EUROGRAPHICS 2014 / B. Lévy and J. Kautz
(Guest Editors)

Volume 33 (2014), Number 2

Optimizing Stereo-to-Multiview Conversion for Autostereoscopic Displays

Alexandre Chapiro[1,2], Simon Heinzle[2], Tunç Ozan Aydın[1], Steven Poulakos[1,2], Matthias Zwicker[3], Aljosa Smolic[1], Markus Gross[1,2]

[1] Disney Research Zurich, [2] ETH Zurich, [3] University of Bern

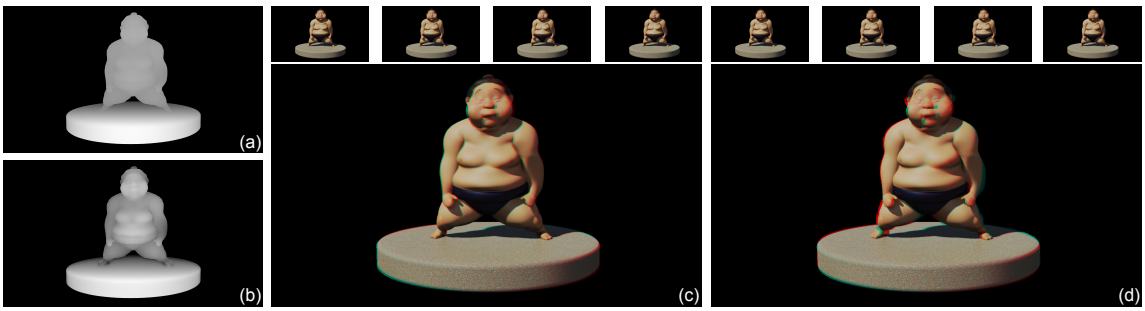

Figure 1: Our method produces depth-enhanced multiview content from stereo images while preserving the original artistic intent. (a) and (b) show the linearly mapped disparities as well as enhanced disparities computed using our method. (c) and (d) show the result of stereo-to-multiview conversion using (a) and (b), respectively. Our method avoids the cardboarding effect that can be seen in the linearly mapped version.

Abstract

We present a novel stereo-to-multiview video conversion method for glasses-free multiview displays. Different from previous stereo-to-multiview approaches, our mapping algorithm utilizes the limited depth range of autostereoscopic displays optimally and strives to preserve the scene's artistic composition and perceived depth even under strong depth compression. We first present an investigation of how perceived image quality relates to spatial frequency and disparity. The outcome of this study is utilized in a two-step mapping algorithm, where we (i) compress the scene depth using a non-linear global function to the depth range of an autostereoscopic display, and (ii) enhance the depth gradients of salient objects to restore the perceived depth and salient scene structure. Finally, an adapted image domain warping algorithm is proposed to generate the multiview output, which enables overall disparity range extension.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Display Algorithms I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Stereo

1. Introduction

Multiview autostereoscopic displays (MADs) are expected to make their way into the households in the near future, and major display manufacturers are intensively working towards consumer-grade screens. A significant limitation of the current autostereo technology is the display's depth range. While the emergence of very high resolution displays (4k and beyond) can alleviate this problem to a certain degree, the constraints on the display depth range will remain as an inherent limitation of the MAD techologies.

In contrast to the recent progress on the display side, autostereoscopic content creation still lacks the tools and standards for the mainstream deployment of MAD technologies. In fact, content creation for 2-view stereo (S3D) for glasses-based systems is just developing and maturing. Even with the emergence of MAD technologies, stereo will remain in use for the foreseeable future, as content creators cannot change rapidly and completely. Consequently, support for legacy stereo content through stereo-to-multiview conversion will likely be a key feature for ensuring a graceful and

backward compatible transition from 2-view stereo to multiview autostereo.

The main technical challenge in faithful stereo-to-multiview conversion is that the disparity range of many S3D scenes often exceeds the limitations of MADs. However, current stereo-to-multiview conversion methods such as depth-image based rendering (DIBR) [SMD*08] and image domain warping (IDW) [SWL*13] directly interpolate between the two input views and do not take the inherent depth limitations of autostereoscopic screens into account. Moreover, unlike in the early days of S3D where the technology was used mainly as a "wow factor", more recently the depth layout is being used as an artistic element to support the content's narrative and action. Thus, any autostereoscopic content creation workflow should not only reduce the content's depth range to the limits of the MAD technology, but also preserve the artistic intent and *perceived* depth layout as much as possible.

In their basic work, Zwicker et al. [ZMDP06] evaluate the bounds on content creation for multiview displays and propose filtering the content as solution for the limited depth range. Didyk et al. recently proposed a framework for depth remapping based on just noticeable differences (JND) of depth perception [DRE*12b]. They identify content creation for multiview as one of the use cases, and also propose blurring the content in addition to depth compression. However, from a creative point of view, filtering the content in this way is unsuitable. For instance those objects that are far off screen are in many cases the most important by artistic intent, and blurring a character that is in center of attention is undesirable. Furthermore, previous methods have not been evaluated for video and live action footage so far. Inspired by this previous work, we address stereo-to-multiview conversion from the point of view of content creation. Rather than JND, we introduce a notion of saliency to capture and characterize artistic intent. Filtering important image content is avoided and instead we rather sacrifice noticeable disparity differences in non-salient regions.

We start by investigating the influence of disparity and texture frequency on the perceived picture quality through a subjective study. Based on the study, we choose a range of disparities that are perceived as pleasant but exceed the theoretical limit of multiview displays. We then compute a disparity mapping that retains the overall depth layout but strives to keep the volume of salient objects to avoid cardboarding. To achieve this, we perform our mapping in two steps. A global non-linear mapping operator first transforms the overall depth range to the range of the display. In a second step, we locally enhance the depth gradients to reduce the effect of cardboarding. Both global mapping and local gradient enhancement are based on saliency. We then generate multiview content directly from the input views using an extended version of image domain warping (IDW) [SWL*13], which is applicable for disparity range extension. We investigate the suitability of the different mapping strategies for synthetic content (where perfect disparity is given) and for live action content (where imperfect disparities pose additional challenges). In a final user study we validate our approach on a variety of live action and synthetic video sequences.

In summary, our paper makes the following contributions

- Subjective user study on perceived quality versus disparity on a multiview autostereoscopic display.
- Global and local disparity mapping algorithms based on saliency for stereo-to-multiview conversion.
- Extended IDW algorithm for optimized disparity mapping, which supports overall disparity range extension.
- Validation of the approaches using a variety of live action and synthetic video content.

In the remainder of this paper we review the related work (Section 2), discuss the subjective experiment on image quality on autostereoscopic displays (Section 3), detail our algorithms for global and local disparity mapping and view interpolation (Section 4), and finally present results (Section 5) and validation (Section 6).

2. Related Work

The area of **glasses-free multiview displays** has been researched extensively in the last decade, and [Lue12, WLGH12, MWDG13] provide an overview on the huge body of previous work. Most commercial displays are based on parallax barriers [Ive03] and integral imaging [Lip08]. Since then, much work has been devoted to improve on these glasses-free displays, with a recent trend towards computational displays [WLHR11, WLHR12, RHS*12, THKM13].A new method for showing stereo video on multi-layer displays was introduced in [SS13]. Unfortunately, their approach cannot deal with multi-view dispalys.

Sampling and depth of field. Similar to 2D displays, multiview displays provide a sampled approximation to continuous light fields. Chai et al. [CTCS00] presented the first analysis on sampling requirements for light field signals. Durand et al. [DHS*05] extended their work to a fundamental analysis of light transport and its sampling requirements. Based on both analyses, Zwicker et al. [ZMDP06] determine the limits of light field displays in terms of depth of field. One of their key findings shared by all multiview displays is the very shallow, device-specific depth of field. Scenes exceeding these boundaries will lead to aliasing artifacts, which can only be avoided by pre-filtering these scenes. [KJ07, RHZN11, MWA*13] extended this work to include aliasing on light field displays in the presence of visual crosstalk.

Content creation for multiview displays still poses an unresolved challenge. These displays require multiple input views, whereas the number of views and depth of field limitations are often not known during production time. A much

more promising approach is to generate multiview images from stereo footage or video+depth, using techniques such as depth-image based rendering (DIBR) [SMD*08] or image domain warping (IDW) [SWL*13]. These techniques determine how to warp the input images to new viewing positions, between the input views. However, they don't consider appropriate mapping of disparity ranges, which can lead to flattening of the perceived image, and thus reduce the depth experience. Very recent work [DSAF*13] addresses content creation for MAD using phase-based motion magnification. Compared to our work, their method does not require disparity information but only supports small disparity ranges, does not allow for local disparity manipulations, and may prefilter visually important content.

Depth adaptation has been proposed to adjust existing stereo images based on various remapping operators [LHW*10, DRE*12b, DRE*12a]. Our approach is similar to Lang et al. [LHW*10] in the sense that we use IDW, which they introduced initially, as well as the notion of saliency to control the warping. However, they did not target MAD and the particular specifics of view interpolation with overall disparity range expansion. Further, they did not cover local disparity gradient enhancements. Didyk et al. [DRE*12b] targets MAD among other applications, but filtering images is not always acceptable. Our method puts artistic intent over perception, as we try to preserve volume of important scene elements, while accepting to lose some JND of depth perception in less important image regions.

3. Subjective Experiment

Multiview displays usually exhibit a very shallow depth of field, but content is often displayed using substantially bigger depth ranges. Despite the violation of the sampling requirements, only a small amount of aliasing artifacts is usually perceived. We therefore investigate the relationship between image disparity and perceived quality with a subjective user study. The goal of the study is to determine the sensitivity of spectators to such depth ranges that exceed the display's depth of field. The outcome of this study is then used as a guideline for disparity mapping in our content creation pipeline.

In our experiment, the *stimuli* consist of the 8 synthesized views of a simple disc with the radius of 100 pixels at a certain distance, displayed against a background positioned at the display plane (Figure 2).

Both the disc and the background were covered with a number of different grayscale textures that varied in spatial contrast frequency and stereoscopic disparity. The textures were generated by applying various low-pass filters to random per-pixel noise in the frequency domain utilizing the Discrete Cosine Transform.

Our *setup* was chosen to resemble a regular viewing experience at a home theater system. All stimuli were presented on an 8-view, 47″ Alioscopy display with an approximate depth of field of ±98mm ([ZMDP06]), which corresponds to a disparity maximum of ±2.66 pixels between two consecutive views. During the experiment the subjects were comfortably seated on a chair 4.3 meters away from the display. Each subject was given a task that consisted of rating the perceived crosstalk and angular aliasing on a scale of 0 to 9 using a computer keyboard. Our subjects were 10 males and 6 females from age 25 to 36. In order to prevent the commonly encountered anchoring problems in rating studies, each subject performed the entire experiment twice, and only the results of the second iteration were used. The subjects were free to spend as much time as they needed at each trial, and most subjects finished the experiment in 20-25 minutes.

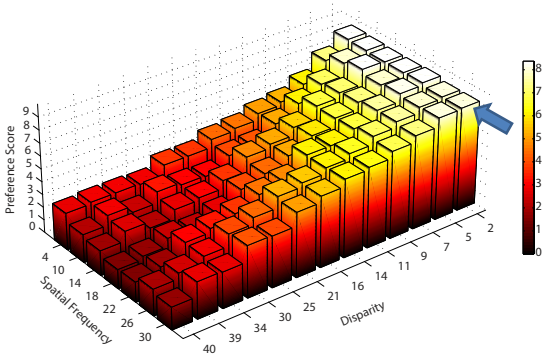

Figure 3: Subjective data showing the relation of spatial frequency and disparity to mean preference score. The blue arrow denotes the depth of field of our display.

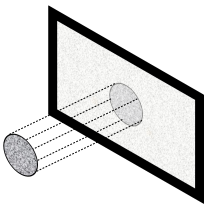

Figure 2: A depiction of how our stereoscopic stimuli is perceived by the subjects.

The mean preference scores over all subjects are shown in Figure 3. The main finding of this study is that disparity has a significant influence on preference score, which is a direct result of depth of field of the multiview display (see bandwidth analysis of Zwicker et al. [ZMDP06]), and not due to other effects such as vergence-accomodation conflict which has a much larger comfort zone of about 306 pixels [SKHB11]. We also found that, to a lesser degree, spatial frequency has also a statistically significant influence on preference score, especially for the middle frequency range. Furthermore, our study shows that disparity ranges of ×2 the display depth of field do only create noticeable artifacts for higher texture frequencies. The quality then degrades almost linearly for even higher disparities. Other lenticular or parallax-barrier based multiview displays will most likely exhibit similar characteristics.

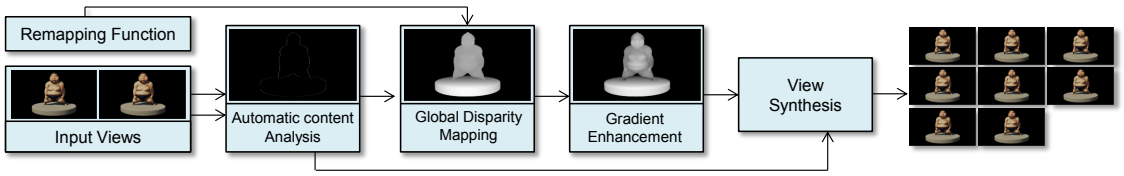

Figure 4: Overview of our stereo-to-multiview conversion pipeline. The input stereo and disparity video is analysed for saliency, and edges in a first step. Next, a global non-linear mapping transforms the input disparity space into the disparity range of the target display. The subsequent local gradient enhancement step then recovers flattened image regions of important objects in a third step. Finally, our optimized image domain warping is used to synthesize the output views for the multiview display.

Using the data shown in Figure 3 we can estimate pleasant disparity ranges by taking the spatial frequency of the content into account and choosing a suitable threshold preference score. In practice, we chose a value of maximum ± 5 pixels disparity for our display to achieve good image quality while allowing for twice the supported depth range.

4. Method

Our algorithm converts stereo 3D input into multiview video output optimized for autostereoscopic displays. The overall pipeline is illustrated in Fig. 4. In a first step, the overall disparity space is globally transformed to a new disparity range suitable for the display device limits. Next, the globally transformed disparities are locally enhanced for salient objects. Then, the transformed disparity map is used to perform view interpolation to generate the final multiview output. A detailed description of our inputs can be seen in Section 5. In the following, we will give more details on the individual steps.

4.1. Global Disparity Mapping

The disparity range of professional stereo content is usually not very well suited for MADs, which tend to support a significantly smaller disparity range. Due to inherent difficulties of disparity estimation, conversion of live action content creates specific challenges. Estimated disparity maps may contain many kinds of artifacts and imperfections, of which cardboard effects and estimation failures are most severe. In the case of cardboarding, gradients across objects are often missing, which can result in flat disparity regions partitioned into multiple layers. Furthermore, estimation failures can lead to drastic changes in disparity or holes in the estimation [†]. Gradient-based approaches such as described in the next section won't work well alone with such non-continuous content, and we therefore propose to use a two-step mapping.

Our pipeline starts by globally transforming the input disparity space into a new piece-wise linear disparity space that better suits the device-dependent limits of MADs. Our mapping works equally well for live-action input (with piecewise linear disparity maps) as well as rendered content (with continuous disparity maps). Our piece-wise linear mapping uses saliency characteristics of the input content to keep important regions as uncompressed as possible. The unavoidable distortion is hidden in areas which are less important.

In the following, we will describe the global mapping. Assuming the original disparity map contains values in a space $[d_{\min}, d_{\max}]$, the mapping is then a function $f : [d_{\min}, d_{\max}] \rightarrow [d'_{\min}, d'_{\max}]$. For our piecewise linear approach, we divide the domain of f into n equally sized bins which are linearly mapped to bins in the co-domain. Thus the linear function $f_i : [d^i_{min}, d^{i+1}_{max}] \rightarrow [d'^i_{min}, d'^{i+1}_{max}]$ is of the form $f_i(x) = \Delta_i x + \alpha_i$. If we define $R = d^{i+1}_{max} - d^i_{min}$ and $R' = d'^{i+1}_{max} - d'^i_{min}$, a linear function would be equivalent to a single bin with $\Delta = R'/R$. We would like our Δ_i to satisfy the following conditions:

$$\sum_{i=1}^{n} \Delta_i = \Delta\,, \quad \Delta_i \geq 0, \forall i. \tag{1}$$

This ensures that we map our disparities exactly into the target space and that the three dimensional position of pixels is never reversed, i.e. a pixel will never be mapped to a position in front of another pixel if it was behind it originally. Given these conditions and naming s_i the sum of the saliency values of all pixels in the bin i, we propose the solution:

$$\Delta_i = \frac{s_i}{\sum_{j=1}^{n} s_j} \Delta k + (1-k)\frac{\Delta}{n}. \tag{2}$$

The coefficient $k \in [0,1]$ controls by how much a given bin can be compressed. A value of 0 defaults the mapping into a simple linear mapping and a value of 1 means bins with no salient pixels will have their disparity completely removed. An illustration showing the result of this algorithm can be seen in Fig. 5. The method described above is directly related to the saliency provided for the scene, and as such is very sensitive to temporal instability in saliency. To prevent the global mapping function from becoming temporally unstable, saliencies are filtered out over several frames which ensures that the disparity re-mapping is similar for consecutive images. A related approach was proposed in [LHW*10], which integrates over saliency instead of computing a piece-

[†] Please compare input disparity maps of synthetic vs. live action content in our supplemental video.

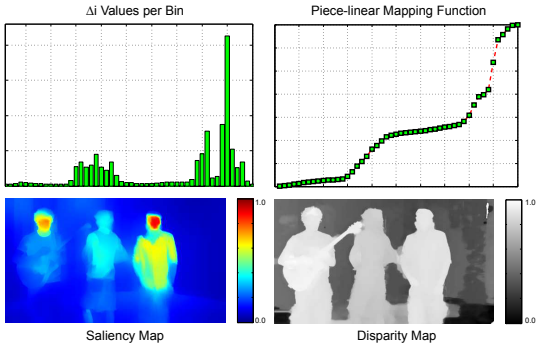

Figure 5: The top right image shows our piecewise linear mapping function with the respective Δ_i values per bin shown on the top left. The bottom images depict the saliency map (left) and the corresponding disparity map (right).

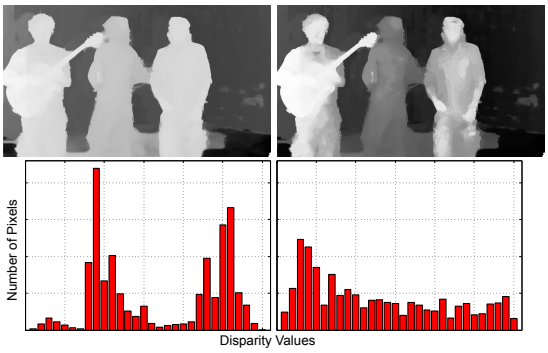

Figure 6: Results of linear (left) and our saliency-based piece-wise linear (right) global mapping. The histograms show how our approach compresses unimportant space, while retaining volume of salient objects as good as possible.

wise linear function, and does not provide a compression safeguard parameter k.

Fig. 6 shows a result of such piecewise linear global mapping. The left side shows results of a linear mapping. Our results are shown on the right. Both are mapped into a fraction of the input disparity range. Our mapping function nicely compresses empty space, while retaining disparity in more salient regions leading to an enhanced depth experience.

Other operators. Similar to [LHW*10] our pipeline supports arbitrary global mappings which can be specified by the user or predefined for a certain system. In principle, all C0-continous and monotonically increasing functions are allowed, such as operators proposed by [DRE*12b], or nonlinear operators proposed in [LHW*10].

4.2. Local Disparity Gradient Enhancement

After the global disparity mapping we perform an additional, local mapping step. Our main goal is to locally enhance disparity gradients in important image regions for an increased depth perception. We formulate our goal as set of constraints, that can then be solved for the locally enhanced disparity map D_L with a least-squares energy minimization. A result of this mapping can be seen in Figure 7. In the following, we will use the ensuing notation. Let $\mathbf{x} \in \mathbf{R}^2 = (x,y)$ be an image position, and $D(\mathbf{x}) \in \mathbf{R}$ be a disparity map.

Gradient constraints. As our central constraint, we enforce the mapped disparity gradients of salient image regions to be similar to the gradients of the input disparity map D_I:

$$\frac{\partial}{\partial x} D_\mathrm{L}(\mathbf{x}) = \alpha \frac{\partial}{\partial x} D_\mathrm{I}(\mathbf{x}), \tag{3}$$

$$\frac{\partial}{\partial y} D_\mathrm{L}(\mathbf{x}) = \alpha \frac{\partial}{\partial y} D_\mathrm{I}(\mathbf{x}). \tag{4}$$

The global parameter α is then a constant factor to control the overall disparity enhancement, and is dependent on the disparity compression from the previous global mapping. In general, we propose to use a factor of $\alpha = 2^{(d_\mathrm{range}/d'_\mathrm{range})}$, where d_range and d'_range are the disparity ranges before and after the global mapping, respectively.

Global mapping constraints. In addition, we enforce the overall mapping to follow the global mapped disparity D_G as closely as possible:

$$D_\mathrm{L}(\mathbf{x}) = D_\mathrm{G}(\mathbf{x}). \tag{5}$$

Least squares energy minimization. The constraints defined in the above equations can then be rewritten as constraints of a linear least squares energy minimization. Let $S(x,y) : \mathbf{R}^2 \to (0,1]$ be a saliency map that classifies important image regions. A small amount of saliency is added to all pixels to prevent null weights in the constraints. Equations (3) and (4) can then be rewritten as

$$E_\mathrm{g}(D_\mathrm{L}) = \sum_\mathbf{x} S(\mathbf{x}) \left|\left|\nabla D_\mathrm{L}(\mathbf{x}) - \alpha \nabla D_\mathrm{I}(\mathbf{x})\right|\right|^2, \tag{6}$$

where ∇ is the vector differential operator, and $||\cdot||$ defines the vector norm. The global mapping constraints (5) are reformulated as

$$E_\mathrm{l}(D_\mathrm{L}) = \sum_\mathbf{x} \left(D_\mathrm{L}(\mathbf{x}) - D_\mathrm{G}(\mathbf{x})\right)^2. \tag{7}$$

The optimum linear least squares solution for $D_\mathrm{L}(\mathbf{x})$ can then be found by minimizing

$$\underset{D_\mathrm{L}}{\operatorname{argmin}} \left(w_\mathrm{g} E_\mathrm{g}(D_\mathrm{L}) + w_\mathrm{l} E_\mathrm{l}(D_\mathrm{L})\right). \tag{8}$$

Note, that this minimum can be computed by solving a linear system, see [GRG*13] for a good overview. The system defined in (8) will try to enhance the gradients of the salient regions, while trying to enforce all other disparity values towards their globally mapped version D_G. Disparity edges, i.e. strong disparity gradients between objects at different depths, can lead to a high contribution to the squared error, and thus such disparity edges would be enforced strongly as

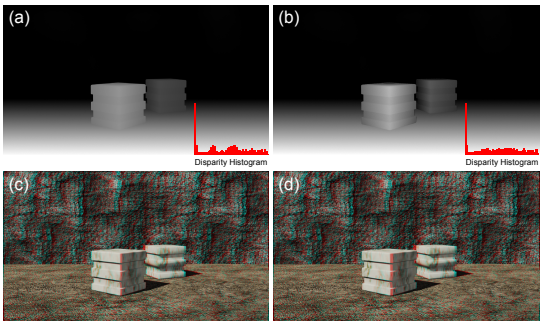

Figure 7: Local disparity gradient enhancement. (a) shows a linearly mapped disparity map of a scene; (b) shows the disparity map adapted by our algorithm; in (c) and (d) two generated views are displayed in anaglyph, in a disparity range similar to two adjacent views of a MAD display. Notice how the cardboarding effect flattens out the cube in (c).

well. As we are only interested in gradients within the objects, these disparity edges need special treatment which we will discuss in the following. This step is different from previous methods, such as [LHW*10].

Disparity edges. The gradient constraint can lead to artifacts around disparity edges due to very high disparity gradients between objects. We thus remove the influence of such disparity edges to enforce the gradient enhancement within objects only. Luckily, disparity edges can usually be detected quite robustly on the disparity map. We use a combination of a simple threshold function and a more sophisticated Canny edge detector on the input disparity D_{I} to determine the set of edge pixels E. Subsequently, we enforce the salience value to be zero at these edge pixels $S(\mathbf{x}) = 0$ for $\mathbf{x} \in E$.

Fig. 7 shows a result of local disparity gradient enhancement. In our result the cubes have more volume, while the linearly mapped version appears more flat.

4.3. View Interpolation

We developed an extension of image domain warping [LHW*10, SWL*13] for stereo-to-multiview interpolation. The previously computed optimized disparity maps are used as main input to control this process. Based on optimized disparity, we formulate a constrained energy minimization problem, which is solved by linear least squares optimization (similar to previous section). The results are warping functions, which deform the input views to generate the novel in-between views. In addition to disparity, we apply conformal constraints that penalize local deformations, and line constraints that disadvantage line bending.

In the following, we will describe the particular constraints in more detail. The optimized disparity map $D(\mathbf{x}) : \mathbf{R}^2$ is used to compute a warp $w(\mathbf{x}) : \mathbf{R}^2 \rightarrow \mathbf{R}^2$, where the deformations should be hidden in visually less important regions. The warp $w(\mathbf{x})$ will then describe the optimal transformation of the input view corresponding to $D(\mathbf{x})$.

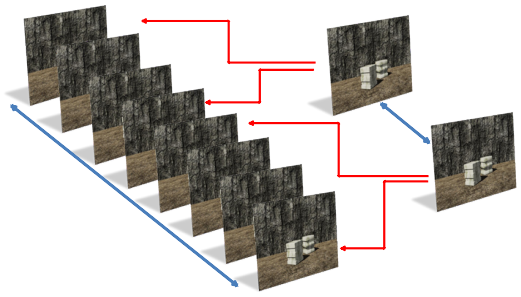

Figure 8: A pair of input figures is warped to a set of multiview results. Notice that the results are now mapped to a completely new disparity range. The original figures are not among the results, which are created by interpolating between two warps, represented here with red arrows.

Disparity constraints. The disparity constraints can be viewed as positional constraints: every point in the image should be translated to the position described by its disparity:

$$w(\mathbf{x}) = \begin{bmatrix} x + D(\mathbf{x}) \\ y \end{bmatrix}. \tag{9}$$

Conformal constraints. The conformal constraints penalize deformations, and are mainly evaluated on visually salient image regions. A constraint of the form $\frac{\partial}{\partial x} w(\mathbf{x})^{(x)} = 1$ prescribes to avoid any compression or enlargement along the x-direction, whereas a constraint of the form $\frac{\partial}{\partial x} w(\mathbf{x})^{(y)} = 0$ penalizes deformations that result in a pixel-shear operation. All four constraints are then formulated as:

$$\frac{\partial}{\partial x} w(\mathbf{x}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \frac{\partial}{\partial y} w(\mathbf{x}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{10}$$

Depth ordering constraints. Because disparity edges often have gradients that are very different from their neighbors, we use edge detection to reduce their saliency values, often preventing artifacts. For this we use the same edge map as the one previously calculated for Section 4.2. Additionally, pixel overlaps where the correct order of pixels along a line is reversed often occurs for large warps. This may generate large distortions in the warp optimization step. To resolve such conflicts we perform a simple check where, in case of an overlap, the occluded pixel is moved to the coordinate of its occluder.

Temporal constraints. Applying all constraints results in output images that have correct disparity values and hide distortions in visually non-salient areas. However, when applying this method for each frame of a video sequence, small changes in the input might result in larger changes within the optimization, which may lead to disturbing temporal artifacts. We try to remedy this by introducing a temporal constraint that takes into account the warps calculated for pre-

vious frames as an additional disparity constraint. This effectively makes these constraints three-dimensional, linking temporally separated pixels, as shown below:

$$w_t(\mathbf{x}) = w_{t-1}(\mathbf{x}). \tag{11}$$

Multiview generation for remapped disparity. Our final goal is to generate multiview content related to the new, optimized disparity maps. This is done in a 2-step approach, as outlined in Figure 8. The first step maps the input images to new virtual images corresponding to the optimized disparity. The second step then does the actual interpolation. This distinction is only conceptional. In practice both warps are done at once. Typically the overall disparity range from leftmost to rightmost view of an MAD is larger than the disparity range of the input stereo pair. Our optimized disparity maps carry the information about this necessary expansion of the overall disparity range, which is illustrated by the blue arrows in Figure 8. Within the expanded range we then linearly interpolate from left and right input view as illustrated by the red arrows in Figure 8. The disparity range between each image pair of the resulting multiview image set is then a fraction of the input disparity range. Such expansion of the overall disparity range with intermediate view rendering would not be easily possible with DIBR, due to dis-occlusions, which require in-painting. For this reason we use IDW instead, which does not create dis-occlusions and can handle disparity range expansions without noticeable artefacts. This step is an extension of the algorithm in [LHW*10].

Assume we are generating the first set of multiview images based on the left input image only. In the first step, the adjustments to the input image according to the disparity change have to be determined. To achieve this, we compute a first warp $w_{\text{ext}}(\mathbf{x})$ using the disparity map $D_{\text{ext}} = D_{\text{I}} - D_{\text{L}}$. This warp then describes the transformation of the left image to its adjusted new left image that corresponds to the disparity map D_{L}. In a second step, a warp $w_{\text{cen}}(\mathbf{x})$ is computed that determines the transformation from the left input image to the center view between the left camera and right camera.

Both warps $w_{\text{ext}}(\mathbf{x})$ and $w_{\text{cen}}(\mathbf{x})$ can then be used to compute warps $w(a)$ that transform the left input image to a first set of multiview images

$$w(a) = a w_{\text{ext}}(\mathbf{x}) + (1-a) w_{\text{cen}}(\mathbf{x}) \;\text{ for } a = [0..1], \tag{12}$$

whereas $a = 0$ corresponds to the left most image, and $a = 1$ corresponds to the center image. The second set of multiview images can then be generated in the same manner based on the right input view.

5. Experiments and Results

We evaluated our pipeline on a variety of synthetic and filmed stereoscopic video sequences. For synthetic scenes, we use ground truth disparity maps and saliency maps rendered from object annotations. This allows the artist to decide which objects should retain as much depth as possible by assigning an importance value to these objects. The importance values are then rendered into a saliency map. For the filmed scenes, we either use automatically generated depth maps [ZRM*] (Musicians, Band, Poker) or computed and additionally hand-tuned depth maps [WFY*10] (Ballons, Kendo). All filmed scenes use an extended version of a contrast-based saliency algorithm [PKPH12] that employs an edge-aware spatiotemporal smoothing [LWA*12] to achieve temporal consistency. Most steps of our pipeline are implemented in Matlab, only the actual warp rendering to generate the interpolated views has been implemented in OpenGL in C++. Multiview image sequences can be generated in 15 - 600 seconds per frame, depending on the input size and resolution of the image warp grid.

For all scenes, we evaluated the simple linear mapping and our saliency-based mapping. The view to view disparity range for our target display is determined using the results of our user study as ± 5 pixels. Figure 9 shows anaglyph results accompanied with the associated disparity maps and histograms. Our method clearly enhances the depth for the salient image regions, and effectively compresses less salient image regions as well as empty disparity ranges. The results generated using our method show rounder, more voluminous objects, and are thus able to convey a deeper depth experience even for such small depth ranges. Figure 10 shows generated multiview images for additional scenes. As can be seen, our adapted warping method is able to hide distortions in visually unimportant regions, and avoids distracting artifacts even for scenes with inaccurate estimated depth maps.

Figure 11 shows a comparison between linear mapping, our mapping algorithm, and another perceptually-based disparity compression algorithm [DRE*12b]. Linearly mapping the input range results in flattening the whole scene uniformly, which results in loss of depth perception and cardboarding. Both our method and Didyk et al.'s method [DRE*12b] compresses to the same overall disparity range, but provide more depth perception. In contrast to our method, Didyk's method uses a perceptual model for noticeable differences based on disparity, luminance and contrast, whereas our model focuses on salient image regions. While both methods lead to an increased depth perception, our method enhances the depth on the front-most persons better while flattening less salient parts. Didyk's method on the other hand is able to retain "just enough" disparity to perceive depth uniformly across the image.

While our method generates improved results compared to a simple linear mapping, there are also some drawbacks. First, our method relies completely on saliency and will not be able to produce improved results if the saliency computation fails. Fortunately, our method will fall back to a simple linear mapping in the worst case, due to our compression safeguard. Second, our method is computationally expensive and not yet ready for real-time applications. In addition, our

rendering method tries to minimize distortions by distributing the error over possibly large, unimportant backgrounds. As a result, our constraints might lead to a small jump between the two middle views, which could be resolved at the expense of other artifacts.

6. Subjective Validation

We validated our method using subjective testing, where we showed multiview video content on an 8-view 47″ Alioscopy display to our subjects. Our stimuli comprised of result pairs using naïve linear mapping as well as our method, presented in random order. In total, 7 video sequences where displayed. After watching the two stimuli in each trial, the subjects were queried on (i) which of the two stimuli has more depth, and (ii) which one has more artifacts. Our validation experiment had 20 participants naïve to the purpose of the study.

Figure 12 shows the responses of each subject averaged over the video scenes. The top figure shows that among the tested subjects there was a strong opinion that our results have more depth. Pearson's chi-square goodness-of-fit analysis demonstrated a statistically significant opinion that our method has more depth, $\chi^2(1, 140) = 37.03$, $p<.01$. In total, 76% of test subjects stated our method to have more depth. The bottom of Figure 12 shows the result for the second question. There was no statistically significant preference, $\chi^2(1, 140) = 1.83$, $p>.1$. Among all votes 56% indicated our method has more artifacts, and 46% indicated that the naïve mapping had more artifacts.

We performed *Anova* analysis to determine if there is a main effect due to either subjects or video sequences. For the depth assessment task, the p values were found 0.9717 for subjects and 0.3036 for video sequences, indicating that both factors do not have a significant effect on our results (both $\gg 0.05$). Same was found to be true for the artifact assessment task, where the p values were 0.8361 and 0.7352 respectively.

In conclusion, our subjective data shows that our method consistently produces results with more perceived depth compared to the naïve mapping, without causing a significant difference in image quality.

7. Conclusion

We presented a saliency-based stereo-to-multiview conversion method that generates optimized content for autostereoscopic multiview displays. In a first step, we perform a global disparity mapping that flattens out unimportant regions while trying to retain important image regions. In a second step, we locally enhance the disparity gradients for visually salient regions. Finally, we employ an extended image domain warping algorithm to render the output views according to the modified disparity maps.

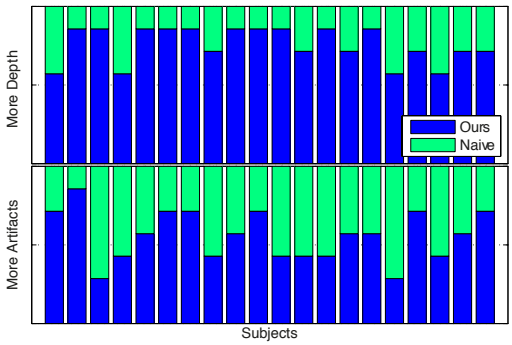

Figure 12: Our validation study revealed a strong opinion among the tested subjects that our method resulted in more perceived depth compared to the naïve mapping (top). The study also showed no clear trend on either methods producing more artifacts than the other.

As shown in our final validation study, our method clearly improves the amount of perceived depth compared to a simple linear mapping. Compared to other state-of-the-art methods, our approach is more faithful and retains the artistic intent. In addition, our extended image domain warping is robust to temporally unstable and inaccurate disparity maps. In our initial user study we validated theoretical limitations on disparity ranges of autostereoscopic displays, while showing that those can be relaxed in practice to some extent. Nevertheless for conversion of typical stereoscopic input, significant disparity remapping is necessary.

8. Acknowledgements

We would like to thank Maurizio Nitti for providing content for this paper. Additional thanks go to all the participants of our user studies and the researchers at DRZ and ETH for their help. The research that led to this paper was supported in part by the European Commission under the Contract FP7-ICT-287723 REVERIE.

References

[CTCS00] CHAI J.-X., TONG X., CHAN S.-C., SHUM H.-Y.: Plenoptic sampling. In *SIGGRAPH* (2000), pp. 307–318. 2

[DHS*05] DURAND F., HOLZSCHUCH N., SOLER C., CHAN E., SILLION F. X.: A frequency analysis of light transport. *ACM Transactions on Graphics 24* (2005), 1115–1126. 2

[DRE*12a] DIDYK P., RITSCHEL T., EISEMANN E., MYSZKOWSKI K., SEIDEL H.-P.: Apparent stereo: The cornsweet illusion can enhance perceived depth. In *Human Vision and Electronic Imaging XVII, IS&T/SPIE's Symposium on Electronic Imaging* (Burlingame, CA, 2012), pp. 1–12. 3

[DRE*12b] DIDYK P., RITSCHEL T., EISEMANN E., MYSZKOWSKI K., SEIDEL H.-P., MATUSIK W.: A luminance-contrast-aware disparity model and applications. *ACM Transactions on Graphics 31*, 6 (2012), 184:1–184:10. 2, 3, 5, 7, 10

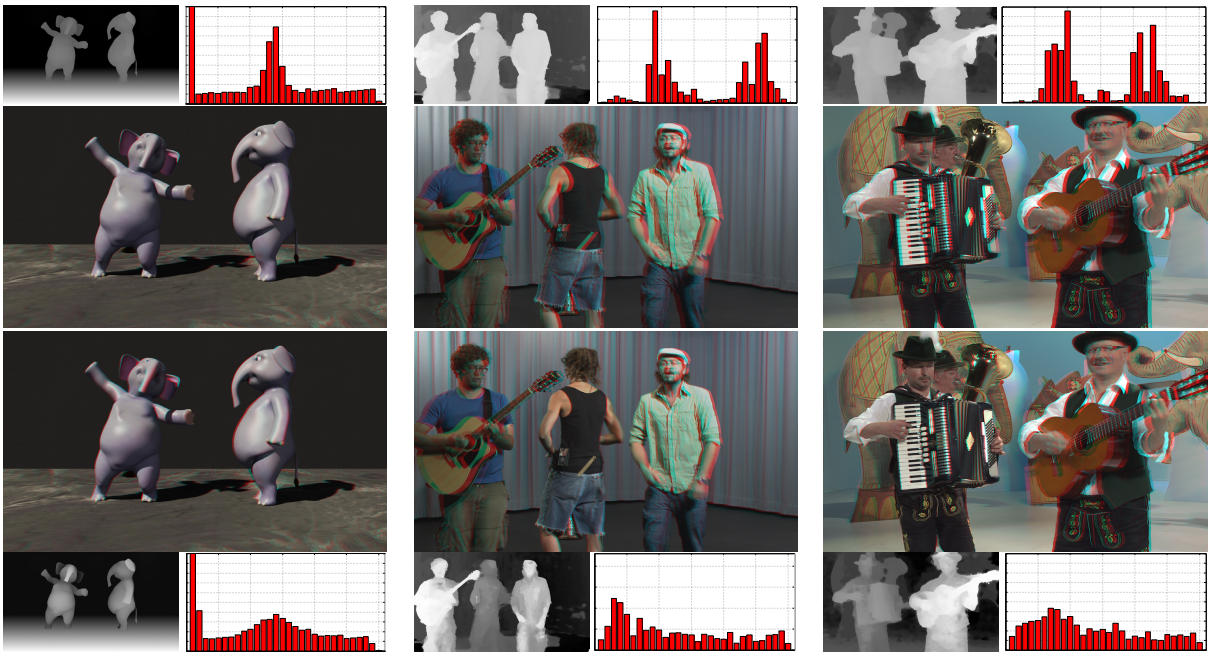

Figure 9: Comparison between linear mapping (top) and our saliency based mapping (bottom), shown in anaglyph with their associated disparity map and histogram. Our method retains more depth volume for the important parts of the scene while flattening out less important parts as well as empty space. Our mapping effectively creates more apparent depth within the same overall depth limits.

[DSAF*13] DIDYK P., SITTHI-AMORN P., FREEMAN W. T., DURAND F., MATUSIK W.: Joint view expansion and filtering for automultiscopic 3d displays. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2013, Hong Kong) 32*, 6 (2013). 3

[GRG*13] GREISEN P., RUNO M., GUILLET P., HEINZLE S., SMOLIC A., KAESLIN H., GROSS M.: Evaluation and FPGA implementation of sparse linear solvers for video processing applications. *Transactions on Circuits and Systems for Video Technology 23*, 99 (2013). 5

[Ive03] IVES F.: Parallax stereogram and process for making same. U.S. Patent No. 725,567, 1903. 2

[KJ07] KONRAD J., JAIN A.: Crosstalk in automultiscopic 3-d displays: blessing in disguise? *Stereoscopic Displays and Virtual Reality Systems XIV 6490*, 1 (2007). 2

[LHW*10] LANG M., HORNUNG A., WANG O., POULAKOS S., SMOLIC A., GROSS M.: Nonlinear disparity mapping for stereoscopic 3D. *ACM Transactions on Graphics 29*, 4 (2010), 75:1–75:10. 3, 4, 5, 6, 7

[Lip08] LIPPMANN G. M.: La photographie integrale. *Comptes-Rendus 146* (1908), 446–451. 2

[Lue12] LUEDER E.: *3D Displays*. Wiley, 2012. 2

[LWA*12] LANG M., WANG O., AYDIN T., SMOLIC A., GROSS M.: Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics 31*, 4 (2012), 34:1–34:8. 7

[MWA*13] MASIA B., WETZSTEIN G., ALIAGA C., RASKAR R., GUTIERREZ D.: Display adaptive 3d content remapping. *Computers & Graphics 37* (2013), 983–996. 2

[MWDG13] MASIA B., WETZSTEIN G., DIDYK P., GUTIERREZ D.: A survey on computational displays: Pushing the boundaries of optics, computation, and perception. In *Computers & Graphics* (2013), vol. 37, pp. 1012–1038. 2

[PKPH12] PERAZZI F., KRÄHENBÜHL P., PRITCH Y., HORNUNG A.: Saliency filters: Contrast based filtering for salient region detection. In *IEEE CVPR* (2012), pp. 733–740. 7

[RHS*12] RANIERI N., HEINZLE S., SMITHWICK Q., REETZ D., SMOOT L. S., MATUSIK W., GROSS M.: Multi-layered automultiscopic displays. *Computer Graphics Forum 31*, 7pt2 (2012), 2135–2143. 2

[RHZN11] RAMACHANDRA V., HIRAKAWA K., ZWICKER M., NGUYEN T.: Spatioangular prefiltering for multiview 3D displays. *TVCG 17*, 5 (2011), 642–654. 2

[SKHB11] SHIBATA T., KIM J., HOFFMAN D. M., BANKS M. S.: The zone of comfort: redicting visual discomfort with stereo displays. *Journal of Vision 11*, 8 (2011), 8:1–8:29. 3

[SMD*08] SMOLIC A., MÜLLER K., DIX K., MERKLE P., KAUFF P., WIEGAND T.: Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems. In *Proc. ICIP 2008, IEEE International Conference on Image Processing* (2008). 2, 3

[SS13] SINGH D. S. K., SHIN J.: Real-time handling of existing content sources on a multi-layer display, 2013. 2

[SWL*13] STEFANOSKI N., WANG O., LANG M., GREISEN P., HEINZLE S., SMOLIC A.: Automatic view synthesis by image-domain-warping. *Image Processing, IEEE Transactions on 22*, 9 (2013), 3329–3341. 2, 3, 6

[THKM13] TOMPKIN J., HEINZLE S., KAUTZ J., MATUSIK W.: Content-adaptive lenticular prints. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2013)* (July 2013), vol. 32. 2

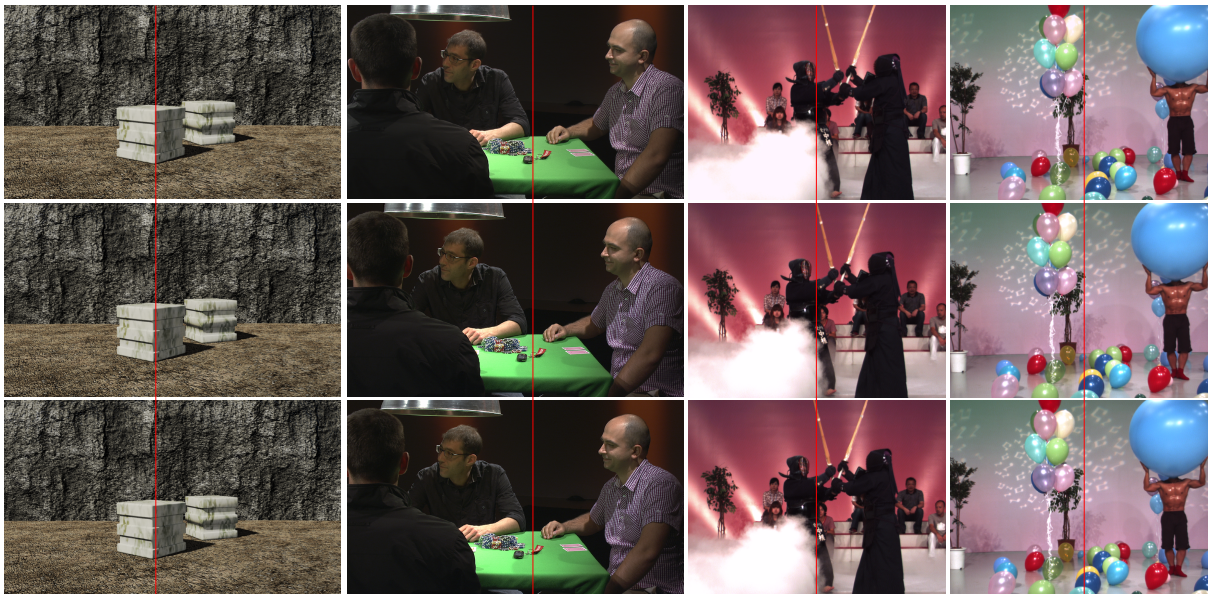

Figure 10: Three views for one frame, generated using our pipeline. Despite the challenging disparity maps, our method is able to hide distortion in visually less important regions and is able to generate novel views without many noticable artifacts.

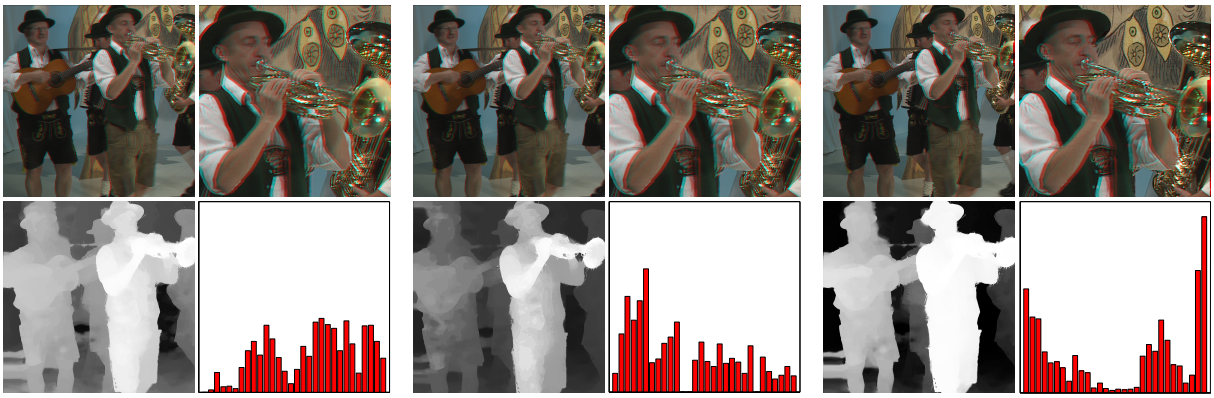

Figure 11: Comparison between linear mapping, our saliency-based mapping, and the mapping of Didyk et al. [DRE*12b] (from left to right). Our mapping is able to increase the perceived depth best, while flattening out lesser important regions. Didyk's method on the other hand retains a noticeable depth difference across the image. Notice the carboarding effect happening in the insets on the left and right.

[WFY*10] WILDEBOER M., FUKUSHIMA N., YENDO T., TEHRANI M., TANIMOTO M.: A semi-automatic multi-view depth estimation method. In *Visual Communications and Image Processing* (2010). 7

[WLGH12] WETZSTEIN G., LANMAN D., GUTIERREZ D., HIRSCH M.: Computational displays. In *ACM Siggraph Course Notes* (2012). 2

[WLHR11] WETZSTEIN G., LANMAN D., HEIDRICH W., RASKAR R.: Layered 3D: Tomographic image synthesis for attenuation-based light field and high dynamic range displays. *ACM Transactions on Graphics 30*, 4 (2011). 2

[WLHR12] WETZSTEIN G., LANMAN D., HIRSCH M., RASKAR R.: Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting. *ACM Transactions on Graphics 31*, 4 (2012), 1–11. 2

[ZMDP06] ZWICKER M., MATUSIK W., DURAND F., PFISTER H.: Antialiasing for automultiscopic 3D displays. In *Eurographics Symposium on Rendering* (2006). 2, 3

[ZRM*] ZILLY F., RIECHERT C., MÜLLER M., EISERT P., SIKORA T., KAUFF P.: Real-time generation of multi-view video plus depth content using mixed narrow and wide baseline. *Journal of Visual Communication and Image Representation*. 7