Available online at www.sciencedirect.com
ScienceDirect
Signal Processing: *Image Communication* 22 (2007) 203–216
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Point-sampled 3D video of real-world scenes ☆

Michael Waschbüsch*, Stephan Würmlin, Daniel Cotting, Markus Gross

Computer Graphics Laboratory, ETH Zurich, Switzerland

Received 21 November 2006; accepted 29 November 2006

Abstract

This paper presents a point-sampled approach for capturing 3D video footage and subsequent re-rendering of real-world scenes. The acquisition system is composed of multiple sparsely placed 3D video bricks. The bricks contain a low-cost projector, two grayscale cameras and a high-resolution color camera. To improve on depth calculation we rely on structured light patterns. Texture images and pattern-augmented views of the scene are acquired simultaneously by time multiplexed projections of complementary patterns and synchronized camera exposures. High-resolution depth maps are extracted using depth-from-stereo algorithms performed on the acquired pattern images. The surface samples corresponding to the depth values are merged into a view-independent, point-based 3D data structure. This representation allows for efficient post-processing algorithms and leads to a high resulting rendering quality using enhanced probabilistic EWA volume splatting. In this paper, we focus on the 3D video acquisition system and necessary image and video processing techniques.

Keywords: 3D video; Free-viewpoint video; Scene acquisition; Point-based graphics

1. Introduction

The 3D video is a novel technology for capturing the dynamics and motion of a real-world scene during recording while providing the user with the possibility to change the viewpoint at will during playback. It is seen as one of the many promising emerging media technologies for next generation home entertainment and spatio-temporal visual effects. Free navigation regarding time and space in streams of visual media directly enhances the viewing experience, degree of immersion and interactivity. However, in most existing systems such virtual viewpoint effects have to be planned precisely and changes are no more feasible after the scene has been shot. As an example, freeze-and-rotate effects have been demonstrated in numerous feature films like *The Matrix*. It can only be realized by using a huge number of digital cameras which have to be placed very accurately. As another example, Digital Air's *Movia*® systems applies high speed, HD cameras which can be controlled precisely such that no software view interpolation

☆This paper is a revised and extended version of the manuscript 'M. Waschbüsch, S. Würmlin, D. Cotting, F. Sadlo, M. Gross; Scalable 3D Video of Dynamic Scenes' published in *The Visual Computer* 21(8–10):629–638, 2005, available at www.springerlink.com, DOI 10.1007/s00371-005-0346-7, ©Springer-Verlag 2005.

*Corresponding author.
E-mail addresses: waschbuesch@inf.ethz.ch (M. Waschbüsch), wuermlin@inf.ethz.ch (S. Würmlin), dcotting@inf.ethz.ch (D. Cotting), grossm@inf.ethz.ch (M. Gross).

0923-5965/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2006.11.009

is needed for the desired effect. But as a consequence for both approaches, changes to the view trajectory are not possible during post-production.

Recently, several multi-view video systems have been presented which allow for realistic re-renderings of 3D video from arbitrary novel viewpoints. However, for producing high-quality imagery, the capturing systems are restricted to configurations where cameras are placed very close together. As an example, the system by Zitnick et al. [41] uses eight cameras covering a horizontal field-of-view of 30°, where only linear arrangements are possible. To allow for more flexibility, i.e. configurations which cover an entire hemisphere with a small number of cameras, either model-based approaches need to be employed (e.g. Carranza et al. [8] with eight cameras) or degradation in visual quality has to be accepted (e.g. Gross et al. [13] with 16 cameras). The latter two systems are also limited by the employed reconstruction algorithms to the capture of foreground objects or even pre-defined objects only.

In Waschbüsch et al. [32] we introduced a scalable framework for 3D video of dynamic scenes. Our work is motivated by the drawbacks of the aforementioned systems and by the vision of bringing 3D video to a new level where capturing, editing and subsequent high-quality re-rendering is cost-effective and convenient as with the 2D counterpart. For this purpose, special hardware solutions for real-time depth estimation, such as 3DV Systems' *ZCam*™ (http://www.3dvsystems.com), and Tyzx's *DeepSea* High-speed Stereo Vision System (http://www.tyzx.com) have recently become available. A central part of our research is the view-independent representation of the captured 3D geometry streams, with the goal to allow for similar authoring and editing techniques as carried out in common 3D content creation and modeling tools. Thereby, creation of spatio-temporal effects becomes straightforward and one has no longer to cope with the common limitations of image-based representations. In this paper we focus on the capturing and geometry processing part of our point-sampled 3D video framework.

2. Related work

This paper extends or integrates previous work in areas like point-based computer graphics, depth from stereo and 3D video. For the sake of conciseness, we refer the reader to the ACM SIGGRAPH 2004 course on point-based computer graphics [1] and to relevant depth from stereo publications [26]. In the following, we will confine ourselves to related work in the area of 3D video.

In 3D video, multi-view video streams are used to re-render a time-varying scene from arbitrary viewpoints. There is a continuum of representations and algorithms suited for different acquisition set-ups and applications. Purely image-based representations [18] need many densely spaced cameras for applications like 3D TV [21]. Dynamic light field cameras [34,37] which have camera baselines of a couple of centimeters do not need any geometry at all. Camera configuration constraints can be relaxed by adding more and more geometry to image-based systems, as demonstrated by Lumigraphs [12]. Voxel-based representations [31] can easily integrate information from multiple cameras but are limited in resolution. Depth image-based representations [2,27] use depth maps which are computed predominantly by stereo algorithms [41]. Stereo systems still require reasonably small baselines and, hence, scalability and flexibility in terms of camera configurations are still not achieved. Redert et al. [25] used depth images acquired by Zcams [15] for 3D video broadcast applications. Appropriate representations for coding 3D audio/visual data are currently investigated by the MPEG-4 committee [28]. On the other end of the continuum, there are model-based representations which describe the objects or the scene by time-varying 3D geometry, possibly with additional video textures [8,16]. Almost arbitrary camera configurations become feasible, but most existing systems are restricted to foreground objects only.

Besides data representations, one has to distinguish between online and offline applications. Matusik et al. [19,20] focused on real-time applications, e.g. 3D video conferencing or instant 3D replays. However, they are restricted to capturing foreground objects only due to the nature of their silhouette-based depth reconstruction algorithms. Gross et al. [13] used a 3D video system based on a point sample representation [35] for their telecollaboration system *blue-c* and share the same limitation of only being able to reconstruct foreground objects. Mulligan et al. [22] also target telepresence. They compute geometric models with multi-camera stereo and transmit texture and depth over a network. Carranza et al. [8] presents an offline 3D video system which employs an a priori shape model which is adapted to the observed outline of a human. However, this system is only able to capture