Eurographics Symposium on Rendering 2013
Nicolas Holzschuch and Szymon Rusinkiewicz
(Guest Editors)

Volume 32 (2013), Number 4

Optimizing Disparity for Motion in Depth

Petr Kellnhofer Tobias Ritschel Karol Myszkowski Hans-Peter Seidel

MPI Informatik

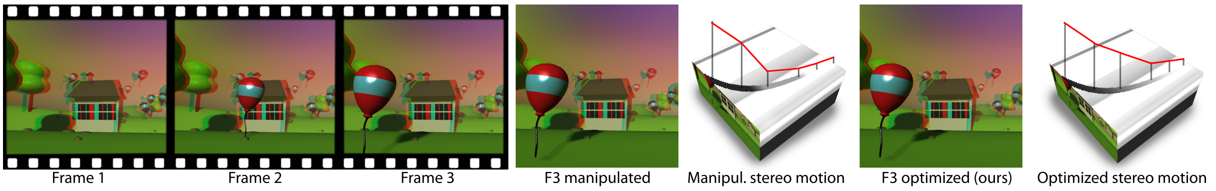

Figure 1: *Three frames of an animation showing a motion of a balloon in depth* (Left to right). *The outcome of disparity manipulation and our optimization, which reduces distortions in the balloon motion perception, are shown for frame* F3. *The cross-section through the space-time cube shows disparity as a function of time and space. When the balloon enters the zone between the house and the tree, the disparity manipulation results in abrupt disparity changes over time that are not present before the manipulation. Our optimization prevents the manipulation from distorting this originally smooth change.*

Abstract

Beyond the careful design of stereo acquisition equipment and rendering algorithms, disparity post-processing has recently received much attention, where one of the key tasks is to compress the originally large disparity range to avoid viewing discomfort. The perception of dynamic stereo content however, relies on reproducing the full disparity-time volume that a scene point undergoes in motion. This volume can be strongly distorted in manipulation, which is only concerned with changing disparity at one instant in time, even if the temporal coherence of that change is maintained. We propose an optimization to preserve stereo motion of content that was subject to an arbitrary disparity manipulation, based on a perceptual model of temporal disparity changes. Furthermore, we introduce a novel 3D warping technique to create stereo image pairs that conform to this optimized disparity map. The paper concludes with perceptual studies of motion reproduction quality and task performance in a simple game, showing how our optimization can achieve both viewing comfort and faithful stereo motion.

1. Introduction

Watching stereoscopic media such as movies and computer games can be an exciting experience, often described by statements such as "…and then this character really jumped *out* of the screen!" Remarkably many such statements refer to a *change* of perceived depth rather than a static condition. Until now, computational modeling of temporal changes of disparity (stereo motion) has only received little attention. Regrettably, manipulation of disparity, which is routinely performed to improve viewing comfort or to achieve artistic objectives, can impede perceived motion in depth. For example disparity compression in local scene regions may induce acceleration of motion-in-depth for objects traversing such regions. Such uncontrolled binocular cues can be perceived as annoying motion artifacts, which in navigation simulators or remote manipulators can affect the performance in precision-demanding tasks such as collision avoidance. In this work we show how to process binocular disparity to reproduce both faithful depth perception and stereo motion.

Let us consider the balloon approaching the viewer in Fig. 1 as an example. Here state-of-the-art solutions will improve viewing comfort by selective disparity compression in empty scene regions and preserve the balloon shape only at individual instants of time. If the manipulation causes compression or expansion of some depth regions with respect to others, the balloon will exhibit sudden change of speed i. e., acceleration, as it moves in depth from one such region to another. Temporal coherence of such disparity manipulation can be maintained by smoothing and propagating the resulting disparity changes along the motion flow, but the objective here is not to preserve the motion appearance fidelity per se. Moreover, since the disparity change is a strong cue for motion itself [GR98], such uncontrolled disparity manipulation may introduce a cue conflict with respect to important pictorial cues such as the balloon size change due to perspective scaling. Clearly, a more holistic approach to the object stereo motion is required that accounts for local scene configurations as well. In this work we address this problem in the following four contributions:

- A perceptual analysis of stereo motion
- An optimization to detect and preserve stereo motion cues
- A perceptual study of stereo motion task performance
- A 3D warping to create a stereo image pair with arbitrary, spatially-varying disparity from a polygonal 3D scene.

2. Background

Dynamic changes of binocular disparity naturally arise through any form of object motion in the surroundings. Even when the eyes perfectly converge on an object moving in depth, which results in the null absolute disparity for this object, the relative disparity with respect to other objects creates a strong cue for detecting motion-in-depth (MID), and estimating its direction and velocity [EC85]. While monoscopic cues such as changing object size and its visibility/occlusion configurations, perspective deformations of inclined surfaces, and lens accommodation may contribute to the motion judgment as well, dynamic binocular disparity greatly improves the precision of motion perception [GR98].

Such reliable motion judgment is required in many everyday tasks such as estimating the time when the approaching object will reach a specified position, called also the time-to-contact (TTC), determining the object impact direction, or performing the interception task of one moving object by another. Clearly, these tasks are of high relevance in many computer game and training simulator scenarios as well, where the participant performance may critically depend on the precision of perceived motion.

It is believed that two binocular mechanisms might contribute to the MID perception, but their precise role is still an open research problem [HNG08]. A changing disparity over time (CDOT) mechanism (called also the stereo-first mechanism) determines relative disparities between scene elements and monitors their changes. An interocular velocity differences (IOVD) mechanism (called also the motion-first mechanism) relies on combining two monocular velocity signals that are derived based on temporally coherent motion patterns separate for each eye.

The sensitivity studies for the MID detection, which have been performed for various temporal frequencies of disparity modulation, revealed the peak sensitivity within the range 0.5–2 cycles-per-degree [Tyl71, Ric72], and the high-frequency cutoff at 10.5 Hz [NBPC05], which is significantly lower than 60 Hz as measured for temporal modulation of luminance contrast.

Harris and Watamaniuk [HW95] and Portfors-Yeomans and Regan [PYR96] investigated the sensitivity of human visual system (HVS) to speed changes in MID, which arise due to the CDOT and IOVD mechanisms. They found the sensitivity to follow Weber's law, where the ratio (the Weber fraction k) of discriminated speed change to a reference speed typically varied between 0.08 up to 0.25. Interestingly, the Weber fraction does not significantly depend on the magnitude of disparity [BS04, Fig. 6], and whether the object moves away or approaches the observer. Based on those findings, we derive a model of perceived disparity velocity changes in Sec. 4.2.

Interaction of binocular MID and monocular cues (in particular the change of size) typically leads to the overall improvement of motion judgement. Gray and Regan [GR98] found that for separately considered monocular and binocular cues consistently underestimated or overestimated values of absolute TTC are obtained, while the accuracy improved significantly when both cues are available. As the linear horizontal width of a moving object decreases the reliability of monocular information drops [GR98], and then the precision of TTC task might fully rely on the quality of binocular information. Surprisingly, binocular vision seems to be important in the TTC task for distances relevant for highway driving up to 75 m [CL88]. As observed by Regan and Beverley [RB79] with increasing motion speed or inspection times (lower framerates) the changing-disparity cue becomes more effective in conveying the MID sensation than the changing-size cue. This is also the case when MID is accompanied by more complex shape changes than simple isotropic rescaling, which may arise for deformable or rigid, but non-rotationally symmetric objects. Also, the detection thresholds for just noticeable MID are typically lower for the binocular cues than for their monocular counterparts. Regan and Beverley demonstrated that a change in size can be cancelled by an antagonistic change in relative disparity, and proposed a simple weighted-sum model to combine both cues.

Heuer [Heu87] reported that for contradictory cues, rivalry can be observed instead of summation, which may lead to the instability of dominating cue. Brenner et al. [BvdBvD86] suggest that conflicting cues might be responsible for large differences between subjects in the motion judgment, and propose that most likely scene interpretation that is selected by subjects should minimize the cue conflicts. Gray and Regan [GR98] observe that the human performance in the TTC task is decreasing for distorted stereo configurations.

All the above indicates that the high accuracy of dynamic disparity information is required to enable reliable MID judgement, which is instrumental in numerous practical tasks. In many object motion scenarios dynamic disparity is the only reliable or the most effective MID cue to perform those tasks.

Even in the presence of other strong MID cues, their effectiveness can be seriously degraded when combined with distorted binocular disparity. Our perceptual study in Sec. 5 is an example where distorted stereo has a significant effect on task performance. As we discuss in the following section such distorted disparity information is quite common in stereo 3D imaging and computer graphics applications.

3. Previous Work

Besides the proper set-up of real and virtual cameras when generating stereo content [Men09, HGG*11, OHB*11], the post-manipulation of disparity has recently received considerable attention [JLHE01, LHW*10, DRE*11, KZC*11, YLXH13]. Typically disparity is manipulated in individual frames, and temporal processing is limited to "smoothing" between different disparity ranges at scene transitions [LHW*10, YLXH13]. Also, it is ensured that for moving scene elements local warping distortions are propagated along the motion flow to the successive frames [LHW*10]. A limited temporal extent of per-frame temporal smoothing and low-pass filtering characteristic of first order smoothing terms used do not allow to maintain high-frequency temporal features of disparity dynamic for complex motions. Our work goes beyond such local smoothing and enables explicit global control of disparity changes over time. This way we can preserve both spatially local disparity manipulations and temporarily global disparity dynamics.

In video retargeting applications rigidity in temporally salient image regions is often enforced [KLHG09, WLSL10]. This can be performed in disparity-driven image warping as well [LHW*10], but here the goal is to avoid geometric deformation of moving objects, which are strong gaze attractors, rather than to prevent deformations of perceived motion trajectory. Wang et al. [WLSL10] used a second-order smoothing term to minimize creation of "virtual" camera motion. Hoffman et al. [HKB11] analyzed the impact of image refresh rate and stereo 3D display technology (precisely, the eye-view separation method) on the visibility of flicker, motion smoothness, and distortions in perceived depth.

While many factors may impact visual comfort when looking at stereoscopic displays the conflict between the eye convergence and accommodation is usually identified as the most prominent one [LIFH09, SKHB11]. Similar to other disparity manipulation techniques [JLHE01, LHW*10, DRE*11, KZC*11, YLXH13], our approach ensures that the comfortable range of binocular disparity is always maintained. Yano et al. [YIMT02] and Speranza et al. [STRH06] investigated the impact of object motion on the visual comfort. They found that the rate of disparity changes over time, which is determined by the object velocity, may strongly affect the visual comfort. Also, frequent changes from crossed disparity with vergence point in front of the horopter to uncrossed disparity with vergence point behind the horopter may significantly reduce the comfort.

Stereo 3D typically evokes higher positive emotions and stronger feeling of immersion in games compared to the 2D mode and often leads to a better accuracy of performed tasks, especially those that involve spatial 3D interaction [KSL12]. For example, Hubona et al. [HWSB99] have found that stereoscopic viewing improves both precision and speed in the object positioning and resizing tasks, while object shadows are far less effective cues. In VR applications, which involve self-motion based on optical flow, binocular 3D information facilitates the judgement of moving object direction through the flow parsing in the HVS into the self-motion and object motion components [MA09]. All these applications involve the movement of objects in some form, and potentially can benefit from our disparity optimization, which should lead to a more natural motion appearance.

4. Our Approach

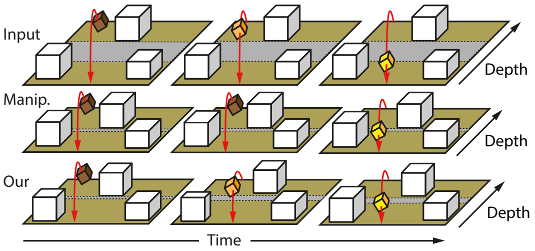

Figure 2: *Starting from an input stereo content* (Top row) *with motion* (red arrow)*, here shown in different frames* (Columns) *a typical manipulation is disparity compression to achieve viewing comfort* (Second row)*. In this example, the flying cube will be slowed down in the proximity of the other cubes and will tend to jump over the empty space between the cubes (marked in grey), where the manipulated disparity is compressed. Our approach* (Third row) *finds a compromise that allows the manipulation where possible and restores motion in depth.*

Our approach takes the temporal disparity field processed by any arbitrary disparity manipulation method and then restores the possibly altered motion-in-depth represented by the disparity change over time (Fig. 2). It matches it to the disparity change over time in the original temporal disparity field but it preserves the spatial disparity characteristic introduced by the manipulation. A post-process design of our method enables its application to a general disparity manipulation, e. g., disparity retargeting [LHW*10, YLXH13] or disparity compression [DRE*11], without its modification or knowledge of implementation details.

To this end we devise a cost function for a potential mapping (Sec. 4.1) that is minimized, leading to a mapping that preserves disparity kinematics (Sec. 4.3). In our optimization we perform a perceptual scaling of disparity velocity changes to better account for the actually perceived changes of object

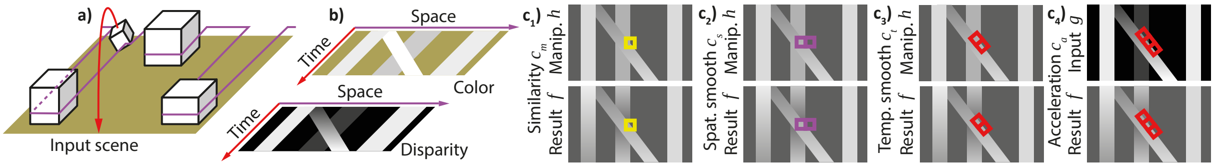

Figure 3: *The cost function computation:* (a) *Input scene as in Fig. 2, where a small cube is moving along a* red trajectory. *Slicing the scene* (violet line) *and adding time as the second dimension results in a temporal light field-like RGB field for two eyes* (b)-top *and a temporal disparity field where brightness depicts depth* (b)-bottom *of the input scene. Our optimization* (c) *finds the time-varying disparity image f such that it preserves the manipulated stereo content in h and the input stereo motion in g. The similarity cost $c_m(f)$ matches disparity at the same space-time positions* (c_1: yellow box) *in f and h. The smoothness costs $c_s(f)$ and $c_t(f)$ match the disparity difference at the same positions in space* (c_2: violet box) *and time* (c_3: red box) *in f and h. Note, how temporal smoothness is aligned with the motion flow. The acceleration cost $c_a(f)$ matches the second order central disparity difference at the same positions* (c_4: red box) *in f and g.*

velocity (Sec. 4.2). Spatio-temporal subsampling (Sec. 4.4) and GPU processing (Sec. 4.5) are required to achieve the real-time performance of our optimization solver. Finally, we use a novel 3D warping approach (Sec. 4.6) to synthesize a new stereo image pair that conforms to the optimized disparity map.

4.1. Cost function

Let $\Omega = \mathbb{R}^2 \times \mathbb{R}^+$ be the space-time domain and $\mathcal{S} = \Omega \to \mathbb{R}$ be the set of all time-varying disparity images defined on it. Disparity in this work refers to vergence angles or pixel disparity measured in arc minutes, which requires a known observer-to-screen distance and screen size. Now, consider $g \in \mathcal{S}$ as well as $h \in \mathcal{S}$, which denote an original and a manipulated time-varying disparity image. A typical change from g to h could be disparity retargeting [LHW*10, YLXH13] or disparity compression [DRE*11]. Our approach finds a third time-varying disparity image $f \in \mathcal{S}$ that optimally combines manipulation and stereo motion preservation with respect to certain costs as shown in Fig. 3.

Our cost function is designed to balance the following four factors. We want the optimized results f to remain similar to the manipulated stereo content h, and the disparity changes introduced, to be smooth in a local space-time neighborhood. At the same time we want to preserve the velocity of the original content g, and strongly penalize any acceleration changes. To this end we have to change the manipulated input once more. Doing this, we need to ensure that performed changes are spatially and temporally coherent. To this end we use four cost functions to be defined now.

First, the optimized time-varying disparity image f should be similar to the manipulated one h (Fig. 3c_1)

$$c_m(f) = \int_\Omega (f(\mathbf{x},t) - h(\mathbf{x},t))^2 .$$

where t denotes time and $\mathbf{x}$ 2D position in the screen space. In what follows, the bold notation is used for vectors.

Second, the change between the disparity images f and h should be spatially smooth (Fig. 3c_2). In order to allow changes in compression manipulation independently for individual moving objects, the spatial smoothness term is weighted non-uniformly based on the inverted local spatial gradient magnitude of the manipulated disparity:

$$c_s(f) = \int_\Omega \left\| \begin{bmatrix} a_x(\mathbf{x},t) & 0 \\ 0 & a_y(\mathbf{x},t) \end{bmatrix} \cdot \left[\nabla_{\mathbf{x}} \big(f(\mathbf{x},t) - h(\mathbf{x},t) \big) \right]^T \right\|^2 ,$$

where $\nabla_{\mathbf{x}}$ is the gradient with respect to $\mathbf{x}$, and $a_x(\mathbf{x},t)$, $a_y(\mathbf{x},t)$ are power functions ensuring that the cost function is little affected on surfaces while smoothing at object silhouettes is strongly suppressed as in [FFLS08]:

$$a_x(\mathbf{x},t) = e^{-10|\frac{\partial h(\mathbf{x},t)}{\partial x}|} \qquad \text{and} \qquad a_y(\mathbf{x},t) = e^{-10|\frac{\partial h(\mathbf{x},t)}{\partial y}|} .$$

We observed that by setting the exponent to -10 any visible degradation of sharp disparity transitions on boundaries of objects is prevented, while sufficient freedom to disparity changes is provided otherwise.

Third, an introduced additional modification of disparity should be smooth along the object motion (Fig. 3c_3):

$$p_f(\mathbf{x},t) = p\big(\nabla f(\mathbf{x},t) \cdot \mathbf{u}(\mathbf{x},t)\big)$$

$$p_h(\mathbf{x},t) = p\big(\nabla h(\mathbf{x},t) \cdot \mathbf{u}(\mathbf{x},t)\big)$$

$$c_t(f) = \int_\Omega \Big(\underbrace{p_f(\mathbf{x},t)}_{\text{Optim. Vel.}} - \underbrace{p_h(\mathbf{x},t)}_{\text{Manip. Vel.}} \Big)^2 ,$$

where $\mathbf{u}(\mathbf{x},t)$ is the 3D screen space-time normalized motion vector at position $\mathbf{x}$ at time t, ∇ the gradient with respect to $\mathbf{x}$ and t, $\cdot$ the 3D dot product, and $p \in \mathbb{R} \to \mathbb{R}$ is a function that maps physical disparity velocity to perceptual units (Sec. 4.2).

Finally – and most different from previous work on energy-based image or disparity manipulation – the original acceleration in g should be preserved (Fig. 3c_4). We therefore construct the acceleration term to match the second derivative of manipulated time-varying disparity image f to the second derivative of original disparity g along the motion path $\mathbf{u}(\mathbf{x},t)$:

$$p_{\mathrm{g}}(\mathbf{x},t) = p\big(\nabla g(\mathbf{x},t)\cdot\mathbf{u}(\mathbf{x},t)\big)$$

$$c_{\mathrm{a}}(f) = \int_{\Omega}\left|\left|\underbrace{\nabla\big(p_{\mathrm{f}}(\mathbf{x},t)\big)}_{\text{Optim. Acc.}} - \underbrace{\nabla\big(p_{\mathrm{g}}(\mathbf{x},t)\big)}_{\text{Orig. Acc.}}\right|\right|^2.$$

Other methods for modification of image sequences often rely on per-frame temporal smoothness enforced by minimization of first temporal derivative [LHW*10, DRE*11, YLXH13, KLHG09, WLSL10]. That reduces the change of optimized image property, disparity in our case, over time so it becomes as constant as possible. Such approaches have two key issues. First, they often operate on per-frame basis which might result in limited temporal extent of the optimization depending on the solver used. We instead use sparse samples in the temporal domain to capture the global temporal characteristic of the motion-in-depth in the range of several seconds. Second, first-order smoothing also removes high frequencies in the temporal disparity signal, turning every original motion into a smooth motion. Instead we use the second temporal derivative (acceleration) of the disparity and match the manipulated time-varying disparity image f to the original acceleration in g. This reintroduce the original motion characteristic independent of its scale that could have been both locally and globally altered by changes in disparity range and distribution. Depth-map guided temporal upsampling later guarantees that we recover high temporal frequencies of disparity lost by sparse sampling (Sec. 4.4). Our method is therefore equivalent to the simple temporal smoothing only in the case of constant speed motion.

Our smoothness term $c_{\mathrm{t}}(f)$ only states that the optimization to restore the motion should respect the manipulated disparity image h and should not alter it rapidly. Therefore it does not contradict the acceleration term $c_{\mathrm{a}}(f)$.

4.2. Perceived Disparity Velocity Changes

We want to model the HVS sensitivity to a change of disparity over time which, as we discussed in Sec. 2, contributes to the perceived velocity of MID through the CDOT mechanism. Such a model should ensure that perceptually important motion characteristics are well-reproduced and otherwise the optimization can safely ignore imperceptible motion distortions.

We assume that the sensitivity follows the Weber-Fechner law and based on the measurement in [PYR96, PYR97] we conservatively set the Weber fraction $k = 0.08$ irrespectively of the motion direction. Let $\dot{\alpha} = d\alpha/dt$ be the change of disparity over time. Then the perceived disparity velocity of $\dot{\alpha}$ is $p(\dot{\alpha}) = \frac{1}{k}(\ln(\dot{\alpha}+1) - \ln(\varepsilon+1))$, where ε is the smallest disparity velocity that can be detected. To compare two disparity velocities $\dot{\alpha}$ and $\dot{\beta}$, using $\Delta p = p(\dot{\alpha}) - p(\dot{\beta}) = \frac{1}{k}(\ln(\dot{\alpha}+1) - \ln(\dot{\beta}+1))$, it is not required to know ε. Effectively, Δp is scaled in sensory just noticeable difference units (JND), which also means that the velocity differences $\dot{\alpha} - \dot{\beta}$ for which $\Delta p < 1$ JND are not perceivable. Since $k = 0.08$, one needs to change the disparity velocity $\dot{\alpha}$ by at least 8 % to discriminate any difference.

The model can be directly applied to the temporal smoothness term $c_{\mathrm{t}}(f)$, where the perceived disparity velocity change is computed for the time-varying disparity images f and h. This is also the case for the acceleration term $c_{\mathrm{a}}(f)$, where the physical disparity velocity is converted into the sensory response units $p(\dot{\alpha})$, prior to the disparity acceleration computation (in the discrete formulation the acceleration is approximated by the second order central differences based on Δp).

4.3. Minimization

Finding the best disparity $f = \arg\min_{\hat{f}} c(\hat{f})$ with

$$c(f) = w_{\mathrm{m}}\,c_{\mathrm{m}}(f) + w_{\mathrm{s}}\,c_{\mathrm{s}}(f) + w_{\mathrm{t}}\,c_{\mathrm{t}}(f) + w_{\mathrm{a}}\,c_{\mathrm{a}}(f)$$

is a constrained (to the target disparity range) optimization problem. The values $w_{\mathrm{m}} = 0.05$ for the data, $w_{\mathrm{s}} = 1.0$ for the spatial, $w_{\mathrm{t}} = 0.1$ for the temporal, and $w_{\mathrm{a}} = 0.1$ for the acceleration weighting are used in all our results. In the following, we will discretize and linearize the problem, before solving it numerically.

Discretization The solution space is discretized into n_{s} spatial and n_{t} temporal elements, which altogether requires $n = n_{\mathrm{s}}n_{\mathrm{t}}$ new disparity values to be found. Thereby the solution is a real vector $\mathbf{f} \in \mathbb{R}^n$. Let $\mathbf{g} \in \mathbb{R}^n$ and $\mathbf{h} \in \mathbb{R}^n$ be discrete versions of g, the original and h, the manipulated time-varying disparity images.

Optimization The first two costs can be written as a discrete differential. The data term cost is $||\mathbf{f}-\mathbf{h}||^2$. The spatial smoothness cost is $||\mathsf{A}_{\mathrm{n}}(\mathbf{f}-\mathbf{h})||^2$, where the matrix A_{n} in row i is 4 in column i, -1 at all four elements with index j that are a spatial neighbor to i and zero otherwise. The two other costs are non-linear due to the perceptual model. Let A_{f} and $\mathsf{A}_{\mathrm{b}} \in \mathbb{R}^{n\times n}$ denote discrete versions of the forward and backward motion flow u, respectively. This motion flow permutation matrix encodes in row i and column j, how much the i-th space-time pixel is a result of forward or backward motion flow of space-time pixel j. The temporal smoothness is

$$||p(\mathsf{A}_{\mathrm{f}}\mathbf{f}-\mathbf{f}) - p(\mathsf{A}_{\mathrm{f}}\mathbf{h}-\mathbf{h})||^2,$$

where $p(\dot{\alpha})$ is applied element-wise. Finally, the acceleration cost is

$$\big|\big|\big(p(\mathsf{A}_{\mathrm{f}}\mathbf{f}-\mathbf{f}) - p(\mathbf{f}-\mathsf{A}_{\mathrm{b}}\mathbf{f})\big) - \big(p(\mathsf{A}_{\mathrm{f}}\mathbf{g}-\mathbf{g}) - p(\mathbf{g}-\mathsf{A}_{\mathrm{b}}\mathbf{g})\big)\big|\big|^2.$$

4.4. Upsampling

Solving the above minimization problem at the full space-time resolution of common stereo content can consume an intractable amount of time and memory. However, we find that the *coarse-to-fine* optimization works well in a subsampled space-time domain, followed by on-the-fly upsampling to the original resolution (Fig. 4). To capture all properties of motion, temporal sampling frequency must be high enough, so that a majority of points on any surface are visible at least in three consecutive frames. In all video sequences considered in this paper we used a uniform subsampling of 1 : 10 in time and 1 : 5 in space. Only the spatio-temporal change of disparity is upsampled and applied to the current frame, to keep fine details. Two different strategies address the upsampling in time and space.

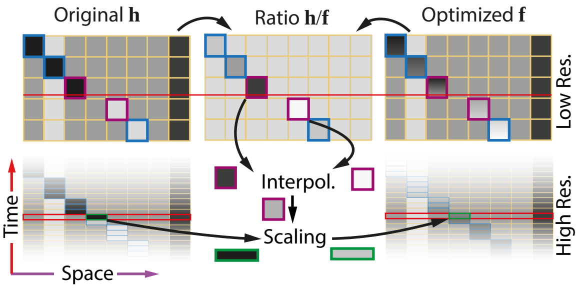

Figure 4: *Temporal upsampling of a spatio-temporal disparity field. An object is moving in both image and depth space* (blue). *The ratio of the temporarily nearest original and optimized disparities* **h** *and* **f** *is interpolated with respect to the motion flow to get the disparity scaling at a given high-resolution frame time* (red). *Scaling is applied to the high-resolution disparity* (green) *to get the high-resolution output disparity.*

Time In time, dense motion flow to the position in the previous keyframe and to the position in the subsequent keyframe are used. We detect occlusion in motion flow by comparing the depth value of the current pixel and the corresponding pixel along the motion flow. If those values differ significantly or the flow points outside the image, we consider the flow to be a disocclusion or occlusion and ignore the corresponding depth value in the interpolation. Instead of interpolating disparity in time, we interpolate disparity gradients and add them to the current high-resolution original frame. This preserves motion changes beyond the temporal sampling frequency of the optimization discretization.

Space In space, joint bilateral upsampling [KCLU07] with the current high-resolution disparity image as the guidance is applied to the output of the temporal upsampling. Again, we interpolate gradients that are applied to the high-resolution frame. This allows for spatial details, finer than the spatial sampling frequency of the discretization of the solver.

4.5. Implementation

Optimization We use a gradient descent to find the best time-varying disparity image. The system cannot be optimized in closed form, due to the perceptual non-linearities and the boundary conditions of positive disparity. Starting from $\mathbf{f}^{(0)} = \mathbf{h}$, in every step i a correction vector $\mathbf{f}'^{(i)}$ is constructed from derivation of all costs and the solution is updated $\mathbf{f}^{(i+1)} = \mathbf{f}^{(i)} - \lambda \mathbf{f}'^{(i)}$. A $\lambda = 0.5$ and approximately 8 iterations were found to be sufficient for convergence to the solution for animations we tested. Adding more iterations and/or usage of smaller λ did not introduce any visible difference in final image sequence nor individual frames when compared visually.

GPU implementation The solver is implemented on a GPU and performs the updates of the mapping function in realtime. We maintain a window of previous frames combined with a prediction of future frames. The solver uses the last solution as the new initial guess. The deformation field is stored into a read-only 3D GPU buffer and each solver iteration is parallelized over all elements. If we consider the resolution of the subsampled space as a constant chosen based on content structure details rather than screen resolution the optimization runs in constant time independent of the target resolution. Please note, that a constant iteration count worked well in our experiments. Both spatial and temporal upsampling require processing linear in the number of output frame pixels. This is computationally equivalent to an application of a simple post-process filter, e. g., motion blur, which is a technique commonly used in rea-ltime applications.

4.6. 3D Warping

Both, the manipulated and the optimized disparity maps do not necessarily correspond to any single pinhole camera projection. Therefore, such disparity patterns cannot be directly produced by conventional ray-tracing or rasterization. Instead, we have to modify the image locally. Typically the scene is rendered from a monocular center point of view and then image warping is performed [LHW*10, DRE*10]. Occlusions are resolved using a depth map, if available. However, disocclusions might still appear if some originally occluded region becomes visible, resulting in typical artifacts (Fig. 5).

To overcome this problem, we propose to render geometry twice. The first pass creates only a linear depth map that is manipulated and optimized in the pipeline described above. The second pass produces a color image pair, but moves vertices in the rasterization such that the resulting disparity conforms to the desired disparity. In order to allow for fine disparity mapping regardless of 3D geometry quality, we utilize tessellation on modern GPUs and adaptively subdivide triangles in object space up to the size of one pixel when projected on the screen. Larger thresholds can be used to favor performance. The rest of the rendering pipeline remains unchanged. We use a cascade of three increasingly approximate ways to reconstruct the appropriate vertex motion: *Direct fetching* followed by *neighbor fetching*, and finally *global disparity curves*.

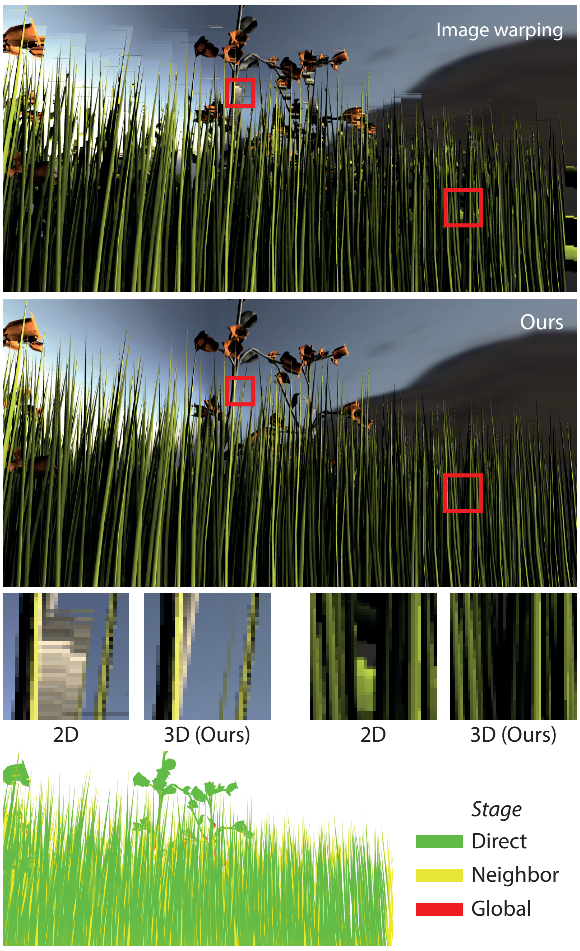

Figure 5: *Common image warping and our approach applied to the left image of a stereoscopic image pair. The insets show close-up views of typical artifacts in common image warping. Color encoding in the bottom image illustrates, which of the three proposed vertex warping methods has been applied for a given image region.*

Direct fetching In almost all cases, direct fetching is sufficient: Let $\mathbf{v}_\mathrm{p}$ be the pixel coordinates of the projection of a vertex $\mathbf{v}$. We read the depth map at $\mathbf{v}_\mathrm{p}$ and check if the difference between the depth of $\mathbf{v}$ and the depth map value is smaller than a threshold ε. If this is the case, we read the disparity map and move the vertex to achieve the desired disparity between the left and right frames. This approach fails in the presence of occlusions or disocclusions, as the disparity map at $\mathbf{v}_\mathrm{p}$ does not contain the disparity that is to be assigned to $\mathbf{v}$.

Neighborhood stage If direct fetching fails, we first find the nearest depth in a 3 × 3 pixel-neighborhood of $\mathbf{v}_\mathrm{p}$ to increase robustness on object edges. If the minimal depth difference is smaller than ε we use the disparity value for the same position in the disparity map. Otherwise, we assume that the vertex was occluded in the original rendering. It might, however, not be occluded after disparity optimization or manipulation. Therefore we need to resolve what disparity it would have had if it was not occluded. We do that by searching the neighboring non-occluded depth map pixels for similar depth values. We assume that spatially close objects with similar depth are likely to have similar disparity.

We use 64 samples from the 2D Halton sequence to generate polar coordinates for sampling in a wider neighborhood of each vertex. The Halton sequence makes the sampling more robust to aliasing compared to regular sampling. The i-th sample position $\mathbf{s}_i$ is

$$\mathbf{s}_i = \mathbf{v}_\mathrm{p} + r_i^2(\cos(2\pi\alpha_i), \sin(2\pi\alpha_i)),$$

where $(r_i, \alpha_i) \in (0,1)^2$ is the i-th element of the 2D Halton sequence. We use square of radius in order to sample the close neighborhood more densely. Once again, the depth difference smaller than ε indicates equality.

If there was no suitable value in the neighborhood either, it is still unclear where to move the vertex $\mathbf{v}$. As a last resort, we revert to a global disparity curve as explained next.

Global curve stage If the depth map sampling failed, there either is no visible object with similar depth in the disparity map or we failed to find it. In that case we cannot recover the correct disparity for the vertex but we try to minimize the error that would be observed as a rendering artifact. To this end, we reconstruct an approximation of the global curve mapping depth to disparity in a pre-process before the 3D warping. The mapping is constructed using radial basis functions with bandwidth prediction. In a first pass we predict bandwidth i. e., how many different disparity values map to a certain neighborhood of depth values. In a second pass, we reconstruct the mapping from depth to disparity with an adapted bandwidth.

In the first pass we build a standard histogram of the depth map. Populated bins will require a higher bandwidth, i. e., smaller kernel for reconstruction, less populated bins need a wider kernel. In particular, bins can be empty. In the second pass, we iterate over all depth-disparity value pairs. We use information about empty bins from the first pass to set the support of a hat reconstruction kernel of disparity values to the left and right neighborhood in the final histogram. The support matches to the distance to the next non-empty bin in a given direction. This way one depth-disparity pair may influence more than one histogram bin and therefore fill missing mapping intervals but does not cause blurring in other parts. Hat function filtering provides linear interpolation for empty intervals of mapping curve. We consider linear interpolation for these regions of depth range to be a conservative choice.

Discussion Conventional image warping [LHW*10, DRE*10] cannot deal with disocclusions. An alternative would be layered depth images (LDI) [SGHS98], that do contain all intersections of a viewing ray per pixel, not only the first. Similar to all image-based warping techniques, LDIs are prone to the undersampling problem, which might degrade the quality of synthesized images, in particular, for surfaces that originally have been seen under grazing angles. We avoid discretization altogether by first warping and discretizing later. Kim et al. [KHH*11] suggest rendering from multiple perspectives to generate a 3D lightfield. Non-physical views can then be created as slices through this field. That, however, involves rendering of many data that will not be used. Our approach instead modifies the rasterization phase itself so that it only produces the final image with desired disparity. Our method only modifies the vertex projection phase of rendering and therefore is easily applicable wherever deferred shading [DWS*88] is used. We achieve that using vertex based warping directly on GPU. In our method, all disocclusions are resolved before the rasterization is performed, therefore, we do not lose any image information.

The resulting rendering might still produce artifacts if the global mapping curve does not match the local disparity mapping. We expect that the disparity manipulation roughly preserves some key properties such as depth ordering and therefore a global curve is a reasonable estimation of disparity at a given depth. While one could consider a more localized reconstruction of the mapping curve, in our scenes, we have not experienced any problems with the global curve approach and we observed a significant reduction of rendering artifacts due to the elimination of disocclusions (Fig. 5).

The rendering is done on per frame basis and the approach is therefore independent on temporal optimization. This makes it applicable to any other stereo content rendering problem.

Per-vertex warping makes the rendering performance highly dependent on the number of vertices in the scene. As the warping happens before culling, even vertices behind the camera will be processed. For real time applications an extension predicting what can be visible after the warping using simplified geometry could be implemented. This would prevent disparity sampling for occluded vertices which is the main performance issue. Another improvement would utilize temporal coherence to decrease the number of disparity map samples per vertex.

5. Results

We performed two perceptual studies in order to evaluate the visual quality of object motion and to measure the performance in hit point-prediction for a ballistic target. Please refer to Fig. 6 and the supplemental video for the stimuli.

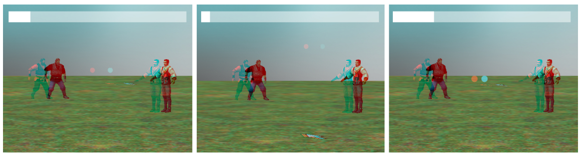

Figure 6: *Three example trials from our performance study.*

Setup In our experiments 10 and 8 different observers naïve in regard to the purpose of the experiment and with normal or corrected-to-normal (stereo) vision were observing a Zalman ZM-M240W polarized stereo display from a distance of 80 cm.

Preference study In the first experiment we displayed stereo video with compressed disparity (similar to the frame-by-frame application of global operator of Lang et al. [LHW*10] without temporal smoothing) and the same video with our additional stereo-motion optimization side-by-side. Three computer generated scenes with motion in depth can be seen in Figs. 1 and 8. The saliency used to guide the compression was given either to the moving or the static object to simulate the artist's intention or to emphasize an important object based on the scene's semantics. The compression was set to be larger than what the display would require, in order to make the stimulus comparable to a reference video with non-manipulated disparity shown in the middle. Subjects were asked to indicate which of the two test sequences is more similar to the reference in terms of object motion. Our solution has been strongly preferred in 85.6% of the cases ($p < 0.01$, binomial test).

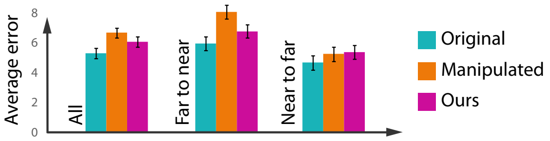

Figure 7: *Motion direction errors with 95% confidence intervals.*

Performance study In the second experiment the performance of hit point-prediction in terms of precision for a flying ball has been investigated (Fig. 6). We considered straight-line and ballistic motion trajectories of a ball moving in a random direction. The ball was shown only for a short initial interval after which observers used the mouse cursor to indicate its hit point on a ground plane. The closest world-space distance between the correct hit point and a ray through the clicked pixel was recorded. The stereo content was either unmodified, compressed, or compressed and processed using our approach. Fig. 7 summarizes the experiment outcome. An analysis of variance (ANOVA) revealed a statistically significant effect ($F(2,333) = 3.99, p < 0.02$) of reduced performance for the compressed disparity with respect to the unmodified one. When directions were analyzed independently there was the same effect for the far-to-near direction ($F(2,165) = 5.57, p < 0.01$), but not for the near-to-far direction ($F(2,165) = 0.64, p = 0.53$). We believe that impact of distortion is smaller in the near-to-far direction due to perspective scaling and smaller disparity change due to manipulation. This results in overall lower magnitude of error and smaller error differences between the original and manipulated disparity scenarios. As can be seen in Fig. 7, overall our method performed worse than unmodified disparity and better than the modified one, but in both cases we could not prove the significance of these effects.

6. Conclusion

This paper introduced an approach to retarget disparity such that stereo motion can be reproduced faithfully. The problem was recast into a time-space-deformation problem that was solved using a numeric optimization procedure that allows for real-time performance. Our solution is independent of the particular stereo manipulation performed which makes it general. For the case of optimization-based manipulation, our perceptual disparity motion terms can be included in a combined optimization. Our perceptual study demonstrates that our disparity retargeting is strongly preferred over disparity manipulations that do not explicitly optimize for faithful motion in depth. The performance study clearly indicates that any disparity manipulation requires special attention in tasks that involve visual tracking of moving objects and precise judgment upon their possible collisions. Finally, we described a novel 3D warping approach to synthesize stereo image pairs that conform to a manipulated disparity map from polygonal 3D scenes. Application of this 3D warping is not limited to disparity maps produced by our system but is applicable to other manipulations as well.

Our approach is subject to several limitations. The perception of stereo motion might be affected by other factors, which have not been considered in this work, such as tracking and verging on the particular moving object, the object's luminance, texture, as well as its possible deformations. We relegate as future work a more in-depth investigation of those issues. Our current experiments were limited to computer-generated animations. Future work will need to show how approximations in the optical flow and scene depth reconstruction can affect our techniques.

References

[BS04] Brooks K. R., Stone L. S.: Stereomotion speed perception: Contributions from both changing disparity and interocular velocity difference over a range of relative disparities. *J. Vis. 4*, 12 (2004). 2

[BvdBvD86] Brenner E., van den Berg A., van Damme W.: Perceived motion in depth. *Vis. Res. 36* (1986), 699–706. 2

[CL88] Cavallo V., Laurent M.: Visual information and skill level in time-to-collision estimation. *Perception 17*, 5 (1988), 623–32. 2

[DRE*10] Didyk P., Ritschel T., Eisemann E., Myszkowski K., Seidel H.: Adaptive image-space stereo view synthesis. In *Proc. VMV* (2010), pp. 299–306. 6, 8

[DRE*11] Didyk P., Ritschel T., Eisemann E., Myszkowski K., Seidel H.: A perceptual model for disparity. *ACM Trans. Graph. (Proc. SIGGRAPH) 30*, 4 (2011), 96:1–96:10. 3, 4, 5

[DWS*88] Deering M., Winner S., Schediwy B., Duffy C., Hunt N.: The triangle processor and normal vector shader: a vlsi system for high performance graphics. In *Proc. of ACM SIGGRAPH* (1988), pp. 21–30. 8

[EC85] Erkelens C., Collewijn H.: Motion perception during dichoptic viewing of moving random-dot stereograms. *Vis. Res. 25*, 4 (1985), 583–588. 2

[FFLS08] Farbman Z., Fattal R., Lischinski D., Szeliski R.: Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Trans. Graph. (Proc. SIGGRAPH) 27*, 3 (2008), 67:1–67:10. 4

[GR98] Gray R., Regan D.: Accuracy of estimating time to collision using binocular and monocular information. *Vis. Res. 38*, 4 (1998), 499–512. 2

[Heu87] Heuer H.: Apparent motion in depth resulting from changing size and changing vergence. *Perception 16*, 3 (1987), 337–50. 2

[HGG*11] Heinzle S., Greisen P., Gallup D., Chen C., Saner D., Smolic A., Burg A., Matusik W., Gross M.: Computational stereo camera system with programmable control loop. *ACM Trans. Graph. (Proc. SIGGRAPH) 30*, 4 (2011), 94:1–94:10. 3

[HKB11] Hoffman D. M., Karasev V. I., Banks M. S.: Temporal presentation protocols in stereoscopic displays: Flicker visibility, perceived motion, and perceived depth. *J Soc Inf Disp 19*, 3 (2011), 271–297. 3

[HNG08] Harris J. M., Nefs H. T., Grafton C. E.: Binocular vision and motion-in-depth. *Spatial Vision 21*, 6 (2008), 531–547. 2

[HW95] Harris J. M., Watamaniuk S. N.: Speed discrimination of motion-in-depth using binocular cues. *Vis. Res. 35*, 7 (1995), 885–896. 2

[HWSB99] Hubona G. S., Wheeler P. N., Shirah G. W., Brandt M.: The relative contributions of stereo, lighting, and background scenes in promoting 3d depth visualization. *ACM Trans. Comput.-Hum. Interact. 6*, 3 (1999), 214–242. 3

[JLHE01] Jones G., Lee D., Holliman N., Ezra D.: Controlling perceived depth in stereoscopic images. In *Proc. SPIE* (2001), vol. 4297, pp. 42–53. 3

[KCLU07] Kopf J., Cohen M., Lischinski D., Uyttendaele M.: Joint bilateral upsampling. *ACM Trans. Graph. (Proc. SIGGRAPH) 26*, 3 (2007), 96:1–96:6. 6

[KHH*11] Kim C., Hornung A., Heinzle S., Matusik W., Gross M.: Multi-perspective stereoscopy from light fields. *ACM Trans. Graph. 30*, 6 (2011), 190:1–190:10. 8

[KLHG09] Krähenbühl P., Lang M., Hornung A., Gross M.: A system for retargeting of streaming video. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 28*, 5 (2009), 126:1–126:10. 3, 5

[KSL12] Kulshreshth A., Schild J., LaViola Jr. J. J.: Evaluating user performance in 3D stereo and motion enabled video games. In *Foundations of Digital Games* (2012), pp. 33–40. 3

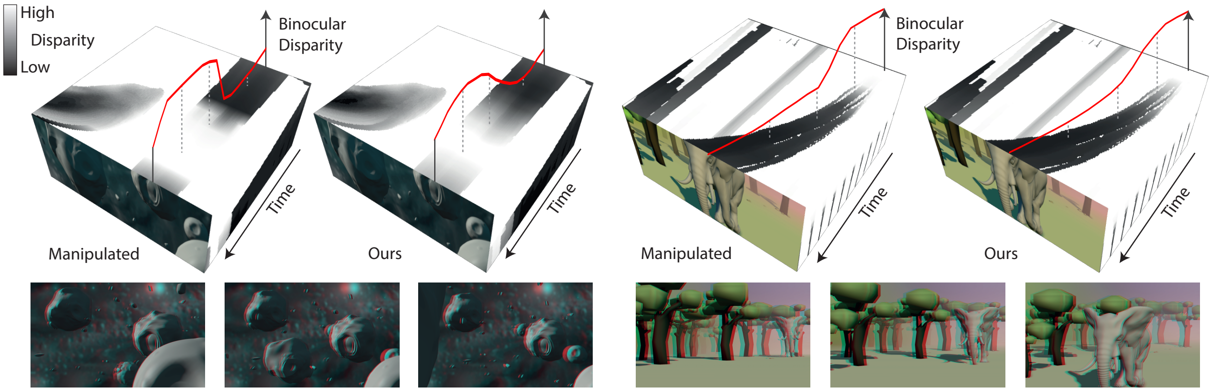

Figure 8: *Two rendered scenes with complex deformation and camera motion used as stimuli in our study.* Top: *Slices through spatio-temporal disparity cubes after manipulation and our optimization. Time of animation proceeds from the back to front side of the cube. Plot shows values of disparity in time for single object in the scene as it moves in 3D space. Plot values were sampled from the cube, scaled and projected to match the cube orientation.* Bottom: *Three frames from each scene.*

[KZC*11] Koppal S. J., Zitnick C. L., Cohen M. F., Kang S. B., Ressler B., Colburn A.: A viewer-centric editor for 3D movies. *IEEE Comp. Graph. and Appl. 31*, 1 (2011), 20–35. 3

[LHW*10] Lang M., Hornung A., Wang O., Poulakos S., Smolic A., Gross M.: Nonlinear disparity mapping for stereoscopic 3D. *ACM Trans. Graph. (Proc. SIGGRAPH) 29*, 4 (2010), 75. 3, 4, 5, 6, 8

[LIFH09] Lambooij M., IJsselsteijn W., Fortuin M., Heynderickx I.: Visual discomfort and visual fatigue of stereoscopic displays: A review. *J. Imaging Science and Technology 53*, 3 (2009), 1–12. 3

[MA09] Matsumiya K., Ando H.: World-centered perception of 3d object motion during visually guided self-motion. *J Vis. 9*, 1 (2009). 3

[Men09] Mendiburu B.: *3D movie making: stereoscopic digital cinema from script to screen.* Focal Press, 2009. 3

[NBPC05] Neinborg H., Bridge H., Parker A., Cumming B.: Neuronal computation of disparity in V1 limits temporal resolution for detecting disparity modulation. *J. Neurosci 25* (2005), 10207–19. 2

[OHB*11] Oskam T., Hornung A., Bowles H., Mitchell K., Gross M.: OSCAM-optimized stereoscopic camera control for interactive 3D. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 30*, 6 (2011), 189:1–189:8. 3

[PYR96] Portfors-Yeomans C., Regan D.: Cyclopean discrimination thresholds for the direction and speed of motion in depth. *Vis. Res. 36*, 20 (1996), 3265–3279. 2, 5

[PYR97] Portfors-Yeomans C., Regan D.: Just-noticeable difference in the speed of cyclopean motion in depth and the speed of cyclopean motion within a frontoparallel plane. *J Exp. Psych.: Human Perception and Performance 23*, 4 (1997), 1074–1086. 5

[RB79] Regan D., Beverley K.: Binocular and monocular stimuli for motion in depth: Changing-disparity and changing-size feed the same motion-in-depth stage. *Vis. Res. 19*, 12 (1979), 1331–1342. 2

[Ric72] Richards W.: Response functions for sine-and square-wave modulations of disparity. *J OSA 62*, 7 (1972), 907–911. 2

[SGHS98] Shade J., Gortler S., He L.-w., Szeliski R.: Layered depth images. In *Proc. of ACM SIGGRAPH* (1998), pp. 231–242. 8

[SKHB11] Shibata T., Kim J., Hoffman D., Banks M.: The zone of comfort: Predicting visual discomfort with stereo displays. *J Vision 11*, 8 (2011), 11:1–11:29. 3

[STRH06] Speranza F., Tam W. J., Renaud R., Hur N.: Effect of disparity and motion on visual comfort of stereoscopic images. In *SPIE* (2006), vol. 6055, pp. 94–103. 3

[Tyl71] Tyler C.: Stereoscopic depth movement: Two eyes less sensitive than one. *Science 174* (1971), 958–961. 2

[WLSL10] Wang Y.-S., Lin H.-C., Sorkine O., Lee T.-Y.: Motion-based video retargeting with optimized crop-and-warp. *ACM Trans. Graph. 29* (2010), 90:1–90:9. 3, 5

[YIMT02] Yano S., Ide S., Mitsuhashi T., Thwaites H.: A study of visual fatigue and visual comfort for 3D HDTV/HDTV images. *Displays 23*, 4 (2002), 191 – 201. 3

[YLXH13] Yan T., Lau R., Xu Y., Huang L.: Depth mapping for stereoscopic videos. *International Journal of Computer Vision 102* (2013), 293–307. 3, 4, 5