Automatic View Synthesis by Image-Domain-Warping

Nikolce Stefanoski, Oliver Wang, Manuel Lang, Pierre Greisen, Simon Heinzle, Aljosa Smolic

***Abstract*—Today, stereoscopic 3D (S3D) cinema is already mainstream, and almost all new display devices for the home support S3D content. S3D distribution infrastructure to the home is partly already established in form of 3D Blu-ray discs, video on demand services, or television channels. However, the necessity to wear glasses is often considered as an obstacle, which hinders broader acceptance of this technology in the home. Multiview autostereoscopic displays enable a glasses free perception of S3D content for several observers simultaneously, and support head motion parallax in a limited range. In order to support multiview autostereoscopic displays in an already established S3D distribution infrastructure, a synthesis of new views from S3D video is needed. In this paper, a view synthesis method based on Image-domain-Warping (IDW) is presented which synthesizes new views directly from S3D video and functions completely automatically. IDW relies on an automatic and robust estimation of sparse disparities and image saliency information, and enforces target disparities in synthesized images using an image warping framework. Two configurations of the view synthesizer in the scope of a transmission and view synthesis framework are analyzed and evaluated. A transmission and view synthesis system that uses IDW was recently submitted to MPEG's call for proposals on 3D Video Technology, where it was ranked among the four best performing proposals.**

***Index Terms*—Three dimensional TV, autostereoscopic displays, format conversion, automatic image synthesis, content creation for multiview autosterescopic displays, energy minimization, sparse disparities, disparity constraints.**

I. Introduction

Stereoscopic 3D (S3D) cinema and television are in the process of changing the landscape of entertainment. Primarily responsible for the change is the fact that technologies ranging from 3D content creation, to data compression and transmission, to 3D display devices are steadily improving and adapted to enable a rich and higher quality 3D experience. However, the necessity to wear glasses is often regarded as a main obstacle of today's mainstream stereoscopic 3D display systems. Multi-view autostereoscopic displays (MAD) overcome this obstacle. They enable glasses free stereo viewing by emitting several images at the same time. The MAD technology ensures that each viewer in front of the display sees only that stereo pair which is appropriate for his particular viewing position. MADs support also motion parallax viewing in a limited range, i.e. occluded scene parts become visible while other parts are again occluded when a viewer moves his or her head. To achieve these advanced functionalities, MADs require not two but many different views as input. Typical MADs which are on the market today require 8-views [1], 9-views [2] or even 28-views [3] as input. Because of the different number of input views required by different MADs, no unique display format exists for such displays. This fact has an impact on the other data formats and format conversion processes involved in the distribution chain from content creation to display (Fig. 1).

Fig. 1. High level view on the usual format conversions required from content production to display.

A naïve approach to produce content for a MAD would be to directly create content in the appropriate N-view display format. However, such an approach is impractical because of to the large number of views which would have to be captured and the high transmission bit rate which would be required to transmit N views to the end-user. Additionally, another drawback of this approach is the fact that the number of views to be captured and transmitted depends on the particular MAD device available at the end-user side. Obviously, there is a need to decouple the production format from the display format. According to the formats involved in the distribution chain (Fig. 1), the transmission format has to enable such a decoupling. Hence, a good transmission format has to fulfill the following requirements:

- An automatic conversion from production to transmission format has to be possible. For live broadcast applications also real-time conversion is required
- The transmission format has to allow an automatic and real-time conversion into any particular N-view display format.
- The transmission format has to be well compressible to save transmission band width.

Today, professional and consumer 3D content production is dominated by S3D content, i.e. 2-view content which is watchable on stereoscopic displays. It is believed that S3D as a production format will dominate over years. On the other hand, display formats for MADs consist of N>2 views. Hence, to close the gap between production and display formats, technologies are required which

All authors are with Disney Research Zurich. Manuel Lang and Pierre Greisen are also affiliated with ETH Zürich.

- Enable an automatic (and real-time) conversion of a 2-view production format into a well compressible transmission format.
- Enable a real-time conversion from the transmission format into a display format. Due to the fact that in the display format the number of required views N may vary, view synthesis technology is required which generates news views based on the particular transmission format.

Research communities [4][5][6][7] and standardization bodies [8][9] continue to investigate formats which are well compressible and enable efficient generation of novel views as required by MADs. Such formats can be divided into two classes: video only formats like S3D and multiview video in general, and depth enhanced formats like singleview video plus depth and multiview video plus depth. The per-pixel depth information included in the depth enhanced formats provides information on the depth structure of the 3D scene. Such depth data can be used for view synthesis by depth-image-based rendering (DIBR) [10][11]. However, high quality view synthesis with DIBR requires high quality depth data. There exist stereo algorithms which can automatically compute depth maps from stereo images, or depth sensors which can capture depth maps. However, these depth maps are usually of insufficient accuracy to allow a high quality synthesis as required e.g. in professional content productions. Consequently, today highly accurate depth maps are usually generated in a semi-automatic process, where stereo algorithms or depth sensors are used to estimate or capture initial depth maps which are then improved in an interactive process. Such manual interactions increase the content production and distribution costs and make a real-time and high quality depth map generation, as needed for live productions like concerts or sports, impossible. Thus, an automatic and real-time conversion of S3D as a production format to a depth enhanced transmission format is not possible in general.

The most simple and cost efficient conversion from S3D as a production format to a transmission format consists of conducting only low level conversion steps (like color space, bit depth, frame-rate, or image resolution conversions). Such a transmission format can be efficiently compressed and is compatible to the existing S3D distribution and stereoscopic display infrastructure. Thus, using a transmission format without supplementary depth data prevents an increase of content production and distribution costs. However, to support MADs, view synthesis technology is needed which generates new views based on 2-view video content only. Such view synthesis technology is in the focus of this article.

In Section II, a fully automatic view synthesis method based on Image-domain-Warping is presented and experimental results are discussed. The presented synthesis method can be used at the decoder side to synthesize new views directly from transmitted S3D content. Also a hardware architecture for real-time view synthesis is presented and analyzed. In Section III, the impact of shifting a part of the complexity of the synthesizer to the encoder side is examined in terms of transmission bit rate and decoder run-time. The article is concluded with a short summary and conclusions. The presented work is based on our previous work published in [12][13][14][15][30].

II. Image-Domain-Warping

Automatic view synthesis technology which synthesizes new views from 2-view video is highly desirable. Such synthesis technology would be compatible to any S3D distribution infrastructure to the home. Image-domain Warping (IDW) is a view synthesis method which is able to automatically synthesize new views based on stereoscopic video input. In contrast to synthesis methods based on DIBR, which relies on dense disparity or depth maps, IDW employs only sparse disparities to synthesize a novel view. It exploits the facts that our human visual system is not able to very accurately estimate absolute depth and that it is not sensitive to image distortions up to a certain level as long as images remain visually plausible, e.g. image distortions can be hidden in non-salient regions. Motivated by these insights, the IDW approach automatically estimates sparse disparities and image saliency maps from the input stereo video. They are used to compute an image warp which enforces desired sparse disparities in the final synthesized image while distortions are hidden in non-salient regions.

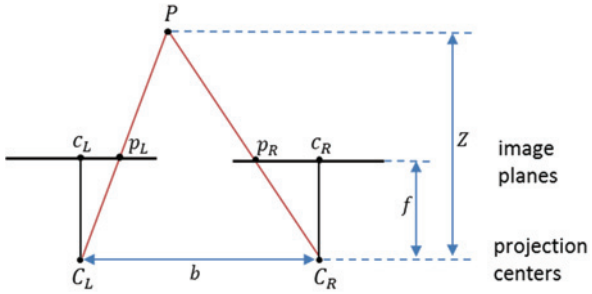

Fig. 2: Stereo pinhole camera model illustrating the projection of a 3D point *P* into the image planes of both cameras.

To find out which disparities have to be enforced in a synthesized image, let's observe Fig. 2. It shows a bird's-eye view of a pin-hole model, parallel stereo camera setup. Projection centers of both cameras are located at positions C_L and C_R at a baseline distance of b and both cameras have a focal length of f. Projecting a 3D point P, which is located at a distance Z from the projection centers, into the respective projection planes gives the projected points p_L and p_R. Hence, the projected points have an image disparity of

$$d = (p_L - c_L) - (p_R - c_R) = \frac{fb}{Z}$$

Obviously, a synthesis of a new view at a position

$$C_{\text{new}} = (1-\lambda)C_{\text{L}} + \lambda C_R,$$

corresponds to having a baseline distance $b_{\text{new}} = \lambda b$ with respect to the left view. Hence, that would give disparities $d_{\text{new}} = \lambda d$, i.e. a linear rescaling of all previous disparities d is required to synthesize the new view. A synthesis with a dense disparity map with rescaled disparities leads to

unavoidable synthesis problems if combined with a geometrically correct forward mapping algorithm, e.g. holes in the synthesized image will appear due to disoccluded areas. In general, a stereo image pair doesn't contain sufficient information to completely describe an image captured from a slightly different camera position. IDW uses image saliency information to deal with this problem.

A. Image Warps

We define a warp as a function that deforms the parameter domain of an image

$$w : [0, W] \times [0, H] \rightarrow \mathbb{R}^2$$

where W and H are the width and height of the image, respectively.

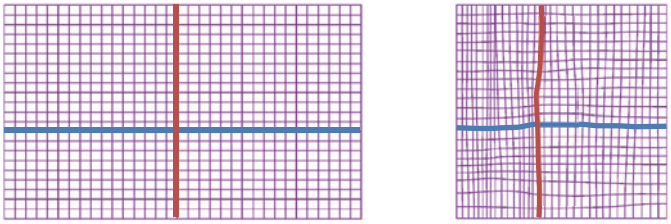

Fig. 3: An example of a warping function that deforms an input image. The warp is parameterized as a regular grid.

Image warps have a long history of use in computer vision and graphics based problems. One area that has seen a large amount of development in image warping is in the problem of aspect ratio retargeting, for which we refer to a recent survey paper [16].

The goal of the IDW method is to compute a warping of each of the initial stereo images that can be used to produce an output image meeting predefined properties (e.g. scaled image disparities). To do this, we formulate a quadratic energy functional *E(w)*. A warp w is then computed by minimizing E. However, to compute a warp, the solution space is reduced to warps defined at regular grid positions (Fig. 3)

$$w[p, q] \coloneqq w(\Delta_x p, \Delta_y q).$$

Obviously, at non-gird positions the warp can be defined by bilinear interpolation of the regular grid [19]. A linear solve can be then used to minimize E, i.e. to compute the warp at the grid nodes (p, q). This warp can then be used to render a warped image [19], i.e. to synthesize an image I_{synth} with pixel coordinates i, j according to

$$I_{\text{synth}}[i, j] := \Psi(I, w)[i, j] \coloneqq I(w^{-1}[i, j]).$$

B. View Synthesis

The IDW algorithm computes N-view video from 2-view video where N>2. It can be separated into four modules, which are shown in Fig. 4.

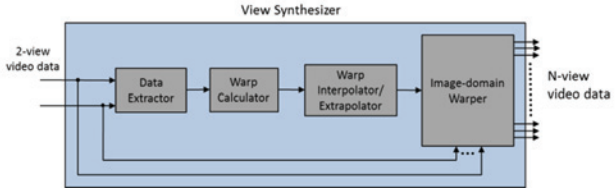

Fig. 4: Block diagram of the view synthesizer which converts 2-view video to N-view video

First, sparse disparities between input views and image saliency from each view are extracted. This information is used in the next module, the Warp Calculator, to formulate the quadratic error functional E. In the Warp Calculator only that warps are computed which are necessary for the synthesis of a view that is located in the middle between two input views. These warps are calculated by minimizing their respective error functionals. Calculated warps are then used in the Warp Interpolator/Extrapolator module to interpolate or extrapolate the warps that are necessary to synthesize the N output views, as required by a particular multi-view autosterescopic display. Finally, in the Image-domain Warper module, images are warped to synthesize the output images.

1) Data Extraction

First, a sparse set of disparity features is extracted (Fig. 5). These sparse disparities are estimated in an automatic, accurate and robust way. Two methods are applied. The first method [17] relies on detecting features, computing descriptors and finding matches between features in both input images. Although this method is characterized by its high robustness and accuracy, a drawback is a potential clustering of features in a few image regions.

Disparities of vertical image edges are particularly important for the stereopsis, i.e. the perceived depth. For this reason, additional features and corresponding disparities are detected such that features lay uniformly distributed on nearly vertical image edges. Disparities of detected features are estimated using the Lucas-Kanade method. The availability of such features is also important to prevent salient synthesis errors with IDW like bending edges in the synthesized image. In the end, disparity outliers are detected and removed using RANSAC.

Fig. 5: Disparities (blue) estimated at sparse feature positions (red).

Additionally, image saliency maps (Fig. 6) are extracted and used to prevent noticeable artifacts introduced by the image warping. They are automatically estimated by computing a saliency map with the method of Guo et al. [18] and blending it with a simple edge map computed with a Sobel filter. The saliency map indicates the level of visual significance for each image pixel. Information about image saliency is then explicitly used during the warp calculation.

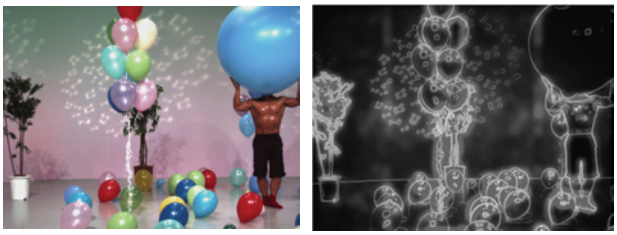

Fig. 6: Original image and its saliency map.

2) Warp calculation

Two warps w_L and w_R are computed which warp images I_L and I_R, respectively, to a camera position located in the center between the two input cameras. For a given sparse set of disparity features $(x_L, x_R) \in \mathbb{F}$, disparities

$$d = \frac{x_R - x_L}{2}$$

have to be enforced. Each warp is computed as the result of an energy minimization problem. An energy functional E is defined, and minimizing it yields a warp w that creates the desired change of disparities after view synthesis. The energy functional is defined with help of the extracted data and consists of three additive terms which are related to a particular type of constraint as described below. Each term is weighted with a parameter λ

$$E(w) := \lambda_{\mathrm{d}} E_{\mathrm{d}}(w) + \lambda_{\mathrm{s}} E_{\mathrm{s}}(w) + \lambda_t E_{\mathrm{t}}(w).$$

It consists of a sparse disparity term E_{d}, a spatial smoothness term E_{s} and a temporal smoothness term E_{t}. While the disparity constraints enforce predefined disparities at a sparse set of image locations, the spatial smoothness constraints force image distortions into less important areas. The temporal smoothness constraints are responsible for keeping the smoothness between temporally consecutive frames.

Disparity Constraints. To synthesize an image at a camera position located in the center between the two input cameras, the following disparity constraints are enforced in warps w_L and w_R

$$w_L(x_L) - x_L = (x_R - x_L)/2$$
$$x_R - w_R(x_R) = (x_R - x_L)/2$$

These constraints are enforced as weak constraints, which leads to the following sparse disparities terms

$$E_d(w_L) = \sum_{(x_L, x_R) \in \mathbb{F}} \left\| w_L(x_L) - \frac{x_R + x_L}{2} \right\|^2$$
$$E_d(w_R) = \sum_{(x_L, x_R) \in \mathbb{F}} \left\| w_R(x_R) - \frac{x_R + x_L}{2} \right\|^2$$

Depending on whether w_L or w_R have to be calculated, the corresponding term E_d is used in E.

Spatial Smoothness Constraints. Let's first define the finite dereference operators

$$\partial_p w[p,q] := w[p+1,q] - w[p,q],$$
$$\partial_q w[p,q] := w[p,q+1] - w[p,q],$$

and let's define a uniform warp as

$$u[p,q] := (\Delta_x p, \Delta_y q)$$

such that

$$\partial_p u[p,q] = (\Delta_x, 0) \text{ and } \partial_q u[p,q] = (0, \Delta_y)$$

The spatial smoothness term E_s measures the geometrical distortion of quads cells of the warp w with respect to the corresponding quads cells of u. It penalizes local deformations by increasing the cost if quad edges of w change their angle or length with respect to quad edges of u

$$E_s(w) = \sum_{p,q} s[p,q] \Big[\| \partial_p (w-u)[p,q] \|^2 + \| \partial_p (w-u)[p,q+1] \|^2 + \| \partial_q (w-u)[p,q] \|^2 + \| \partial_q (w-u)[p+1,q] \|^2 \Big].$$

The cost for each quad is weighted with the average saliency of this quad $s[p,q]$. Consequently, the warp is stiffer in salient regions and forces distortions in less salient regions. In particular, with help of these constraints, target disparities are enforced by keeping salient texture undistorted, while distortions are introduced in less salient regions. Because of the stiffness of salient regions (strength of the smoothness term), isolated incorrect feature correspondences in these regions have only a limited impact on the final synthesis result. Wrong disparities in less salient regions can lead to visible artifacts. However, in practice, the number of features detected in less salient regions is low compared to the total number of detected features.

Temporal Smoothness Constraints. Temporal constraints are applied to minimize temporal artifacts. If w^t denotes the warp at time instant t, then the energy term to be minimized is

$$E_t(w^t) = \sum_{p,q} \| w^t[p,q] - w^{t-1}[p,q] \|^2$$

Energy Minimization. After specifying the three terms of the energy functional

$$E(w) := \lambda_{\mathrm{d}} E_{\mathrm{d}}(w) + \lambda_{\mathrm{s}} E_{\mathrm{s}}(w) + \lambda_t E_{\mathrm{t}}(w)$$

the warp w is computed by finding the minimum of the functional using a solver for sparse least squares systems. This functional always represents an over-constrained equation system. The number of spatial and temporal smoothness constraints is dense in the number of degrees of freedom of the warp w to be solved, while the number of disparity constraints depends on the number of detected features, which is content dependent. Even for the first frame of a scene, where no temporal constraints are applicable (i.e. $\lambda_t = 0$), a unique solution for w exists, if one or more disparity constraint are provided. In practice, the number of disparity constraints is in the range of 5k-10k, which allows numerically stable solutions. The degree in which the postulated constraints are fulfilled is controlled by the weights $\lambda_{\mathrm{d}}, \lambda_{\mathrm{s}}$ and λ_t. In all our experiments we use a fixed set of weights, i.e. weights are sequence independent. We identified them by performing subjective tests with videos that were synthesized with different parameters. The parameters which we identified are

stable, i.e. they lead to good synthesis results in practice also with sequences which were not in the test set. Nevertheless, we observe that in some scenes with fast camera motion warping artifacts can occur. In these cases, an automatic adjustment of the temporal weight λ_t could be incorporated. However, due to the fast camera motion, these artifacts are often not observable in practice. In all of our experiments, we solve for warps with a resolution of 180x100 in the case of stereoscopic HD sequences. Subjective tests indicate that solving for warps with a resolution of 180x100 is sufficient in terms of synthesis quality for many stereoscopic HD sequences. Nevertheless, a further study of the impact of the warp resolution on the synthesis quality in dependence of the video resolution and video content is necessary.

In each time instant, warp calculation computes 2 warps, i.e. for the input image pair (I_L, I_R) two corresponding warps (w_L, w_R) are calculated. These warps enforce disparities as they are required for the synthesis of an image at a central camera position.

3) Warp Interpolation/Extrapolation

Multiview autostereoscopic displays require many views from different camera positions as input. Although the presented warp calculation could be used to compute dedicated warps to for each desired output camera position, we restrict the warp computation process to a computation of only two warps per time instant, i.e. warps which map the two input views to a central camera position (Fig. 7). The main reason for this restriction is to reduce the overall computational complexity of the warp calculation. Furthermore, using this approach, the complexity of the warp computation does not depend on the number of output views required by a particular display system. To compute warps which map the input views to desired arbitrary output camera positions, a simple, low complexity method for warp interpolation-extrapolation is used. Thus, warp interpolation-extrapolation computes all warps which are needed for the synthesis of as many output views N as required by a display system.

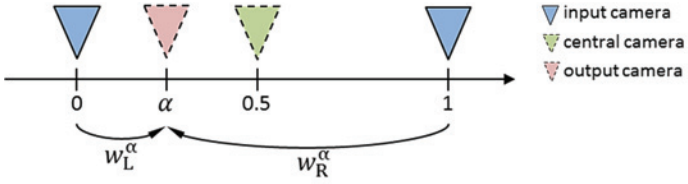

Fig. 7: Cameras, camera positions, and associated warps.

The projection centers of the two input cameras define an axis. We label the position of the projection centers on this axis as 0 and 1 (Fig. 7). To obtain warps w_L^α and w_R^α which enforce the appropriate disparities for an output view at a position α on this axis, we modify appropriately the warps $w_L^{0.5}$ and $w_R^{0.5}$, which were calculated in the Warp Calculation module. More precisely, the offsets of warped positions $w_L^{0.5} - u$ and $w_R^{0.5} - u$ are rescaled (u represents a uniform warp). The new warps are computed as

$$w_L^\alpha = 2\alpha(w_L^{0.5} - u) + u$$

$$w_R^\alpha = 2(1-\alpha)(w_R^{0.5} - u) + u$$

which corresponds to a blending between w_L^α and u, and w_R^α and u, respectively, i.e. an interpolation or extrapolation. Note that for $\alpha = 0$, the warp w_L^α is a uniform warp, i.e. the corresponding synthesized image is equal to the input image, while for $\alpha = 0.5$ warps w_L^α and w_R^α are equal to $w_L^{0.5}$ and $w_R^{0.5}$, respectively. For each output position α, which is required by a display system, dedicated warps w_L^α and w_R^α are interpolated or extrapolated.

In Fig. 7, we consider camera positions which are normalized by the real baseline distance between the input cameras, i.e. the normalized baseline distance between the input cameras is always 1. However, in professionally produced stereoscopic video content, the real baseline distance between the input cameras is adapted from shot to shot or even continuously within a shot. This is usually done in order to express an artistic intent but also to keep image disparities within a certain pixel range to prevent visual discomfort. Thus, in professionally produced stereoscopic content, we can expect that image disparities are in this limited range called the comfort zone. In our experiments with professional content, we obtain good synthesis results for normalized positions α which are between -0.5 and 1.5. Outside of this range, warping artifacts like bending edges or artifacts due to incorrect texture in disoccluded areas start to become observable. Please note that this range does not depend on the real baseline distance of the input cameras if image disparities are in the comfort zone.

Fig. 8: Image warped at a central position between the input image pair. Black region at the right image border represents a not textured region.

4) Image-domain Warping

An output image at a position α is synthesized according to

$$I_\alpha = \begin{cases} \Psi(I_L, w_L^\alpha) & : \quad \alpha \le 0.5 \\ \Psi(I_R, w_R^\alpha) & : \quad \alpha > 0.5' \end{cases}$$

i.e. I_α is synthesized based on the input image which is closer to the desired output position. Because warps are continuous, no holes can occur in the synthesized image. In particular, disoccluded regions are implicitly inpainted by stretching unsalient texture from the neighborhood into the region. We noticed that this kind of implicit inpaiting provides good synthesis results in practice as long as views are synthesized which are in the range $-0.5 \le \alpha \le 1.5$ (Fig. 7). However, if only one image is used for the synthesis, empty regions can occur on the left or right border of the output image (Fig. 8).

Fig. 9: Images synthesized at a central position between an input image pair (left column), and at a position on the right of the right input image (right column).

Hence, for output images located at a position between the input images, texture from a synthesis with the second input image is used to fill the empty border region. In Fig. 9, final synthesis results are shown.

C. *Experimental Results*

Recently, the Moving Pictures Experts Group (MPEG) issued a Call for Proposals (CfP) on 3D Video Coding technology [9] with the goal to identify i) a 3D video format, ii) a corresponding efficient compression technology, and iii) a view synthesis technology which enables an efficient synthesis of new views based on the proposed 3D video format. The authors of this article proposed in a joint proposal [20] the transmission and view synthesis system shown in Fig. 10.

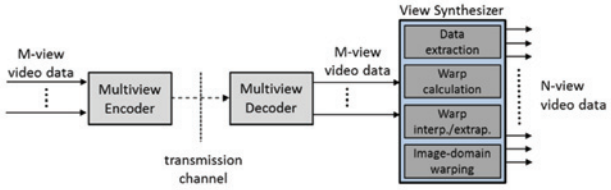

Fig. 10: Proposed transmission and view synthesis system.

Hence, a 3D video format is proposed that consists only of video data captured from a limited number of camera views. It is compressed by a multiview video coder [21][22] which is based on an early version of the newly developed High Efficiency Video Coding standard [23]. At the decoder-side, a view synthesis based on IDW is used, which includes Data Extraction, Warp Calculation, and Warp Interpolation/Extrapolation. Please note that such a 3D format with 2 views is already supported by existing consumer and professional stereo cameras.

With each proposal to the CfP, compressed bit streams, a decoder, and view synthesis software had to be provided. Bit streams had to be compressed at predefined target bit rates Proposals were evaluated by assessing the quality of the synthesized views through formal subjective testing on both stereoscopic and multiview autostereoscopic displays [24]. Fig. 11 shows the quality assessed on a multiview autostereoscopic display. Qualitatively similar results were also assessed on a stereoscopic display; the corresponding stereo sequences can be found here for download [25].

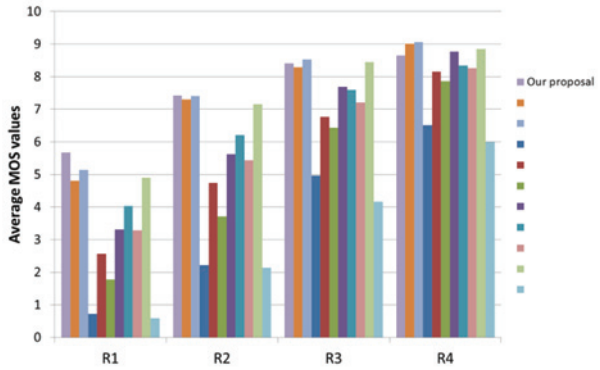

Fig. 11: Assessed quality of the MPEG proposals in the multiview autosterescopic display test scenario. R1 to R4 indicate increasing target bit rates. A Mean Opinion Score (MOS) of above 8 indicates transparent visual quality.

Assessment results show that our proposed system (Fig. 10) is among the four best performing proposals. We want to emphasize that our system is the only system among all proposals which doesn't encode any supplementary data (like depth maps) in order to use it at the decoder side for view synthesis. It is demonstrated that a system consisting of efficient multi-view video coding in combination with a view synthesis based on IDW is capable to generate high quality synthesis results.

We measured the run-time of our implementation of the view synthesizer on a PC based on Intel Core i7-920, 4 x 2.67GHz, and 12 GB RAM; only one CPU core was used. In Table 1, average run-times are reported. Obviously, Data Extraction and Warp Calculation dominate the overall run-time. However, their run-time remains unchanged if the number of output views N is changed. Only the run time of the Warp Interp./Extrap. and Image-domain-warping depends linearly on N.

Table 1: Run-time required for the synthesis of one new view from a stereoscopic image pair.

Module name	seconds
Data Extraction	2.5
Warp Calculation	1.8
Warp Interp./Extrap.	0.0000008
Image-domain Warping	0.5

D. Dedicated Hardware Architecture

Automatic view synthesis using IDW requires real-time processing to be usable in end-user consumer electronics devices. To this end, we devise a hardware architecture of a full view synthesis pipeline and provide numbers of FPGA synthesis results. The obtained hardware performance numbers of the IDW pipeline provide valuable insights on where the computationally challenging parts are and that a full HD real-time pipeline on a chip is within reach. In the following, we present the hardware-adapted algorithmic flow and hardware architecture, together with FPGA synthesis and performance results. Most of the individual components have been described in previous work and details are therefore omitted here.

1) Overview

The algorithmic flow is conceptually similar to the general IDW flow. However, the selection and adaptation of the different algorithms has been made based on hardware-efficiency and real-time capabilities, which has not been a major concern so far. For example, one difference is the disparity estimation step which uses a simpler but much more computationally efficient approach to feature extraction.

Fig. 12 provides a high-level block diagram of our view synthesis system. The input and output are DVI/HDMI high-definition videos. The overall design is conceptually divided into two parts: an infrastructure part handling all FPGA-board and I/O specific controllers and the core view synthesis components which are, in principle, device-independent and can be ported to other FPGAs, boards, and possibly to dedicated hardware.

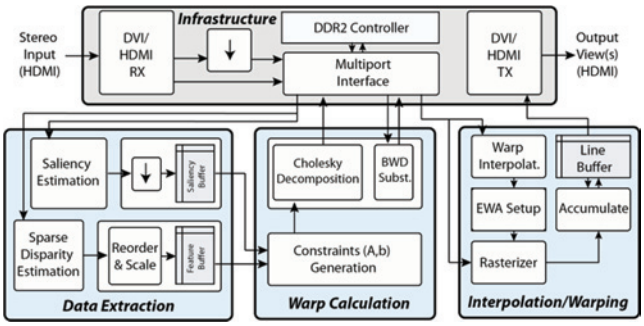

Fig. 12: High-level block diagram of the IDW-based view synthesis system.

2) Disparity Estimation

Sparse disparity estimation using robust feature matching techniques are computationally expensive. For the hardware architecture we therefore replace feature matching with a simpler disparity estimation algorithm. Dense, local window matching has been shown to be very efficient in real-time streaming video applications [30]. However, dense methods contain a lot of outliers and are thus far from being robust or reliable. In order to alleviate the lack of robustness, we add two confidence metrics that enable us to discard all features with low confidence. First, we use the standard left-right consistency check and second, we compare the minimal cost value γ at disparity d to an absolute threshold and to the second lowest cost value γ' at disparity d' to get a measure of how unique the specific cost value is:

$$\text{conf} \propto \frac{|d - d'|}{|\gamma - \gamma'|}$$

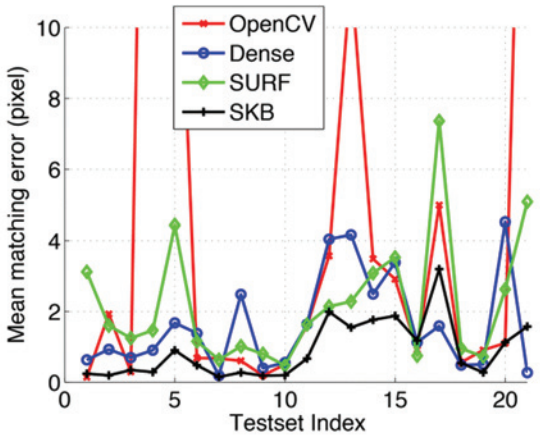

Fig. 13: Comparison of feature matcher performances using the Middlebury 2006 test set.

Fig. 13 shows a comparison of different robust disparity estimation strategies for the Middlebury 2006 test set. As can be seen, extracting sparse disparities from dense matching (denoted as 'Dense' in Fig. 13) is not considerably worse than using standard feature matching techniques. The caveat is that this only holds true for mostly rectified stereo footage.

The hardware architecture is an extension of the hardware architecture of standard dense block matching setup (e.g. [30]), with a block that looks for the first *and* the second lowest cost value.

3) Other Components

The other parts of the view synthesis pipeline relay on components described in prior work. The saliency estimation is done using the algorithm described in [18] and uses the hardware architecture described in [28]. For the energy minimization, we use a sparse linear solver [29]. In particular, we use a direct solver based on the Cholesky decomposition, which has a computational complexity proportional to the bandwidth of the matrix. Since image width is higher than its height, solving the transposed problem reduces the bandwidth and hence computational resources. That is why additional buffers are required to reorder saliency and features. Since the IDW framework entirely relates on forward warps, we use a forward warping approach and use elliptical weighted average (EWA) splatting. EWA is more complex than traditional bilinear backward mapping but does not require warp inversion or explicit anti-aliasing. Several hardware implementations for EWA splatting exist [26][27][28]. For the results figures in this work we use the simplified version described in [27].

Table 2: Utilization of hardware resources and hardware performance on an ALTERA Stratix IV FPGA. The resources are logic cells (LUTs), registers (regs), arithmetic units (DSPs), and on-chip RAM (bl. RAM).

	LUTs	Regs	DSPs	bl. RAM	FPS	Res.
Sal. Est.	8.6k	13k	100	0.8M	60	512x288
Sal. Buff.	0.3k	0.5k	15	0.25M	60	512x288
Feat. Match.	41k	21k	0	82k	25	1024x576
Feat. Buff.	0.3k	0.2k	0	0.2M	25	1024x576
Constr. Gen.	19k	0.8k	50	21k	25	126x224
Solver	36k	53k	164	1.1M	25	126x224
Warp Interp.	0.5k	0.4k	40	14k	60	1080p
Rendering	4k	2.8k	80	0.5M	60	1080p
Core Tot.	93k	92k	449	3M	60	1080p
% of Str. IV	22%	22%	44%	15%		

4) FPGA Synthesis Results

Table 2 provides a summary of the hardware resource utilization and performance of the different steps of a view synthesis core. In this particular implementation, one view is generated out of a stereo input video. Generating several views in parallel can be achieved by instantiating more rendering cores and interpolating the warps accordingly (Section II.B.3)). Furthermore, we assume that we have rectified stereo video as input, i.e., we solve only for node positions warped in horizontal direction. Adding the vertical direction would halve the throughput of the solver.

The computational bottleneck is clearly the linear solver, which achieves a grid of size 224x126 at approximately 25 frames/s. Therefore, disparity and saliency are also estimated on lower temporal and/or spatial resolution. The full system has been implemented using VHDL and verified against MATLAB models. A real-time system demonstration is currently developed as well as an efficient multiview rendering architecture.

III. Transmission of Warps as Supplementary Data

To reduce the computational complexity at the receiver side, we modify the transmission system which was proposed in Fig. 10. The modified system is shown in Fig. 14. Thus, it is proposed to shift the warp extraction and warp calculation part to the sending side, and, in addition to the multiview data, to efficiently compress and transmit the warp calculation result, i.e. a restricted set of warps. The two modules shifted to the encoder represent the computationally most expensive parts of the view synthesizer, as it is shown in Table 1. Thus, the view synthesis at the receiver side is now a process which requires significantly less computational complexity in comparison to the system presented in Fig. 10. Furthermore, a generation of warps at the encoder side also allows to generate warps offline, e.g. in order to store them with stereoscopic video on storage media. Content producers can now directly control the synthesis quality at the receiver. These benefits come at the cost of an increase of the transmission bit rate by the rate required for the warp coding. Also a new transmission format for stereoscopic video plus warp data has to be supported.

In this Section, two efficient warp coding methods, a dedicated warp coder and a warp coding method based on a video coder, are shown, and corresponding evaluation results are presented.

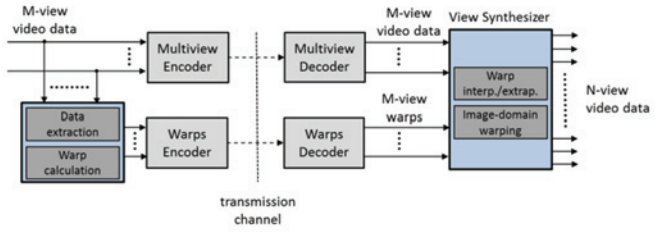

Fig. 14: Modified transmission and view synthesis system.

A. Warp coding with a dedicated warp coder

Warps of all time instances and views are encoded successively. For each view, they are encoded separately and multiplexed into a single bit stream. Without loss of generality, the coding of a warp sequence assigned to one view is described in the following. We denote the warp at time instant f as w^f. Each warp w^f is represented as a regular quad grid (Fig. 3) of fixed resolution where each node of the grid is indexed with integer coordinates i, j and has a 2D location $w^f[i,j] \in \mathbb{R}^2$ assigned.

In Fig. 15, a block diagram of the warp coder is shown. To exploit the regular structure of a warp, each warp is spatially partitioned using a quincunx resolution pyramid. Each partition is then predictively encoded using a closed loop DPCM [31] in combination with a spatio-temporal predictor. CABAC [32] is employed for entropy coding of residuals. A Coder Control is used to adjust the quantization step size and the number of partitions to be encoded in each frame. It has the goal to achieve the best compromise between number of bits needed for coding a warp and quality of the image which is synthesized using a decoded and reconstructed warp.

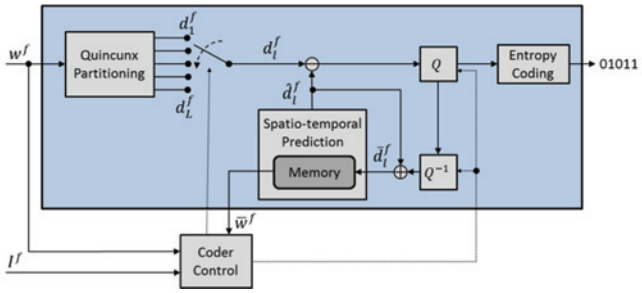

Fig. 15: Block-diagram of warp coder.

1) Spatial partitioning

A partition consists of a set of 2D locations which we call a group of locations (GOLs). Each warp w^f is partitioned into GOLs using a quincunx resolution pyramid, which is illustrated in Fig. 16. The lowest resolution grid of the resolution pyramid and the difference sets between successive resolutions specify a partitioning of the locations of w^f into GOLs d_l^f. Locations of each GOL are then successively encoded from the top d_1^f to the bottom d_L^f of the pyramid, where L denotes the total number of GOLs.

2) Intra-Warp and Inter-Warp Prediction

Similar to video coding standards, three warp coding types and corresponding prediction modes are supported: INTRA, INTER_P and INTER_B. After prediction, residuals $w[i,j] - \hat{w}^f[i,j]$ are computed, uniformly quantized, and entropy coded. Quantized residuals of each GOL are entropy coded independently from other GOLs.

In the INTRA prediction mode, all locations of d_1^f are scanned row-wise and predicted in a closed loop DPCM from previously scanned spatially neighboring locations, as it is shown in Fig. 17. Locations of all other GOLs d_l^f are predicted by the respective centroids computed from spatially neighboring locations in $\bigcup_{k=1}^{l-1} d_l^f$ as indicated in Fig. 17. Please note that the quincunx pyramid guarantees that for each interior location in $d_{l\geq 2}^f$ always four spatial neighbors exist in $\bigcup_{k=1}^{l-1} d_l^f$, which can be used for intra-warp prediction. Due to the regular structure of the neighborhood, which is induced by the quincunx resolution pyramid, it makes no difference if warp locations or offsets of warped locations are used for prediction.

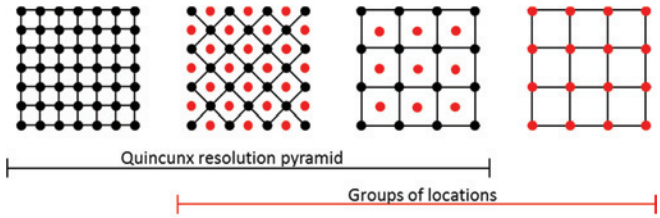

Fig. 16: Black nodes represent a quincunx resolution pyramid of 3 layers derived from a warp with a resolution of 7x7. Red nodes represent derived groups of locations.

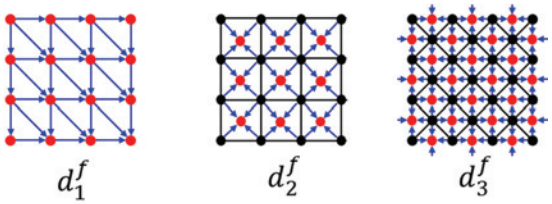

Fig. 17: Illustration of intra-warp prediction dependencies.

Prediction modes INTER_P and INTER_B use INTRA residuals from already encoded warps (reference warps) to predict INTRA residuals of the current warp. The corresponding predictors for node locations are defined as

$$\hat{w}^f[i,j]_{\text{INTER\_P(r)}} = \hat{w}^f[i,j]_{\text{INTRA}} + \bar{w}^r[i,j] - \hat{w}^r[i,j]_{\text{INTRA}}$$

$$\hat{w}^f[i,j]_{\text{INTER\_B(r,s)}} = \alpha\hat{w}^f[i,j]_{\text{INTER\_P(r)}} + (1\text{-}\alpha)\hat{w}^f[i,j]_{\text{INTER\_P(s)}}$$

where r and s indicate time instances of reference warps, $\alpha = |f-s|/|r-s|$, and $\bar{w}[i,j]$ and $\hat{w}[i,j]$ represent a reconstructed and predicted location, respectively

3) Image Quality vs. Rate Optimization

The bit rate and the quality are controlled during encoding by the Coder Control module. Coder Control determines for each time instant f, the quantization step size $\tilde{\Delta}$ and the total number of GOLs $\tilde{L}$ to be encoded. Hence, instead of coding all GOLs per time instant, Coder Control can decide to encode only the first $\tilde{L} \in \{0,\dots,L\}$ GOLs, if that gives the best compromise in terms of bit rate and image quality. Locations of not encoded GOLs are always reconstructed by employing the prediction mode assigned to the current time instant and assuming zero valued residuals. We maximize the following objective function to determine the parameter set $(\tilde{\Delta},\tilde{L})$ for a given time instant

$$Q(w,\tilde{L},\tilde{\Delta}) - \lambda R(w,\tilde{L},\tilde{\Delta}).$$

Here Q computes the peak-signal-to-noise-ratio (PSNR) between the image synthesized with original warp $\Psi(I,w)$ and the image synthesized with decoded and reconstructed warp $\Psi(I,\bar{w})$. Both images are synthesized in the coding loop. Rate R represents the number of bits required for coding the first $\tilde{L}$ GOLs with step size $\tilde{\Delta}$. The Lagrangian multiplier λ [33] represents the slope of the image quality vs. warp rate function (QRF) $Q(R)$, which is obtained by maximizing the objective function [34]. Consequently, a maximization of the objective function prevents the coding of GOLs which don't lead to an appropriate increase in image quality in relation to the necessary increase in bit rate, where the parameter λ controls this relation. This approach guarantees that the maximum view synthesis quality is achieved in terms of PSNR with the bits spent for a warp.

4) Experimental results and discussion

To support the application scenario presented in Fig. 14, we first compute warps from original 2-view video data. These warps are coded with the method presented in this Section. Warp coding is performed with hierarchical groups of warps structures of sizes 12 and 15 to enable a random access at each 0.5 seconds for the 25 and 30 Hz sequences, respectively (as it was specified in the CfP [9]). Fig. 18 shows the impact of the bit rate used for coding the warps on the image quality of the synthesized views. The image quality is measured between the synthesis results obtained with i) the original warps and ii) the coded and reconstructed warps.

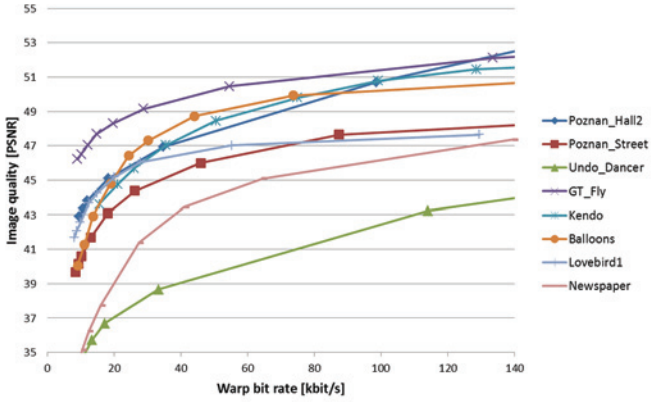

Fig. 18: Warp rate vs. image quality curves.

Each curve is obtained by coding with a fixed quantization parameter $\Delta = 0.5$ and by varying the Lagrange parameter $\lambda \in \{0.038, 0.035, 0.03, 0.025, 0.02, 0.015, 0.01, 0.005\}$.

Informal viewing showed that visually lossless quality with respect to a synthesis with original warps is achieved at a PSNR of about 45dB for almost all sequences. In the case of

the Undo_Dancer sequence, which is an animated sequence containing large high frequency patterns, a lower PSNR of about 39 dB is perceived as visually lossless. Table 3 shows i) the exact rates of the warps for both views at visually lossless quality as well as ii) the bit rates of the coded 2-view videos at the highest target bit rates as they were submitted in [20] as answer to the CfP. Obviously, the warp bit rate for both views represents only 3.6% on average of the total bit rate needed for transmitting the warps and the 2-view video together. Note that in the other three of the four winning proposals of the CfP [9], depth maps were part of the 3D video format besides the 2-view video data. Thereby the coded depth maps represented a portion between 5.8% and 22.1% of the total bit rate in the respective proposals. The last column in the table shows the warp compression ratio, i.e. the ratio between the rate of compressed warps and the rate of uncoded warps in binary representation. Note that each node position of a warp is represented by 2 floating point values. In our experiments, we noticed that a quantization of offsets of warped positions to 16 bit/node is sufficient to represent warps that allow visually lossless synthesis quality. That gives for a 100x180 resolution warp sequence at 25 frames per second a raw bit rate of 7.2 Mbit/s. Hence, the proposed warp coding scheme leads to a significant compression ratio of 1:302 on average with respect to warps quantized at 16 bit/node. That corresponds to a compression to 0.05 bit/node.

Table 3: Warp coding results with the dedicated warp coder.

Sequence name	Frames per second	Warp Rate [kbit/s]	Video Rate [kbit/s]	Warp rate to total rate ratio	Warp compression ratio
Poznan_Hall2	25	18.2	520	3.4%	395
Poznan_Street	25	26.2	1307	2.0%	275
Undo_Dancer	25	33.3	998	3.2%	216
GT_Fly	25	19.6	1098	1.8%	367
Kendo	30	35.5	690	4.9%	243
Balloons	30	24.4	800	3.0%	354
Lovebird1	30	20.0	828	2.4%	432
Newspaper	30	64.5	719	8.2%	134
Average		30.2	870	3.6%	302

A comparison between the run-times of the synthesis algorithms, which are executed at the receiver sides in the application scenarios shown in Fig. 10 and Fig. 14, shows that the view synthesis run-time is reduced by a factor of 9 on average if data extraction and warp calculation are shifted to the encoder side. Based on our previous work [26][27], we also devised a hardware architecture for the decoder-side view synthesis shown in Fig. 14, which runs in real-time with 2x 1080p30 video input and synthesizes 8 views.

B. Warp coding with help of a video coder

To take advantage of already existing and highly sophisticated video coding technology, we propose the warp coding system show in Fig. 19 as an alternative to the dedicated warp coding method shown in Section III A.

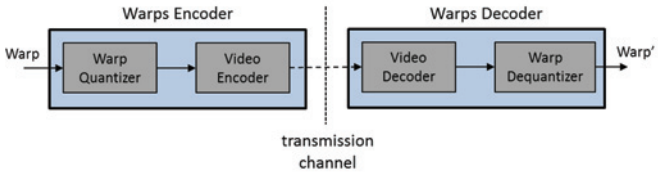

Fig. 19: Warp coding system using a video coder.

1) Coding system

Similar to the coding with the dedicated warp coder, warps are encoded separately for each view and then multiplexed into a single bit stream. To encode a warp, first, a Warp Quantizer is used to convert each warp into an 8-bit grayscale image representation. Grayscale images are then encoded with HEVC [23], the upcoming video coding standard of MPEG and VCEG, which is a successor of the famous H.264|AVC. At the Warp Decoder, grayscale images are decoded with HEVC and warps are reconstructed in the Warp Dequantizer.

Fig. 20: A grayscale image computed from a warp.

Warp Quantization first computes the offset of each warped node $d[i,j] = w[i,j] - u[i,j]$. These offsets are then uniformly quantized to 8-bit resolution and stored in a grayscale image. Since stereoscopic video is usually captured with parallel camera setups to prevent undesired vertical disparities, corresponding warps deform images only in horizontal direction. Thus, in this case, only components of the quantized x-coordinates have to be compressed. Fig. 20 shows an example of x-components stored in a grayscale image. In case of convergent camera setups also y-coordinate components are stored together with the x-coordinate components in the same grayscale image using a top-bottom representation. Minimum and maximum coordinates of warp locations are saved as meta-data to be later used at the decoder side in the Warp Dequantizer to reconstruct warps again from grayscale images. HEVC is used to encode and decode grayscale images.

2) Experimental results and discussion

The presented warp coding system was proposed to JCT-3V, the joint group of MPEG and VCEG dedicated to the development of 3D video coding standards, where a subjective test was conducted [35]. The test compared the synthesis quality achieved by IDW, which used coded warps and coded video data, in comparison to the synthesis quality achieved by the best DIBR renderer available to JCT-3V, which used coded depth and coded video data. Thereby the same coded video data was used in both cases while warp and depth data had almost the same bit rate. Table 4 shows the rates of the used warp and video data, where each sequence was compressed at two different rates points, a high and a low rate point. The

ratio between depth (or warp) and video bit rate used in the experiments is recommended by the JCT-3V group to achieve best quality with DIBR.

Table 4: Warp coding results obtained with help of a video coder.

Sequence name	Frames per second	Warp Rate [kbit/s]	Video Rate [kbit/s]	Warp rate to total rate ratio	Warp compression ratio
PoznanHall2	25	53	456	10%	135
		17	133	11%	433
PoznanStreet	25	80	1182	6%	91
		19	285	6%	380
UndoDancer	25	134	2376	5%	54
		35	513	6%	207
GhostTownFly	25	129	1848	7%	56
		30	419	7%	237
Kendo	30	89	596	13%	97
		20	201	9%	431
Balloons	30	66	631	9%	132
		17	215	7%	502
Newspapercc	30	113	647	15%	76
		23	204	10%	379
Average (high rate point):		95	1105	9%	91
Average (low rate point):		23	281	8%	367

Fig. 21 shows the average of the voting results, where subjects were allowed to vote 'IDW' if they preferred the synthesis by IDW, 'DIBR' if they preferred the synthesis by DIBR, 'same' if they considered that the quality was the same, and 'don't know' if they could not decide which synthesis result to prefer. The voting results show that at high bit rates IDW achieves the same synthesis quality as DIBR, i.e. a majority of 59% vote for same quality, while 19% vote for IDW and 19% vote for DIBR. At low bit rates a preference for IDW is observable, i.e. 37% prefer IDW, 11% prefer DIBR and 44% vote for same quality.

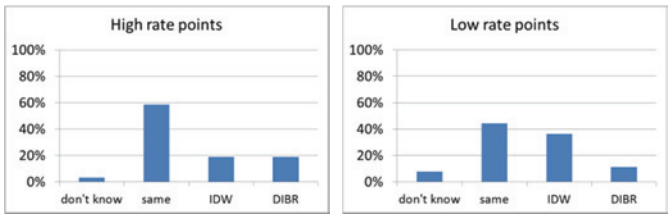

Fig. 21: Voting results of a subjective test comparing the synthesis quality between IDW and DIBR using compressed data.

A comparison between the warp-rate-to-total-rate ratios shown in Table 3 and Table 4 indicates that warps can be compressed more strongly with the dedicated warp coder while keeping subjective quality visually lossless. While a dedicated warp coder can have a stronger coding efficiency, the advantage of reusing video coding technology for warp coding lies the reduced development and production costs, i.e. in the reuse of available video coding chips. For this reason, based on the evaluation results presented in this chapter, JCT-3V planes to extend the upcoming 3D-HEVC standard by warp coding based on HEVC. This will enable the transmission of multi-view video plus warp data with an international standard, which will allow the use of IDW for view synthesis at the receiver side.

IV. Conclusions

In this article, we presented a view synthesis method based on Image-domain-Warping. Our approach automatically synthesizes new views from stereoscopic 3D video. It relies on an automatic estimation of sparse disparities and image saliency information, and enforces target disparities in the synthesized images using an image warping framework. A large subjective study coordinated by MPEG showed that multi-view video coding of 2-view or 3-view video in combination with a decoder-side view synthesis based on Image-domain Warping leads to high quality synthesis results without requiring depth map estimation and transmission. The corresponding MPEG proposal was considered as one of the four winning proposals.

We also devised a hardware architecture for view synthesis based on Image-domain-Warping. We implemented it in VHDL and evaluated single hardware modules. We have strong indications that an efficient hardware implementation of the complete view synthesis architecture will run in real-time in end-user devices.

To reduce the computational complexity (and with that also the energy requirements) of the view synthesis at the decoder-side in the scope of a transmission system, we shifted parts of the view synthesis algorithm to the encoder-side and evaluated the impact of this modification on the transmission bit rate and view synthesis run-time. We report that through the shift of complexity to the encoder-side, the view synthesis run-time is reduced by a factor of 9 on average at the cost of a small increase in bit rate. Warps, which are computed at the encoder-side, are encoded with two methods, a coder explicitly developed for coding warps, and a coder which uses HEVC, the upcoming video coding standard developed jointly by MPEG and VCEG. While with a dedicated coder warps can be compressed on average to 3.6% of the total bit rate, the coder using HEVC compresses warps to 9% of the total bit rate at subjectively lossless quality. Thus, a dedicated warp coder can achieve a higher coding efficiency. However a reuse of existing video coding technology has the advantage of reduced development and production costs. For this reason, JCT-3V planes to extend the upcoming 3D-HEVC standard by warp coding based on HEVC, which will allow receivers to perform a synthesis of new views based on Image-domain-Warping.

References

[1] www.alioscopy.com
[2] www.toshiba.com
[3] www.dimencodisplays.com
[4] Müller, K.; Merkle, P.; Wiegand, T., "3-D Video Representation Using Depth Maps," Proceedings of the IEEE , vol.99, no.4, pp.643-656, April 2011
[5] Aljoscha Smolic, "3D video and free viewpoint video—From capture to display", Pattern Recognition, Volume 44, Issue 9, September 2011
[6] S. Würmlin, E. Lamboray, M. Gross, "3D video fragments: dynamic point samples for real-time free-

viewpoint video", Computers and Graphics, 28 (1) (2004), pp. 3–14 Elsevier Ltd

[7] S.B. Kang, R. Szeliski, P. Anandan,"The geometry-image representation tradeoff for rendering", ICIP 2000, Vancouver, Canada, September 2000.

[8] Smolic, K. Müller, P. Merkle, A. Vetro, "Development of a new MPEG standard for advanced 3D video applications", ISPA 2009. Salzburg, Austria, Sep. 2009.

[9] ISO/IEC MPEG, "Call for Proposals on 3D Video Coding Technology," MPEG N12036, March 2011.

[10] L. Yu, T. Masayuki, Y. Zhao, C. Zhu, "3D-TV System with Depth-Image-Based Rendering: Architecture, Techniques and Challenges", Springer New York, 1st Edition, 2012.

[11] Christoph Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV", Proc. SPIE 5291, 93, 2004.

[12] M. Farre, O. Wang, M. Lang, N. Stefanoski, A. Hornung, A. Smolic, "Automatic Content Creation for Multiview Autostereoscopic Displays Using Image Domain Warping", Hot3D Workshop 2011, Barcelona, Spain, July 2011.

[13] Smolic, Y. Wang, N. Stefanoski, M. Lang, A. Hornung, M. Gross, "Non-linear Warping and Warp Coding for Content-adaptive Prediction in Advanced Video Coding Applications", ICIP 2010, Hong Kong, China, Sep. 2010.

[14] Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, Markus Gross, "Nonlinear Disparity Mapping for Stereoscopic 3D", In ACM SIGGRAPH 2010, Article 75, 10 pages, 2010.

[15] N. Stefanoski, M. Lang, A.Smolic , "Image Quality vs Rate Optimized Coding of Warps for View Synthesis in 3D Video Applications", ICIP 2012, Orlando, USA, Sep. 2012.

[16] Rubinstein, Michael and Gutierrez, Diego and Sorkine, Olga and Shamir, Ariel. A Comparative Study of Image Retargeting. SIGGRAPH Asia 2010.

[17] Frederik Zilly and Christian Riechert and Peter Eisert and Peter Kauff (2011). Semantic Kernels Binarized -- A Feature Descriptor for Fast and Robust Matching. CVMP 2011

[18] Guo, C. and Ma, Q. and Zhang, L. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. CVPR 2008.

[19] G. Wolberg, (1990). Digital Image Warping, IEEE Computer Society Press, Los Alamitos, Calif.

[20] N. Stefanoski, P. Espinosa, O. Wang, M. Lang, A. Smolic, S. Bosse, M. Farre, K. Müller, H. Schwarz, M. Winken, T. Wiegand (2011). Description of 3D Video Coding Technology Proposal by Disney Research Zurich and Fraunhofer HHI. MPEG, Doc. M22668, Geneva, Switzerland, Nov. 2011.

[21] H. Schwarz, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, D. Marpe, P. Merkle, K. Müller, H, Rhee, G. Tech, M. Winken, and T. Wiegand, "3D Video Coding Using Advanced Prediction, Depth Modeling, and Encoder Control Methods", under submission at Picture Coding Symposium 2012, Krakow, Poland, May 2012.

[22] K. Müller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, H. Rhee, G. Tech, M. Winken, and T. Wiegand, "3D High Efficiency Video Coding for Multi-View Video and Depth Data", submitted to IEEE Transactions on Image Processing, Special Issue on 3D Video Representation, Compression and Rendering, October, 2012.

[23] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard", IEEE Transactions on Circuits and Systems for Video Technology, December 2012.

[24] ISO/IEC MPEG, "Report of Subjective Test Results from the Call for Proposals on 3D Video Coding," MPEG N12347, 2012.

[25] http://zurich.disneyresearch.com/videodata/niko/3DV_CfP

[26] P. Greisen, M. Schaffner, S. Heinzle, M. Runo, A. Smolic, A. Burg, H. Kaeslin, M. Gross, "Analysis and VLSI Implementation of EWA Rendering for Real-time HD Video Applications", IEEE Trans. on TCSVT, Vol. 22, No. 11, pp. 1577-1589, November 2012.

[27] P. Greisen, R. Emler, M. Schaffner, S. Heinzle, F. Gurkaynak, "A General-Transformation EWA View Rendering Engine for 1080p Video in 130nm CMOS", in Proceedings of VLSI-SoC, 2012.

[28] P. Greisen, M. Lang, S. Heinzle, A. Smolic, "Algorithm and VLSI Architecture for Real-Time 1080p60 Video Retargeting", in Proceedings of High Performance Graphics, 2012.

[29] P. Greisen, M. Runo, P. Guillet, S. Heinzle, A. Smolic, H. Kaeslin, M. Gross, "Evaluation and FPGA Implementation of Sparse Linear Solvers for Video Processing Applications", IEEE Trans. on TCSVT, accepted, 2013

[30] P. Greisen, S. Heinzle, A. Burg, M. Gross, "An FPGA-based processing pipeline for high definition stereo video", EURASIP Journal on Image and Video Processing, 2011.

[31] N. S. Jayant, Peter Noll, "Digital Coding of Waveforms: Principles and Applications to Speech and Video", Prentice-Hall, Englewood Cliffs NJ, USA, 1984

[32] D. Marpe, H. Schwarz, and T. Wiegand, "Context-Based Adaptive Binary Arithmetic Coding in the H.264 / AVC Video Compression Standard", IEEE Trans. on CSVT, Vol. 13, No. 7, pp. 620-636, July 2003.

[33] Hugh Everett III,"Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources", Operations Research, Vol. 11, No. 3, pp. 399-417, 1963.

[34] T. Wiegand, B. Girod, "Lagrange multiplier selection in hybrid video coder control", ICIP 2001, Thessaloniki, Greece, Oct. 2001.

[35] N. Stefanoski, "Subjective testing results on comparing warp-based with depth-based synthesis from coded data at same bit-rate", JCT-3V m28347, Geneva, Switzerland, 2013