Display-Independent 3D-TV Production and Delivery Using the Layered Depth Video Format

Bogumil Bartczak, Patrick Vandewalle, *Member, IEEE*, Oliver Grau, *Member, IEEE*, Gérard Briand, Jérôme Fournier, Paul Kerbiriou, Michael Murdoch, Marcus Müller, Rocco Goris, Reinhard Koch, and René van der Vleuten

***Abstract*—This paper discusses an approach to 3D-Television that is based on the Layered Depth Video (LDV) format. The LDV format contains explicit depth and occlusion information, allowing for the generation of novel viewpoints for stereoscopic and auto-stereoscopic multi-view displays. Thus, the format is effectively invariant to the display type and also allows the depth impression to be easily changed to best meet viewers' preferences for visual comfort. The major aspects of a content delivery chain based on the LDV format are discussed in this paper. The requirements placed on data acquisition are introduced, and a multi-camera system, which is well suited for LDV compliant data capture, is presented. Also discussed is the conversion of different input data streams, like standard stereo videos, multi-view data supplemented by depth data, and videos from wide baseline setups, to the LDV format. Moreover, the advantages of the LDV format in editing and mixing are examined. The paper also presents a transmission system based on currently available coding and transmission standards. Optimization of the bandwidth via different approaches to the compression of the LDV signal is analyzed, and the results of conducted experiments in this field are discussed. Finally, the aspects of perceptual human factors for the proper evaluation of 3D-TV services and the implemented LDV system are examined. This contribution reflects the efforts of the EU-funded project 3D4YOU to unify all aspects of 3D-TV production.**

***Index Terms*—Compression, depth reconstruction, depth-based formats, layered depth video, perceptual human factors, 3D television.**

Manuscript received August 06, 2010; revised January 21, 2011; accepted February 07, 2011. Date of publication April 05, 2011; date of current version May 25, 2011. This work was supported by the Project 3D4YOU under Grant 215075 of the Information Society Technologies area of the EU's 7th Framework programme.

B. Bartczak and R. Koch are with the Institute of Computer Science at the Christian-Albrechts-University of Kiel, 24118 Kiel, Germany (e-mail: bartczak@mip.informatik.uni-kiel.de).

P. Vandewalle, M. Murdoch, and R. van der Vleuten are with Philips Research Laboratories, 5656 AE Eindhoven, The Netherlands.

O. Grau is with the BBC Research & Development, Centre House, London W12 7SB, UK.

G. Briand and P. Kerbiriou are with the Technicolor R&D France, 35576 Cesson-Sevigné, France.

J. Fournier is with the Orange Labs TECH/OPERA 4, 35512 Cesson-Sévigné, France.

M. Müller is with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut Image Processing, 10587 Berlin, Germany.

R. Goris is with the Philips Consumer Lifestyle High Impact Innovation Center, 8-040 5656 AE Eindhoven, The Netherlands.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2011.2120790

I. Introduction

A 3D-TV system consists of four building blocks: capture, post-processing, distribution, and display. The complexity of each block and how well these can be coordinated will determine the long-term feasibility of 3D-TV. It is difficult to assess which of these blocks requires the most attention; however, it is agreed that the display technology used will determine the acceptance by the general public. The choice of display technology together with the constraints induced by the characteristics of the human visual system will have consequences on the other building blocks of a 3D-TV system. When considering the introduction of display technology to the mass market, different key elements have to be taken into account. Displays not only have to allow for pleasant stereoscopic viewing, but also have to be affordable and to fit into the average living room environment. An optimal solution would not deviate far from the look-and-feel of today's regular TV sets and would also be backward compatible to today's television format. The same basically holds true for the display's impact on the content production process. If a display's proper operation requires much more information or computational power, or if the guidelines for creating content are significantly changed, these will hamper the introduction of 3D-TV due to the sheer costs such technology would induce.

Judging from the currently available broadcast and display technologies (see [1], [2] for details), only stereoscopic and autostereoscopic displays come close to fulfilling these requirements. The flat screen of a stereoscopic display is used to display two different images in accordance with the required binocular cues. Glasses are utilized to separate the images for each eye. The need for glasses is eliminated by autostereoscopic displays through the use of parallax barriers or lenticular arrays. However for the proper operation of autostereoscopic displays more than two views on a scene have to be shown [3].

Since 2003, 3D cinemas have become increasingly popular and many stereoscopic productions have been released successfully. The operational principles of stereoscopic displays and 3D cinemas are similar, so that the repurposing of cinema content for 3D-TV at home seems straightforward. However, because of the natural linkage in the human visual system between accommodation and vergence, care must be taken to ensure our ability to fuse two horizontally displaced views into a single stereoscopic image [4]–[7]. For this the viewing conditions have to be taken into account when displaying binocular cues. An optimal solution would be to generate the left and right views of a scene in unison with the intentions of the content creator, the viewer's preferences and the physical viewing conditions. The two views that are captured in today's 3D cinema productions with two horizontally displaced cameras can provide this situation only for a particular combination of these conditions. Hence, the direct display of cinema content on a much smaller screen typically does not match the requirements for proper stereoscopic viewing.

An alternative to this direct approach is to deliver appropriate information to a display system so that this system is able to calculate (render) the desired images from it. In general, this requires knowledge of the geometric structure in the scene. It is not a trivial task to acquire and to use this information for the proper generation of images in 3D-TV applications. A viable approach is found in image based rendering techniques (IBR) [5], which use color images as a primary data structure to reduce the need of abstract content descriptions but nevertheless allow photo realistic image rendering. The performance of these rendering approaches with respect to computation speed and accuracy is significantly improved when depth information is provided. For this, the color images are assigned depth maps so that the position in 3D space of each pixel is effectively known. This information allows the rendering of novel views of the scene by warping color values via their 3D position in accordance with the novel view's camera position [9]. Implementing this technique into a 3D-TV system not only allows the adaptation of stereoscopic content to the viewing environment, but also the rendering of more than two views of a scene without increasing the overhead in the content production pipeline. This way the proper operation of glass-free autostereoscopic displays for 3D-TV at home becomes possible. Additionally, other display technologies like volumetric displays [5], which might reach the market in the future, will be operable with the IBR approach from color plus depth data without increasing the content production overhead.

The structure of the paper follows the flow of data through the content production chain; after further discussion on prior work on depth-based 3D-TV, the important technical aspects of image and data capture for 3D-TV are discussed. This is followed by the description of the proposed Layered Depth Video (LDV) format generation itself, which is built on top of depth extraction from the captured content. In Section V, post-production steps of editing the video material and mixing with computer generated content are explained. These steps typically follow after the captured data have been transformed into LDV. Section VI discusses the different findings in respect to compressing the data and presents an experimental setup which demonstrates the transmission of encoded LDV content via standard broadcast methods to different display types. Following the details of the technical aspects of the content delivery chain, Section VII describes the evaluation of the combined technologies' implemented quality in terms of perceptual human factors.

II. Depth-Based 3D-TV

Different formats for content representation via color plus depth have been proposed, which shall be discussed in the following together with prior work in this field.

A. *Review of Depth-Based Formats in 3D-TV*

Implementing a 3D-TV system that uses explicit depth information and image based rendering has various advantages. It not only makes the system independent of the display used, it also relieves the content generation process of many tedious and error prone manual tasks which are necessary in today's conventional stereo. Explicit depth information allows the formulation of computer algorithms to detect sources of discomfort like border collisions and the crossing of disparity limits in a single shot or between scene cuts [10]. Because the final images are rendered on demand, it is not necessary to react to such violations during content acquisition. In fact, the rendering can automatically be adapted by either adding a floating window [6] or by scaling and shifting the depth as proposed in [10]. Also, introducing computer-generated content like logos, subtitles, or large scale composited, keyed sequences is facilitated. Here the explicit depth information provides an intuitive interface to introduce virtual objects, leaving the creation of stereoscopically consistent images to the rendering. This also provides a way to avoid mismatches in the camera geometries or optics that can cause vertical disparities or other unnatural effects [6].

It is a challenge to provide reliable combinations of color and depth information and to implement a reliable rendering from these data. Different proposals have been made to approach these challenges. In 1998, a European project called PANORAMA [11] demonstrated the feasibility of using a single color view plus depth to achieve scalable stereoscopic rendering and viewing in a videoconferencing system. The suitability of the same representation for 3D-TV was investigated by the later European project ATTEST [12]. ATTEST was able to show that the representation is well suited for 3D-TV applications, and that it can be transmitted over existing channels with a low overhead. Despite the applicability of this format, the absence of occlusion information quickly reduces the quality of reproduced images when the baseline between rendered views increases, as is required for multi-view autostereoscopic displays. In [13] therefore an approach to 3D-TV is discussed that uses multiple color views and annotates each with depth information. This format is known as the Multiple Video plus Depth (MVD) format and is inspired by the convincing results that were achieved by Zitnick *et al.*, published in 2004 [14], who implemented a multi video based depth reconstruction and rendering system for high quality free-viewpoint video. The work presented in [14] still serves as the basis for most of today's improved proposals for depth reconstruction and rendering algorithms in respect to free-viewpoint video and 3D-TV [15], [16].

MVD obviously contains a lot of redundant information, as foreground and occluded scene parts will be visible from multiple viewpoints. Using the MVD format, occlusions only become apparent when the targeted viewpoint is specified, and therefore the rendering of new views has to include occlusion handling, i.e. detection and filling of disocclusions, which complicates the rendering process. Through the choice of a reference view and a maximal amount of viewpoint change from this reference, MVD data can be converted into the Layered Depth Video (LDV) format (also see Fig. 1). LDV is the natural extension of Layered Depth Images [17] to video data. In

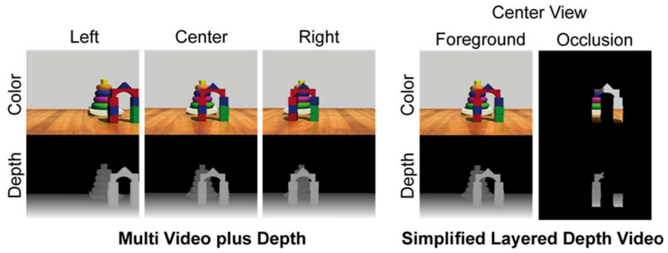


Fig. 1. Differences between MVD and LDV. On the left side, MVD is shown for three cameras in simple scene. On the right hand side, a simplified LDV representation is displayed, which basically holds the same occlusion information.

the LDV format, the scene is represented by a sorted stack of corresponding color and depth images, all relative to the chosen reference viewpoint. Each pixel in the LDV contains depth and color information for the foreground and for each scene element occluded behind it. Redundant disocclusions can be determined and removed, reducing the amount of information required. Furthermore, the rendering from LDV works without complicated occlusion handling and thus can be executed very efficiently.

In its simplest form, the LDV will contain only a single occlusion layer, which will not allow a faithful reproduction of scenes with arbitrary complexity. However, the size of secondary disocclusions is often small so that this simplification will only slightly degrade the reproduction quality. A further simplification of the LDV format can be used if the rendering is constrained to only create horizontally displaced images. In this case, the encoded depth information can be interchanged with horizontal disparity values relative to the average human eye distance, which significantly reduces the complexity of the rendering algorithm. This simplified LDV format was also integrated into the Philips autostereoscopic WoW displays [18] for real-time rendering of different views. Thus the supply of content to these displays is very convenient as no additional image processing has to be performed before viewing.

A drawback of the non-redundant scene representation via LDV is that the rendered image quality decreases with the reproduced viewpoint's distance from the chosen reference view, due to suboptimal resampling of the original input data. In contrast, the amount of data resampling can be reduced in the MVD approach by selecting the closest input views during rendering. In order to combine the best of both approaches, hybridization of MVD and LDV can be considered, like a simple augmentation of the LDV format with additional color views (LDV-R [19]) or even the use of multiple views in full LDV representation as proposed in [20]. Considering such extensions to multiple LDVs, one has to be aware that the number of required input viewpoints for a proper occlusion generation will increase significantly, so that the use of a single LDV or an MVD with two views (MVD2) will often constitute the best compromise between reproduction quality and required input data overhead.

B. 3D4YOU Project

The use of depth-based formats for 3D-TV has several advantages, but it also introduces novel conditions and requirements into the delivery chain that are different than the more familiar two-camera approaches of today's stereo productions. Investigations of a depth based format for 3D-TV uses were conducted in 2002 and 2004 in the European project ATTEST [12]. ATTEST's investigations included the capture of depth data with an active depth camera, the semi-automatic assignment of depth to 2D material, the development of autostereoscopic displays, and the distribution of data to these displays. ATTEST was successful in demonstrating the feasibility of the single color plus depth approach with respect to capture and distribution. In Feb. 2008, another European project called 3D4YOU started. Extending the goals of ATTEST, 3D4YOU's focus was the development of a pragmatic content delivery chain in the LDV format. For this purpose, expert partners from industry and academic institutions in the fields of content production, data acquisition, broadcasting, display technology, computer graphics, and video editing, as well as data encoding, compression, and computer vision, have been brought together. Namely, these partners are Philips, BBC, Fraunhofer HHI, Orange, Technicolor, University of Kiel, and KuK GmbH.

In order to gain insights into the technical aspects of generating LDV data and its handling in editing, mixing, coding, and distribution, the 3D4YOU project made a number of production tests to cover a wide set of scenarios and program genres. In studio captures, a number of scenarios and systems were tested. Further, a test production about peregrine falcons was carried out together with the Natural History Unit of the BBC, involving stereoscopic camera rigs and extensive post-production, including outdoor and studio shoots in combination with 3D graphics. In addition, material of a rugby sport scenario was provided from a regular monocular, but multi-camera, setup in order to test concepts of deriving 3D information from wide-baseline setups. This paper gives an overview of the experiences gained and describes proposed solutions to challenges that were discovered during evaluation of the data captured during these production trials within 3D4YOU.

III. Novel Methods for 3D-TV Data Acquisition

In this section, the acquisition of data that can be turned into high quality LDV data is discussed. The challenge here is that direct capture of associated color and depth information is not possible in decent quality today. Research in computer vision, however, has developed algorithms that allow the reconstruction of depth information from captured color images with different viewpoints. It is sensible to consider these algorithms for the purpose of 3D-TV content generation, as in this scenario multiple viewpoints of a scene will have to be captured in any case. The algorithms are then able to assign depth maps to these color data with matching resolution. Although these algorithms have their limitations, and their processing speed in general is slow, the actual data acquisition can be done in real-time so that the capture of dynamic scenes is feasible. An alternative to the application of these stereo reconstruction algorithms is the use of active depth measurement devices like laser scanners or so called Time-of-Flight (ToF) cameras [21]–[23]. Laser scanners in general are considered to deliver depth data with very high accuracy; however, their bulk and operation principle restricts them to the reconstruction of static scenes. The quality and resolution of data captured with ToF-cameras is not as high as provided by laser scanners, but their mode of operation allows application in dynamic scenes. The depth measurements of ToF-cameras have different error sources than stereo reconstruction algorithms, so there is a gain in information when combining both approaches. This section therefore discusses hybrid camera rigs consisting of multiple color cameras and ToF-cameras. In order to define a pragmatic and flexible capture setup, the working scheme of stereo reconstruction algorithms has to be considered. A short introduction to the general technique of image based depth reconstruction is given as well.

A. Background on Depth Reconstruction Techniques

Depth reconstruction from color images is based on the assumption that a scene point will generate similar color values when measured from different viewpoints. This assumption is used to look for the positions of corresponding colors across different images. Given the geometric camera parameters of the images, this correspondence can then be turned into a depth value. Obviously this approach has several shortcomings, as there are many occasions where the comparison of color values will be ambiguous. This is the case in homogeneous image regions, like uniformly colored objects, and at occlusions, where one image depicts parts of an object that are not contained in the other image.

B. LDV Compliant Capture

From the insights into the technique of reconstructing depth from color images, it can be concluded that in order to be able to reliably reconstruct the scene depth at a specific image position, the corresponding scene point has to be visible in at least two color images taken from different viewing positions. This cannot be achieved with only two cameras like in the simple stereo setup. At least three viewpoints that have mutually exclusive occlusion regions are required to fully reconstruct a view's depth. As explained in Section I, the rendering of novel viewpoints requires scene information from regions occluded from the perspective of a single viewpoint; therefore, the depth reconstruction from a single viewpoint will not deliver sufficient information. As a consequence, it is at least necessary to fully reconstruct the depth for two images from different viewpoints. This will then allow the rendering of novel images in between these two views without lack of data. As stated above, three cameras are required to avoid problems with occlusions during depth reconstruction. Arranging four cameras in a horizontal line provides a compact camera setup where the two central cameras each have two neighboring cameras, mutually covering the respective occlusion areas. This constitutes the minimal setup fulfilling the discussed requirements and was therefore used by 3D4You during a shoot in December 2008. Pictures of this rig are shown in Fig. 2 (left). It was built on top of an existing mirror-rig designed and constructed by KuK Film GmbH by adding two additional "satellite" cameras on each side of the mirror-box. In this way, the rig naturally incorporates the simple stereo setup and thus is backward compatible with today's standard 3D film production pipelines. Additionally, a ToF-camera was integrated into the rig. These devices currently have a low resolution, but unlike the stereo-matching approach, they do not rely on scene texture to properly deliver depth information, so the capture system is able to handle homogeneous scene elements. Although the chosen arrangement is pragmatic and requires little hardware resources, it also shows some limiting characteristics. First of all, a single ToF camera is not able to cover the occlusions for two different color viewpoints and thus can only support the reconstruction of foreground depth and one occlusion region. Secondly, the setup is disadvantageous for use with the LDV format. LDV uses a scene description from a single viewpoint but with multiple layers. These layers allow the storage of occlusion information without particular preferences on the novel viewpoint's pose. Thus the viewpoint used for the LDV definition is best positioned in between the two center cameras, so that it can capture the occlusion information from both viewpoints. Since no data is captured for this center viewpoint, all the data in the LDV representation has to be reconstructed.

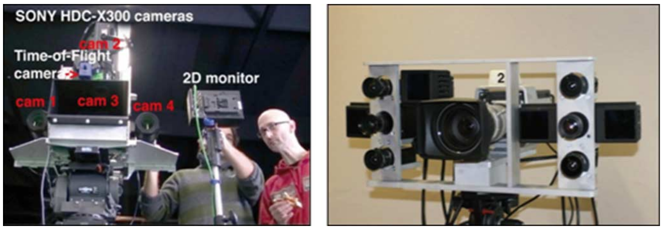


Fig. 2. (Left) Realization of four camera rig plus one ToF-camera used during 3D4You shooting in Dec. 2008. (Right) 5 + 2 Rig setup with LDV reference camera in the center. Placed to the left and the right of the center view are two vertical rigs consisting of two color cameras (top-bottom) and correlating time-of-flight cameras in the middle.

Adding another color and another ToF-camera to the hardware setup creates a configuration that improves the previously discussed rig with respect to baseline configurations, occlusion coverage by ToF-measurements, and LDV compatibility. However, it requires a departure from the standard stereo-centric capture approach. The proposed system is constructed from two vertically arranged rigs, which each consist of two color cameras on top and bottom and a ToF-camera in the center. Between these two vertical rigs, a reference camera is placed. Fig. 2 (right) shows this rig. Obviously, the reference camera is well suited to function as a viewpoint for LDV delivery. This no longer guarantees a fully-covered field of view for two color viewpoints in the rig; however, it maximizes the ToF-camera's coverage of the reference cameras field of view without sacrificing the ability to measure depth in horizontally-occluded scene elements. This is beneficial in the LDV centered reconstruction approach, where color information for the occlusion layers has to be warped into the reference viewpoint in any case.

1) Preview Capabilities: In order to allow the effective operation of a 3D-TV capture system, content producers should have some intuitive feedback from this system on what impression will be conveyed to the viewers. For this, an early preview of the depth impression must be supplied by the capture system. Using a standard stereo setup, this can easily be done by directly displaying the images on a stereoscopic display. This preview, however, will not be able to correctly reflect the final depth impression unless the correct display geometry is used. In the case where no standard stereo configuration

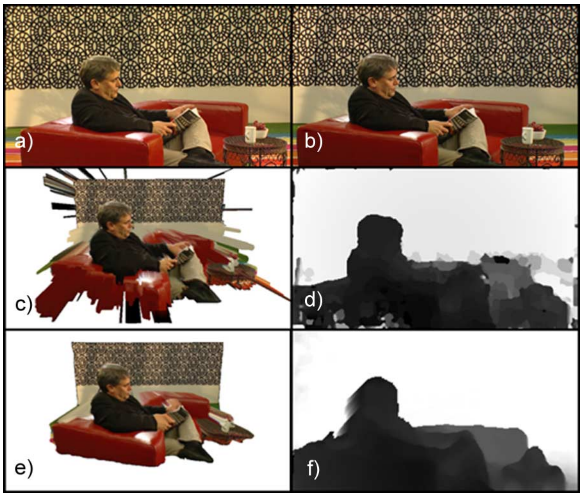


Fig. 3. (b) Comparison of real-time depth assignment to a color view (d) from local stereo matching and (f) from ToF-measurements via triangle mesh warping. (a) and (b) show the input images used in stereo matching. (d) is the result of stereo matching and depth assignment to view (b), while (c) shows the rendering of a textured triangle-mesh generated from (d). Analogous (f) shows the depth measured by a ToF-camera and warped into view (b). (e) is rendered via a textured triangle-mesh generated from the depth map (f).

is used, or where a multi-view display or different scene geometries need to be assessed, fast image based rendering techniques that require depth information have to be applied. Assuming that rendering algorithms are available, like for instance those integrated in the Philips WoW displays, the task is to deliver LDV data in real-time. One approach would be the use of real-time local stereo matching algorithms. Here, GPU implementations exist that are able to estimate disparity maps at up to 100Hz, assuming the images are sufficiently down-sampled [24]. This approach, however, will have difficulties in homogeneous and occluded image regions, generally leading to significant artifacts. Down sampling the input images furthermore reduces the distinguishable depth levels, which leads to unrealistic cardboard effects. In order to demonstrate the performance characteristics of real-time stereo algorithms, Fig. 3(c) and (d) show the results of applying such a scheme to the input images shown in Fig. 3(a) and (b). These results were calculated using 3 × 3 NCC-matching as a base dissimilarity measure and aggregating costs over a 21 × 21 window using the asymmetric and separated variant of the bilateral filtering scheme proposed in [25], which accurately preserves discontinuities. In order to achieve real-time performance, the scheme was implemented on the GPU, and the input images were downsampled by a factor of 5. Fig. 3(d) shows severe artifacts in homogeneous color regions like the wall and the sofa. Other errors are clustered at mutually occluded image regions, which in this case are primarily on the right hand side of the person. Fig. 3(c) shows the rendering of a textured 3D triangle-mesh generated from this depth map. It reflects the impact of the errors on the rendering of new viewpoints.

An alternative to stereo estimation is given by the use of ToF-cameras. Although these devices have low spatial image resolution, their depth resolution is quite high. Moreover, their depth measurement is insensitive to the scene's texturing and does not have occlusion conflicts. These characteristics are exploited in the preview system developed in 3D4YOU [26], which delivers results as shown in Fig. 3(e) and (f). Although these results are not free of errors, the errors are fewer and significantly less severe, so this approach is better suited for assessing the proper setup of a scene.

2) Acquiring Camera Geometry: The extraction of depth information and the conversion into the LDV format requires knowledge of the imaging geometry of the camera rig, which helps in the consistent transfer of depth information between different cameras and constrains the search range for corresponding image positions. It can then also be used to automatically rectify images (remove vertical and rotational disparities), which in typical stereo production is a tedious yet important step during post-production.

Calculation of this geometry can be attempted from the captured content itself via auto-calibration methods [27]. Another approach is to capture images of a known calibration target in different positions and orientation, before or after acquisition of the target content. Auto-calibration methods are demanding in regard to the input data and are therefore less reliable; furthermore, no auto-calibration approach for ToF-cameras is known today. Use of a calibration pattern is thus often preferred for calibration of small baseline camera rigs. Well suited for this is the procedure proposed in [28]. It uses a statistical optimization scheme for parameter estimation. By means of an analysis-by-synthesis approach, the number of observations that are used in the parameter estimation can be increased to make the approach reliably estimate the parameters for the ToF-cameras.

IV. LDV Generation

In the previous section, LDV compliant data acquisition was discussed. However, a multitude of content exists that was not generated following this approach. Also, in the future alternative capture methods will be investigated. This section therefore examines the generation of LDV data from different input channels as sketched in Fig. 4<<AQ1>>. It first discusses a flexible approach to generating occlusion information for the simplified LDV format when narrow baseline viewpoints and depth maps are available for each. Two automatic depth map extraction schemes from narrow baseline camera setups are presented next. The first approach is designed to generate consistent depth maps from the standard stereo data available from today's common 3D cinema productions with two cameras. The second uses the data captured with the camera setups proposed in Section III and extracts reliable depth maps by combining ToF-depth measurements and color information. Finally, this section outlines the generation of LDV data from wide baseline scenarios which are typical in sports like rugby and soccer.

A. Occlusion Layer From Narrow-Baseline Set-Ups

As depicted in Fig. 4, the input to the proposed LDV generation approach is corresponding color and depth images either from multiple views or from a single viewpoint. Once high quality depth maps have been obtained and a reference view has been selected, precise occlusion information must be generated. This occlusion information is required when rendering side views, where scene elements emerge from behind the foreground objects. In order to determine which parts in the LDV

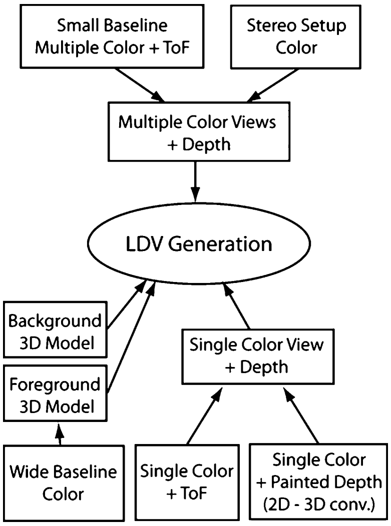

Small Baseline Multiple Color + ToF
Stereo Setup Color
Multiple Color Views + Depth
LDV Generation
Background 3D Model
Foreground 3D Model
Single Color View + Depth
Wide Baseline Color
Single Color + ToF
Single Color + Painted Depth (2D - 3D conv.)

Fig. 4. Data input paths to generate the LDV format in 3D4You.

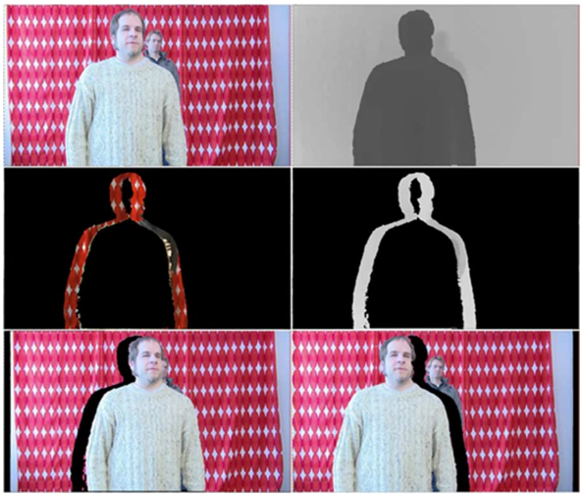


Fig. 5. (Top) Color and depth information of a reference view. Warping the color with respect to the depth information towards a left and right view produces (black) masks where (bottom) disocclusions occur. (Middle) Filling this occlusion with techniques described here delivers the missing occlusion information for depth and color.

view have to be filled in, occlusion masks are generated by rendering the chosen reference foreground depth map into virtual side views [29]. Fig. 5 shows a result of this process.

Most troublesome for the generation of occlusion information is the case in which corresponding color and depth maps are only available for a single view. Such a monoscopic data constellation can occur if a single ToF camera is coupled with a single color camera, as in the ATTEST project, or if 2D film material is converted to 3D material and depth maps are painted manually [6]. Also, in the case where multiple viewpoints are used, uncovered occlusions might still be present in the input data. A special case is given in the wide baseline scenario where background model information is exploited to extract the foreground. This changes the conditions for LDV generation because color and depth are explicitly separated between background and foreground. The handling of this situation is discussed in the section *LDV from Wide-Baseline Data* below. In the other situations, multiple approaches exist for occlusion estimation, also known as inpainting [17], [30]–[33]. These techniques are similar to the techniques used in the restoration of paintings or of old videos from film.

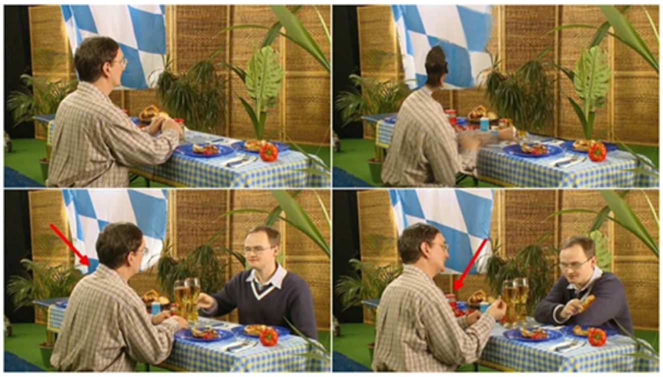


Fig. 6. (Top right) Result using temporal inpainting. (Top left) Occlusion information behind the person for the input frame is filled in using data from other frames. Specific examples are indicated in the bottom left and bottom right frame.

Each of the techniques takes one or multiple images with depth maps as input, in combination with the occlusion mask indicating which areas have to be filled in. However, a distinction based on the origin of the data used for filling in the disocclusion will be made here. Temporal inpainting uses the temporal axis to look for the missing data in other frames [31], [33], while multi-view inpainting makes use of the data from cameras with a slightly different viewpoint to see behind objects [17], [31]. Finally, single-frame inpainting uses structural analysis of the surrounding background to fill in the unknown part [32].

Temporal inpainting: In order to fill in occlusion information via temporal inpainting [31], a video for a single view is analyzed to find frames in which the missing information is visible. This is easiest for scenes with a static background that are shot with a fixed camera. Assuming that the depth map accurately indicates which depth values belong to the background, the occlusion information can then be created by gathering background data throughout the sequence.

If the camera moves or the background is not static, the background motion has to be compensated. First, a motion estimation algorithm is applied. The occlusions are then filled in using an onion peeling strategy, starting from the outside of the occlusion such that the most uncertain part is located at the center. For each depth or color value to be filled in, a nearby background motion vector is used to determine the location in the other frame from where data has to be fetched. Fig. 6 shows an exemplary result of this approach.

Multi-view inpainting: If a scene is captured with multiple cameras as described in Section III, the other view(s) can be used for occlusion estimation. Depending on the number of cameras used and their position with respect to the 'central' camera, different occlusion parts can be filled in: cameras to the left or right can be used for filling in occlusions to the left or right around foreground objects, respectively. In order to fill in the occlusions, background information from the side view has to be warped to the correct location in the center view. This can be done using the depth map for the side view. Accuracy of this

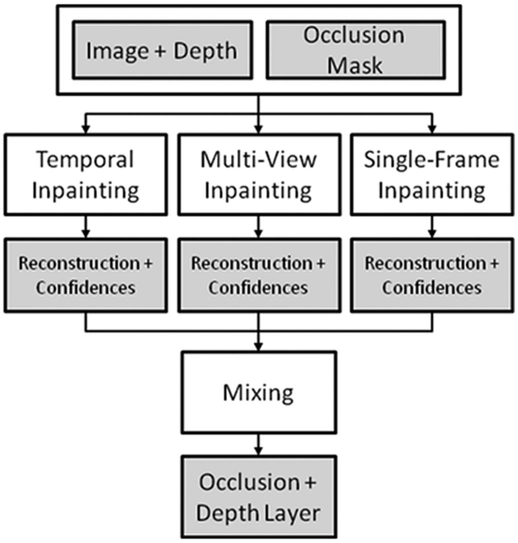


Fig. 7. Confidence-based approach for occlusion generation. Multiple inpainting methods are applied and combined using the respective confidence levels.

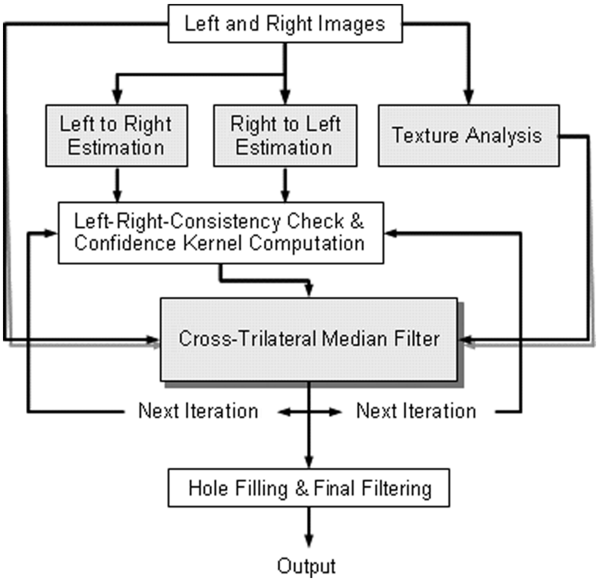

Left and Right Images
Left to Right Estimation
Right to Left Estimation
Texture Analysis
Left-Right-Consistency Check & Confidence Kernel Computation
Cross-Trilateral Median Filter
Next Iteration
Next Iteration
Hole Filling & Final Filtering
Output

Fig. 8. Processing scheme for the acquisition of highly consistent depth maps from an image pair.

side view depth map is very important, as it determines the precision of the inpainted data.

Single-frame inpainting: If no reliable temporal or multi-view data are available, single-frame inpainting is performed [32]. In this case, we opt for a very reliable and temporally stable inpainting method: repetition of the last known color and depth value of the background outside the occlusion. This is also a method with extremely low complexity, so it can also be used if computational power is too limited for the other methods.

Combined approach: In order to have a good, reliable general inpainting approach, we combine the above three methods in a confidence-based system. Each algorithm is applied separately, returning occlusion data along with a confidence estimate for all the data. Confidence measures are based on the matching errors, homogeneity of the inpainted data, and a prior confidence (temporal results are most reliable, while single-frame inpainting results have the lowest prior confidence). The results are then combined using these confidences to produce a stable and reliable occlusion layer (see also Fig. 7 the work by Oliveira *et al.* [33] and Klein Gunnewiek *et al.* [31]).

B. Depth Generation From Narrow-Baseline Data

1) Consistent Depth Maps From Stereo Data: Most of the content that is produced for stereoscopic viewing is captured with the simple setup of two horizontally displaced color cameras. In order to convert this input data to the more flexible LDV format, this section discusses a scheme that assigns consistent depth maps to each view. The process starts by estimating an initial set of depth maps, which are acquired by Hybrid Recursive Matching (HRM) [34]. HRM unites block-recursive disparity matching and pixel-recursive optical flow estimation in one common scheme to generate depth maps that are almost smooth and temporally consistent. However, due to noise, homogeneous areas, occlusions, etc., there are mismatches in the initial depth maps that will lead to artifacts during the rendering of virtual views. To eliminate these mismatches, an iterative post-processing approach inspired by bilateral filtering is applied. Fig. 8 sketches this processing scheme. Two independent matching processes first generate a pair of disparity maps via left to right and right to left estimation. By checking the consistency of the two depth maps and looking at the correlation values, the confidence of each estimate is measured. Based on these confidence values, a trilateral median filter [35] is applied to increase the consistency of the two depth maps. This trilateral filter extends a bilateral filter [36] which consists of geometric and photometric kernels, by means of third kernel that penalizes unreliable disparities. Moreover, the parameters of the filter are adapted to the local image texture. The scheme is applied iteratively, updating the consistency and confidence calculation at each iteration step. After a few iterations, the number of consistent estimates no longer increases, and a final occlusion inpainting procedure as discussed earlier is applied. Fig. 9 shows the processing results of a stereo pair from the MPEG sequence “Cafe.”

2) Fusion of ToF Depth and Multi-View Stereo: Providing only two views to the depth reconstruction does not allow for proper reconstruction of image elements occluded in one of the images; instead, inpainting techniques have to be used. Another problem with image based reconstruction is the presence of ambiguous image elements like homogeneous regions or repetitive structures. To cope with occlusions and ambiguities, the capture setups discussed in Section III have been developed. These contain more than two viewpoints and integrate Time-of-Flight cameras. The multiple viewpoints allow stereo reconstruction in occluded regions, and the ToF cameras deliver depth regardless of the scene’s texturing. Often the availability of depth from ToF measurements makes further stereo matching unnecessary. It is, however, necessary to upsample the low resolution ToF depth maps and to align them with the captured color images. For this purpose, Yang *et al.* [37] proposed a robust approach that was successfully adapted to the needs of 3D-TV data generation [29], [38]. In circumstances where ToF measurements

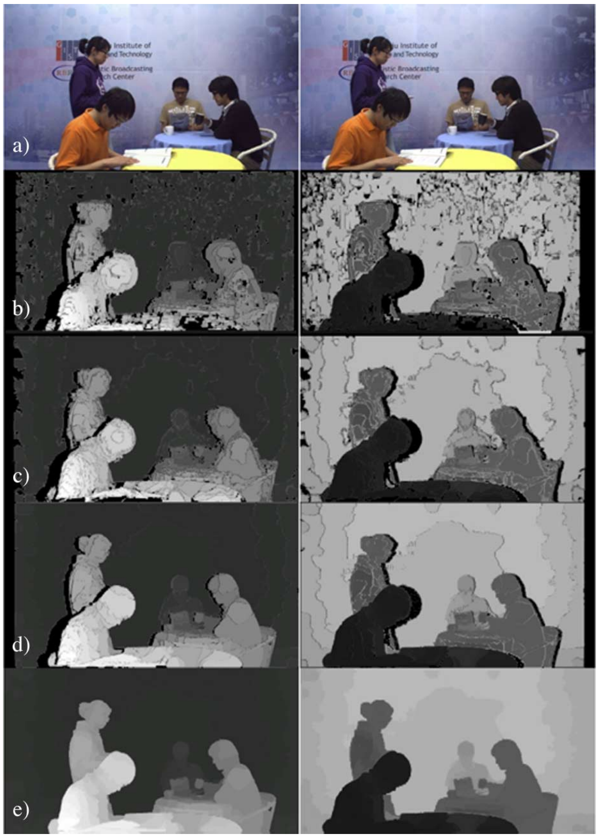


Fig. 9. Iterative processing to increase the consistency of disparity maps for a stereo input pair. (a) Original Input images. (b) Initial disparity maps. (c) Result after 1 iteration. (d) Result after 5 iterations. (e) Final disparities with simple occlusion filling. Left to right disparities are shown in the left column, and right to left disparities are shown in the right column. In the right to left disparities, dark values correspond to close objects. This relation is inverted in the left to right disparity case.

lack precision or contain gross errors, it is possible to combine them with stereo measurement to automatically extract reliable depth maps. Since the input data for stereo is fundamentally different from the data delivered by ToF-Cameras, it is difficult to define an algorithm that is not biased towards one of the data sources without discarding important information. A novel algorithm was proposed in 3D4YOU. This algorithm uses depth maps from local stereo estimation after removing all unreliable measurements. In this way, the low resolution ToF depth-measurements are supplemented by reliable high-resolution depth measurements. In contrast to previous proposals in this field, the different nature of the ToF-measurements and stereo reconstruction algorithms is thus considered. Hence, a bias towards the low resolution ToF maps is avoided, allowing fine structures only visible in the high resolution color data to be successfully reconstructed, while ambiguous image regions are resolved by the ToF measurements Fig. 10. [39] describes the algorithm in full detail.

C. LDV From Wide-Baseline Data

In diverse TV productions, like gameshows or sports, the scenery must be simultaneously captured from several widely-spaced viewing angles in order to provide appropriate coverage of the filmed action. Often, the space for capture equipment is limited, meaning that the use of large camera setups like stereo rigs or the system described in the Section *LDV Compliant Capture* are not feasible or can become very expensive. Fortunately, in such scenarios the background scenery, such as a sports arena, is known *a priori* and can be provided in digital form. This allows for an alternative approach to LDV generation that shall be discussed in the following.

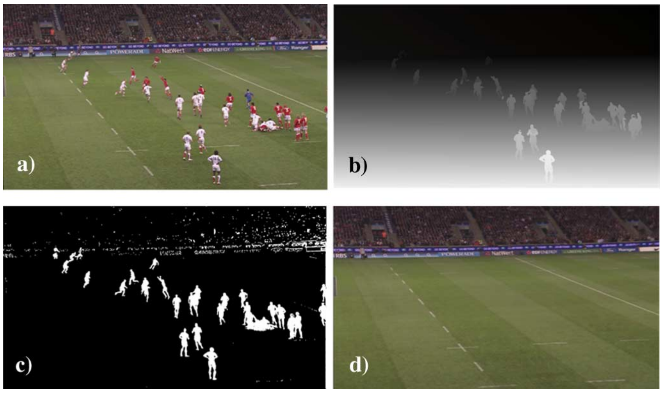


Fig. 10. (a) Color image and (b) foreground depth map generated with the wide baseline reconstruction approach. The depth of the players is found by applying a visual hull technique, while the playing field's depth is calculated from the given background model. The separation of foreground and background is coded in the mask shown in (c). This mask is generated by means of difference keying between (d) a prepared color plate and (a) the reference image. This color plate is also used for the occlusion color in the final LDV.

The wide-baseline LDV generation approach proposed here utilizes a static background model and requires a number of cameras with known parameters in a wide-baseline configuration. For broadcast cameras mounted on a pedestal, the camera parameters are computed in real-time from image-features alone [40]. The digital model of the static background's photometric and geometric properties is manually prepared before the actual LDV generation. This model then allows for foreground segmentation via chroma-keying and difference keying techniques (see). Finally, a 3D volumetric visual hull is computed, followed by an iso-surface generation to give a 3D surface model for each frame of the foreground action, as described in detail in [41].

In order to generate the final LDV format, the original camera images are augmented with a depth channel and an occlusion layer with depth and color. The required depth information for the foreground layer is acquired by rendering the generated 3D visual hull model of the foreground together with the static background model using a scan-line renderer.

The occlusion layer contains image information from the same camera angle as the original camera image but without the foreground action. While the required depth information is easily retrieved by again rendering a depth map via the static background model, the color information, which is subject to lighting conditions, is best used from images that were taken at approximately the same time. One option to generate this occlusion color is to fill in the obscured background from images taken by other cameras. One problem found with this procedure is that the areas often look slightly different in color, due to different color characteristics of the cameras (mismatched color balance) or anisotropic effects. Another option to fill in occluded areas is to use color images generated from

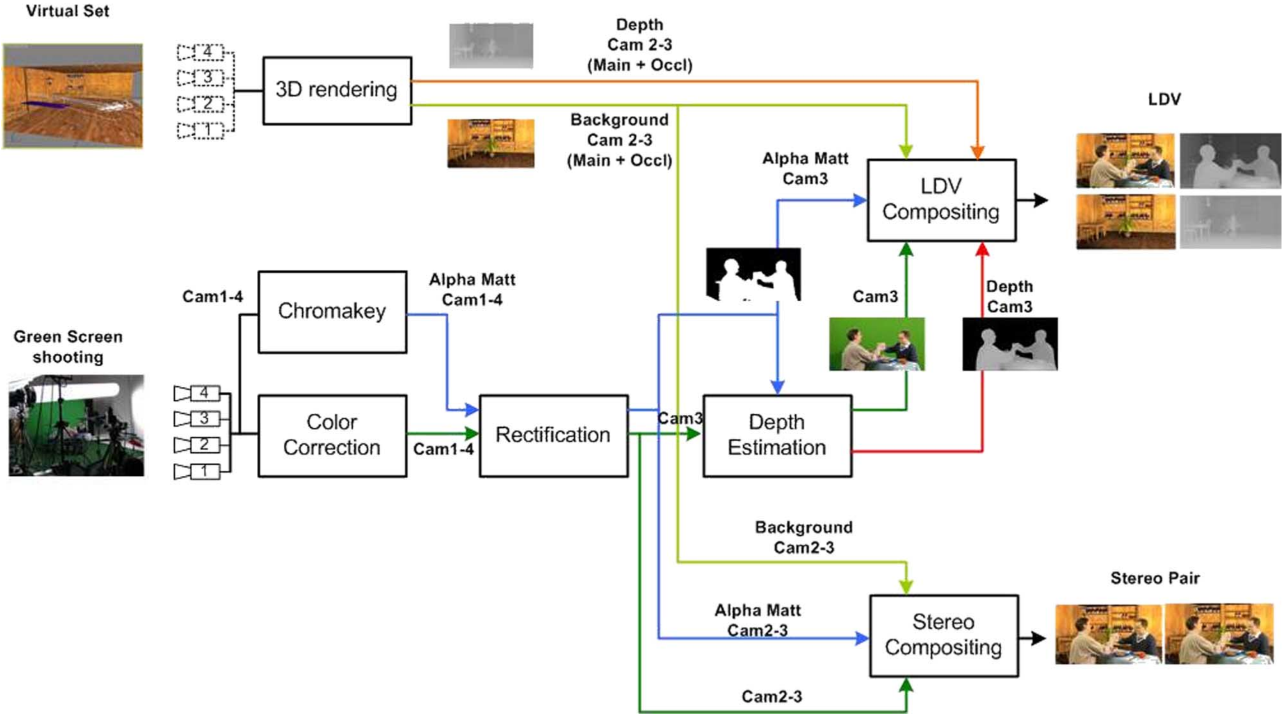


Fig. 11. Processing workflow of LDV and stereo compositing.

the background plates used in the difference keying technique described above. Although the generated images are slightly less detailed and miss shadows, the results of this approach look less disturbing.

V. Editing and Mixing

Beyond the flexibility provided by the LDV format to target different display types and sizes, this format also brings new capabilities in the editing and mixing domain. The depth map associated with the video represents a third mixing dimension that can integrate efficiently with conventional 2D mixing. The depth of the real scene, while not always fully reliable, can be exploited to overlay computer generated imagery (CGI) and can even be used to create advanced visual effects.

With LDV, the overlay of graphics is more straightforward compared to the same operation in stereo, where such graphics must be placed coherently from one view to the other. CGI can efficiently enrich natural LDV content, and as long as the mixing is properly performed, no new artifacts are introduced. For the CGI component, the depth information or its LDV representation can directly be extracted from the modeling tools used for the design.

Two use cases of editing and mixing are described next. First, virtual studio techniques used to help the generation of LDV content are discussed. The second use case highlights the exploitation of LDV depth information for inserting subtitles and logos.

1) Virtual Studio Techniques in LDV Production: Well known virtual studio concepts are used for mixing synthetic content with real content. The foreground of the final compositing is shot in front of a green or blue screen, and a synthetic background is then added Fig. 11. The fact that the background is generated from a 3D model brings many advantages. Its depth can be obtained directly from the rendering process that is usually provided with modeling tools. This ensures that the depth map is perfect for the background and does not require an reconstruction process. Furthermore, foreground depth estimation can exploit the alpha matte information from the chromakey, either for clipping the depth at the border of the foreground or simply as post-processing for cleaning up the depth maps.

The overall workflow is depicted in Fig. 11. This workflow covers the generation of both the LDV content (aligned with Camera 3) and a stereo pair (aligned with Camera 2 and 3). The 3D model of the background has to be manually aligned to the real scene. Then the rendering is processed with the camera parameters provided by the calibration step described in *Acquiring Camera Geometry*. Under the assumption that occluded areas contain primarily background, the occlusion layer (textures + depth) are directly and exclusively extracted from the synthetic background from the 3D rendering process. The alpha matting is performed for each view independently and undergoes geometric rectification to match the corresponding rectified videos. The alpha matte is then reused for mixing the synthetic background with the real foreground for both the LDV signal and the stereo pair.

In this technique, the background must be as coherent as possible with the natural foreground, both in terms of context and rendering quality. Especially when the resulting sequences are used for subjective tests, as described in the Section *Perceptual Human Factors*, any inconsistency can introduce a bias in the user test results.

2) Enhanced Subtitling and Logo Insertion for LDV Content: The insertion of subtitles and logos has become a challenging

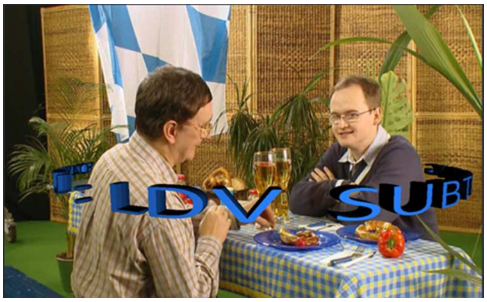


Fig. 12. Result of LDV subtitling. Due to the given depth information, mutual occlusions between the inserted elements and the scene are handled automatically.

topic in 3D-TV in comparison with current practices for conventional video. The positioning of added CGI in the Z direction (screen to user axis) with regard to the original signal must be adjusted on viewing criteria. Furthermore, with the LDV format, new possibilities for mixing are offered by the depth information, such as new ways of inserting subtitles or logos. In addition to X and Y mixing, a Z mixing can be employed using the depth when appropriate and/or reliable enough. As an example of new effects which can be proposed, Fig. 12 shows a subtitle rotating around the table, appearing or disappearing depending on its position with respect to the objects and actors in the scene.

It is essential that the Technical Director in charge of editing can preview or quickly check the effect and quality of the mixing. In 3D4YOU, a real-time depth based compositing system was developed, allowing a live insertion of subtitles and logos. Graphic content is designed offline using a 3D modeling tool and exported to the editing system. The rendering of this content to a LDV signal is performed first, and it is combined with the original LDV signal on pixel basis, taking into account both signals' depth information. The quality of the resulting pictures directly depends both on the reliability of the depth maps and the requirements of the mixing itself. The proposed real-time LDV compositing system will help the Technical Director decide on the appropriate mixing effects depending on the depth map's quality.

VI. Coding and Distribution

Having generated 3D-TV content in LDV format, the next important step is the encoding and distribution of the data to the living room, where it then can be viewed on different displays. In contrast to standard TV, the amount of data transmitted, as well as the requirements for the data's quality after transmission, has increased. Investigations of compression algorithms were therefore conducted, which are discussed in the next section. This is followed by a short description of an experimental transmission setup, which demonstrates the distribution of LDV data via Digital Video Broadcasting-Terrestrial (DVB-T) and Internet Protocol Television (IPTV).

A. Encoding and Compression

A number of depth-enhanced formats for 3D video are currently investigated in the research community: MVD with 2 and 3 views, LDV, and also LDV associated with a second view (see also Section *Review of Depth-based formats* in 3D-TV). Such formats are also investigated by standardization organizations, like ISO MPEG, in the context of a new 3D video coding technology [20], [42], [43]. This subsection contributes to this work by describing different investigations conducted for finding a LDV representation that provides higher coding efficiency. Standards in MPEG already exist to efficiently distribute single view color video, stereo video signals [44], and even depth maps [45], [46]. These standards were used as a starting point to code and distribute LDV data, as discussed in Section *Distribution and Reception*, and to investigate possible improvements in the coding efficiency.

The primary target for improved efficiency is the proper coding of depth data and of occlusion information. The quality of this information is vital for a good synthesis of stereoscopic views at the receiver site, thus the optimization of bandwidth used via compression becomes a demanding task. For the assessment of the impact of novel approaches to coding and transmission of LDV, quality measures were studied. From these studies it can be concluded that while the perceived quality of the rendered stereoscopic images can be well represented by the SSIM index [47], the PSNR between an original and an encoded depth map is most correlated to the synthesized images' quality (see [48] for details). Therefore, the PSNR was used as an objective quality measure to assess the improvement of depth encoding approaches.

For the compression of depth information carried in the LDV format, it can be exploited that depth maps in general contain only a few discontinuities and therefore largely consist of low frequency content. This suggests that depth maps with lower resolution can be used for encoding. Experiments show that the results of this compression strategy depend on the content's complexity. This approach is most promising when low bitrates have to be achieved; however a certain loss in quality has to be accepted.

It was moreover observed that in comparison to other depth based formats, the coding and transmission efficiency of LDV due to its sparse occlusion information is quite high. However, the content in the occlusion layer may be scattered spatially. This is disadvantageous with respect to state-of-the-art coders like MPEG-4 AVC/H.264 or MPEG MVC, which operate block-wise. Further processing of the occlusion layer is thus proposed. First, pixel groups consisting of only a few elements are removed. These small holes are easily interpolated in the later image rendering process, and moreover the temporal alignment of the image data is increased. Pixel groups that contain sufficient occlusion information are then filled up to form square blocks of pixels in the occlusion layer. With this change, experiments performed demonstrate an increase of coding efficiency of up to 25% in bit rate while preserving the reproduction quality. In this context, a coding parameter that controls the significance of occlusion information by means of a minimal disparity threshold was introduced. In the case where the change in disparity at a disocclusion boundary is small (i.e., the size of the required occlusion information is small), the threshold initializes a local muting of the occlusion layer, which further increases the coding efficiency.

Further experiments were designed to exploit the Flexible Macro Block Ordering (FBO) technique, which is a novel coding technique in the MPEG-4 AVC/H.264 standard. This technique allows the assignment of image blocks to so called slices. The advantage of this assignment is that each slice can be coded independently, so that image elements with different characteristics can be coded differently. Moreover, advanced error correction mechanisms become applicable, like the interpolation of lost blocks from other received slice content. On the other hand, this technique requires the encoding and transmission of the Macro Block Allocation map. Unfortunately, due to this requirement the experiments conducted could not show a gain in efficiency by means of FBO.

Finally, it must be remarked that with respect to compression, the estimation techniques applied to extract depth information as described in Section *LDV Generation* are generally not stable enough. Due to the imperfect estimation process, content may lack spatial and temporal consistency. As a consequence, natural sequences and CGI content show big differences in behavior. For instance, a depth signal represents 20% of global bandwidth for CGI and 50% for natural sequences.

B. Distribution and Reception

In order to show the practicability of 3D-TV in the LDV format with distribution standards available today, a demonstrator was built. This demonstrator consists of a transmission part and a reception part. Both DVB-T and IPTV are supported transmission media. The receiver is based on a PC platform and supports reception of data in the simplified LDV format in both MPEG2 and MPEG4 encoded formats. The receiver first does a demodulation of the DVB-T or IPTV modulated streams, followed by demultiplexing into four elementary streams according to the four layers of the simplified LDV format: texture, depth, occlusion and occlusion depth (see also Section *Review of Depth-based formats* in 3D-TV and Fig. 1). Each elementary stream is decoded, and all decoded streams are packed together into the LDV format, which is supported by the Philips autostereoscopic display. Passive glasses-based Stereoscopic displays only showing two views in a frame interleaved pattern are also supported by means of an additional renderer implemented on a FPGA. Due to the nature of the chosen format, the derivation of standard TV video data is also possible. A simplified schematic of the demonstrator is shown in Fig. 13.

VII. Perceptual Human Factors

As with the assessment of any visual technology, perceptual human factors must be given ample attention. The two most important aspects to consider are the overall viewing experience, which includes image quality as well as the added value of depth, and the visual comfort, which includes eye strain and fatigue that are potentially exacerbated by 3D viewing.

To understand these aspects, it is necessary to consider all the technological details, distributed throughout the audiovisual chain, that influence the user's perception of 3D TV services. Inherent sources of degradation, according to the type of architecture [49], include the capture and production of video sequences, the compression of 3D video formats, the network type and associated protocols, and the image formats and rendering technology used in 3D displays. 3D4YOU aims to provide experimental results regarding the perceptual impact of the 3D video formats, coding and compression methods, and long-duration viewing, as well as guidance for capture and production of 3D content. A description of these ongoing activities follows. Eventually, these studies and methodologies should be integrated into the standardization process, e.g. ITU-T and ITU-R activities, to have a fair assessment as well as to compare results from different laboratories in the world.

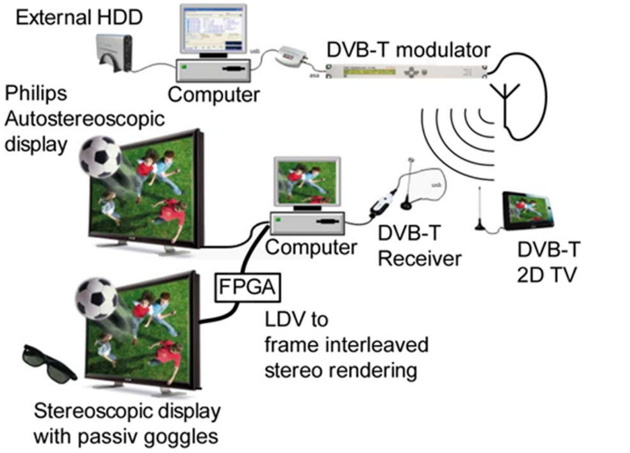


Fig. 13. Simplified schematics of the hardware setup used to demonstrate the transmission of LDV data to different display types.

A. Evaluation of Viewing Experience

In order to ease the specification process and the setup of the end-to-end architecture, a performance evaluation has to be carried out with subjective tests aimed at understanding the end-user's opinion in terms of visual quality, depth rendering, visual comfort, etc. Most important for 3D4YOU is the evaluation of the technologies developed in the project. For this purpose, subjective tests on the 3D representation formats under investigation, including not only LDV, but also LDV-R, MVD2, and DES, and the 3D compression strategies related to the studied video formats are conducted.

Perceptual studies related to the assessment of viewing experience have two general requirements. The first concerns the production of 3D sequences for testing purpose, which must have sufficient variety in content and be sufficiently artifact-free. Original test sequences must have full temporal and spatial definition, with no compression, except when assessing the impact of compression methods. To study the effect of reconstruction, sequences must have various scene complexities in terms of motion, texture and depth range, e.g. indoor and outdoor contents. A full factorial of three levels for each complexity criterion would lead to twenty-seven cases of video complexities, but in practice, four to five video complexities are enough to determine the impact of scene contents on results. This requirement has been addressed in 3D4YOU in the shooting of new live-action sequences and the generation of synthetic sequences.

The second requirement is for new subjective assessment methodologies for short duration sequences using multidimensional scales (visual experience, video quality, depth rendering, visual comfort, etc.). Previous studies [50], [51] have shown that it is essential to develop multidimensional approaches using perceptual labels such as visual experience, naturalness, presence or depth rendering to describe more precisely the viewer's opinion. For this purpose, a new subjective test methodology, based on the SAMVIQ methodology (described in ITU-R BT.1788) has been defined and implemented to take into account multidimensional perceptual scales, specifically viewing experience, depth rendering, and visual comfort.

A first test, on 3D representation formats, was aimed at comparing the visual performance of rendered views from LDV and to compare this new format with native stereo i.e. the original stereo content captured at the same time. Three main objectives have been defined: to determine the value of depth scalability functionality, to evaluate the efficiency of the occlusion layer in the LDV format, and to assess the stereo backward compatibility of this new video format in terms of video quality. The test has been carried out considering four sequences from the 3D4YOU shooting sessions. Unfortunately, the prototypic depth reconstruction and rendering algorithms developed in 3D4YOU available at the time of the experiments did not achieve the highest quality possible. Therefore, the evaluation of these experiments indicates that directly captured imagery has a higher visual quality than the imagery rendered from the LDV representation. It however cannot be deduced what potential the LDV representation has towards achieving comparable visual quality.

B. Evaluation of Visual Comfort

To evaluate the visual comfort of a 3D-TV system, it is necessary to specify objective and subjective measurements [52]–[54] as well as to develop methodologies for long duration viewing. Recent literature [55] makes a distinction between visual fatigue, which is a measurable decrease in optical performance, and visual discomfort, which is its subjective correlate. It provides a variety of measurement methodologies and shows that some viewers, approximately 20% of the population, can be considered more susceptible to discomfort and visual fatigue. Because long duration LDV content was not available in the 3D4YOU project, contemporary, high quality stereo video can be used, as it is assumed that the mechanisms behind visual comfort and fatigue are independent of the encoding format.

In 3D4YOU, a study was completed which looked at the visual discomfort associated with the effects of motion, depth, and subtitling. The study combined a continuous assessment of comfort during a 30-minute viewing session with questionnaire-based assessments of physical discomfort and contributing factors. The video content consisted of two movies varying in motion magnitude: a dynamic sports sequence and a relatively static travelogue, both of which had similar, high amounts of depth. The sequences were shown with and without subtitles, across participants. Visual fatigue, as measured with a vergence facility test before and after viewing, was not significantly encountered, although some mild discomfort was noted in questionnaire results. The questionnaire also confirmed that participants incorporated on-screen motion, depth, and changes in depth in their continuous assessment of visual comfort. Modeling the continuous assessment of visual comfort as a linear combination of objective video characteristics related to motion and depth was successful for individual scenes. However, because of the inter-dependency between video characteristics and their non-linear relationship with visual comfort, creating a single model of comfort for all of the video content was not possible. It was found that additional effort was required while reading subtitles to keep them visually fused and sharp. This induced additional discomfort, yet did not impact the benefits of 3D viewing in terms of naturalness and viewing experience. From these results it can be concluded that the possibilities granted to post-production by the LDV format, e.g. the possibility to change the depth impression and the convenient mixing facilities with early previews described in Section V-A-2) *Enhanced subtitling and logo insertion for LDV content*, are very important in proper 3D-Film productions.

VIII. Conclusion and Future Work

This paper discusses a 3D-TV delivery chain that is based on the Layered Depth Video (LDV) format. This depth-based content representation provides functionality that is essential for display independent delivery of 3D-TV to the home and difficult to provide by other means. The paper contributes detailed discussions of all major elements of this content generation approach and presents state-of-the-art solutions to challenges encountered with this scheme. Specifically, a novel and pragmatic capture setup was described incorporating color cameras as well as ToF cameras. Furthermore, different novel algorithms were presented that extract depth and occlusion information from different input sources and fuse them into the LDV format. Also described was the exploitation of the depth information provided in the LDV format to edit and mix captured content with computer generated data. New insights into compressing LDV signals were discussed, and the feasibility of transmitting content in LDV format to different display types was demonstrated. Another focus of the paper was human perceptual factors with respect to production rules and the assessment of the developed technology's visual quality. Here, novel test methodologies and early results were explained. The insights and results presented in this work are based on real production trials that were conducted within the European 3D4YOU project and are thus representative for real-life applications.

Nevertheless, further investigations and developments are required to make this approach practical in real productions. First of all, the prototype implementations made in 3D4YOU have to be integrated into a solid workflow by reworking control interfaces and by the construction of dedicated hardware solutions. Many of the algorithms presented in the paper already consider parallelization and the use of GPU hardware. In order to achieve live broadcast capabilities, however, run-time optimization of the whole processing pipeline, presented in this work, will be needed. Further scientific work that provides improved solutions to specific parts of the LDV based 3D-TV production pipeline will have to be carried out to ensure high quality service in concert with future consumer demands. This paper provides the necessary background information for this, as well as a valuable starting point for further investigations. In particular, auto-calibration techniques have to be investigated that are also able to calibrate ToF cameras, so that the proposed capture setup becomes simpler to use. Moreover, it is necessary to investigate novel coding, compression, and transmission standards. Also, the rendering of stereoscopic images from LDV has great potential for quality improvements. Furthermore, the adaptation of the many proposed stereo reconstruction techniques that only focus on the high quality reconstruction of foreground depth to the proposed LDV reconstruction is worth investigating. Finally, proper production rules for capturing and cinematography with respect to human factors have to be refined.

Despite the remaining work on quality-improvements and production techniques, this paper shows that 3D4YOU was able to develop key technologies for the realization of a depth-based 3D-TV production chain in the LDV format. It presents a proof of concept for the feasibility of this approach for many production scenarios ranging from documentaries to sports. Because this approach is display independent, supports multi-view autostereoscopic displays, and eases post-production, 3D4YOU's approach has the potential to become a reasonable alternative to the 3D film production techniques applied today.

References

[1] L. Onural, T. Sikora, J. Ostermann, A. Smolic, M. Civanlar, and J. Watson, "An assessment of 3DTV technologies," in *Proc. Nat. Assoc. Broadcast. Show*, 2006.

[2] S. Reichelt, R. Häussler, G. Fütterer, and N. Leister, "Depth cues in human visual perception and their realization in 3D displays," in *Proc. SPIE*, 2010, pp. 76900B–76900B-12.

[3] N. A. Dodgson, "Autostereoscopic 3D displays," *Comput.*, vol. 38, pp. 31–36, 2005.

[4] J. Cutting, "How the eye measures reality and virtual reality," *Behavior Res. Methods, Instrum., Comput.*, vol. 29, no. 1, pp. 27–36, 1997.

[5] D. Hoffman, A. Girshick, K. Akeley, and M. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *J. Vis.* vol. 8, no. 3, p. 33, Sep. 2010 [Online]. Available: http://www.journalofvision.org/content/8/3/33.long, (Mar. 2008)

[6] B. Mendiburu, *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Burlington, MA: Focal Press, 2009.

[7] R. Klein Gunnewiek and P. Vandewalle, "How to display 3D content realistically," in *Proc. Int. Workshop Video Process. Quality Metrics Consum. Electron. (VPQM)*, 2010.

[8] C. Buehler, M. Bosse, L. Mcmillan, S. Gortler, and M. Cohen, "Unstructured lumigraph rendering," in *Proc. Comput. Graphics (SIGGRAPH)*, 2001, pp. 425–432.

[9] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proc. SPIE*, vol. 5291, pp. 93–104, 2002.

[10] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3D," in *ACM Trans. Graph. (SIGGRAPH)*, 2010, vol. 29, pp. 75:1–75:10.

[11] J. Ohm, K. Grüneberg, E. Hendriks, E. Izquierdo, D. Kalivas, M. Karl, M. D. Papadimatos, and A. Redert, "A realtime hardware system for stereoscopic videoconferencing with viewpoint adaptation," *Signal Process.: Image Commun.*, vol. 14, pp. 147–171, 1998.

[12] C. Fehn, P. Kauff, M. Op De Beeck, F. Ernst, W. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton, "An evolutionary and optimised approach on 3D-TV," in *Proc. Int. Broadcast Conf.*, 2002, pp. 357–365.

[13] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, O. A. Smolic, and R. Tanger, "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," *Image Commun.*, vol. 2, pp. 217–234, 2007.

[14] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *ACM Trans. Graphics (SIGGRAPH)*, 2004, pp. 600–608.

[15] A. Smolic, K. Müller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View synthesis for advanced 3D video systems," *EURASIP J. Image Video Process.* p. 11, 2008 [Online]. Available: http://www.hindawi.com/journals/ivp/2008/438148.abs.html, (2008, Nov.) [Sep., 2010]

[16] Y. Huang and C. Zhang, "A layered method of visibility resolving in depth image-based rendering," in *Proc. Int. Conf. Pattern Recog.*, 2008, pp. 1–4.

[17] J. W. Shade, S. Gortler, L.-W. He, and R. Szeliski, "Layered depth images," in *Proc. 25th Annu. Conf. Comput. Graphics Interactive Tech.*, 1998, pp. 231–242.

[18] 3D Solution, "3D Interface specification," Jan. 14, 2011 [Online]. Available: http://www.business-sites.philips.com/shared/assets/3dsolutions/downloads/3DInterfaceWhitePaper.pdf., (Apr 08, 2009)

[19] W. H. A. Bruls and R. Klein Gunnewiek, "Options for a new efficient, compatible, flexible 3D standard," in *Proc. Int. Conf. Image Process. (ICIP)*, 2009, pp. 3461–3464.

[20] A. Smolic, K. Müller, P. Merkle, P. Kauff, and T. Wiegand, "An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution," in *Proc. Picture Coding Symp. (PCS)*, 2009, pp. 389–392.

[21] Z. Xu, R. Schwarte, H. Heinol, B. Buxbaum, and T. Ringbeck, "Smart pixel - photonic mixer device (PMD)," in *Int. Conf. Mechatronic Mach. Vis. Practice*, 1998, pp. 259–264.

[22] R. Lange, P. Seitz, A. Biber, and R. Schwarte, "Time-of-Flight range imaging with a custom solid-state image sensor," in *Laser Metrology Inspection (EOS/SPIE)*, 1999.

[23] H. Kraft, J. Frey, T. Moeller, M. Albrecht, M. Grothof, B. Schink, H. Hess, and B. Buxbaum, "3D-Camera of high 3D-framerate, depth-resolution and background light elimination based on improved PMD(Photonic Mixer Device)-technologies," in *Int. Conf. Opt. Technol., Opt. Sensors Meas. Tech.*, 2004.

[24] L. Wang, G. Gong, M. Gong, and R. Yang, "How far can we go with local optimization in real-time stereo matching," in *Proc. 3D Process. Vis. Transmiss.*, 2006, pp. 129–136.

[25] K. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 650–656, 2006.

[26] A. Frick, B. Bartczak, and R. Koch, "Real-time preview for layered depth video in 3D-TV," in *Proc. SPIE*, 2010, pp. 77240F–77240F-10.

[27] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge University Press, 2004.

[28] I. Schiller, C. Beder, and R. Koch, "Calibration of a PMD-Camera using a planar calibration pattern together with a multi-camera setup," in *Proc. Soc. Photogrammetry Remote Sens.*, 2008, pp. 297–302.

[29] A. Frick, B. Bartczak, and R. Koch, "3D-TV LDV content generation with hybrid ToF-multicamera rig," in *Proc. IEEE 3DTV Conf. : True Vis.—Capture, Transmiss. Display 3D Video*, 2010, pp. 1–4.

[30] B. Barenbrug, R.-P. M. Berretty, and R. Klein Gunnewiek, "Robust image, depth, and occlusion generation from uncalibrated stereo," in *Proc. SPIE 6803*, 2008.

[31] R. Klein Gunnewiek, R.-P. M. Berretty, B. Barenbrug, and J. P. Magalhaes, "Coherent spatial and temporal occlusion generation," in *Proc. SPIE 7237*, 2009.

[32] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Trans. Image Process.*, vol. 12, pp. 882–889, 2003.

[33] M. Oliveira, B. Bowen, R. McKenna, and Y. Chang, "Fast digital image inpainting," in *Proc. Int. Conf. Vis., Imaging, Image Process.*, 2001, pp. 261–266.

[34] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo-analysis by hybrid recursive matching for real-time immersive video conferencing," *IEEE Trans. Circuits Syst. Video-Technol.*, vol. 14, pp. 321–334, 2004.

[35] M. Mueller, F. Zilly, and P. Kauff, "Adaptive cross-trilateral median filter," in *Proc. IEEE 3DTV Conf.: True Vis.—Capture, Transmiss. Display 3D Video*, 2010, pp. 1–4.

[36] C. Tomasi and R. Manduchi, "Bilateral ?ltering for gray and color images," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 1998, pp. 839–846.

[37] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[38] A. Frick, F. Kellner, B. Bartczak, and R. Koch, "Generation of 3D-TV LDV-content with time-of-flight camera," in *Proc. IEEE 3DTV Conf.: True Vis. - Capture, Transmiss. Display 3D Video*, 2009, pp. 1–4.

[39] B. Bartczak and R. Koch, "Dense depth maps from low resolution time-of-flight depth and high resolution color views," in *Proc. Int. Symp. Adv. Vis. Comput.*, 2009, pp. 228–239.

[40] G. A. Thomas, "Real-time camera tracking using sports pitch markings," *J. Real Time Image Process.*, vol. 2, no. 2/3, pp. 117–132, Nov. 2007.

[41] O. Grau and V. Vinayagamoorthy, "Stereoscopic 3D sports content without stereo rig," in *Proc. IBC*, 2009.

[42] *Vision on 3D Video*, ISO/IEC JTC1/SC29/WG11 (MPEG) N10357, Video and Requirements Group, 2008, Lausanne, Switzerland.

[43] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P. H. N. de With, and T. Wiegand, "The effects of multiview depth video compression on multiview rendering," *Signal Process.: Image Commun.*, vol. 24, no. 1+2, pp. 73–88, 2009.

[44] A. Vetro, Y. Su, H. Kimata, and A. Smolic, "Joint multi view video model," Joint Video Team, Hangzhou, China, Doc. JVT-U207, Oct. 2006.

[45] ISO/IEC JTC1/SC29/WG11, "Text of ISO/IEC FDIS 23002-3 representation of auxiliary video and supplemental information," Marrakech, Morocco, Doc. N8768, Jan. 2007.

[46] ISO/IEC JTC1/SC29/WG11, "Text of ISO/IEC 13818-1:2003/FDAM2 carriage of auxiliary data," Marrakech, Morocco, Doc. N8799, Jan. 2007.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, Apr. 2004.

[48] P. Kerbiriou and G. Boisson, "Looking for an adequate quality criterion for depth coding," in *SPIE Electron. Imaging, 3D Image Process. (3DIP) Appl. Conf.*, 2010.

[49] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "New requirements of subjective video quality assessment methodologies for 3D-TV," in *5th Int. Workshop Video Process. Quality Metrics Consum. Electron. (VPQM)*, 2010.

[50] L. Meesters, W. IJsselsteijn, and P. Seuntiens, "A survey of perceptual evaluations and requirements of three-dimensional TV," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, pp. 381–391, 2004.

[51] P. Seuntiëns*, Visual Experience of 3D TV*. Eindhoven: Eindhoven University of Technology, 2006.

[52] A. Murata, A. Uetake, W. Otsuka, and Y. Takasawa, "Proposal of an index to evaluate visual fatigue induced during visual display terminal tasks," *Int. J. Human Comput. Interaction*, vol. 13, pp. 305–321, 2001.

[53] A. Uetake, A. Murata, M. Otsuka, and Y. Takasawa, "Evaluation of visual fatigue during VDT tasks, Systems, Man, and Cybernetics," in *IEEE Int. Conf. Syst., Mach. Cybern.*, 2000, pp. 1277–1282.

[54] F. L. Kooi and A. Toet, "Visual comfort of binocular and 3D displays," *Displays*, vol. 25, pp. 99–108, 2004.

[55] M. T. M. Lambooij, W. A. IJsselsteijn, M. Fortuin, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *J. Imaging Sci. Technol.*, vol. 53, no. 3, p. 030201-(14), 2009.

Bogumil Bartczak received the diploma in computer engineering from the University of Kiel, Germany, in 2005.

Since 2005 he has been with the Multimedia Information Processing Group at the Computer Science Institute of the University of Kiel. In this time he has been working in several scientific projects focusing on the integration of computer vision and computer graphics methods into information processing systems. His field of expertise is the dense depth reconstruction from color images also in combination with active time-of-flight sensors, Augmented Reality and general purpose GPU programming.

Patrick Vandewalle (M'03) received the M.S. degree in electrical engineering from Katholieke Universiteit Leuven, Belgium, in 2001 and the Ph.D. degree from Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland, in 2006.

From 2006 to 2007, he was a Postdoctoral Researcher at EPFL. He joined Philips Research in September 2007, where he is currently a Senior Scientist. His research interests are in signal and image processing, sampling, digital photography, and 3D.

Oliver Grau (M'10) received the Diploma (Master) and the Ph.D. from the University of Hanover, Germany.

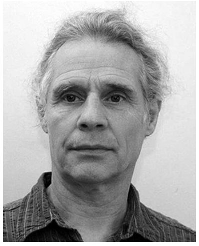
From 1991-2000 he worked as a Research Scientist at the University of Hanover and was involved in several national and international projects, in the field of industrial image processing and 3D scene reconstruction for computer graphics applications. In 2000 he joined the BBC Research & Development Department in the UK. He was working on a number of national and international projects on 3D scene reconstruction and visualization. His research interests are in new innovative tools for visual media production using image processing, computer vision and computer graphic techniques and he published a number of research papers and patents on this topic. He was and is active as reviewer for scientific journals, research bodies like EPSRC, EC-FP7 and as a programme committee member of several international conferences. Further he was the initiator and chair of CVMP, the European Conference on Visual Media Production in London.

Gerard Briand was graduated from the "Ecole Supérieure d'Electricité" located in Paris.

He started his career at Thomson CSF in 1986 as a R&D Engineer in the field of video processing and video compression. He successively managed projects dealing with Content Production and Content Delivery. He is currently working at Technicolor R&D France (formerly Thomson R&D France) at the Research & Innovation division. Gerard Briand's main interests are in the field of computer graphics, mixed reality technologies as well as new interaction systems, he particularly works on related innovative applications targeting both professional and consumer domains. In this context he is responsible for a collaborative project called ReV-TV investigating the association of virtual reality techniques with TV programs.

Jérôme Fournier received the Ph.D. in signal and image processing from the University of Rennes in 1995.

The main aim of his thesis was the subjective evaluation of stereoscopic television. This led to his participation in the European project, RACE DISTIMA. After that, he worked on video compression for video communications at LEP (Philips Research Laboratory in France) and participated in the standardization process of H.263. In 1997, he joined France Telecom R&D and worked on the subjective evaluation of video sequences and on the implementation of standardized video codecs like MPEG-4 Part 2 and H.264. From 2004-2006, he was in charge of HDTV activities until the deployment of France Telecom HDTV service in June 2006. Since 2006, Jérôme has been working on the deployment of the Orange stereoscopic 3DTV as well as on the assessment of innovative 3DTV depth-based video formats as part of the European project, 3D4YOU.



Paul Kerbiriou received the Ph.D. (Dr.-Ing.) in biomedical engineering from the University of Technology of Compiègne (UTC) in 1985.

Paul has worked for the Caption Company (Telmat Group) in the 3D graphics domain in which he played the role of system architect. He joined Technicolor R&D France (formerly Thomson R&D France) in 1999 to bring his expertise in multimedia system architecture for MPEG-4 Systems applications development. He is currently working on 3DTV applications, including video processing and coding. He has also been involved in French and European collaborative projects, the most recently one being 3D4YOU.

Michael J. Murdoch received the M.S. in computer science from Rochester Institute of Technology, NY, in 2005, and the B.S. in chemical engineering from Cornell University, Ithaca, NY, in 1997.

He worked for 11 years for Eastman Kodak in Rochester, NY, modeling and evaluating system image quality of OLED displays, inkjet printing, film, and motion picture systems. Since 2008, he has worked for Philips Research in Eindhoven, The Netherlands, where his work centers on visual perception for display and lighting applications.