Three-Dimensional Video Postproduction and Processing

Components and processes in the 3-D content creation workflow and advanced algorithms applied to captured signals to create high-quality content are elaborated on in this paper.

By Aljoscha Smolic, Peter Kauff, Sebastian Knorr, Alexander Hornung, Matthias Kunter, Marcus Müller, and Manuel Lang

ABSTRACT | This paper gives an overview of the state-of-the-art in 3-D video postproduction and processing as well as an outlook to remaining challenges and opportunities. First, fundamentals of stereography are outlined that set the rules for proper 3-D content creation. Manipulation of the depth composition of a given stereo pair via view synthesis is identified as the key functionality in this context. Basic algorithms are described to adapt and correct fundamental stereo properties such as geometric distortions, color alignment, and stereo geometry. Then, depth image-based rendering is explained as the widely applied solution for view synthesis in 3-D content creation today. Recent improvements of depth estimation already provide very good results. However, in most cases, still interactive workflows dominate. Warping-based methods may become an alternative for some applications in the future, which do not rely on dense and accurate depth estimation. Finally, 2-D to 3-D conversion is covered, which is an important special area for reuse of existing legacy 2-D content in 3-D. Here various advanced algorithms are combined in interactive workflows.

KEYWORDS | Postproduction; stereoscopic video; view synthesis; 2-D to 3-D conversion; 3-D video

I. INTRODUCTION

Year 2010 was the year of breakthrough for 3-D video. Numerous 3-D movies were released to theaters and created large revenues. The most successful film ever Avatar is now a 3-D movie. The 3-D Blu-ray entered the shops and stereo displays the living rooms. Major events were broadcast in 3-D and first commercial 3DTV channels are on air now. Global players in consumer electronics assume that in 2015 more than 30% of all high-definition (HD) panels at home will be equipped with 3-D capabilities. Gamers enjoy their favorite entertainment in a new dimension. Mobile phones, personal digital assistants (PDAs), laptops, and similar devices provide us the extended visual sensation anytime, anywhere. Three-dimensional camera systems are already available for professional and private users.

Unlike in previous attempts to establish 3-D video in wide markets, there are good reasons to believe that this time it will be sustainable. Stereoscopic 3-D is this time mature enough for the market in terms of quality of technology and content. There is a clear demand from user side and attractive business opportunities are visible. The content and value chains are in place for various application scenarios.

Despite the high standards that stereoscopic 3-D has reached today, there is still room for improvement of technology and a number of central problems remain unsolved. Further, new display technology such as autostereoscopic multiview displays, integral imaging, or holographic displays as well as advanced functionalities such as free viewpoint navigation require further research and development.

In the context of the 3-D video processing chain from capture to display, this paper is focused on postproduction and processing. This covers any algorithmic manipulation to the signals after they are captured and before the data
are encoded. Related processing such as 3-D display adaptation is covered here as well. In general, such processing can be done in real time or offline, fully automatic or user assisted. Besides real-world footage as captured by cameras, also animated, synthetic content will be covered. We give an overview of the state of the art and outlook to remaining challenges and future research in the area of 3-D video postproduction and processing.

Section II gives an introduction to functionalities, use cases, and main applications covered in this paper. Then, Section III covers a couple of basic algorithms that fall in the area of low level computer vision. Methods using depth-image-based rendering (DIBR) are the most important algorithms for 3-D video processing today. This is described in Section IV. A recently developed alternative are warping-based approaches, which are then presented in Section V. A highly important specific area is 2-D/3-D conversion of existing content. Such algorithms allow us to reuse available content in 3-D, which is a highly important economic factor for content providers. Section VI elaborates on 2-D/3-D conversion. Finally, Section VII summarizes the paper and gives an outline of future research directions.

II. FUNDAMENTALS AND TARGET FUNCTIONALITIES

The history of stereoscopy dates back to 1838 when Sir Charles Wheatstone published his fundamental paper about binocular vision [94]. Although the basic principle of stereopsis is fairly simple, improper stereo 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 3-D display is a fake of the human visual system, and if not done properly, results can be annoying. Production of stereo 3-D content is therefore a difficult art that requires a variety of technical, psychological, and creative skills and has to consider perception and display capabilities. Postproduction and processing includes a variety of advanced signal processing algorithms. Often this means manipulation of captured signals and conversion into other representations and corresponding data. Specifically important is rendering of virtual views other than those captured by cameras.

In Sections II-A–K, we give an overview of functionalities, issues, and problems in the context of 3-D postproduction and processing. Then, the following sections will focus on technical aspects.

A. Stereoscopic Comfort Zone

A good summary of the practice of stereography is given, for instance, in [95]. Stereographers have to consider a variety of conditions, guidelines, and rules to create good stereo, which requires significant experience. Careful planning and execution of shootings is necessary to guarantee that the captured content will look as desired and not cause any headaches. Retakes are very expensive and in live broadcast scenarios even impossible. In addition to the required specific camera equipment, all of this makes live action 3-D productions very expensive. In animated content creation the artists have, of course, full freedom of the design, but they follow the same conditions, guidelines, and rules to create good stereo. In addition to live action productions, they have the freedom to create artistic depth effects that are not possible with real stereo cameras, but can only be realized in cumbersome postproduction today.

In principle, stereography has to ensure that all action remains within the stereoscopic comfort zone as illustrated in Fig. 1. The 3-D experience is generally comfortable, if salient objects stay close to the screen (bright areas in Fig. 1). Darker areas should be visited only occasionally for specific purposes, for instance, to stress important scenes of a movie by a lot of depth atmosphere, or to let objects come far out of the screen to create a catchy effect. Areas of retinal rivalry have to be avoided, or may be used only for very specific effects (e.g., fast moving objects, rain). The available depth volume is restricted compared to the real 3-D world, so the main job of a stereographer is “to bring the whole real world inside this virtual space called the comfort zone” [95].

The parameters to do so are mainly the inter-axial distance or baseline and the convergence of the cameras. The camera baseline controls the overall range or scale of disparities, i.e., the depth volume of objects. Convergence controls the absolute disparity values, i.e., the position of objects relative to the screen. In the following, we list typical issues that have to be considered to create a good 3-D experience.

B. Control of Absolute Disparity

Convergence can be easily controlled by shifting the views with respect to each other. Therefore, productions can generally be done with parallel cameras in digital times today and proper convergence can be added in postproduction as desired. However, sizing effects related to depth placement still have to be considered. Changing depth position of objects without changing their horizontal and vertical size at the screen will change the 3-D size

<table>
<thead>
<tr>
<th>Fig. 1. Illustration of the stereoscopic comfort zone [95], [107].</th>
</tr>
</thead>
</table>

© 2010 Disney Enterprises
perception, which limits the amount of shift convergence that can be practically applied. Further, too much shifting of disparities without scaling (compression) may cause that the allowed disparity range is exceeded, e.g., forcing the eyes to diverge, which has to be avoided.

C. Scene Depth Adaptation

Depending on screen size and resolution, maximum disparity ranges (positive and negative) can be calculated, which should not be exceeded for pleasant 3-D experience [95]. These are related to the human eye distance. Usually directors use the full range only occasionally, for example, to stress important and emotional scenes, and to give the audience enough time to relax the visual system during scenes with limited depth. Such depth budgets are part of careful planning during 3-D productions. Transitions have to be considered carefully as well. Drastically changing depth should be avoided. Specific design of scenes and editing of transitions are usually applied, such as reconverging the shots towards each other to have salient scene elements at similar depth over the transition. Among others such as focal length, the disparity range is mainly defined by the inter-axial distance (baseline) of the stereo cameras. In case of real stereo content, the stereo baseline is fixed during capturing and cannot be changed afterwards. However, the adaptation of the scene depth after the content was captured is of utmost importance for stereo processing. In practice, this means virtual view interpolation from the given stereo pair.

D. 3-D Display Adaptation

A similar problem arises again, when the content has to be reformatted for different viewing conditions (screen size and distance), e.g., from cinema to TV, which requires again an adaptation of the disparity range. Three-dimensional content that was optimized for certain viewing conditions consisting of display size and viewing distance will look different for other viewing conditions. Stereographers often select the baseline during shooting in dependence on the presence factor (ratio of screen width to viewing distance) of the targeted media platform. In order to keep a good 3-D impression and to preserve the artistic intention, depth composition has to be modified, when viewing conditions are changed. Both 3-D reproduction and binocular 3-D perception highly depend on the viewing conditions, i.e., the viewing distance \( r \) and the screen size or better, the on-screen disparity \( d \). This is illustrated in Fig. 2 and expressed by

\[
z = \frac{r \cdot d}{e - d}.
\]

The reproduced depth \( z \) is a linear function of the viewing distance \( r \). This means that if we change \( r \) also the z-extension of objects changes whereas their horizontal and vertical extensions at the screen are kept unchanged. A cube is not reproduced as a cube anymore but stretched if we increase the viewing distance. Hence, depth impression varies depending on the distance in a cinema theater and even more if conditions are changed completely, e.g., using a TV set and a living room setting or a handheld device. In theory, the relation between real 3-D scene geometry and 3-D reproduction at stereoscopic screens is even more complex and known from literature as stereoscopic 3-D distortion [97]. More details can be found in [108]. In principle, however, these distortions can be compensated and 3-D impression can be adapted to the viewing distance, if the depth information of the scene is known [109]. In this case, virtual view interpolation can be used to render a new stereo pair with adapted baseline fitting to given screen size and viewing distance. The same algorithms can be used to allow the user of, e.g., a 3DTV set to control depth impression as desired, by virtually manipulating the baseline.

E. Local Disparity Adaptation

Positioning of objects within the stereoscopic window (screen) also has to be considered carefully. Cropped objects that are floating in front of the screen and intersect with the image borders will cause retinal rivalry (the so-called stereoscopic window violation). In postproduction, this can be corrected using the floating window technique (also called proscenium arch), which is a virtual shift of the screen plane towards the viewer [95]. However, in live broadcast applications, automatic detection and correction mechanisms are necessary. The artistic freedom to redesign the depth structure of a scene on object basis in postproduction is in general a highly desired feature and should allow to keep and adjust volume, size, and shape of objects while changing depth. To achieve this flexibility, scenes are sometimes captured with multiple camera rigs, each appropriate for a certain depth range (regarding object volume) and then composed into a final stereo pair (multirigging). Such cumbersome and expensive
acquisition and postproduction can be avoided if object-based depth editing is possible.

**F. Automatic Correction and Manipulation of Stereo Live Broadcast**

Live broadcast is a highly important application scenario for 3DTV services. For instance, live sport events have often been driving forces for adoption and distribution of new TV technology. Unfortunately, the sport stereo live broadcast production is also very difficult. Since no corrections are possible in such a real-time scenario, any violation of the stereoscopic comfort zone will be directly visible to the audience and may create a very bad impression.

Consider, for instance, a football game, where a midangle shot is used to cover some action on the field. Stereo parameters of the camera system (baseline and convergence) are adjusted accordingly. If now the camera rotates fast due to fast action in the game, it may happen that players, referees, spectators, etc., who are too close to the camera, get into the picture. This will cause bad stereo including window and range violation. The 3-D impression will be lost and leave a bad and unpleasant experience. Such cases happen quite frequently and are a real problem for live broadcast of sporting events. It is very difficult to avoid such situations in live capturing of fast and unpredictable action that takes place over a wide depth range.

In order to keep dramaturgic freedom similar to what we know from 2-D productions today, and to capture and experience all action and excitement of live sports with a new dimension, technology is required that would assist the crews on set to avoid such situations and to correct any stereographic error automatically in real time. This is a very difficult task and again would require some kind of novel view synthesis to correct the errors.

**G. Mixing and Composition of 3-D Material, Real and Animated Content**

Practically all animated movies are released to theatres in 2-D and 3-D versions today. In this case, artists have all the freedom to design 3-D scenes that are visually pleasing and exciting without violating the rules of stereography. Also, most other 2-D cinema and TV content today does not contain only pixels that are captured by one single camera. Images are composed out of content from different sources such as different real cameras, graphical elements, animated parts, special effects, etc. This processing can also be automatic or user assisted, real time, or offline. Algorithms, tools, and workflows are established but also continuously extended and improved. Some of this technology is outlined in Section V.

Third dimension adds a new dimension to this and complicates things further. For instance, graphic overlays cannot be simply pasted over other footage. Depth composition must be considered. If a graphic overlay is placed, for instance, over some scene element but behind it in depth, the results will be annoying, conflicting, and unrealistic. To avoid this, graphics or subtitles are placed far out of the screen in many productions today to make sure that all other action stays behind them. Still in live broadcast collisions may happen. In consequence, knowledge about depth composition of all source material is necessary for mixing and composition in 3-D. Reliable depth estimation and view synthesis is crucial for stereography.

**H. Content Creation for Autostereoscopic Multiview Displays**

So far, discussion was focused on stereoscopic 3-D using exactly two views targeted for the user’s eyes. Technology and content creation for stereoscopic 3-D has reached a high level of maturity, which is why we see corresponding markets growing fast. However, today’s stereoscopic 3-D technology also has limits. The necessity to wear glasses and the inability to reproduce motion parallax are often considered as major obstacles.

More advanced display systems such as autostereoscopic multiview displays [98] or even displays based on integral imaging [99] and holography [100] are capable to overcome these obstacles in principle. However, this is paid by the price of potentially many more than two views to be displayed. Content creation for such displays is still widely a research task. For many applications it seems unrealistic to capture all the views necessary for display. A common idea is instead to use a subset of real views and to generate the remaining content by view synthesis [102]. This principle is illustrated in Fig. 3.

Perception and depth reproduction issues in the context of autostereoscopic multiview and other advanced 3-D displays are different from glasses-based stereoscopic systems. For instance, the available depth range that can be reproduced is much more limited. Otherwise, annoying ghosting artifacts may become visible. This results in different rules of stereography, which also have to be considered, when reformatting stereo content for autostereoscopic multiview displays. Often compression of depth based on view synthesis will be necessary.

**I. Enabling Free Viewpoint Video Functionalities**

The term free viewpoint video (FVV) refers to a functionality known from computer graphics where the user has the freedom to watch a 3-D scenery from arbitrary viewpoints and directions, within a certain operating range [103]. In contrast to computer graphics, FVV is related to real-world scenery as captured by real cameras. In principle, the advanced displays described in Section II-H already provide a limited FVV functionality, as the views change when moving in front of the screen. FVV relies on some kind of a 3-D data representation reconstructed from real-world footage [101]. Then, view synthesis is the key to provide the functionality.

Today some FVV systems are used in production to create special effects. Providing FVV to end users is still widely a research topic.
J. 2-D/3-D Conversion
Conversion of available 2-D content for rerelease in 3-D is a highly important topic for content providers and for success of 3-D video in general. It naturally completely relies on virtual view synthesis of a second view given the original 2-D video. Due to the importance, the whole Section VI is devoted to 2-D/3-D technology.

K. Virtual View Synthesis and Generic 3-D Video Format
In consequence, the possibility to manipulate the depth of stereo content after it was captured, regarding global range and on local object basis, will solve and ease a lot of problems that we face in stereography and 3-D display today. Virtual view synthesis is the key algorithmic functionality for 3-D postproduction and processing. Robustness and quality of synthesized views are decisive factors for design of solutions for specific applications, which may be automatic or user assisted, real time, or offline.

Research is ongoing to define a generic 3-D video format for production and distribution that would ideally support all extended 3-D video functionalities in an efficient way [96]. Such a format may include, e.g., at least a stereo pair (or more views) and additional associated data such as depth maps, occlusion layers, camera calibration information, etc. An example is illustrated in Fig. 3. Depth-enhanced stereo (DES) enables a variety of advanced functionalities as outlined in Sections II-A–J via view synthesis, including baseline adaptation, support of autostereoscopic displays, free viewpoint functionalities, etc. MPEG is currently preparing a new related coding standard.

III. BASIC ALGORITHMS
A lot of basic signal processing is required in conventional stereo production to avoid some of the above mentioned perception conflicts such as binocular rivalry and to allow comfortable stereo viewing. This section will give an overview of this basic processing. In contrast to the algorithms described in the following sections, they do not need special information about the depth structure of the captured scene or about point correspondences between the two stereo images. They mainly concentrate on correction of geometric distortions, color matching, and assistant system to obtain the right stereo geometry.

A. Correction of Geometrical Distortions
One main requirement of a good stereo production is that stereo images are free of any geometrical distortions and that they contain horizontal disparities only. This assumes a classical stereo viewing setting where the viewer is sitting upright and centered relative to the screen. However, having horizontal disparities only is an ideal solution that cannot be achieved in practice. Every stereo rig contains parts of finite mechanical accuracy such that the stereo images will not necessarily be aligned correctly in vertical direction. In addition, some stereographers

![Fig. 3. Depth-enhanced stereo (DES), extending high-quality stereo with advanced functionalities based on view synthesis.](image-url)
prefer convergent camera settings instead of parallel stereo geometry what causes additional keystone distortions in the stereo images. Finally, real lenses usually impose radial distortions. Moreover, the focal length might differ slightly between the two cameras and, when changing the focus, the internal lens parameters such as the focal length might be affected. Furthermore, when using zoom lenses, the principal point shifts, the focal length is changed over a wide range of values. The motors for zoom level and focus do not synchronize perfectly, such that slightly different focal lengths may also occur in this case.

Related distortions can be corrected in two subsequent operational steps. In the first step, distortions of the used lenses can be measured separately by using professional photometric tools [11]. That especially refers to lens distortions, deviations in focal length, and the position of the principal point. As in some cases distortions depend on lens settings (e.g., focus or zoom), measurements have to be done under different operational conditions in this case, and measurements results have to be stored in lookup tables for later processing.

As soon as the distortion parameters are known for a given operational shooting mode (e.g., by taking them as operational metadata from the look-up table), they can be used for corrections and normalization of the stereo views during postproduction. Furthermore, applying the process to RGB channels separately also allows for the correction of chromatic aberrations (see Fig. 4).

A special problem in 3-D shooting is to correct lens distortions consistently for both cameras. Usually, a separate application of photometric or other optical measurement tools to single lenses is sufficient for a first-order correction. Many lens manufacturers have therefore announced for the future to measure lens distortions directly during fabrication and to store them as a digital footprint at an onboard chip. Nevertheless, it might be useful to measure remaining second-order distortions jointly for a particular stereo pair after rigging. This especially applies to stereo rigs with synchronized zoom lenses where remaining deviations from the corrected state can then be stored in lookup tables in dependence on focal length and focus. After geometrical normalization, the two stereo views are free of lens distortions and chromatic aberrations, they have the same focal lens, and the principal point (intersection of optical axis with image plane) is always in the center of the images. Hence, further distortions such as keystones caused by convergent camera settings or vertical misalignments by improper stereo rigging can be corrected by a regular rectification process as known from literature in a second step. This can either be done by using data from a stereo calibration with suitable calibration charts or, what is even better and usually preferred for onset processing, by using implicit calibration based on robust detection and tracking of feature point correspondences [12]–[15].

In specific scenarios, more than two cameras are used to simultaneously capture content. In such cases, multicamera rig calibration has to be applied, which further complicates systems and algorithms [105].

B. Color Matching

Color correction and color grading are already important processing steps in conventional 2-D postproduction and all postproduction suites usually offer sophisticated tools for these purposes. In stereo postproduction, however, this process is much more complicated. Apart from the fact that the process has to be done twice and the two processes have to be coordinated in a satisfying manner, the new problem of color matching arises.

One main reason for binocular rivalry is asymmetry in luminance, contrast, and color. Significant colorimetric differences between the two stereo views of a stereoscopic image pair can lead to eyestrain and visual fatigue [16]. Although the impact of colorimetric differences is not fully investigated and understood yet, there is a common consensus in literature that the two stereo images should be as identical as possible, especially in contrast, but with some lower criticality also in luminance and color. To this end, if displayed side by side no discrepancy should be visible between stereo images [16], [17].

Color matching can be done at different places of the production chain. As much as possible should already be done directly at the stereo rig by suitable colorimetric adjustment of the two cameras before capturing, either manually or with some assistance from special visualization modes (RGB parade, vectorscope, checkerboard, etc.) or automatically by analyzing color deviations.

A simple method for automatic adjustments is histogram filtering [18]. For this purpose, RGB histograms are measured for a sufficiently large number of matching feature points representing the same parts of the scene. The colorimetric fine adjustment is then based on fitting quantiles between the RGB histograms of both cameras and deriving related lookup tables from quantile fitting for nonlinear color correction. Fig. 5 shows the results of such a histogram-based color matching.

![Fig. 4. Examples of measuring and correcting geometrical lens distortions and chromatic aberration.](image-url)
Sophisticated color matching can be done during primary and secondary color grading of professional post-production. Many postproduction suites already offer simultaneous processing of two stereo sequences during color grading. Visualization tools such as RGB parade, vectorscope, or waveforms are used for color matching in this case. Fig. 6 shows an example.

C. Adjustment of Stereo Geometry

As already mentioned in Section II, the stereo geometry is mainly defined by the inter-axial distance and the convergence of the stereo cameras. Both have to be selected and balanced carefully during production to achieve good stereo content.

Most of all stereo productions are captured by a parallel stereo geometry, i.e., with parallel camera axes and the same camera orientation. Even in case of convergent cameras, the geometry will be converted into a corresponding parallel setup by means of rectification as explained in Section III-A. Note that shooting with parallel camera geometry means that the convergence plane (i.e., the zero-disparity plane) would appear infinite, if no further postprocessing is applied. Thus, if stereo captured with a parallel setup would be directly displayed on a 3-D screen, the entire scene would appear in front of the screen. To avoid this undesired situation, convergence is usually adapted during postproduction by shifting the rectified images horizontally in contrary directions. As left and right images do not fit any longer in this case, they have to be cropped and up-scaled to fill again the entire image frame. Therefore, this process is often called “shift–crop–scale” in stereo postproduction (see Fig. 7). The resulting stereo pairs should avoid stereoscopic window violations (proscenium arch, see Section II-E).

Whereas convergence adjustment by shift–crop–scale can be done during postproduction, the stereo baseline has to be fixed during shooting. The selection of the right baseline is of similar importance as the selection of the right convergence plane. Both parameters have to fit together to take full advantage of the available depth budget of given viewing conditions and to control the stereoscopic comfort zone in an optimal manner.

Therefore, the usage of camera assistance systems has recently been proposed to support stereographers and camera teams at the set. Examples are the StereoBrain 3-D Processor from Inition, the SIP 2100 from 3ality, or the STAN from Fraunhofer HHI [18]. The objective of all these systems is to analyze the disparity range of a given scene at the set before or while capturing, to visualize how it fits into the stereoscopic comfort zone and to preview the effect of the shift–crop–scale in dependence on the selection of the convergence plane \( Z_{\text{conv}} \). Fig. 8 shows an example of a previsualization at the STAN. Another system that predicts specific distortions of a given stereo pairs was presented in [110]. An interactive stereo editing system was recently published in [111].
IV. DEPTH-BASED METHODS

Information on the depth structure of a scene always plays an important role in postproduction. That is already well known from conventional TV and cinema production. For instance, together with camera calibration and tracking, depth information is frequently used for visual effect editing, depth keying, augmented reality, scene composition, mixing of computer graphics data with real content, virtual studio productions, and currently also for advanced chroma keying and blue/green screening in news or weather forecast.

In 3-D postproduction, however, the role of depth information becomes even more important in a manifold manner. Many functionalities in the context of stereo processing outlined in Section II require virtual view synthesis. Today, DIBR is the most important class of algorithms to perform such high-quality virtual view synthesis, as shown, for instance, in [112]. Therefore, this section gives an overview of related algorithms and some examples of depth estimation and view synthesis.

A. Estimation of Depth Structure From Stereo

The extraction of depth information from two or more views of a fixed camera setup capturing a given scene from different perspectives simultaneously—also known as computational stereo or structure-from-stereo (SfS)—has been an important issue of computer vision research for years [41]–[51]. Often, this process results in a dense pixel-by-pixel depth map for each camera view. The computation of these depth maps is, in general, a three-step process consisting of camera calibration, stereo matching, and depth reconstruction. If the internal and external camera parameters are known from initial camera calibration, disparities from the intermediate stereo matching process can be turned directly into depth values during the final depth reconstruction.

While initial camera calibration and final depth reconstruction are well understood, solving the correspondence problem of stereo matching is still under investigation and evaluation in the computer vision community. As it is an ill-posed and under-constrained problem, assumptions in the form of imposed constraints, such as smoothness, uniqueness, consistency, and occlusion handling, are used to make the problem computationally feasible. In this context, literature coarsely distinguishes between two different categories, local and global methods [45]–[47]. Local methods such as block matching or optical flow estimation define the constraints for small regions surrounding the pixel of interest for which correspondences in another view is searched [42], [43]. In contrast, global methods define constraints for the entire image in the form of a cost function, which is then minimized by means of global optimization such as graph cuts, belief propagation, or simulated annealing [45], [48].

Although global methods have made substantial progress during the last decade and often outperformed local methods [56], their use for postproduction of 3-D video has been limited. Until now, their advantage over local methods could only be proven for still stereo images. The most important criterion for depth-based processing of video data, however, is temporal consistency of depth maps—a constraint that has been difficult to achieve by global methods so far. First approaches have been presented in [116] and [117]. Temporally inconsistent depth data may cause perceptible and very annoying flickering artifacts during image-based rendering and novel view synthesis and should be therefore avoided in any case. In addition, global optimization is not well suited for 3-D video processing in general due to its high computational complexity.

As a consequence, spatially local methods that are less complex than global methods are so far usually more suitable for 3-D video processing [42], [52]. Most of them are window-based algorithms (e.g., block matching) that exploit implicit spatial and temporal smoothness constraints and allow high amount of spatial and temporal consistency of the depth maps by applying recursive matching structures in horizontal, vertical, and temporal directions while scanning the 3-D video sequences. The main disadvantage of spatially local methods is the so-called foreground fattening, which leads to poor performance at object borders; i.e., the transitions in the estimated depth maps do not fit very well to the real object boundaries. Further problems are mismatches caused by ambiguities in nontextured regions or periodic textures and/or regions with similar texture. However, these problems can be handled by
applying color segmentation and using the segment information to refine disparities at object borders and to regularize disparities in regions with low or periodic textures.

B. Postprocessing and Regularization of Depth Maps

Several methods have been proposed to postprocess estimated depth maps. Although they are very different approaches, they all share the same objectives—aligning depth discontinuities to the object borders, removing noise and mismatches, and filling occlusions. Many approaches are based on the idea of image segmentation and plane fitting to refine an initial depth estimate [49]–[51]. Others use some kind of neighborhood filters to compute a weighted average of the disparities in the filter window. In [52], a down-sampled depth map is used. It is processed by joint-bilateral filters against a corresponding down-sampled image. The filtered depth map is then up-sampled with a full-resolution image, again by using joint-bilateral filters. The down-sampling reduces the details within objects and removes small outliers in the input depth map. In [53], intensity gradient information is first extracted from the preprocessed intensity image. This gradient information is then used to guide an edge-preserving anisotropic diffusion process on the input depth map. In parallel, a texture analysis detects critical regions with ambiguities (homogeneities, similarities, periodicities, etc.) that might cause mismatches.

C. Depth-Image-Based Rendering (DIBR)

Complementary to depth estimation, the synthesis of new virtual stereo views by image-based rendering also belongs to the research of computer vision. Depending on known geometric properties, different approaches exist in the literature, e.g., image-based rendering (IBR), depth-image-based rendering (DIBR), layered-depth images (LDI), and intermediate view reconstruction (IVR).

DIBR [1], LDI [64], and IVR [49], [65] are the most important in the given context. They need dense pixel-by-pixel depth maps to render a novel view. Additionally, LDI stores several color and depth values for each pixel to compensate possible occlusions. Finally, IVR renders intermediate views between neighboring images with associated dense depth information. Thus, the occlusion problem almost does not exist in IVR.

Recent advances in DIBR focus, for instance, on specific handling of depth discontinuities to reduce artifacts along object borders [102], [104]. Further, matting algorithms are applied to handle mixed pixels and to obtain realistic depth boundaries [49]. An editing system for enhancement of video using information captured by photographs based on depth has been presented in [113]. Examples of pure image-based view interpolation without usage of depth are given in [114] and [115], however, suitability for the given problems in 3-D video processing is limited.

D. Example Application of Depth-Based Methods

As an example, Fig. 9 shows the processing scheme of a depth estimator using hybrid recursive matching (HRM) in combination with adaptive cross-trilateral median filtering (ACTMF) for postprocessing.

The HRM algorithm that is used for initial depth estimation is based on a hybrid solution using both spatially and temporally recursive block matching and pixel-recursive depth estimation [42]. Due to its recursive structure, the HRM algorithm produces almost smooth and temporally consistent pixel-by-pixel disparity maps. In the structure from Fig. 9, two independent HRM processes (stereo matching from right to left and, vice versa, from left to right view) estimate two initial disparity maps, one for each stereo view. The confidence of disparity estimation is then measured by checking the consistency between the two disparity maps and by computing a confidence kernel from normalized cross correlation used by HRM. In parallel, a texture analysis detects critical regions with ambiguities (homogeneities, similarities, periodicities, etc.) that might cause mismatches.

Finally, the two initial disparity maps are postprocessed by using ACTMF

$$D_0 = \text{weighted}_\{p \in \Omega\} \{w_p, D_p\}. \quad (2)$$

**Fig. 9.** Depth estimator using hybrid recursive matching (HRM) and adaptive cross-trilateral median filtering (ACTMF).
The adaptive filter is applied to initial disparity values $D_{ip}$ at pixel positions $p$ within a filter region $n_s$ around the center position $s$. Moreover, $D_{os}$ denotes the value of filtered output disparity at position $s$ and $w_p$ the weighting factors at position $s$ of the adaptive filter kernel

$$w_p = \text{conf}(D_{ip}) \cdot \text{dist}(p - s) \cdot \text{seg}(I_p - I_s).$$ (3)

Following the definition of a weighted median operator, the weighting factors are used to increase the frequency of a particular disparity value $D_{ip}$ before applying a standard median operation. The weighting factors depend on three adaptive smoothing terms, all ranging from 1 to 10. The first one aggregates all results from previous consistency checks, computed confidence kernels, and texture analysis to one overall confidence term $\text{conf}()$ for the initial disparity value $D_{ip}$ at position $p$. A high confidence measure is scored by 10 and, vice versa, a low confidence by 1. The second one is a distance function $\text{dist}()$ with scores, which when they are nearer to 10, the closer the position $p$ is to the center position $s$ within the filter area. The third one describes a segmentation term $\text{seg}()$ that scores high, if the input disparity $D_{ip}$ at position $p$ belongs with high probability to the same image segment as the filtered output value $D_{os}$ at position $s$, and low if not. As usual, in conventional bilateral filtering, this term is driven by the difference between the color values $I_p$ and $I_s$ at positions $p$ and $s$, respectively.

In concept from Fig. 9, ACTMF is used iteratively and the left–right consistency check and the related calculation of the confidence term are repeated at every iteration step. Final results can be achieved after a few iterations. Fig. 10 shows an example for MPEG test sequence “Café.” The picture on the top shows the left image of the original stereo pair, the bottom left picture shows the initial disparity map after HRM processing, and the bottom right picture shows the filtered disparity map after ACTMF.

In addition, Fig. 12 shows the results from view interpolation at three equidistant positions between the two stereo views, calculated from the input stereo images and the estimated depth maps from Fig. 11.

This kind of depth estimation is usually sufficient to produce robust depth maps and to achieve good image quality during DIBR. Nevertheless, in case of critical content and high-quality postproduction, it might be necessary to supervise the entire process and to refine the results by additional manual work. In this case, rotoscoping of critical image elements can be used to enhance the depth maps at object borders.

Further, new approaches combine color cameras with specific depth sensors that directly capture scene depth [106]. Typical time-of-flight (ToF) depth sensors are of 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 cause 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 [106].
V. WARPING-BASED METHODS
In principle, depth-based approaches as presented in the previous section allow for an accurate synthesis of novel output images in postproduction. In practice, however, tasks such as dense depth reconstruction and inpainting of occluded image regions cannot yet be solved in a fully automatic manner, since these are computationally complex and underconstrained problems. Correspondingly, the amount of required manual intervention is often considerable and one of the main reasons why stereoscopic postproduction is very expensive and cumbersome.

An alternative to depth-based methods for novel view synthesis are image-based warping techniques, i.e., methods that deform or morph image content directly in image space without reprojection over a dense 3-D depth proxy [19], [20]. These methods generally work by defining feature correspondences (e.g., pixels, lines, etc.) between a source and a target image. Novel images can then be generated by first computing new (e.g., interpolated) positions of these correspondences, and then warping the remaining image content according to a non-linear deformation field. Typically, this deformation field is computed from the feature constraints, low level visual saliency information, and high level constraints provided by the user [21].

In recent works, this type of image warping has shown to be a practical and powerful solution to a variety of problems, which were difficult to solve using previous depth-based view interpolation techniques. For example, image warping can be used to retarget images and video between different aspect ratios and resolutions [21], for undistorting and optimizing image content [22], and stabilization of 3-D camera paths in postproduction [23]. Although the resulting output images are generally not as geometrically accurate as the depth-based methods, they often produce visually plausible results, since our visual system and perceptual compensation mechanisms are tolerant when interpreting image content with slight visual inconsistencies. In many application scenarios such as the ones mentioned above, this type of visual plausibility is fully sufficient.

A simple example for novel view generation using image-based warping [107] is illustrated in Fig. 13. In the middle image, the statue shown in the original left image has been moved slightly to the right while the background sky remains mostly fixed. Note that neither segmentation nor inpainting of occluded image regions has been applied. Instead the background image content is compressed and stretched by a non-linear warp function (visualized in the right image) in a narrow band around the statue. When viewed cross-eyed, the left and middle images can be fused as a stereoscopic image pair and provide a basic impression of depth.

The considerable advantage is that warping-based methods are conceptually simple and can often be computed fully automatically and in real time, without the need for accurate camera calibration, dense depth estimates, inpainting, or other error prone processing steps required for depth-based view interpolation. More research is still required in this field, but future warping-based methods could be an alternative to depth-based view interpolation techniques. One important open question in this context is how far the tolerance of users for such warping operations is given. Certainly, at some point artifacts will become visible and psycho-visual effects such as fatigue may arise.

VI. 2-D/3-D CONVERSION
The 2-D/3-D conversion is currently a hot topic for 3DTV applications and the 3-D cinema entertainment industry because of the lack of 3-D content for the new era of 3-D displays and 3-D cinema systems. Much research and development has been done to increase the conversion speed and quality during the last years.

Two dominant deployments can be noticed that target two different markets: 3-D home entertainment and 3-D theatrical entertainment. In Fig. 14, an overview of the conversion approaches and their underlying applications is presented.

These approaches can be classified into manual and automatic conversion methods. Table 1 roughly compares these methods regarding their conversion speed and quality.

On the one hand, companies such as JVC, Samsung, and Dynamic Digital Depth focus on real-time automatic conversion algorithms to bundle their technology with the new 3-D-capable display systems. This enables consumers to watch their conventional DVD- or Blu-ray library in stereoscopic 3-D. However, many people believe that this technology is just a gimmick since the conversion quality is quite bad and often causes headaches.

On the other hand, more complex conversion systems both automatic and manual try to convert footage for theatrical release in a most accurate way. Basically, these conversion methods are extremely labor intensive and heavily depend on the workflow and degree of automation to be cost efficient.

So far, there are just a few movie titles converted from 2-D to 3-D, such as Nightmare before Christmas, G-Force, Lions 3-D: Roar of the Kalahari, and parts of Harry Potter to mention some of them.

For 2-D/3-D conversion, depth cues are needed to generate a novel stereoscopic view for each frame of an input sequence. The human brain can decode two kinds of depth cues: monocular and binocular depth cues.
Monocular depth cues are focus and defocus, perspective, aerial perspective, relative size, light and shading, texture, motion parallax, and interposition. The latter are accommodation, convergence, and binocular discrepancy. Only the monocular depth cues are visible in 2-D content and can be utilized in further processing steps.

Sections VI-A–C give a comprehensive overview of the 2-D/3-D conversion workflow and state-of-the-art approaches to increase both the conversion quality and conversion speed. First, we will recapitulate the whole conversion workflow, and second, automation approaches for the several conversion steps will be presented. Finally, complete automatic conversion systems are reviewed.

### A. Manual Conversion Techniques and Workflow

Manual conversion techniques are very time consuming, and thus cost intensive. However, the more efforts one puts on the conversion steps the better is the quality of the 3-D output. For economical reasons, manual conversion techniques are mostly utilized for cinematic productions where the quality of the resulting stereo footage has to be very high. Fig. 15 shows an example of a workplace for 2-D/3-D conversion as a 3-D-postprocessing step for feature films.

<table>
<thead>
<tr>
<th>Table 1 Comparison of Manual and Automatic Conversion Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conversion</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Speed</td>
</tr>
<tr>
<td>Quality</td>
</tr>
</tbody>
</table>

The conversion workflow basically consists of three production steps:

- roto-scoping/segmentation;
- depth assignment;
- inpainting.

In the first step, each frame of a sequence is segmented (rotoscoped) into single objects or fragments depending on their location in 3-D space. Remember that 2-D content has all of the normal depth cues of the real world with the exception of accommodation, convergence, and binocular discrepancy. Thus, an operator can easily select the objects that should be roto-scoped in an image or sequence of images. There are many commercial software programs (such as Mocha, Silhouette, Eyeon Fusion, Autodesk Flame, or Adobe After Effects) available that assist an operator on this labor-intensive work (see Fig. 16). However, roto-scoping is still the most time-consuming part of the conversion pipeline. Thus, a lot of research is done on automatic and user-assisted image and object segmentation to speed up this process.

If object segmentation masks are available, one has to “rebuild” or reconstruct the scene according to the depth of the objects. This production step can be titled as “depth assignment” or “3-D compositing.” The simplest way to
generate a stereoscopic view is to shift the objects either to the left or to the right (according to their depth and zero parallax settings). However, this simple object shift results in a quite bad and unrealistic stereoscopic depth perception, i.e., the objects appear as flat layers in a 3-D environment. This effect is widely known as cardboard effect or puppet theater, which might be intended for cartoons.

To avoid this effect, i.e., to give an object a kind of roundness, the operator or “3-D compositor” needs to create displacement maps, i.e., to assign individual depth to each pixel of an object over time, and thus to the whole sequence. This can also be called “internal object depth.” Then, each pixel can be shifted according to its depth value using DIBR [1], [2] or any 3-D-warping technique.

According to the shape of an object, this depth assignment can be facilitated by using simple geometric primitives such as spheres, cylinders, cubes, or even gradients, e.g., for planes. If using a 3-D modeling and animation tool such as Autodesk Maya or 3ds Max, the object textures can be projected on such a simple geometry and reprojected into a second virtual view (the stereoscopic view).

The main challenge of this conversion step is to deal with transparencies. In [3], a system and method of treating semitransparent objects is described.

The last step of the conversion pipeline is the occlusion filling. Since objects are shifted according to their assigned depth, background information is not available at depth discontinuities. These regions have to be recovered either from surrounding frames in time, directly from spatial information within the current frame, or both. Stamping from the frame itself presents the risk of generating depth artifacts due to repetitive patterns.

The process of occlusion filling is also known as inpainting. A simple manual inpainting process can be the usage of a brush tool to retouch the missing parts. Of course, this process is even more time consuming than the rotoscoping step since the retouching needs to be consistent over time, especially in binocular vision. In many cases, one tries to remove foreground objects to create clean-plates of the background (e.g., as used in the Declipse format of Philips [4]).

However, occlusion filling is one of the most challenging parts in the 2-D/3-D conversion workflow and currently a very attractive research area to increase the quality and automatism.

Only a few companies worldwide provide a 2-D/3-D conversion service so far. Some of them hold several key patents on systems and methods to speed up the conversion workflow, e.g., In-Three, Inc. [5], [6], Conversion Works, Inc. [3], [7], Industrial Light & Magic [8], or Dynamic Digital Depth, Ltd. [9], [10].

B. Automatic and Semiautomatic Conversion Techniques

From Section VI-A, it is clear that all of the conversion steps can somehow be automatized. This section outlines algorithms that can assist the operator at least in a semi-automatic way to perform certain tasks in the aforementioned 2-D/3-D conversion workflow.

1) Segmentation: Especially in the object segmentation and tracking step a lot of research has been done during the last decades [24]–[29]. Depending on the footage, different approaches can be utilized for segmenting frames into objects or fragments [24]. For static or camera-motion compensated videos, background subtraction techniques are commonly used. Here, the background is modeled and matched with each frame of the sequence. Current approaches are based on Gaussian distributions, Gaussian mixture models (GMM), or kernel density estimation (KDE). Additionally, optical-flow-based techniques [25] are very effective if moving objects are in the scene. These techniques analyze the flow vectors to detect objects, which have an individual motion compared to the camera motion.

A further approach to segment objects within frames is image segmentation [26], which can be divided into region-based and contour-based methods. Approaches of the first category basically utilize features such as color or texture to group neighboring pixels into regions. Current techniques are based on clustering, morphology, graph theory, statistical methods, edge-based methods, or neuronal networks.

Contour-based methods apply mostly active contours [27]. After the first initialization process, active contours iteratively fit to the boundary of an object. The adaption of the contour can be controlled by internal (rigidity) and external (gradient, color, texture) factors. GrabCut is an approach to combine region-based and contour-based methods using graph cut optimization [118].

Finally, segmented objects have to be tracked over time to obtain time-consistent segmentation masks for all frames of a video sequence; see, e.g., [119]. A good overview of object tracking algorithms can be found in [28].

Video SnapCut [29] is one of the latest and most powerful algorithms in this research field and also available as plug-in for Adobe After Effects. Segmentation is achieved by the collaboration of a set of local classifiers, each classifier adaptively integrating multiple local features such as color, edge, and trained shape prior.

2) 3-D Structure Recovery (Shape From X): The recovery of 3-D structure from a single image or a sequence of images is an integral part of computer vision. Many fundamental algorithms have been developed so far. These methods can be classified into two categories: methods that recover 3-D structure from a single viewpoint (monocular 3-D structure estimation) and methods that utilize the information of two or more images captured from another perspective (stereo, motion, or multiview 3-D structure estimation).

Monocular 3-D structure estimation algorithms try to recover 3-D structure information from monocular cues,
i.e., shape from focus (SFF) and defocus (SFD) [30], [31], shape from shading (SFSh) [32], [33], and shape from texture (SFT) [34]–[40]. The latter can be further divided into structural texture analysis [34], [35], statistical texture analysis [36], [37], and methods that extract geometrical scene properties such as vanishing points and lines [38]–[40]. An in-depth analysis of a related learning-based approach with various experiments is given in [120].

If two or more views taken from a slightly different perspective are available, stereo, motion, or multiview 3-D structure estimation can be applied: structure from stereo (SfS) [41]–[57], structure from motion (SfM) [58]–[61], and shape from silhouette (SfSi) [62], [63].

SfS techniques were covered in detail in Section IV. They estimate disparities between two or more views of a fixed camera setup capturing the 3-D scene simultaneously. If the internal and external camera parameters are known, disparities can be turned directly into depth values. SfS is less important for 2-D/3-D conversion since in that case the second view is actually already available.

SfM estimates the structure and camera motion (rotation, translation, and internal camera parameters such as focal length) from an image sequence captured with a single camera. These techniques require the camera to move, i.e., to translate. Pure rotation (pan) is not sufficient. Estimation is done by detecting and tracking 2-D features over time. The 3-D coordinates and camera parameters are recovered utilizing triangulation and self-calibration methods. For a final improvement of the reconstruction results, bundle adjustment is often applied.

Finally, SfSi methods reconstruct 3-D object shapes by projecting binary segmentation masks into 3-D space. The object shape is the intersection of all these projections. In general, several views captured simultaneously from different viewpoints are needed to get a precise 3-D reconstruction of an object. SfSi is also less relevant for the case of 2-D/3-D conversion.

3) Inpainting: Due to the recomposition of the scene during the rendering of new views, regions that where originally occluded by foreground objects will become visible. Since the actual texture is not available it has to be estimated using spatial or spatio–temporal approaches. These methods of image completion [66] are often referred to as occlusion filling [80], inpainting [67], [79], or infilling [71]. Unlike in other image completion applications, inpainting for stereo view generation is highly dependent on the scene’s depth structure. Thus, a disoccluded region can only be filled with texture that resembles the background seamlessly.

Usually, the regions to be filled have a vertical structure, i.e., a small horizontal aperture depending on the object’s disparity. Therefore, simple horizontal pixel extrapolation or pixel row mirroring methods can lead to satisfying results depending on the adjacent background structure. Also 2-D interpolation methods are useful but yield in rather blurred results. For the more general case, spatial methods such as texture synthesis [68], [69] or methods of higher order border continuation [70] can be applied. See Fig. 17 for a comparison.

For texture synthesis, also known as image quilting, holes are filled by patching neighboring regions seamlessly. Often user interaction is required to determine the source patch, which then is less attractive for automation.

Continuing the borders of an occlusion hole by analyzing the adjoining geometry is a very promising approach [70], [71]. Here, contours or isophotes of the background are propagated into the hole and the resulting subregions are filled by the corresponding colors. Anisotropic diffusion methods are often applied in order to keep the contours as sharp as possible [72], [73], [80]. A combination of texture synthesis and geometry continuation is presented in [74] and [75].

For 3-D image sequences temporal consistency is of high importance. Thus, spatio–temporal inpainting approaches that enforce global temporal consistency for all patches in and around a hole can be applied [76]–[78]. Additionally, due to motion of the camera and the objects throughout the sequence it is likely that the original background emerges over time. Therefore, motion-based video mosaicing can be used to fill occlusion holes [79]. However, the success of mosaic-based hole patching depends on the type of camera motion.

C. Entire Offline and Real-Time Conversion Systems

Automatic conversion systems that have been introduced during the last years can be divided into offline and real-time conversion approaches. Real-time conversion methods are based on very simple approximations and are commonly used in 3DTV applications, where degradations in stereoscopic depth perception are more tolerable.

Many of these real-time conversion approaches rely on motion parallax [81]–[86]. The simplest way is of course using motion vectors directly from compressed video data, e.g., MPEG [81] or H.264 [82]. However, this technique

![Fig. 17. (a) Original image from [67] with a bungee jumper as a foreground object, (b) interpolation-based inpainting of the removed foreground object using OpenCV versus (c) patch-based inpainting.](image-url)
can only recover relative depth accurately, if the motion of all scene objects is directly proportional to their distance from the camera. Independent object motion would result in false depth values and thus decreases the 3-D effect. Furthermore, the overall quality heavily depends on the degree of compression.

Another very simple effort is to decode depth from color or intensity information directly from the input images, e.g., [85]–[87]. Lighter intensity or higher saturation values indicate that objects are closer to the camera. This assumption only holds in a relatively small proportion of footage encountered (for example, when the light source illuminates foreground objects). However, in [87], subjective tests demonstrate that color-based approaches yield good results on some selected test data.

Especially for outdoor environments, height-depth cues are sometimes applied [88]. The idea behind this is that objects on the bottom of an image are often closer to the camera than objects on the top. This assumption is of course valid in a landscape with a sky, horizon, and a plane in the foreground. In more complex environments, edges are detected and traced to identify objects or regions of similar depth.

Hybrid depth cuing is a more reliable approach for 2-D/3-D conversion and increases the conversion quality. A combination of motion-parallax and color information is introduced in [85]. Furthermore, in [86], motion-parallax and geometric perspective recovered from line detection are fused and bilaterally filtered to generate more accurate depth maps than applying only one of the two depth recovering methods.

Companies such as JVC, Samsung, or DDD that are offering real-time conversion systems utilize the aforementioned techniques in a hybrid manner. Although the 3-D quality seems to be promising, it is not sufficient to watch an entire 90-min movie in 3-D without perceptual discomfort.

More advanced automatic 2-D/3-D conversion systems are specialized on some limited capture conditions. Basically, these approaches will never run in real time but may increase both conversion speed (compared to manual conversion) and conversion quality (compared to real-time conversion).

Since camera motion often boosts the workload in the manual rotoscoping process, it is in the focus of many conversion systems. In [89] and [90], bidirectional motion estimation is applied for the recovery of rigid structure and motion. Furthermore, a Bayesian framework based on extended Kalman filters is introduced to handle the occlusion problem.

In [91], a system for generation of super-resolution stereo and multiview video from monocular video based on realistic stereo view synthesis (RSVS) is presented. It combines both the powerful algorithms of SfM and the idea of image-based rendering to achieve photo consistency without relying on dense depth estimation. In Fig. 18, the image-based rendering step of RSVS is depicted.

![Virtual view reconstruction using SfM and perspective transformations.](image)

Three-dimensional scene points computed with SfM are reprojected into a virtual stereo frame and into surrounding views of the original monocular video sequence. Then, 2-D point correspondences are utilized to estimate perspective homographies between the virtual and neighboring original views. Finally, each pixel of a neighboring original view is warped into the stereo frame according to the perspective transformation equation

\[ m_i = H_i m_{\text{multi}} \]  

(4)

where \( H \) is the homography between the stereo frame and the right original view in Fig. 18.

However, the approach is limited to rigid scene structures. In [92], this system is extended to deal with independently moving objects within a rigid scene.

Finally, in [93], only planar transformations on temporal neighboring views are applied to virtually imitate a parallel stereo camera rig. This approach does not utilize 3-D information from SfM. But, in any case, time consistency along the sequence is heavily dependent on the 3-D scene and camera motion. Thus, a stereo rig has to be modeled correctly to ensure a realistic 3-D performance.

**VII. SUMMARY AND OUTLOOK**

Today, we are witnessing broad and most probably sustainable entry of 3-D video technology, applications, and content into our everyday lives. Unlike in previous attempts, 3-D
technology is now mature enough and content creation is understood well enough to ensure wide acceptance. Nevertheless, stereoscopic 3-D content creation is still a difficult task so that there is still a lot of room for improvement and extension of available solutions.

Stereography sets rules and recommendations for proper 3-D content creation, which have to consider perception as well as display properties. Depth range adaptation, display adaptation, and object-based local depth composition adaptation are fundamental tasks in 3-D postproduction. This is most difficult in live broadcast scenarios, especially sports. Mixing and composition of 3-D content has to take depth composition of all source material into account. Support of advanced displays and free viewpoint navigation still requires substantial research and development for proper content creation. The 2-D/3-D conversion is a highly actual and important area that enables reuse of existing legacy content in a new dimension.

Virtual view synthesis and depth composition manipulation is the key technology to solve many of these issues in 3-D video postprocessing and production. Most algorithms for this purpose today are based on DIBR. Significant progress has been made in the area of depth estimation, though fully automatic, robust, reliable, accurate recovery of scene depth cannot always be guaranteed. User interaction can help improving robustness in non-real-time application scenarios. Basic processing stages including correction of geometrical distortions, color matching, and adjustment of stereo geometry have to ensure basic quality of the stereo content.

Warping-based methods recently achieved a lot of attention in the context of video retargeting. Such methods can also be used for virtual view synthesis. As such they may be an alternative to depth-based approaches for some application scenarios. They avoid the need to compute dense and accurate depth maps, as well as the disocclusion problem. On the other hand, they are only capable to compute visually plausible synthesized views, which may limit applications.

The 2-D/3-D conversion is a highly actual and important area for itself. Various different algorithms are combined to generate a stereo pair given available 2-D video. Interactive workflows are used to ensure acceptable quality. In consequence, content creation for stereoscopic 3-D and beyond is an area that is mature enough, but there is still a lot of room for improvement and extension of available technology. Much research is going on worldwide in the academic sector as well as in companies. Driven from the emerging business opportunities, numerous products and services for 3-D content creation and processing are emerging. We may conclude that this is an area still full of challenges and potential in many ways.

REFERENCES


ABOUT THE AUTHORS

Aljoscha Smolic received the Dr. Eng. degree in electrical and information engineering from Aachen University of Technology (RWTH), Aachen, Germany, in 2001. He joined Disney Research Zurich, Zurich, Switzerland, in 2009, where he leads the “Advanced Video Technology” group. Prior to that, he was a Scientific Project Manager at Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute (HHI), Berlin, Germany. He conducted research in various fields of video processing, video coding, computer vision, and computer graphics, and published more than 100 referred papers in these fields. Since 2003, he had teaching appointments at Technical University of Berlin, ETH Zurich, Universitat Politecnica de Catalunya (UPC), Universidad Politecnica de Madrid (UPM), and Universitat de les Illes Balears(UIB).

Dr. Smolic is an Area Editor for Signal Processing: Image Communication and served as Guest Editor for the PROCEEDINGS OF THE IEEE, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE SIGNAL PROCESSING MAGAZINE, and other scientific journals. He chaired the MPEG ad hoc group on 3DAV pioneering standards for 3-D video. In this context, he also served as one of the Editors of the multiview video coding (MVC) standard.

Peter Kauff received the Diploma degree in electrical engineering and telecommunication from the Technical University of Aachen, Aachen, Germany, in 1984. He is the Head of the “Immersive Media & 3D Video” Group at Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute (HHI), Berlin, Germany. He has been with HHI since 1984, mainly engaged in the area of digital HDTV signal processing and advanced 3-D video processing for immersive media. He has been involved in numerous related international research projects for a long time and is active in the running national and international research projects PRIME, 3D@SAT, 3D4YOU, 3D Presence, 2020 3D Media, MUSCADE, AVATECH, and Fascinate.

Sebastian Knorr received the Dipl.-Ing. and Dr.-Ing. (Ph.D.) degrees in electrical engineering from the Technische Universität Berlin, Berlin, Germany, in 2002 and 2008, respectively. He worked as a Research Assistant and Senior Researcher from 2002 to 2009 in the field of 3-D image processing and computer vision in the Communication Systems Lab, Technische Universität Berlin. During this time, he was involved in several European Networks of Excellence, e.g., VISNET and 3DTV. Currently, he is the Managing Director and Head of the Software Application Development and Know-How Transfer Department of Imcube Media GmbH.

Alexander Hornung received the Diplom and Ph.D. degrees in computer science from the Computer Graphics Group, RWTH Aachen University, Aachen, Germany, in 2003 and 2008, respectively. He is a Research Scientist at Disney Research Zurich, Zurich, Switzerland. Before joining Disney, he was a Postdoctoral Researcher at the Computer Graphics Laboratory, ETH Zurich, Zurich, Switzerland. His research interests lie in computer graphics and vision. In particular, he is interested in the fields of video processing, image-based rendering and 3-D reconstruction, and 2-D animation.

Matthias Kunter received the Dipl.-Ing. and Dr.-Ing. (Ph.D.) degrees in electrical engineering from the Technische Universität Berlin, Berlin, Germany, in 2002 and 2008, respectively. He worked as a Research Assistant and Senior Researcher from 2002 to 2009 in the field of 3-D image processing and computer vision at the Communication Systems Lab, Technische Universität Berlin. During this time, he was involved in several European Networks of Excellence, e.g., VISNET and 3DTV. Currently, he is the Managing Director and Head of the Software Application Development and Know-How Transfer Department of Imcube Media GmbH.

Marcus Müller received the Diploma degree in computer science from the Technische Universität Berlin, Berlin, Germany, in 2004. He is a member of the Immersive Media & 3-D Video Group, Image Processing Department, Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI), Berlin, Germany. He has been with HHI since 2004, where he is involved in several German and European projects related to 3DTV.

Manuel Lang received the M.Sc. degree in computer science from ETH Zurich, Zurich, Switzerland, in 2008, where he is currently working towards the Ph.D. degree. He is a Research Scientist at Disney Research Zurich, Zurich, Switzerland. He works in the field of computer graphics and computer vision. In his current research, he focuses on video production and editing tools, in particular, for stereoscopic 3-D. He worked as an intern for NVIDIA where he could pursue his additional interest in GPU programming. Before joining Disney, he worked as a Research Assistant in the Computer Graphics Lab headed by Prof. M. Gross.