Pattern Recognition 44 (2011) 1958–1968

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

3D video and free viewpoint video—From capture to display

Aljoscha Smolic [1]

Disney Research, Zurich, Clausiusstrasse 49, 8092 Zurich, Switzerland

ARTICLE INFO

Available online 15 September 2010

Keywords:
3D video
Stereo video
Free viewpoint video
3DTV

ABSTRACT

This paper gives an end-to-end overview of 3D video and free viewpoint video, which can be regarded as advanced functionalities that expand the capabilities of a 2D video. Free viewpoint video can be understood as the functionality to freely navigate within real world visual scenes, as it is known for instance from virtual worlds in computer graphics. 3D video shall be understood as the functionality that provides the user with a 3D depth impression of the observed scene, which is also known as stereo video. In that sense as functionalities, 3D video and free viewpoint video are not mutually exclusive but can very well be combined in a single system. Research in this area combines computer graphics, computer vision and visual communications. It spans the whole media processing chain from capture to display and the design of systems has to take all parts into account, which is outlined in different sections of this paper giving an end-to-end view and mapping of this broad area. The conclusion is that the necessary technology including standard media formats for 3D video and free viewpoint video is available or will be available in the future, and that there is a clear demand from industry and user for such advanced types of visual media. As a consequence we are witnessing these days how such technology enters our everyday life

1. Introduction

Convergence of technologies from computer graphics, computer vision, multimedia and related fields enabled the development of advanced types of visual media, such as 3D video (3DV) and free viewpoint video (FVV), which expand the user's sensation beyond what is offered by traditional 2D video [1]. 3DV offers a 3D depth impression of the observed scenery, which is also referred to as stereo. Specific displays are applied that ensure that a user sees different views with each eye. If the views are created properly the brain will fuse the views and a 3D depth impression will be perceived. This basic principle of stereopsis has been known since 1838 when Sir Charles Wheatstone published his fundamental paper about binocular vision [2]. Although the basic principle of stereopsis is fairly simple, improper 3DV can easily result in bad user experience. This can be caused by technical difficulties, e.g. of display systems, or by improper content creation. In fact the depth impression from a 3D display is a fake of the human visual system and, if not done properly, results can be unpleasant. Production of 3DV content is therefore a difficult art that requires a variety of technical, psychological and creative skills and has to consider perception and display capabilities [3]. Today many technical and artistic problems are resolved and 3DV has reached a high level of maturity. 3DV is available in cinemas, on Blu-ray disc, TV, games, mobile phones, PDAs, laptops, etc.

FVV allows the user an interactive selection of viewpoint and direction within a certain operating range, as known from computer graphics in virtual worlds and games [4]. In contrast to virtual objects and environments of computer graphics, FVV is about real world scenes as captured by natural cameras. Usually multiple camera signals are processed and converted into a suitable 3D scene representation (see next section) that allows rendering of arbitrary views. Apart from research prototypes, so far FVV has mainly been used commercially on production side, e.g. for famous stop-motion special effects in the movie "The Matrix" or in the "EyeVision" system [5] for sports effects (see also fundamental work in [6]). The company LiberoVision offers an FVV effects system for sports broadcast [7]. Users can enjoy FVV so far only in a very limited form using advanced displays as outlined in Section 7.

Both functionalities FVV and 3DV do not exclude each other. In contrary, they can be very well combined within a single system, since they are both based on a suitable 3D scene representation. This means that given a 3D representation of a scene, if a stereo pair corresponding to the human eyes can be rendered, the functionality of 3DV is provided. If a virtual view (i.e. not an available camera view) corresponding to an arbitrary viewpoint and viewing direction can be rendered, the functionality of FVV is provided. The ideal future visual media system will provide full FVV and 3DV at the same time.

E-mail address: smolic@disneyresearch.com

[1] Work for this paper was performed during the author's prior affiliation with the Fraunhofer Institute for Telecommunications-Heinrich-Hertz-Institut (FhG-HHI), Berlin, Germany.

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.09.005

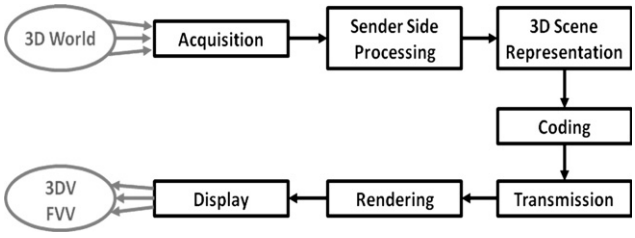

Fig. 1. 3DV and FVV processing chain from capture to display.

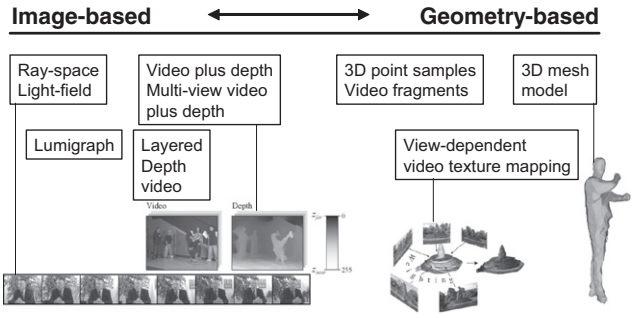

Fig. 2. 3D scene representations for 3DV and FVV.

In order to enable 3DV and FVV applications, the whole processing chain, including acquisition, sender side processing, 3D representation, coding, transmission, rendering and display, needs to be considered. All these technologies are broad research areas on their own. The complete 3DV and FVV processing chain is illustrated in Fig. 1. The end-to-end system design has to take all parts into account, since there are strong interrelations between all of them. For instance, an interactive display that requires random access to 3D data will affect the performance of a coding scheme, which is based on data prediction.

Depending on concrete application scenario and 3D scene representation various different algorithms and systems are available for each of the building blocks, which are the focus of this paper. The next section will first introduce and classify 3D scene representations and explain implications. Then Section 3 will give an overview of acquisition systems and processes, which may include cameras and other types of sensors. Section 4 is devoted to any sender side processing, i.e. conversion of the captured signals into the data of the 3D scene representation format. Efficient coding and transmission of these data are the focci of Section 5. Rendering, i.e. generation of the desired output views, is covered in Section 6. Section 7 outlines display systems and finally Section 8 summarizes and concludes the paper. Since this paper gives an end-to-end overview of technologies related to 3DV and FVV, coverage of each of these individual areas is brief, although they are all broad research areas in computer vision, computer graphics and multimedia communications. The goal is to give a general understanding from capture to display and to highlight specific aspects in all related areas.

2. 3D scene representation

The choice of a 3D scene representation format is of central importance for the design of any 3DV or FVV system [8]. On the one hand, the 3D scene representation sets the requirements for acquisition and signal processing on the sender side, e.g. the number and setting of cameras and the algorithms to extract the necessary data types. On the other hand, the 3D scene representation determines the rendering algorithms (and with that also navigation range, quality, etc.), interactivity, as well as coding and transmission. Therefore, 3D scene representation determines the end-to-end design and capabilities of any 3DV and FVV system.

In the computer graphics literature, methods for 3D scene representation are often classified as a continuum in between two extremes as illustrated in Fig. 2 [9]. These principles can also be applied for 3DV and FVV. The one extreme is represented by formats with full knowledge about the scene geometry. This approach can also be called geometry-based modeling. In most cases scene geometry is described on the basis of 3D meshes. Real world objects are reproduced using geometric 3D surfaces with an associated texture mapped onto them and the appearance properties. Alternative geometry-based formats use 3D point clouds [10] or voxels [11] to represent 3D geometry in FVV applications. Geometry-based methods allow full freedom of virtual view rendering as known from classical computer graphics. However, this is paid by the drawback of the need for accurate and robust 3D reconstruction of objects and scenes, which can often not be guaranteed (see Section 4).

The other extreme in 3D scene representations in Fig. 2 is called image-based modeling and does not use any 3D geometry at all. In this case virtual intermediate views are generated from available natural camera views by interpolation. The main advantage is a potentially high quality of virtual view synthesis avoiding any 3D scene reconstruction. However, this benefit has to be paid for by dense sampling of the real world with a sufficiently large number of natural camera view images. In general, the synthesis quality and the potential navigation range increase with the number of available views. Hence, typically large numbers of cameras have to be set up to achieve high-performance rendering, and a tremendous amount of image data needs to be processed therefore. Contrariwise, if the number of used cameras is too low, interpolation and occlusion artifacts will appear in the synthesized images, possibly affecting the quality. Such representations are also called light field [12] or Ray-Space [13,14].

In between the two extremes there exist a number of methods that make more or less use of both approaches and combine the advantages in some way. Some of these representations do not use explicit 3D models but depth or disparity maps [15]. Such maps assign a depth value to each pixel of an image (see Fig. 8). Virtual views can be rendered from depth and video in a limited operating range by depth image based rendering (DIBR) [16]. In order to broaden the navigation range the concept can easily be extended to multi-view video multiple depth (MVD) [17–20]. Layered depth video (LDV) [21] as an extension of the concept of layered depth images [22] is a more compact alternative representation. Closer to the geometry-based end of the continuum we find for instance representations that use 3D models with multiple textures and view-dependent texture mapping for rendering [23,24].

As we can see from Fig. 2 the different 3D representation formats include a variety of different data types. They all have specific advantages and drawbacks and the design of a specific application and system has to find the right trade-off. In any case the choice of the 3D representation format determines all other modules in the processing chain. The following sections give an overview of those from capture to display.

3. Acquisition

In most cases 3DV and FVV approaches rely on specific acquisition systems. Although automatic and interactive 2D–3D conversion (i.e. from 2D video to 3DV or FVV) is an important

research area for itself (see [25,26] and references therein for more details). Most 3DV and FVV acquisition systems use multiple cameras to capture real world scenery [27]. These are sometimes combined with active depth sensors, structured light, etc. in order to capture scene geometry. The camera setting (e.g. dome type as in Fig. 3 or linear as in Fig. 4) and density (i.e. number of cameras) impose practical limitations on navigation and quality of rendered views at a certain virtual position. Therefore, there is a classical trade-off to consider between costs (for equipment, cameras, processors, etc.) and quality (navigation range, quality of virtual views, etc.). Fig. 3 illustrates a dome type multi-camera acquisition system and captured multi-view video. Such a dome typesetting enables one to represent the whole space in between and an arbitrary viewing angle onto the scenery inside. However, spanning such a large volume requires many cameras or results in a sparse sampling with the mentioned restrictions on quality. Dome type settings are often used for controlled indoor environments and experimental setups. These have practical relevance for studio productions. Circular settings have been used to capture entire sports stadiums with about special 30 cameras [5]. The LiberoVision system is capable of reconstructing broadcast quality views of a football field using only the normally available production camera setup [7], with restrictions imposed by the setting (e.g. no views from opposite direction).

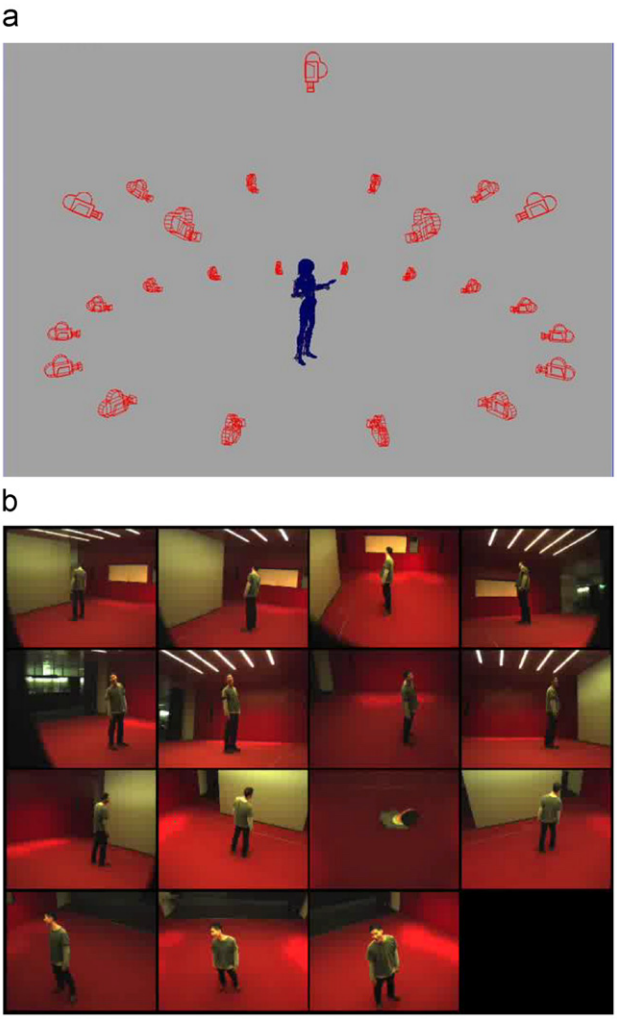

Fig. 3. Multi-camera setup for 3DVO acquisition and captured multi-view video.

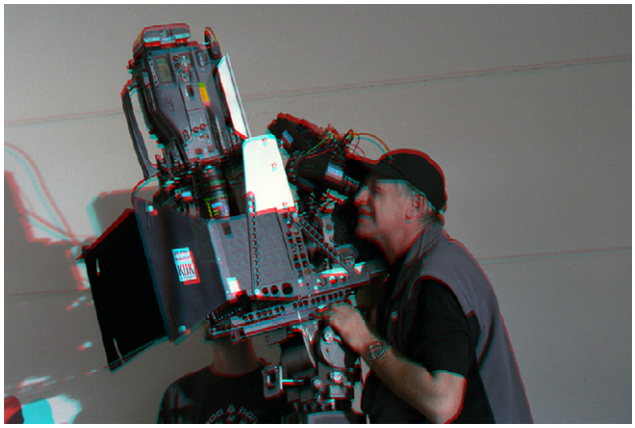

Fig. 4. Stereo video production, motorized camera rig with beam splitter.

A commercially highly important area is acquisition and production of stereoscopic video, i.e. 2 views that correspond to the user's eye positions. As mentioned before this simplest type of 3DV has recently entered broad mass markets. However, production of high quality stereo video is a difficult process. Improper stereo content may cause bad user experience including headaches, eye strain and fatigue. Today stereographers know the rules for production of proper stereo video content [3,26]. This includes adaptation of the baseline and convergence of the cameras to the scene content to be captured. Often, for instance, for far field views, they have to bring the cameras closer together than physically possible. Therefore they often use mirror rigs including a camera pair and a beam splitter as shown in Fig. 4. The rigs are motorized to have full control over baseline and convergence on set. Integrated image processing tools help to control stereo parameters like disparity ranges on set [28]. Although some stereo camera rigs are commercially available (e.g. [29,30]), there is still a lot of room for improvement and extension of such systems.

One line of research is the extension towards multi-camera systems and to include for instance depth sensors. Fig. 5 shows on the left side a 3-camera system with one digital cinema camera in the middle and 2 HD cameras as satellites on the sides. Such a setting can be used for instance to capture data to estimate depth for the middle camera, which is useful for successive post-production steps, like mixing and compositing content from different sources. On the right side Fig. 5 shows an extended stereo rig [31]. It also includes satellites and specific depth sensors. All these streams can then be used for successive depth estimation, etc. as discussed in the next sections.

Finally, light field and similar approaches require dense sampling of the 3D scenery, meaning a dense camera setting. Examples are shown in Fig. 6. This easily extends to 2D array type of camera settings [32]. In any case, acquisition for 3DV and FVV can result in complex settings of cameras and other sensors. These have to be mounted, the signals have to be recorded and stored, cameras have to be synchronized, etc. This includes peripheral equipment like PCs, grabber cards, disks, cables, etc. Thus, design and implementation of such systems may easily become a complex and cumbersome engineering task.

4. Sender side processing

After acquisition, the necessary data as defined by the 3D representation format have to be extracted from the multiple video and other captured data. This sender side processing can

Fig. 5. Extended multi-view camera systems for 3DV and FVV acquisition.

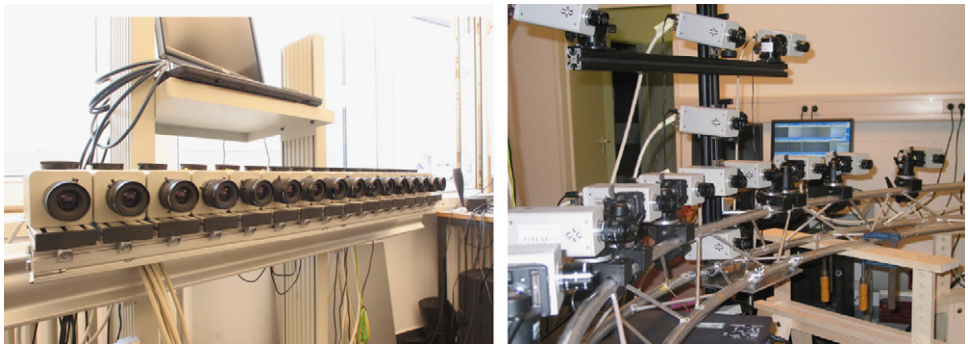

Fig. 6. Dense multi-view camera settings.

include automatic and interactive steps; it may be real-time or offline. Content creation and post-processing are included here. Tasks may be divided into low-level computer vision algorithms and higher-level 3D reconstruction algorithms. 3D reconstruction algorithms include for instance depth estimation and visual hull reconstruction to generate 3D mesh models. A general problem of 3D reconstruction algorithms is that they are estimations by nature. The true information is in general not accessible. Robustness of the estimation depends on many theoretical and practical factors. There is always a residual error probability that may affect the quality of the finally rendered output views. User-assisted content generation is an option for specific applications to improve performance. Purely image-based 3D scene representations do not rely on 3D reconstruction algorithms, and therefore do not suffer from such limitations.

4.1. Low-level algorithms

Low-level vision may include basic algorithms like color correction, white balancing, de-Bayering, normalization, filtering, lens distortion correction, camera calibration, rectification, segmentation, feature extraction and tracking, etc. Solutions for these tasks are well-established; however, in practice they often turn out to be difficult. The first group of algorithms (color correction, white balancing, de-Bayering, normalization, filtering, lens distortion correction, etc.) may be regarded as close to sensor correction and adaptation of very basic image properties [33]. Although all of this almost sounds trivial, incorrect white balancing and color adaptation of large multi-camera setups can be difficult and imperfections can cause problems for consecutive processing stages.

Camera calibration establishes the relation between the pixels in the images and the 3D world geometry. This is a research area for itself. Numerous algorithms and procedures have been developed to estimate extrinsic and intrinsic camera parameters for various camera setups [33]. Still special care has to be taken in practice since any errors may make later recordings unusable and multi-camera calibration of large setups is still a cumbersome and complex task. Rectification is the process of alignment of two or more images towards a common center of projection, which is very useful to ease successive stages of depth/disparity estimation and view synthesis [34]. Given the camera calibration it is a simple equation for 2 images, but for multi-view setups it can only be solved pairwise or if all cameras are perfectly aligned to a line with their centers of projection. In practical linear multi-view setups (e.g. Fig. 6 left) this can be approximated with small warping corrections to the images compensating small displacements off the common baseline.

Segmentation is the process of separation of images and video into objects (in the sense of physical objects) or regions (of certain common properties, not necessarily complete and meaningful objects). Again this is a complete research area in itself, and a very difficult and in general unresolved one (e.g. [35–38]). A vast amount of algorithms have been proposed for this purpose, each having specific advantages and drawbacks. Still any application that relies on some segmentation (see e.g. visual hull reconstruction below) has to find an own setting and solution; still, perfect quality and robustness can rarely be guaranteed. Errors will directly influence the following processing stages. Practical systems often work in controlled environments (studios, blue screen, etc.) or are realized as user assisted workflows.

Finally we want to mention feature extraction and tracking as a basic algorithm. Such algorithms as for instance the famous scale-invariant feature transform (SIFT) tracker [39] may form a basic part of higher level algorithms like camera calibration or reconstruction of 3D point clouds from feature trajectories. Again,

feature tracking and related structure from motion is an important research area in itself.

4.2. High-level 3D reconstruction algorithms

High-level 3D reconstruction algorithms are those that generate some kind of 3D geometry. They are high level in the sense that they use the captured signals and results of low-level algorithms to create the data of the selected 3D scene representation format. This may be for instance 3D mesh models with textures or video plus depth data as described in Section 2. In this section we outline a few examples of this broad research area.

One important class of 3D reconstruction algorithms is shape-from-silhouette or visual hull reconstruction [40]. Such algorithms typically use a multi-camera setup as shown in Fig. 3. The object of interest is segmented in each of the camera views. Any error in this stage will directly result in artifacts in the rendered output views. The 3D volume of the object of interest, or more precisely its convex hull, can be reconstructed using volume carving [41]. The result is a volumetric voxel model as illustrated in Fig. 7(a). Further steps may apply surface extraction using the marching cubes algorithm [42], surface smoothing [43] and mesh complexity reduction [44] as illustrated in Fig. 7(b)–(d). The result is a surface 3D mesh model as widely used in computer graphics providing the same functionalities but describing a real world object. For photorealistic rendering, a s described below in Section 6, the model has to be textured using color pixel data from the original video signals.

Numerous variants and improvements of shape-from-silhouette have been described over the years, e.g. [45]. It has been shown that exploiting a priori knowledge about the object of interest in the form of pre-defined mesh templates of physical deformation models helps to improve the results [46]. Inherent problems arising from segmentation and occlusion remain to be handled.

Other types of 3D structure recovery include shape-from-focus and -defocus (SfD) [47,48], shape-from-shading [49,50] and structure-from-motion (SfM) [51–54]. Some methods extract geometrical scene properties like vanishing points and lines to get a 3D reconstruction [55–57]. More important for 3DV and FVV are structure-from-stereo or depth and disparity estimation algorithms [58–68]. In this case 2 or more views of a scene are available and disparity between them or 3D depth of scene points is estimated using correspondences. This is an important core research area of computer vision and numerous algorithms for this purpose have been proposed [61,63]. A typical result is shown in Fig. 8 [17]. For each pixel in the color image there is a corresponding depth value describing its depth in the scene. In this example depth is quantized with 8 bits on a logarithmic scale between a minimum and maximum distance Z_{near} and Z_{far}. Brighter areas are closer to the camera; darker areas are further in the background. Such a 3D data representation allows rendering of virtual views nearby the available color image (see below in Section 6), with that enabling 3DV and FVV; however, dis-occlusion artifacts increase with distance of the virtual view from the available view. Reliable, accurate, robust and automatic depth estimation is still a difficult task, which is why this is still a very active and important research area. Inherent problems arise from occlusions and ambiguities, which is why reliable correspondence estimation cannot always be guaranteed. Practical problems include noise and other types of inaccuracies of signals and camera setups.

Some approaches combine color cameras with specific depth sensors that directly capture scene depth [31,70]. Typical time-of-flight (ToF) depth sensor has a relatively low resolution, e.g. 204×204 pixel. The generated depth maps have to be registered with the color images, which is not an easy task since both cameras are inherently located at different positions. Further, ToF sensors are very sensitive to noise and temperature and the depth range they can capture is quite limited. Also non-linear distortions can create problems. A promising approach is to combine ToF depth maps with high resolution depth maps computed by classical stereo algorithms to get the best of both worlds [70].

The concept of video plus depth is easily extended to multiview video plus depth (MVD), as illustrated in Fig. 9 [17]. A scene is captured by a multi-camera rig as illustrated in Fig. 6. Depth data are estimated for each of the original camera views. This extends the potential navigation range for 3DV and FVV applications. Also the quality of depth estimation can be improved by combining information from multiple cameras into the estimation algorithms [17,18].

MVD is a powerful representation format for 3DV and FVV; however, this results in a huge amount of highly redundant data. An alternative is to use a representation known as layered depth video (LDV) [21], extending the concept of layered depth images [22]. Image information is stored in different layers that are extracted from the original views by projecting the views onto each other and eliminating duplicated content. An example of LDV is shown in Fig. 10. It has a main layer with associated depth map and also an occlusion layer with associated depth data. The main layer is one of the original views, while the occlusion layer is

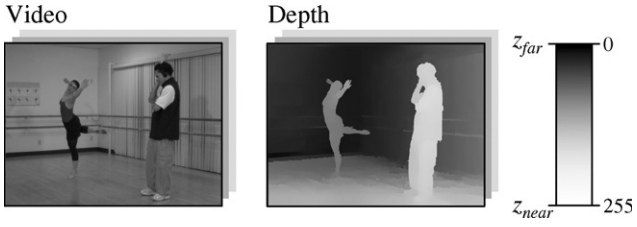

Fig. 8. Video and associated per pixel depth data.

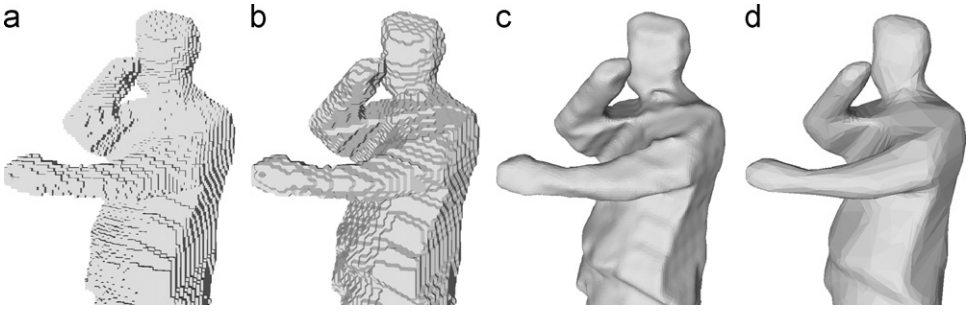

Fig. 7. Different steps of visual hull reconstruction.

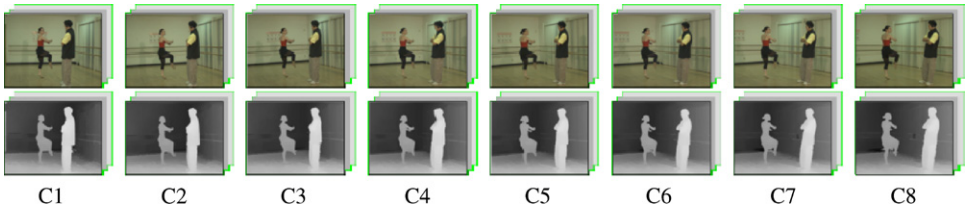

Fig. 9. Multi-view video plus depth. A scene is captured by 8 synchronized cameras (C1–C8) at the same time from different viewpoints. Depth is estimated for each of the camera views.

Fig. 10. Layered depth video with a main and an occlusion color layer, each with associated depth.

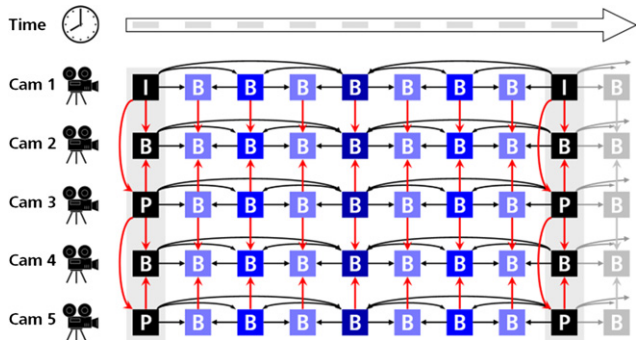

Fig. 11. Multi-view video coding (MVC).

extracted from all other views and contains only the content that is occluded in the main layer view. Note that in principle LDV can contain more than 2 layers. LDV seems attractive compared to MVD due to the more compact representation of the data. However, extraction of LDV is based on view warping (see below in Section 6) and with that on error prone depth data. Furthermore, this simple approach ignores influence from reflections, shadows, etc., which result in different appearance of the same content in different views. These are preserved in an MVD representation. Thus an end-to-end comparison of MVD and LDV taking coding and all other steps of the processing chain into account is yet to be provided.

5. Coding and transmission

For transmission over limited channels 3DV and FVV data have to be compressed efficiently. This has been widely studied in the literature and powerful algorithms are available for many of the resented representation formats [69]. International standards for content formats and associated coding technology are necessary to ensure interoperability between different systems. ISO-MPEG and ITU-VCEG are international organizations that released a variety of important standards for digital media including standards for 3DV and FVV. Classical 2-view stereo has been already supported by MPEG-2 since the mid-1990s. However, since the general application of 3DV did not develop into a significant market yet, there was no use of related standards so far.

Finally, these days 3DV is reaching wide consumer markets. First solutions for 3DTV use the so-called frame compatible 3D. Left and right views are down-sampled horizontally or vertically by a factor of 2 and combined into a single image of original resolution in a side-by-side or top–down manner [71]. Then this combined image is passed for encoding and transmission using the available algorithms and systems. This way the available coding and transmission infrastructure can be used without any change. It is the easiest way for fast and cheap introduction of 3DTV services. However, this is paid for by the price that half of the resolution of the images is lost. Alternatives are to use a quincunx or temporal interleaving of the views also suffering from loss of resolution [71].

More advanced solutions exploit inter-view redundancies for efficient compression by inter-view prediction and provide the means to keep full resolution of both views. Multi-view video coding (MVC) is a recently released extension of H.264/AVC, which is illustrated in Fig. 11 [72]. It is currently the most efficient way to encode 2 or more videos showing the same scenery from different viewpoints. MVC allows the design of a variety of different spatio-temporal prediction structures that can be tailored to a given application scenario. The so-called stereo high profile of MVC provides specific setting for the most important case of stereo (2-view) video. It forms the basis of the "Blu-ray 3D™" specification.

Video plus depth as illustrated in Fig. 8 is already supported by a standard known as MPEG-C Part 3. It is an alternative format for 3DV that requires view synthesis at the receiver (see next section). Video plus depth supports extended functionality compared to classical 2-view stereo such adjustment of depth impression to different displays and viewing preferences [15] in a limited way. Support for more advanced depth-based formats such as MVD and LDV is currently under study in a new activity in MPEG [71]. The goal is to provide an efficient format and coding specification that would support advanced 3DV functionalities which rely on view synthesis. These functionalities include a wide range of auto-stereoscopic displays and individual adjustment of the depth impression as explained in Section 6.

Different model-based 3D representations for FVV are supported by various tools of the MPEG-4 standard, which is in fact a

rich multimedia framework providing dedicated coding tools for a variety of different data types. Fig. 12 illustrates coding and multiplexing of dynamic 3D geometry, associated video textures and auxiliary data using MPEG-4, which enables model-based FVV [73]. The MPEG-4 Animation Framework eXtension (AFX) [74] is a part of the standard that supports for instance efficient compression of static and dynamic meshes, such as Frame-based Animated Mesh Compression (FAMC) [75]. Also FVV representations based on point clouds [10] are readily supported by MPEG-4 AFX tools. Associated video textures can be encoded with any available video codec such as H.264/AVC.

For transport and storage of multimedia data, MPEG standards already provide efficient mechanisms. These Systems specifications are known for instance as MPEG-2 TS (Transport) or MPEG-2/4 file format. They are very flexible and configurable. Embedding of all the data described before (MVC, MPEG-C Part 3, AFX) is already possible. Support for new data types (e.g. MVD or LDV coded data) will be specified once available. With that, transport and storage of 3DV and FVV data can rely on available infrastructure.

6. Rendering

Rendering is the process of generation of the final output views from data in the 3D representation format, after decoding if applicable. Fig. 13 illustrates the functionality of FVV using an example of an interactive 3D scene with an FVV object included. An editor created a scene that further includes a 360° panorama and a dynamic computer graphics object. The FVV object was captured as illustrated in Fig. 3, 3D reconstructed as shown in Fig. 7, represented and coded as shown in Fig. 12. Now rendering is done by classical computer graphics methods. Multi-texturing was applied using the video textures with the FVV object [4]. The user can navigate freely and watch the dynamic scene from any desired viewpoint and viewing direction. Fig. 13 includes 6 different viewpoints at 5 points in time (top middle and right are at the same time from different viewpoints).

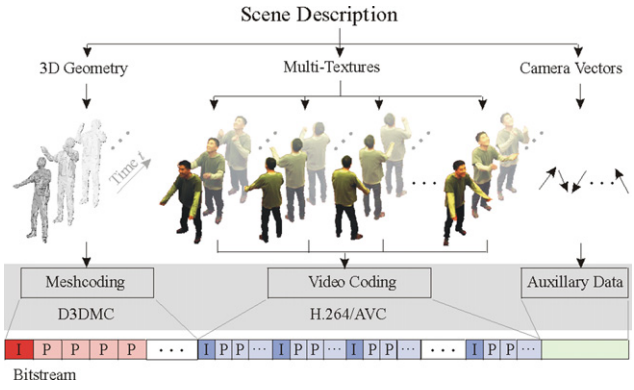

Fig. 12. Coding and multiplexing of dynamic 3D geometry, associated video textures and auxiliary data using MPEG-4 for model-based FVV.

In case of a depth-based 3D scene representation, DIBR is performed to create the output views [16]. Given a color image and an associated depth map as shown in Fig. 8 along with camera calibration information, any pixel of the image can be projected into the 3D space and then projected back onto an arbitrary virtual camera plane, creating a virtual image. This is in short the principle of virtual view synthesis. The quality of the virtual view depends on the accuracy of the depth data. Further, dis-occlusions will appear in the rendered views that have to be filled by inpainting [76] or some other kind of image completion. The amount of dis-occlusions increases with the distance of the virtual view from the original view, thus drastically limiting the potential navigation range using single video plus depth. However, generation of a 2nd virtual stereo view to the given original camera view is possible in most cases, due to the small required distance, which is in the range of the human eye distance. In this case the depth impression is adjustable, since the virtual distance (baseline, inter-ocular) between the 2 views can be selected by the user.

Using more than one view plus depth in an MVD or LDV representation widens the potential navigation range. Fig. 14 illustrates virtual view synthesis using 2 views plus depth as input. Any virtual view in between the available views can be rendered and dis-occlusions in one view can be filled to a wide extent with content from the other view. However, some unavoidable holes may remain for specific scene geometries, where content is occluded in both available views. Specific handling of depth discontinuities helps to reduce artifacts along object borders [17,20,77].

For more than 2 views the navigation range easily extends even further by pairwise switching. In principle then a spatio-temporal video volume is represented by the given data and

Fig. 13. Integrated interactive 3D scene with FVV.

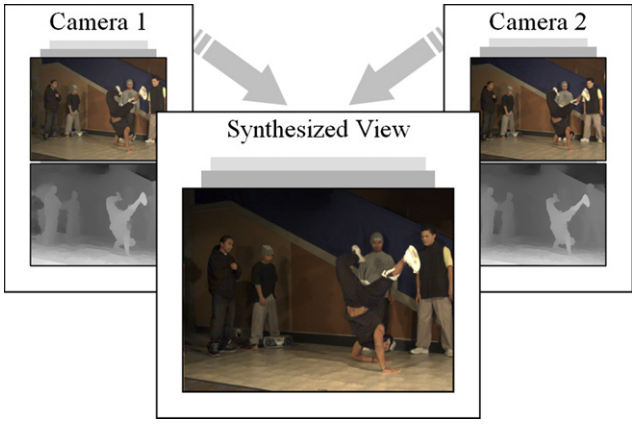

Fig. 14. Intermediate view synthesis from multiple video plus depth data.

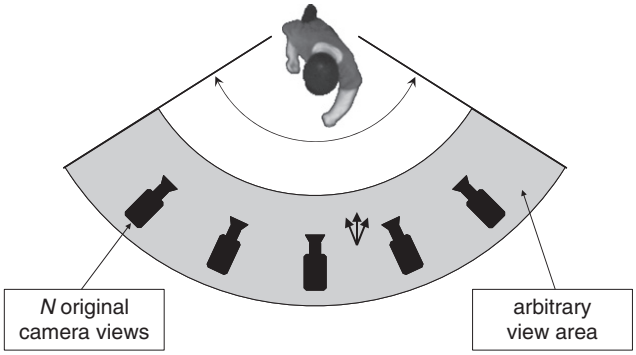

Fig. 15. Spatio-temporal video volume supported by MVD and LDV, given a certain camera setting.

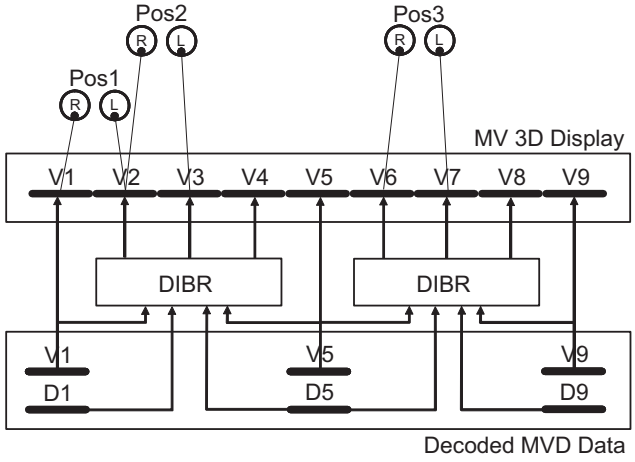

Fig.16. Efficient support of auto-stereoscopic displays; Pos: viewpoint, R: right eye, L: left eye, V: view/image, D: depth.

rendering algorithms as illustrated in Fig. 15. FVV is supported within a certain operating range given by the camera setting. A high quality stereo pair can be rendered with arbitrary virtual baseline allowing individual adjustment of the depth impression. Further, an arbitrary number of output views M can be rendered, which can be for instance bigger than the number of original views N. Such an approach with $M > N$ allows efficient support of auto-stereoscopic displays, which require 9, 16 or even more views (see Section 7). As illustrated in Fig. 16 only a subset of the required display views has to be transmitted along with associated depth data and the other views can be created via DIBR [19]. In consequence, MVD and derived LDV supports all advanced FVV and 3DV functionalities in an operating range given by the camera setting.

7. Display

Finally the rendered output views are presented to the user on a display. FVV requires interactive input from the user to select the viewpoint. This can be done by classical devices like mouse or joystick. Some systems also track the user (head or gaze) employing cameras and infrared sensors [78].

In order to provide a depth impression 2 or more views have to be presented to the user appropriately at the same time using a specific 3D display. Such 3D displays ensure that the user perceives a different view with each eye at a time. If it is a proper stereo pair, the brain will compute a 3D depth impression of the observed scene.

Currently, various types of 3D displays are available and under development [79]. Traditional technology uses classical 2-view stereo with one view for each eye and some kind of glasses to filter the corresponding view into each eye. The old fashioned anaglyph principle relies on color separation and suffers from limited color reproduction capability and visual discomfort due to different spectral contents presented to each eye [79]. Modern stereo display systems rather rely on polarization or shutter technology. A 3D display system (may be a cinema screen as well) based on polarization shows both images at the same time with different polarizations. The glasses act as different polarizing filters to separate the views. A 3D shutter display shows both images in temporally alternating sequence. The glasses are synchronized with the display and open and close appropriately so that each eye sees only the corresponding views. Such stereoscopic displays are already quite mature and readily available for professional and private users.

Multi-view auto-stereoscopic displays are advanced systems, which do not require glasses [80]. Here, 2 or more views are displayed at the same time and a lenticular sheet or parallax barrier element in front of the light emitters ensures correct view separation for the viewer's eyes. If more than 2 views are used, also limited FVV functionality is supported in the sense that if the user moves in front of the screen he can perceive a natural motion parallax impression. This is illustrated schematically in Fig. 16. A user at position 1 sees views 1 and 2 with right and left eyes, respectively, only. Another user at position 3 sees views 6 and 7; hence multi-user 3D viewing is supported. Assume a user moves from position 1 to position 2. Now views 2 and 3 are visible with the right and left eyes, respectively. If V1 and V2 is a stereo pair with proper human eye distance baseline, then V2 and V3 as well and so on, a user moving in front of such a 3D display system will perceive a 3D impression with dis-occlusions and occlusions of objects in the scenery depending on their depth. This motion parallax impression will not be seamless and the number of different positions is restricted to M-1. In practice auto-stereoscopic displays still suffer from a number of limitations. The available depth range is limited compared to glasses-based stereo systems. Ghosting and cross-talk often reduce viewing comfort. Also, the navigation range (number of views and angle) and the resolution of each single view are limited.

Even more sophisticated display systems use principles of light-field rendering [81], integral imaging [82], or holographic

Table 1
Comparison of different 3D scene representations with regard to stages of the 3DV/FVV processing chain.

	Acquisition	Reconstruction	Coding	Rendering	FVV range
Model-based	Relatively few cameras necessary for wide navigation	Full 3D geometry reconstruction, error prone and difficult for complete scenes, often applied for objects of interest, exploiting a-priori knowledge or interactive operation	Moderate bitrate, can be set very low at limited quality	Classical computer graphics	Wide, dome settings (surround) possible
Depth-based	Medium number of cameras necessary	Depth estimation, error prone	Medium bitrate	Depth-based view interpolation	Medium
Image-based	Dense sampling of the scene, many cameras necessary to enable navigation	None	High bitrate if many views are used	View interpolation, light field rendering	Limited

technology [83]. These are currently mainly available as research prototypes or single installations. Content creation for such types of displays and development of the whole corresponding media pipelines are research areas still in their infancy.

8. Summary and conclusions

This paper provided an overview of 3DV and FVV from capture to display. Naturally, different aspects of this broad research area were summarized briefly. For more details the reader is referred to the publications listed below.

3DV and FVV were introduced as extended visual media that provide advanced functionalities compared to standard 2D video. Both can very well be provided by a single system. New technology spans the whole processing chain from capture to display. The choice of a certain 3D scene representation is determining the design of all other modules along the processing chain. Technology for all the different parts is available, and is maturating and further emerging.

The 3D scene representations in Fig. 2 and the 3DV/FVV processing chain in Fig. 1 create a matrix of different technologies. This is illustrated in Table 1 in a simplified version to capture the main characteristics. Each 3D scene representation (model-based, depth-based and image-based) creates a processing chain from capture to display. At each stage of the pipeline different algorithms/technologies are applied with different advantages and drawbacks. As such Table 1 gives a high level mapping of 3DV and FVV technologies and may help in the design of specific applications and systems.

Growing interest for such applications is noticed from industry and users. 3DV is well established in cinemas, with more and more content being created. There is a strong push from industry to bring 3DV also to the home front, e.g. via Blu-ray or 3DTV. FVV is so far established as a post-production technology. FVV end-user mass market applications are still to be expected for the future.

Acknowledgments

I would like to thank the Interactive Visual Media Group of Microsoft Research for providing the Breakdancers and Ballet data sets, and the Computer Graphics Lab of ETH Zurich for providing the Doo Young multi-view data set.

Background, knowledge and material for this paper were developed during my previous employment at the Fraunhofer Institute for Telecommunications-Heinrich-Hertz-Institut (FhG-HH), Berlin, Germany. Text and illustrations were developed in collaboration with colleagues including Karsten Mueller, Philipp Merkle, Birgit Kaspar, Matthias Kautzner, Sabine Lukaschik, Ralf Tanger, Marcus Mueller, Frederick Zilly, Peter Kauff, Peter Eisert, Thomas Wiegand, Christoph Fehn and Ralf Schaefer. I would also like to thank Sebastian Knorr from imcube and the Technical University of Berlin.

References

[1] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, T. Wiegand: 3D video and free view-point video—technologies, applications and MPEG standards, in: ICME 2006, International Conference on Multimedia and Expo, Toronto, Ontario, Canada, July 2006.
[2] C. Wheatstone, Contributions to the physiology of vision—Part the First. On some remarkable, and hitherto unobserved, phenomena of binocular vision, Philosophical Transactions of the Royal Society of London 128 (1838) 371–394 Received and Read June 21.
[3] B. Mendiburu, in: 3D Movie Making—Stereoscopic Digital Cinema from Script to Screen, Elsevier, 2008.
[4] A. Smolic, K. Mueller, P. Merkle, T. Rein, P. Eisert, T. Wiegand, Free viewpoint video extraction, representation, coding, and rendering, in: Proceedings of the ICIP 2004, IEEE International Conference on Image Processing, Singapore, October 24–27 2004.
[5] 〈http://www.ri.cmu.edu/events/sb35/tksuperbowl.html〉.
[6] T. Kanade, P.J. Narayanan, P.W. Rander, Virtualized reality: concepts and early results, in: IEEE Workshop on the Representation of Visual Scenes (in conjunction with ICCV'95), Boston, MA, June 1995.
[7] 〈http://www.liberovision.com/〉.
[8] A. Smolic, P. Kauff, Interactive 3D video representation and coding technologies, in: Proceedings of the IEEE, Special Issue on Advances in Video Coding and Delivery, vol. 93(1), January 2005.
[9] S.B. Kang, R. Szeliski, P. Anandan: The geometry-image representation tradeoff for rendering, in: ICIP 2000, IEEE International Conference on Image Processing, Vancouver, Canada, September 2000.
[10] S. Würmlin, E. Lamboray, M. Gross, 3D video fragments: dynamic point samples for real-time free-viewpoint videoComputers and Graphics, Special Issue on Coding, Compression and Streaming Techniques for 3D and Multimedia Data 28 (1) (2004) 3–14 Elsevier Ltd.
[11] S.M. Seitz, C.R. Dyer, Photorealistic scene reconstruction by voxel colouring, International Journal of Computer Vision 35 (2) (1999) 151–173.
[12] M. Levoy, P. Hanrahan, Light field rendering, in: Proceedings of the ACM SIGGRAPH, August 1996, pp. 31–42.
[13] Masayuki Tanimoto, Free viewpoint television—FTV, in: Proceedings of the PCS 2004, Picture Coding Symposium, San Francisco, CA, USA, December 15–17, 2004.
[14] T. Fujii, M. Tanimoto, Free-viewpoint TV system based on Ray-Space representation, SPIE ITCom 4864 (22) (2002) 175–189.
[15] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, L. Vangool, E. Ofek, I. Sexton, An Evolutionary and Optimised Approach on 3D-TV, Proceedings of the of IBC 2002, Int. Broadcast Convention, Amsterdam, Netherlands, 2002.
[16] C. Fehn, 3D-TV using depth-image-based rendering (DIBR), in: Proceedings of the of Picture Coding Symposium (PCS), San Francisco, CA, USA, December 2004.
[17] C. L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, High-quality video view interpolation using a layered representation, in: Proceedings of the ACM SIGGRAPH and ACM Transactions on Graphics, Los Angeles, CA, USA, August 2004.
[18] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, R. Tanger, Depth map creation and image based rendering for advanced 3DTV services providing interoperability and scalability, Signal Processing: Image Communication. Special Issue on 3DTV (2007).

[19] A. Smolic, K. Müller, K. Dix, P. Merkle, P. Kauff, T. Wiegand, Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems, in: Proceedings of the ICIP 2008, IEEE International Conference on Image Processing, San Diego, CA, USA, October 2008.

[20] K. Mueller, A. Smolic, K. Dix, P. Merkle, P. Kauff, T. Wiegand, View synthesis for advanced 3D video systems, EURASIP Journal on Image and Video Processing, 2008 (2008) doi:10.1155/2008/438148.

[21] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, T. Wiegand, Reliability-based generation and view synthesis in layered depth video, in: Proceedings of the MMSP 2008, IEEE International Workshop on Multimedia Signal Processing, Cairns, Australia, October 2008.

[22] J. Shade, S. Gortler, L.W. He, R. Szeliski, Layered depth images, in: Proceedings of the SIGGRAPH'98, Orlando, FL, USA, July 1998.

[23] A. Smolic, K. Mueller, P. Merkle, T. Rein, P. Eisert, T. Wiegand: Free viewpoint video extraction, representation, coding, and rendering, in: ICIP 2004, IEEE International Conference on Image Processing, Singapore, October 24–27, 2004.

[24] T. Matsuyama, X. Wu, T. Takai, T. Wada, Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video, IEEE Transactions on Circuits and Systems for Video Technology 14 (3) (2004) 357–369.

[25] M. Kunter, S. Knorr, A. Krutz, T. Sikora, Unsupervised object segmentation for 2D to 3D Conversion, in: Proceedings of the of the SPIE: Stereoscopic Displays and Applications XX Conference, San José, USA, 2009.

[26] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Mueller, M. Lang, 3D video post-production and processing, Invited Paper, Proceedings of the IEEE, Special Issue on 3D Media and Displays, in press.

[27] A. Kubota, A. Smolic, M. Magnor, T. Chen, M. Tanimoto, Multi-view imaging and 3DTV—special issue overview and introduction, IEEE Signal Processing Magazine, Special Issue on Multi-view Imaging and 3DTV 24 (6) (2007).

[28] F. Zilly, P. Eisert, P. Kauff, Real-time analysis and correction of stereoscopic HDTV sequences, in: Proceedings of CVMP, London, November 2009.

[29] 〈http://www.3alitydigital.com/〉.

[30] 〈http://www.binocle.com/〉.

[31] 〈http://www.3d4you.eu/index.php〉.

[32] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Levoy, M. Horowitz, High performance imaging using large camera arrays, in: Proceedings of the SIGGRAPH 2005, July 31–August 4, Los Angeles, CA, USA.

[33] Fabio Remondino, Clive Fraser, Digital camera calibration methods: considerations and comparisons, IAPRS vol. XXXVI, Part 5, Dresden, 25–27 September 2006.

[34] C. Loop, Z. Zhang, Computing rectifying homographies for stereo vision, Computer Vision and Pattern Recognition 1 (1999) 131.

[35] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviours, IEEE Transactions on Systems, Man and Cybernetics (2004).

[36] L. Lucchese, S.K. Mitra, in: Color Image Segmentation: A State-of-the-Art Survey, Department of Electrical and Computer Engineering, USCB, 2001.

[37] A. Yilmaz, O. Javed, M. Shah, in: Object Tracking: A Survey, ACM Computing Surveys, 2006.

[38] X. Bai, J. Wang, D. Simons, G. Sapiro, Video snapcut: robust video object cutout using localized classifiers, in: Proceedings of the ACM SIGGRAPH, 2009.

[39] D.G. Lowe, Object recognition from local scale-invariant features, in: Proc.eedings of the International Conference on Computer Vision, Corfu, Greece, 2004, pp. 1150–1157.

[40] A. Laurentini, The visual hull concept for silhouette-based image understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (2) (1994) 150–162.

[41] W. Matusik, C. Buehler, R. Raskar, S.J. Gortler, L. McMillan, Image based visual hulls, in: SIGGRAPH'00 Proceedings, July 2000, pp. 369–374.

[42] T. Lewiner, et al., Efficient implementation of marching cubes cases with topological guarantee, Journal of Graphics Tools 8 (2003) 1–15.

[43] G. Taubin, Curve and surface smoothing without shrinkage, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV 95), 1995.

[44] H. Hoppe, Efficient implementation of progressive meshes, Computers & Graphics 22 (2) (1998) 27–36.

[45] C. Liang, K.K. Wong, 3D reconstruction using silhouettes from unordered viewpoints, Image and Vision Computing 28 (4) (2010) 579–589.

[46] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.P. Seidel, S. Thrun, Performance capture from sparse multi-view video, in: SIGGRAPH'08: ACM SIGGRAPH 2008 papers, 2008, pp. 1–10.

[47] Y.-F. Liu, A unified approach to image focus and defocus analysis, Ph.D. Thesis, State University of New York at Stony Brook, New York, 1998.

[48] T.-C. Wei, Three dimensional machine vision using image defocus, Ph.D. Thesis, State University of New York at Stony Brook, New York, 1994.

[49] R. Zhang, P.-S. Tsai, J.E. Cryer, M. Shah, Shape from shading: a survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (8) (1999) 690–706.

[50] J.-D. Durou, M. Falcone, M. Sagona, Numerical methods for shape-from-shading: a new survey with benchmarks, Computer Vision and Image Understanding 109 (1) (2008) 22–43.

[51] P. Beardsley, P.H.S. Torr, A. Zisserman, 3D model acquisition from extended image sequences, in: Proceedings of the European Conference on Computer Vision (ECCV), 1996, pp. 683–695.

[52] T. Jebara, A. Azarbayejani, A. Pentland, 3D structure from 2D motion, IEEE Signal Processing Magazine 16 (3) (1999) 66–84.

[53] R.-I. Hartley, A. Zisserman, in: Multiple View Geometry in Computer Vision, 2 Edition, Cambridge University Press, 2004.

[54] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, R. Koch, Visual modeling with a hand-held camera, International Journal of Computer Vision 59 (3) (2004) 207–232.

[55] D. Hoiem, A.A. Efros, M. Hebert, Automatic photo pop-up, ACM Transactions on Graphics 24 (3) (2005) 577–584.

[56] S. Lee, D. Feng, Gooch, perception-based construction of 3D models from line drawings, in: Proceedings of the Symposium on Interactive 3D Graphics and Games, 2008.

[57] P. Parodi, G. Piccioli, 3D shape reconstruction by using vanishing points, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (2) (1996) 211–217.

[58] E. Arce, J. Marroquin, High-precision stereo disparity estimation using HMMF models, Image and Vision Computing 25 (5) (2007) 623–636.

[59] N. Atzpadin, P. Kauff, O. Schreer, Stereo analysis by hybrid recursive matching for real-time immersive video conferencing, IEEE Transactions on Circuits and Systems for Video Technology 14 (2004) 321–334.

[60] S.B. Kang, R. Szeliski, J. Chai, Handling occlusions in dense multi-view stereo, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2001, pp. 103–110.

[61] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision 47 (2002) 7–42.

[62] M.Z. Brown, D. Burschka, G.D. Hager, Advances in computational stereo, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (8) (2003) 993–1008.

[63] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, R. Szeliski, A comparison and evaluation of multi-view stereo reconstruction algorithms, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 519–528.

[64] V. Kolmogorov, R. Zabih, Computing visual correspondence with occlusions using graph cuts, in: International Conference on Computer Vision (ICCV), July 2001.

[65] M. Bleyer, M. Gelautz, A layered stereo matching algorithm using image segmentation and global visibility constraints, ISPRS Journal of Photogrammetry and Remote Sensing 59 (3) (2005) 128–150.

[66] O.P. Gangwal, R.P. Beretty, Depth map post-processing for 3D-TV, digest of technical papers, in: International Conference on Consumer Electronics, 2010.

[67] 〈http://vision.middlebury.edu/stereo/〉.

[68] C. Zitnick, T. Kanade, A cooperative algorithm for stereo matching and occlusion detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (7) (2000) 675–684.

[69] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G.B. Akar, G. Triantafyllidis, A. Koz, Coding algorithms for 3DTV—a survey, IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Multiview Video Coding and 3DTV 17 (11) (2007).

[70] B. Bartczak, R. Koch, Dense depth maps from low resolution time-of-flight depth and high resolution color views, in: Proceedings of the ISVC 09, Las Vegas, Nevada, USA, December 1–3, 2009.

[71] A. Smolic, K. Mueller, P. Merkle, A. Vetro, Development of a new MPEG Standard for Advanced 3D Video Applications, in: ISPA 2009, 6th International Symposium on Image and Signal Processing and Analysis, Salzburg, Austria, September 2009.

[72] P. Merkle, A. Smolic, K. Mueller, T. Wiegand, Efficient prediction structures for multiview video coding, IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Multiview Video Coding and 3DTV 17 (11) (2007).

[73] A. Smolic, K. Mueller, P. Merkle, M. Kautzner, T. Wiegand: 3D video objects for interactive applications, in: EUSIPCO 2005, Antalya, Turkey, September 4–8, 2005.

[74] M. Bourges-Sévenier, E.S. Jang, An introduction to the MPEG-4 animation framework extension, IEEE Transactions on Circuits and Systems on Video Technology 14 (7) (2004) 928–936.

[75] K. Mamou, N. Stefanoski, H. Kirchhoffer, K. Mueller, T. Zaharia, F. Preteux, D. Marpe, J. Ostermann, The new MPEG-4/FAMC Standard for Animated 3D Mesh Compression, in: 3DTV-CON'08, May 28–30, 2008, Istanbul, Turkey.

[76] M. Bertalmío, G. Sapiro, V. Caselles, C. Ballester, Image inpainting, in: Proceedings of the of ACM SIGGRAPH, New Orleans, USA, July 2000.

[77] A. Hornung, L. Kobbelt, Interactive pixel-accurate free viewpoint rendering from images with silhouette aware sampling, Computer Graphics Forum 28 (2009) 2090–21038 28 (2009) 2090–2103.

[78] S.-W. Shih, J. Liu, A novel approach to 3-D Gaze tracking using stereo cameras, IEEE Transactions on Systems, Man, and Cybernatics—Part B: Cybernatics 34 (1) (2004).

[79] J. Konrad, M. Halle, 3-D displays and signal processing—an answer to 3-D Ills? IEEE Signal Processing Magazine 24 (6) (2007).

[80] A. Woods et al., State of the art in stereo and auto-stereo 3D displays, Invited Paper, in: Proceedings of the IEEE, Special Issue on 3D Media and Displays, in press.

[81] T. Balogh, T. Forgacs, O. Balet, E. Bouvier, F. Bettio, E. Gobbetti, G. Zanetti, A large scale interactive holographic display, in: Proceedings of the IEEE VR 2006 Workshop on Emerging Display Technologies, Alexandria, VA, USA, March 25–29, 2006.

[82] B. Javidi et al., Integral Imaging 3D Display Technologies, Invited Paper, in: Proceedings of the IEEE, Special Issue on 3D Media and Displays, in press.

[83] L. Onural et al., Holographic 3D Display Technologies, Invited Paper, in: Proceedings of the IEEE, Special Issue on 3D Media and Displays, in press.

Aljoscha Smolic received the Diploma in -Eng. degree in Electrical Engineering from the Technical University of Berlin, Germany in 1996 and the Dr. Eng. degree in Electrical and Information Engineering from Aachen University of Technology (RWTH), Germany, in 2001.

He joined Disney Research Zurich, Switzerland, in 2009, where he is employed as a Senior Research Scientist and Head of the "Video of the Future" group. Prior to that he was the Scientific Project Manager at the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI), Berlin, also heading a small research group. He has been involved in several national and international research projects, where he conducted research in various fields of video processing, video coding, computer vision and computer graphics and published more than 90 referred papers in these fields. In current projects he is responsible for research in 2D video, 3D video and free viewpoint video processing and coding.

In 2003–2009, he was teaching at the Technical University of Berlin full lecture courses on Multimedia Communications and Statistical Communications Theory. Since 2009 he is teaching a full lecture course on Multimedia Communications at ETH Zurich. He was a Visiting Professor at Universitat Politècnica de Catalunya (UPC), Universidad Politècnica de Madrid (UPM) and Universitat de les Illes Balears (UIB) teaching full lecture courses on 3D video and free viewpoint video processing. He supervised more than 25 Master's and Bachelor's Theses at several German universities and was involved on 8 doctorate committees in Germany, France, Spain, Sweden, The Netherlands and Finland.

Dr. Smolic received the "Rudolf-Urtlel-Award" of the German Society for Technology in TV and Cinema (FKTG) for his dissertation in 2002. He is Area Editor for Signal Processing: Image Communication and served as Guest Editor for the Proceedings of the IEEE, IEEE Transactions on CSVT, IEEE Signal Processing Magazine and other scientific journals. He is Committee Member of several conferences, including ICIP, ICME and EUSIPCO and served in several Chair positions of conferences. He chaired the MPEG ad hoc group on 3DAV pioneering standards for 3D video, and further working groups of MPEG and the JVT developing standards for 3D video. In this context he also served as one of the Editors of the Multi-view Video Coding (MVC) standard.