics.

Model-Based Free Viewpoint Video Reconstruction and Rendering

In the beginning of each multi-view video sequence, the subject stands in an initialization pose. Using the silhouette-overlap with the projected body model in each camera view, an initial set of pose, scaling and deformation parameters is found by means of a multi-step optimization procedure. These scaling parameters deform our a priori body model such that it optimally conforms to the recorded person. The body model consists of a hierarchical arrangement of 16 segments, the surface geometry being defined by roughly 21,000 triangles. The model's kinematics is defined via an underlying skeleton made up of 17 joints that provide 35 pose parameters.

The criterion that guides our motion capture and initialization procedures is the overlap between the projected body model and the input silhouettes in each camera view. A quantitative measure for this overlap is the pixel-wise XOR between the projected model silhouette and the input image silhouette in each camera view (Figure

Figure 2: *Top row: XOR error function (left), human body model (middle), textured body (right). Bottom row: Three additional renderings of the same time step of free viewpoint video.*

2). The error metric used during optimization is the sum of the XOR values from each camera view. We exploit consumer-level graphics hardware to efficiently compute this error function.

The motion parameters of the body model are found by performing a non-linear optimization in the pose parameter space at each time step of video. We use a direction set method to numerically solve the optimization problem. The search for an optimal solution is improved by hierarchically solving optimization problems on subsets of the pose parameters. By this means, we prevent convergence to incorrect local minima of the overall optimization problem. In order to deal with fast body motion, a pre-selection step (grid-search) on the lower-dimensional parameter spaces of the limbs is performed.

We have demonstrated [10] that the energy function evaluation can be accelerated in two ways. First, in order to reduce the amount of data transferred between GPU and CPU, the energy function evaluation is only done on sub-windows of the image plane. Second, the rendering overhead during the XOR computation can be further reduced by only rendering those body parts whose pose parameters are currently optimized. The whole problem lends itself to a parallel implementation using five CPUs and GPUs, which significantly improves the performance of the motion capture algorithm.

During playback of the reconstructed free viewpoint video, we render the body model in the sequence of captured body poses. We apply projective texturing using the input camera images to create a realistic time-varying surface appearance of the rendered geometry (Figure 2). To combine the camera images from different viewpoints, the texture colors at each vertex are blended. In order to compute the blending weights, the visibility of each vertex in each camera view is considered. The actual spatial blending weights can be computed in a view-independent and/or view-dependent way. The view-independent weight for each camera is the reciprocal of the angle between the vertex normal and the camera-viewing vector. The view-dependent weight for each camera is the reciprocal of the angle between the input viewing direction and the output viewing direction. We introduce an additional rescaling factor of the view-independent weights that provides us with more control over the visual appearance.

In some frames of video, the model may not be perfectly aligned with the image silhouettes in each camera view. This can lead to disturbing visual artifacts in the textures. We use two methods, a modified visibility computation and a texture expansion, to prevent these artifacts. The modified visibility computation determines the visibility of a vertex from a set of slightly displaced camera views instead of only the actual input camera views. This way erroneous projections of foreground texture on occluded and more distant geometry are prevented. Second, the texture information at silhouette boundaries is expanded into the background by performing image dilation on the background-subtracted video frames.

Enhanced Reconstruction Using 3D Motion Fields

We have demonstrated that a broad range of complex and rapid body motion is robustly captured using silhouette-based techniques. However, improvements are possible in those portions of the body with small-scale details (such as features of the face) whose visual appearance is deteriorated even through small pose inaccuracies. To be maximally effective, we further developed our original silhouette-based tracking algorithm into a hybrid approach that incorporates texture information into the motion estimation process [9]. The enhanced motion estimation scheme follows a predictor-corrector approach. Looking at one time step of video, the parameters found by the silhouette fitting step form a first estimate of the body pose. Now, the model standing in the estimated pose is textured with the multi-view video frames of the previous time step and rendered into each camera view. From the optical flows computed between the actual

Figure 3: *Impressions from a freeze-and-rotate shot.*

video frames at the current time step and the rendered prediction images, a 3D motion field is reconstructed. This motion field contains 3D vectors for the majority of vertices of surface geometry that describe how their position should be updated. Using these vectors, pose update parameters are computed that improve the alignment of the model with the input data.

Results

Using the silhouette-based pose parameter estimation, even rapid motion, such as the motion of a ballet dancer, is captured robustly (Figure 2). Small-scale details, such as wrinkles in clothing, are preserved and lead to a highly realistic physical appearance of the 3D video from any viewpoint (Figure 3). Using a parallel implementation of the fitting system that runs on five CPUs and GPUs, we obtain an average pose parameter estimation time of around 1s per frame for rapid motion. Fitting times below 1s are achieved for slow body motion. The effects of the additional motion field step on 3D video quality are on a small scale but most noticeable in the face area whose accurate reconstruction is essential for a good visual impression.

Future Perspectives

The field of 3D video is a premier example for a research area that draws both from the development of new algorithms and the progress of hardware technology. The growing number of researchers working on the subject and the increasing number of research publications at major international graphics and vision conferences shows the popularity of the field. For our work, the ACM SIGGRAPH annual conference was the ideal event to present our results. There, not only did we have the chance to present our work to the research community but also to practitioners from the industry, and, in turn, obtained valuable feedback from both sides.

We expect that future research in 3D video and 3DTV will focus on the development of a fully automatic pipeline that incorporates the acquisition, the reconstruction and the rendering in real time. Advances in computational technology will allow us to push the algorithmic frontiers even further, thereby attacking novel problems such as the simultaneous capturing of geometric scene models and reflection models, enabling relighting of the 3D video footage.

References

1. ACM SIGGRAPH *Computer Graphics* quarterly. Vol. 33(4), Focus: Applications of Computer Vision to Computer Graphics, 1999.

2. Carranza, J., C. Theobalt, M. Magnor and H.-P. Seidel. Free-viewpoint video of human actors, in *Proceedings of SIGGRAPH 2003,* pp. 569-577, San Diego, USA, ACM, 2003.

3. Gavrila, D. The visual analysis of human movement, *CVIU,* 73(1):82-98, January 1999.

4. Matsuyama, T., T. Takai. Generation, visualization, and editing of 3D video, in *Proc. of 1st International Symposium on 3D Data Processing Visualization and Transmission* (3DPVT'02), p. 234 ff., 2002.

5. Matusik, W., C. Buehler, R. Raskar, S. Gortler and L. McMillan. Image-based visual hulls, in *Proceedings of ACM SIGGRAPH 2000,* pp. 369-374, 2000.

6. Matusik, W. and H. Pfister. 3D TV: A Scalable System for Real-Time Acquisition, Transmission, and Autostereoscopic Display of Dynamic Scenes, to appear in *Proc. of ACM SIGGRAPH 2004*

7. Moeslund, T. B. and E. Granum. A survey of computer vision-based human motion capture, *Computer Vision and Image Understanding,* 81(3):231-268, 2001.

8. Narayanan, P., P. Rander and T. Kanade. Constructing Virtual worlds using dense stereo, in *Proc. of ICCV 98,* pp. 3-10, 1998.

9. Theobalt, C., J. Carranza, M. A. Magnor and H.-P. Seidel. Enhancing silhouette-based human motion capture with 3D motion fields, in *11th Pacific Conference on Computer Graphics and Applications,* pp. 185-193, IEEE, Canmore, Canada, October 2003.

10. Theobalt, C., J. Carranza, M. Magnor and H.-P. Seidel. A parallel framework for silhouette-based human motion capture, in *Proc. of Vision, Modeling and Visualization 2003,* pp. 207-214, Munich, Germany, November 2003.

11. Theobalt, C., M. Li, M. Magnor and H.-P. Seidel. A flexible and versatile studio for synchronized multi-view video recording, in *Proceedings of Vision, Video and Graphics,* pages 9-16, 2003.

12. Zitnick, C. L., S.B. Kang, M. Uyttendaele, S. Winder and R. Szeliski. High Quality Video View Interpolation Using A Layered Representation, to appear in *Proc. of ACM SIGGRAPH 2004.*

About the Contributors

Christian Theobalt, Joel Carranza, Marcus A. Magnor and Hans-Peter Seidel work together at MPI Informatik in Saarbrücken, Germany. You can contact them all through lead author Theobalt.

Theobalt received his M.Sc. degree in artificial intelligence from the University of Edinburgh, Scotland, and his Diplom (M.S.) degree in computer science from the Saarland University, Saarbrücken, Germany. His research interests include motion analysis from video, 3D computer vision and image- and video-based rendering and reconstruction.

Christian Theobalt
Computer Graphics Group
MPI Informatik
Stuhlsatzenhausweg 85
66123 Saarbrücken
Germany
Tel: +49 681 9325 419
Fax: +49 681 9325 499
E-mail: **theobalt@mpi-sb.mpg.de**