Workshop Report 08w5070
Multi-View and Geometry Processing for 3D Cinematography

Rémi Ronfard (Xtranormal Technology and INRIA Rhone-Alpes),
Gabriel Taubin (Brown University)

July 14-18, 2008

By 3D cinematography we refer to techniques to generate 3D models of dynamic scenes from multiple cameras at video frame rates. Recent developments in computer vision and computer graphics, especially in such areas as multiple-view geometry and image-based rendering have made 3D cinematography possible. Important applications areas include production of stereoscopic movies, full 3D animation from multiple videos, special effects for more traditional movies, and broadcasting of multiple-viewpoint television, among others. The aim of this workshop was to bring together scientists and practitioner who have contributed to the mathematical foundations of the field, as well as those who have developed working systems. There were 20 participants from Canada, the United States, Europe and Asia. A total of 20 talks of length 30 minutes were presented during the five-day workshop.

A book comprising extended versions of these presentations is currently under production, and will be published by Springer-Verlag in 2010 [3].

1 Overview of the Field

The name 3D cinematography is motivated by the fact that it extends traditional cinematography from 2D (images) to 3D (solid objects that we can render with photorealistic textures from arbitrary viewpoints) at the same frame rate. A first workshop on 3D Cinematography took place in New York City in June 2006 jointly with the IEEE Conference on Computer Vision and Pattern Recognition [1]. Selected speakers from this workshop were invited to write extended papers, which after review were published as a special section in IEEE Computer Graphics and Applications [2]. At the time some prototypes had demonstrated the ability to reconstruct dynamic 3D scenes in various forms and resolutions. Various names were used to refer to these systems, such as virtualized reality, free-viewpoint video, and 3D video. All of these efforts were multi-disciplinary. These advances had clearly shown the promises of 3D cinematography systems, such as allowing real-time, multiple-camera capture, processing, transmission, and rendering of 3D models of real dynamic scenes. Yet, many research problems remained to be solved before such systems can be transposed from blue screen studios to the real world.

2 Recent Developments and Open Problems

This second workshop on 3D Cinematography was focused on summarizing the progress made in the field during the two years subsequent to the first workshop, and in particyular on real-time working systems and

applications, with an emphasis on recent, realistic models for lights, cameras and actions. Indeed, 3D cinematography can be regarded as the geometric investigation of lights (how to represent complex lighting, how to relight, etc.); cameras (how to recover the true camera parameters, how to simulate and control virtual cameras, etc.); and actions (how to represent complex movements in a scene, how to to edit, etc.). From a geometric viewpoint, it is a hard problem to represent complex, time-varying scenes and their interactions with lights and cameras. One important question wxplored in this workshop was - what is the dimensionality of such scenes ? Space decomposition methods are popular because they provide one approximate answer, although not of very good quality. It has become increasingly evident that better representations are needed. Several partial solutions are proposed in the workshop papers, illustrated with examples. They include wavelet bases, implicit functions defined on a space grid, etc. It appears that a common pattern is the recovery of a controllable model of the scene, such that the resulting images can be edited (interaction). Changing the viewpoint is only one (important) aspect. Changing the lighting and action is equally important. Recording and representation of three-dimensional scenes. This is at the intersection of optics, geometry and computer science, with many applications in movie and entertainment technology. Note that the invention of cinema (camera and projector) was also primarily a scientific invention that evolved into an art form. We suspect the same thing will probably happen with 3D movies. This will still be based on optical systems. But computers have since replaced the mechanics. What motivates our field ? Build 3D cameras (record the scene) and 3D projectors (display the scene); Important addition - build 3D editing suites, to interact with the scene in NOVEL ways (edit the scene).

3 Presentation Highlights

3.1 Towards 4D Capture and 6D Displays: Mask for Encoding Higher Dimensional Reflectance Fields

Presented by Ramesh Raskar, MIT Media Lab This talk describes a capture method that samples 4D reflectance field using a 2D sensor and a display method that encodes 6D reflectance field on 2D film for subsequent viewing. They capture the 4D reflectance field using a lightfield camera. The lightfield camera used optical spatial heterodyning to multiple sub-aperture views inside a camera. They describe reversible modulation of 4D light field by inserting a patterned planar mask in the optical path of a lens based camera. They reconstruct a 4D light field from a 2D camera image without any additional lenses as required by previous light field cameras. The patterned mask attenuates light rays inside the camera instead of bending them, and the attenuation recoverably encodes the ray on the 2D sensor. Their mask-equipped camera focuses just as a traditional camera might to capture conventional 2D photos at full sensor resolution, but the raw pixel values also hold a modulated 4D light field. The light field can be recovered by rearranging the tiles of the 2D Fourier transform of sensor values into 4D planes, and computing the inverse Fourier transform. The lightfield is captured with minimum reduction in resolution allowing a 3D encoding of depth in a traditional photo. They display 6D reflectance field using a passive mask (2D film) and additional optics. Traditional flat screen displays (bottom left) present 2D images. 3D and 4D displays have been proposed making use of lenslet arrays to shape a fixed outgoing light field for horizontal or bidirectional parallax. They present different designs of multi-dimensional displays which passively react to the light of the environment behind. The prototypes physically implement a reflectance field and generate different light fields depending on the incident illumination, for example light falling through a window. They discretize the incident light field using an optical system, and modulate it with a 2D pattern, creating a flat display which is view *and* illumination-dependent. It is free from electronic components. For distant light and a fixed observer position, they demonstrate a passive optical configuration which directly renders a 4D reflectance field in the real-world illumination behind it. Combining multiple of these devices they build a display that renders a 6D experience, where the incident 2D illumination influences the outgoing light field, both in the spatial and in the angular domain. Possible applications of this technology are time-dependent displays driven by sunlight, object virtualization and programmable light benders / ray blockers without moving parts.

3.2 Skeleton Cube: Estimating Time-Varying Lighting Environments

Presented by Takashi Matsuyama, Kyoto University, Japan This is joint work with T. Takai, and S. Iino.

Lighting environments estimation is one of important functions to realize photometric editing of 3D video; lighting can give various effects on 3D video. In his talk, Matsuyama proposed the Skeleton Cube to estimate time-varying lighting environmens: e.g. lightings by candles and fireworks. A skeleton cube is a hollow cubic object and located in the scene to estimate its surrounding light sources. For the estimation, video of the cube is taken by a calibrated camera and then observed self shadows and shading are analyzed to compute 3D distribution of time-varying point light sources. Matsuyama's team developed an iterative search algorithm for computing the 3D light source distribution. Several simulation and real world experiments showed its effectiveness.

3.3 Smooth and non-smooth wavelet basis for capturing and representing light

Presented by Dana Cobzas, University of Alberta, Canada Indirectly estimating light sources from scene images and modeling the light distribution is an important, but difficult problem in computer vision. A practical solution is of value both as input to other computer vision algorithms and in graphics rendering. For instance, photometric stereo and shape from shading requires known light. With estimated light such techniques could be applied in everyday environments, outside of controlled lab conditions. Light estimated from images is also helpful in augmented reality in order to consistently relight an artificially introduced object. While algorithms that recover light as individual point light sources work for simple illumination environments, it has been shown that a basis representation achieves better results for complex illumination. In her talk, Cobzas proposed a light model that uses Daubechies wavelets and a method for recovering light from cast shadows and specular highlights in images. She assumes that the geometry is known for part of the scene. In everyday images, one can often obtain a CAD model of man-made objects (e.g. a car), but the rest of the scene is unknown. Experimentally, she has tested her method for difficult cases of both uniform and textured objects and under complex geometry and light conditions. She evaluated the stability of estimation and quality of scene relighting using smooth wavelet representation compared to a non-smooth Haar basis and two other popular light representations (a discrete set of infinite light sources and a global spherical harmonics basis). She demonstrated good results using the proposed Daubechies basis on both synthetic and real datasets. This is joint work with Cameron Upright and Martin Jagersand.

3.4 Accurate Camera Calibration from Multi-View Stereo and Bundle Adjustment

Presented by Yasutaka Furukawa, University of Illinois at Urbana-Champaign The advent of high-resolution digital cameras and sophisticated multi-view stereo algorithms offers the promises of unprecedented geometric fidelity in image-based modeling tasks, but it also puts unprecedented demands on camera calibration to fulfill these promises. In this talk, Furukawa presents a novel approach to camera calibration where top-down information from rough camera parameter estimates and the output of a publicly available multi-view-stereo system on scaled-down input images are used to effectively guide the search for additional image correspondences and significantly improve camera calibration parameters using a standard bundle adjustment algorithm. The proposed method has been tested on several real datasets, including objects without salient features for which image correspondences cannot be found in a purely bottom-up fashion; and image-based modeling tasks, including the construction of visual hulls where thin structures are lost without this additional step of re-calibration procedure. This work was funded by Industrial Light and Magic, for applications in movie special effects.

3.5 Large scale multiview video capture

Presented by Bennett Wilburn, Microsoft Research Asia Wilburn discussed issues in large scale multiview video capture, with football matches as a motivating example. He briefly reviewed existing multiview video capture architectures, their advantages and disadvantages, and issues in scaling them to large environments. Then he explained that today's viewers are accustomed to a level of realism and resolution which is not feasibly achieved by simply scaling up the performance of existing systems. He surveyed some methods for extending the effective resolution and frame rate of multiview capture systems. He explored the implications of real-time applications for smart camera design and camera array architectures, keeping in mind that real-time performance is a key goal for covering live sporting events. Finally, he commented briefly on some

of the remaining challenges for photo-realistic view interpolation of multi-view video for live, unconstrained sporting events.

3.6 Capturing Live Action for 3D Cinema

Presented by Paul Beardsley, Disney Research Zurich Image capture for live-action 3D cinema is traditionally done using a pair of stereo cameras which provide the left-eye and right-eye sequences that will be projected on the cinema screen. This constrains artistic control of 3D effects because decisions about stereo parameters - such as choice of baseline and vergence - are made during shooting, and cannot easily be manipulated afterwards. Beardley's talk described Disney's current work on a heterogeneous sensor array composed of a cinematographic camera, support cameras, and depth sensors, to shoot live action 3D cinema. The post-production process allows a user to specify a pair of virtual stereo cameras viewing the original scene, with synthetic generation of the left-eye and right-eye images of the virtual rig. Thus stereo parameters, and hence the 3D effects that will be perceived by the viewer in the completed movie, cease to be a fixed and irrevocable choice made when shooting and are instead opened up to artistic control during post-production.

3.7 Introducing FTV

Presented by Masayuki Tanimoto, University of Nagoya, Japan Tanimoto described a new type of television named FTV (Free viewpoint TV). FTV is an innovative visual media that enables us to view a 3D scene by freely changing our viewpoints. FTV is based on the ray-space method that represents one ray in real space with one point in the ray-space. By using this method, Tanimoto and his team constructed the world's first real-time FTV system including the complete chain from capturing to display. He also developed new type of ray capture and display technologies such as a 360-degree mirror-scan ray capturing system and a 360 degree ray-reproducing display. He believes FTV will be widely used since it is an ultimate 3DTV, a natural interface between human and environment, and an innovative tool to create new types of content and art.

4 3D Video: Generation, Compression and Retrieval

Presented by Kiyo Aizawa, University of Tokyo, Japan In his talk, Aizawa explained the issues relating with compressing and broadcasting 3D video, which is a sequence of 3D models. 3D video reproduces a real moving object and provides free view point functionality. Differing to CG animation, models in the sequence of 3D video varies in the number of their vertices, connectivities, etc. Together with NHK and ATR in Japan, Aizawa developped specific compression methods for improving the quality of capture and reproduction of 3D video.

4.1 FTV with Free Listening-Point Audio

Presented by Masayuki Tanimoto In this talk, Tanimoto presented novel media integration of 3D audio and visual data for FTV with free listening-point audio. He captures the multi viewpoint and listening-point data, which are completely synchronized, by camera array and microphone array. This experiment demonstrates that it si possible to generate both free viewpoint images and free listening-point audio simultaneously.

4.2 The filming and editing of stereoscopic movies

Presented by Larry Zitnick, Microsoft Research, Redmont, USA The editing of stereoscopic movies, in which two views are shown to a user to provide the illusion of depth, leads to a variety of novel challenges. For instance when creating cuts between scenes, it is generally desirable to maintain a consistent vergence angle between the eyes. This may be accomplished by careful filming or in post-production using a variety of techniques. In this talk, Zitnick discussed basic video editing tasks in the context of stereoscopic movies, as well as more complex techniques such as the "Hitchcock effect", fade cuts and effects unique to stereoscopic movies.

4.3 Binocular cinematography: 3-D movies for the human eyes

Presented by Frederic Devernay, INRIA, France Most often, what is referred to as 3-D movies are really stereoscopic (or binocular) motion images. In stereoscopic motion images, two 2-D movies are displayed, one for the left eye and one for the right eye, and a specific device guarantees that each eye sees only one movie (common devices are active or passive glasses, parallax barrier displays or lenticular displays). 3-D content can be displayed as stereoscopic motion images, but the movie itself does not hold 3-D content, thus the name binocular cinematography. Although shooting a stereoscopic movie seems to be as simple as just adding a second camera, viewing the resulting movie for extended durations can lead to anything from a simple headache to temporary or irreversible damage to the oculomotor function. Although the film industry pushes the wide distribution of 3-D movies, visual fatigue caused by stereoscopic images should still be considered as a safety issue. In his talk, Devernay described the main sources of visual fatigue which are specific to viewing binocular movies, which can be identified and classified into three main categories: geometric differences between both images which cause vertical disparity in some areas of the images, inconsistencies between the 3-D scene being viewed and the proscenium arch (the 3-D screen edges), and discrepancy between the accommodative and the convergence stimuli that are included in the images. For each of these categories, he proposes solutions to either issue warnings during the shooting or correct the movies in the post-production phase. These warning and corrections are made possible by the use of state-of-the-art computer vision algorithms.

4.4 From 3D Studio Production to Live Sports Events

Presented by Adrian Hilton, University of Surrey In his talk, Hilton reviewed the challenges of transferring techniques developed for multiple view reconstruction and free-viewpoint video in a controlled studio environment to broadcast production for football and rugby. Experience in ongoing development of the iview free-viewpoint video system for sports production in conjunction with the BBC will be presented. Production requirements and constraints for use of free-viewpoint video technology in live events will be identified. Challenges presented by transferring studio technologies to large scale sports stadium will be reviewed together with solutions being developed to tackle these problems. This highlights the need for robust multiple view reconstruction and rendering algorithms which achieve free-viewpoint video with the quality of broadcast cameras. The advances required for broadcast production also coincide with those of other areas of 3D cinematography for film and interactive media production.

4.5 Photo-realistic Rendering from Approximate Geometry

Presented by Marcus Magnor, TU Braunschweig, Germany For 3D cinematography from sparse recording setups, estimating full 3D geometry of the dynamic scene is essential. If the geometry model and/or camera calibration is imprecise, however, multi-view texturing approaches lead to blurring and ghosting artifacts during rendering. In his talk, Magnor presented novel on-the-fly GPU-based strategies to alleviate, and even eliminate, rendering artifacts in the presence of geometry and/or calibration inaccuracies. By keeping the methods general, they can be used in conjunction with many different image-based rendering methods and projective texturing applications.

4.6 New Methods for Video-based Performance Capture

Presented by Christian Theobalt, Stanford University, Max Plank Institute Performance capture means reconstructing models of motion, shape and appearance of a real-world dynamic scene from sensor measurements. To this end, the scene has to be recorded with several cameras or, alternatively, cameras and active scanning devices. In this talk, Theobalt presented his recent work on mesh-based performance capture from a handful of synchronized video streams. His method does without a kinematic skeleton and poses performance capture as mesh deformation capture. In contrast to traditional marker-based capturing methods, the approach does not require optical markings and even allows to reconstruct detailed geometry and motion of a dancer wearing a wide skirt. Another important feature of the method is that it reconstructs spatio-temporally coherent geometry, with surface correspondences over time. This is an important prerequisite for post-processing

of the captured animations. Performance capture has a variety of potential applications in visual media production and the entertainment industry. It enables the creation of high quality 3D video, a new type of media where the viewer has control over the camera's viewpoint. The captured detailed animations can also be used for visual effects in movies and games. Theobalt briefly talked about ways to post-process the captured data such that they can be modified with off-the-shelf animation software.

4.7 Dense 3D Motion Capture from Synchronized Video Streams

Presented by Yasutaka Furukawa, University of Illinois at Urbana-Champaign In his talk, Furukawa described a novel approach to nonrigid, markerless motion capture from synchronized video streams acquired by calibrated cameras. The instantaneous geometry of the observed scene is represented by a polyhedral mesh with fixed topology. The initial mesh is constructed in the first frame using the publicly available PMVS software for multi-view stereo. Its deformation is captured by tracking its vertices over time, using two optimization processes at each frame: a local one using a rigid motion model in the neighborhood of each vertex, and a global one using a regularized nonrigid model for the whole mesh. Qualitative and quantitative experiments using seven real datasets show that this algorithm effectively handles complex nonrigid motions and severe occlusions.

4.8 Automatic Virtual Cinematography

Presented by Remi Ronfard, Xtranormal, Montreal, Canada Current research in 3D cinematography is concerned with automating the tasks of placing cameras and lights in a virtual world to create cinematic shots and editing of those shots into a movie. This has applications in *real-time cinematography* for computer games and *scripted cinematography* for movie pre-production. Focusing on the latter case, Ronfard presented a quick overview of both traditional and virtual cinematography, including script analysis, shot selection, camera placement and editing, and discussed the issues and opportunities facing this new research area.

4.9 New Directions for Active Illumination in 3D Photography

Presented by Douglas Lanman, Brown University, USA In his talk, Lanman presented recent work on novel 3D capture systems using active illumination at Brown University. Specifically, he focused on two primary topics: (1) Multi-Flash 3D Photography and (2) Surround Structured Illumination. Extending the concept of multi-flash photography, Lanman demonstrates how the surface of an object can be reconstructed using the depth discontinuity information captured by a multi-flash camera while the object moves along a known trajectory. To illustrate this point, Lanman presented experimental results based on turntable sequences. By observing the visual motion of depth discontinuities, surface points are accurately reconstructed - including many located deep inside concavities. The method extends well-established differential and global shape-from-silhouette surface reconstruction techniques by incorporating the significant additional information encoded in the depth discontinuities. Lanman continued his discussion by exploring how planar mirrors can be used to simplify existing structured lighting systems. In particular, he described a new system for acquiring complete 3D surface models using a single structured light projector, a pair of planar mirrors, and one or more synchronized cameras. He projects structured light patterns that illuminate the object from all sides (not just the side of the projector) so that he is able to observe the object from several vantage points simultaneously. This system requires that projected planes of light be parallel, and so he constructed an orthographic projector using a Fresnel lens and a commercial DLP projector. A single Gray code sequence is used to encode a set of vertically-spaced light planes within the scanning volume, and five views of the illuminated object are obtained from a single image of the planar mirrors located behind it. Using each real and virtual camera, he is able to recover a dense 3D point cloud spanning the entire object surface using traditional structured light algorithms. This configuration overcomes a significant hurdle to achieving full 360x360 degree reconstructions using a single structured light sequence by eliminating the need for merging multiple scans or multiplexing several projectors.

4.10 Hierarchical Model for Capturing and Texturing of 3D Models from 2D Images

Presented by Martin Jagersand, University of Alberta, Canada Jagersand described a three scale hierarchical representation of scenes and objects, and explained how this representation is suitable for both computer vision capture of models from images and efficient photo-realistic graphics rendering. The model consists of (1) a conventional triangulated geometry on the macro-scale, (2) a displacement map, introducing pixelwise depth with respect to each planar model facet (triangle) on the meso level. (3) A photo-realistic micro-structure is represented by an appearance basis spanning viewpoint variation in texture space. To demonstrate the three-tier model, Jagersand implemented a capture and rendering system based entirely on budget cameras and PC's. For capturing the model, he uses conventional Shape-From-Silhouette for the coarse macro geometry, variational shape and reflectance estimation for the meso-level, and a texture basis for the micro level. For efficient rendering the meso and micro level routines are both coded in graphics hardware using pixel shader code. This maps well to regular consumer PC graphics cards, where capacity for pixel processing is much higher than geometry processing. Thus photo-realistic rendering of complex scenes is possible on mid-grade graphics cards. He showed experimental results capturing and rendering models from regular images of humans and objects.

4.11 3D Video of Human Action in a Wide Spread Area with a Group of Active Cameras

Presented by Takashi Matsuyama, Kyoto University, Japan 3D video is usually generated from multi-view videos taken by a group of cameras surrounding an object in action. To generate nice-looking 3D video, the following three constraints should be satisfied simultaneously: (1) the cameras should be well calibrated, (2) for each video frame, the 3D object surface should be well covered by a set of 2D multi-view video frames, and (3) the resolution of the video frames should be enough high to record the object surface texture. From a mathematical point of view, it is almost impossible to find such camera arrangement and/or camera work that satisfy these constraints. Moreover, when an object performs complex actions and/or moves widely, it would be a reasonable way to introduce active cameras to track the object and capture its multi-view videos; otherwise a large number of (fixed) cameras are required to capture video data satisfying the constraints. Then, the fourth constraint is imposed: (4) the group of active cameras should be controled in real time so that each video frame satisfies the above three constraints. In his talk, Matsuyama described a *Cellular Method* to capture 3D video of human action in a wide spread area with a group of active cameras. The problem to find the camera work that satisfies the above four constraints is formulated as an optimization process and then an algorithm to find an optimal solution is presented with experimental results. This is joint work with H. Yoshimoto, and T. Yamaguchi.

4.12 Multi-View Stereo beyond the Lab Setting

Presented by Michael Goesele, TU Darmstadt, Germany Goesele presented a multi-view stereo algorithm that addresses the extreme changes in lighting, scale, clutter, and other effects found in large online community photo collections and other data sets not captured specifically for reconstruction purposes. The basic idea of the algorithm is to intelligently choose images to match, both at a per-view and per-pixel level. Goesele demonstrates that such adaptive view selection enables robust performance even with dramatic appearance variability. The stereo matching technique takes as input sparse 3D points reconstructed from structure-from-motion methods and iteratively grows surfaces from these points. Optimizing for surface normals within a photo-consistency measure significantly improves the matching results. While the focus of the approach is to estimate high-quality depth maps, it can also be extended to merge the resulting depth maps into compelling scene reconstructions. Goesele demonstrated the algorithm on standard multi-view stereo data sets and on casually acquired photo collections of famous scenes gathered from the Internet. This is joint work with Brian Curless, Hugues Hoppe, Noah Snavely and Steve Seitz.

5 Scientific Progress Made

The next frontier is the synthesis of virtual camera movements along arbitrary paths extrapolating cameras arranged on a plane, a sphere, or even an entire volume. This raises difficult issues. What are the dimensions of the allowable space of cinematographic cameras that professional cinematographers would want to synthesize? In other words, what are the independent parameters of the virtualized cameras that can be interpolated from the set of existing views? Further, what is the range of those parameters that we can achieve using a given physical camera setup? Among those theoretically feasible parameter values, which are the ones that will produce sufficient resolution, photorealism, and subjective image quality? These questions remain open for future research in this new world of 3D cinematography.

6 Outcome of the Meeting

A book comprising extended versions of this workshop presentations is currently under production, and will be published by Springer-Verlag in 2010 [3]. This book will also include an introductory survey of the field, and an extensive list of references.

References

[1] IEEE Workshop on Three-Dimensional Cinematography (3DCINE'06), R. Ronfard and G. Taubin, co-chairs, New York City, Thursday, June 22, 2006 (in conjunction with CVPR), `http://perception.inrialpes.fr/3dcine/`

[2] R. Ronfard and G. Taubin, Introducing 3D Cinematography. In *IEEE Computer Graphics and Applications*, **27(3)**, 18-20, May/June 2007.

[3] R. Ronfard and G. Taubin (eds.), Proceedings of the BIRS Workshop on Multi-View Image and Geometry Processing for 3D Cinematography, 13-18 July 2008, Springer-Verlag, 2010 (to appear).