What makes 2D-to-3D stereo conversion perceptually plausible?

Petr Kellnhofer Thomas Leimkühler Tobias Ritschel Karol Myszkowski Hans-Peter Seidel
MPI Informatik

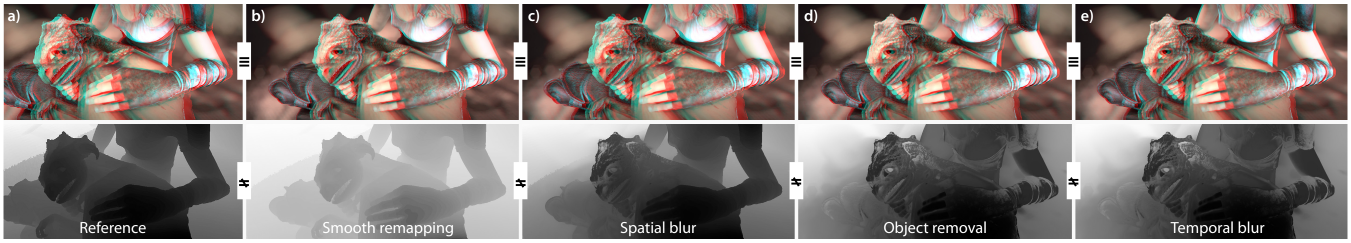

Figure 1: *We intentionally introduce the depth distortions typically produced by 2D-to-3D conversion into close-to-natural computer generated images* (a, top) *such as the one from the MPI Sintel dataset [Butler et al. 2012] where ground truth depth is available* (a, bottom). *User response to stereo images* (b–e, top) *showing typical disparity distortions* (b–e, bottom) *gives an indication whether a certain amount of distortion results in functional equivalence for natural images or not. According to numerical measures such as PSNR or perceptual disparity metrics, the depth is considered very different* (inequality sign, bottom)*, whereas it is functionally equivalent* (equivalence sign, top).

Abstract

Different from classic reconstruction of physical depth in computer vision, depth for 2D-to-3D stereo conversion is assigned by humans using semi-automatic painting interfaces and, consequently, is often dramatically wrong. Here we seek to better understand why it still does not fail to convey a sensation of depth. To this end, four typical disparity distortions resulting from manual 2D-to-3D stereo conversion are analyzed: i) smooth remapping, ii) spatial smoothness, iii) motion-compensated, temporal smoothness, and iv) completeness. A perceptual experiment is conducted to quantify the impact of each distortion on the plausibility of the 3D impression relative to a reference without distortion. Close-to-natural videos with known depth were distorted in one of the four above-mentioned aspects and subjects had to indicate if the distortion still allows for a plausible 3D effect. The smallest amounts of distortion that result in a significant rejection suggests a conservative upper bound on the quality requirement of 2D-to-3D conversion.

CR Categories: I.3.3 [Computer Graphics]: Three-Dimensional Graphics and Realism—Display Algorithms

Keywords:

1 Introduction

The majority of images and videos available is 2D, and automatic conversion to 3D is a long-standing challenge [Zhang et al. 2011]. The requirements imposed on the precise meaning of “3D” might differ: For applications such as view synthesis, surveillance, autonomous driving, human body tracking, relighting or fabrication, accurate physical depth is mandatory. Obviously, binocular disparity can be computed from such accurate physical depth, allowing for the synthesis of a stereo image pair using image-based rendering. However, it is not clear what depth fidelity is required to produce plausible disparity in natural images, which include other monocular cues.

In this paper we argue that physically accurate depth is not required to produce plausible disparity. Instead, we provide evidence that as long as four main properties of the disparity hold, it is perceived as plausible. First, the absolute scale of disparity is not relevant, and any reasonable smooth remapping [Jones et al. 2001; Lang et al. 2010; Didyk et al. 2012] is perceived equally plausible and may even be preferred in terms of viewing comfort and realism. Therefore, we can equally well use disparity that is the same as the physical one under a smooth remapping. Second, not every detail in the scene can be augmented with plausible depth information, resulting in isolated objects that remain 2D or lack disparity relative to their content. We will see that, unless those objects are large or salient, this defect often remains largely unnoticed. Third, the natural statistics of depth and luminance indicate that depth is typically spatially smooth, except at luminance discontinuities [Yang and Purves 2003; Merkle et al. 2009]. Therefore, not reproducing disparity details can be acceptable and is often not even perceived, except at luminance edges [Kane et al. 2014]. Fourth and finally, the temporal perception of disparity allows for a temporally coarse disparity map, as fine temporal variations of disparity are not perceivable [Howard and Rogers 2012; Kane et al. 2014]. Consequently, as long as the error is 2D-motion compensated [Shinya 1993], depth from one point in time can be used to replace depth at a different, nearby point in time.

2 Previous work

In this section, we review the three main approaches for 2D-to-3D (manual, automatic and real-time), the use of luminance and depth edges in computational stereo, as well as perceptual modeling of binocular and monocular depth cues.

2D-to-3D conversion Manual conversion produces high-quality results but requires human intervention, which can result in substantial cost. They are based on painting depth annotations [Guttmann et al. 2009] with special user interfaces [Ward et al. 2011] and propagation in space and time [Lang et al. 2012]. The semi-supervised method of Assa and Wolf [2007] combines cues extracted from an image with user intervention to create depth parallax.

Automatic conversion does not need manual effort, but does require lengthy computation to produce results of medium quality. The system of Hoiem et al. [2005] infers depth from monocular images by a low number of labels. Make3D [Saxena et al. 2009] is based on learning appearance features to infer depth. Both approaches show good results for static street-level scenes with super-pixel resolution but require substantial computation. Non-parametric approaches rely on a large collection of 3D images [Konrad et al. 2012] or 3D videos [Karsch et al. 2014] that have to contain an exemplar similar to a 2D input. For cel animation with outlines, T-junctions have been shown to provide sufficient information to add approximate depth [Liu et al. 2013].

Real-time methods to produce disparity from 2D input videos usually come at low visual quality. A simple and computationally cheap solution is to time-shift the image sequence independently for each eye, such that a space-shift provides a stereo image pair [Murata et al. 1998]. This requires an estimate of the camera velocity and only works for horizontal motions. For other rigid motions, structure-from-motion (SfM) can directly be used to produce depth maps [Zhang et al. 2007]. SfM makes strong assumptions about the scene content such as a rigid scene with camera motion. Additionally, individual cues such as color [Cheng et al. 2010], motion [Huang et al. 2009] or templates [Yamada and Suzuki 2009] were combined into a disparity estimate in an ad-hoc fashion.

Commercial 2D-to-3D solutions [Zhang et al. 2011] based on custom hardware (e. g., JVC's IF-2D3D1 Stereoscopic Image Processor) and software (e. g., DDD's Tri-Def-Player), reveal little about their used techniques, but anecdotal testing shows the room for improvement [Karsch et al. 2014].

Perception of luminance and depth Since luminance and depth edges often coincide, e. g., at object silhouettes, full-resolution RGB images have been used to guide depth map upsampling both in the spatial [Kopf et al. 2007] and the spatio-temporal [Richardt et al. 2012; Pajak et al. 2014] domain. Analysis of a database with range images for natural scenes reveals that depth maps mostly consist of piecewise smooth patches separated by edges at object boundaries [Yang and Purves 2003]. This property is used in depth compression, where depth edge positions are explicitly encoded, e. g., by using piecewise-constant or linearly-varying depth representations between edges [Merkle et al. 2009]. This in turn leads to significantly better depth-image-based rendering (DIBR) [Fehn 2004] quality compared to what is possible at the same bandwidth of MPEG-style compressed depth, which preserves more depth features at the expense of blurring depth edges.

The spatial disparity sensitivity function determines the minimum disparity magnitude required to detect sinusoidal depth corrugations of various spatial frequencies [Howard and Rogers 2012]. The highest resolvable spatial frequency is about 3–4 cpd (cycles per degree), which is almost 20 times below the cut-off frequencies for luminance contrast [Wandell 1995]. Similar investigations in the temporal domain indicate that the highest sinusoidal disparity modulation that can be resolved is about 6–8 Hz [Howard and Rogers 2012], which is significantly lower than the 70 Hz measured for luminance [Wandell 1995]. As analyzed by Kane et al. [2014], the picture is different for disparity step-edges in space and time, which are important in real-world images. They found that, for step-edge depth discontinuities, observers might still notice blur due to removal of spatial frequencies up to 11 cpd, indicating that while overall disparity can be smoothed significantly, this is not the case for depth discontinuities. They could further show that filtering temporal frequencies higher than 3.6 Hz from a step signal remains mostly unnoticed. Their findings indicate that the temporal disparity signal might be sparsely sampled and even more aggressively low-pass filtered, without causing visible depth differences. In this work, we conduct similar experiments for natural scenes involving monocular cues.

Surprisingly, depth edges appear sharp, even though human ability to resolve them in space and time is low. One explanation for this is that the perceived depth edge location is determined mostly by the position of the corresponding luminance edge [Robinson and MacLeod 2013].

In previous work, perception was taken into account for stereography when disparity is given [Didyk et al. 2012], but it was routinely ignored when inferring disparity from monocular input for 2D-to-3D conversion. Interestingly, depth discontinuities that are not accompanied by luminance edges of sufficient contrast poorly contribute to the depth perception and do not require precise reconstruction in stereo 3D rendering [Didyk et al. 2012].

3 Experiment

In this experiment we would like to find how typical 2D-to-3D stereo conversion distortions affect the plausibility of a stereo image or movie. To this end, we intentionally reduce physical disparity in one of four aspects and collect the users' response.

3.1 Materials

Stimuli Stimuli were distorted variants of a given stereo video content with known, undistorted disparity. We used four video sequences from the MPI Blender Sintel movie dataset [Butler et al. 2012], and the Big Buck Bunny movie by The Blender Foundation, which provide close-to-natural image statistics combined with ground-truth depth and optical flow. Additionally, we used four rendered stereo image sequences with particularly discontinuous motion that are especially susceptible to temporal filtering. Stimuli were presented as videos for temporal distortions and as static frames for spatial distortions to prevent threshold elevation by presence of motion that would underestimate the effect for a theoretical completely static scene. The scenes did not show any prominent specular areas that required special handling [Dabala et al. 2014].

Distortions were performed in linear space with a normalized range of $(0, 1)$. For stereo display, this normalized depth was mapped to vergence angles corresponding to a depth range of $(57, 65)$ cm surrounding a display at 60 cm distance. This distribution around the display plane reduces the vergence-accommodation conflict, however, in some cases a window violation occurred. As the salient content was located at the image center, the influence of this artifact likely was low. Finally depth-image-based rendering [Fehn 2004] was used to convert the monocular image with a distorted depth map into a stereo image pair. Stimuli were subject to exactly one out of five distortions: identity (placebo), remapping, object removal, spatial blur and temporal blur.

Placebo Firstly, the original content without any distortion is used as a control group. This is required to understand how often subjects report a difference when there is none, establishing an upper bound on what to expect if there really are distortions.

Remapping of the linear depth map D was performed for each location $\mathbf{x}$ by means of a power curve $D'(\mathbf{x}) = D(\mathbf{x})^\gamma$ with a γ value of $r_1 \in \{0.9, 0.8, 0.6, 0.3, 0\}$. A value close to 1 indicates no change. Small values indicate a more compressive function. A value close to zero indicates a 2D stimulus with very little global disparity. We have chosen the power function as it is the most basic signal compression method that accounts for weaker abilities of depth discrimination with increasing depth by the human visual

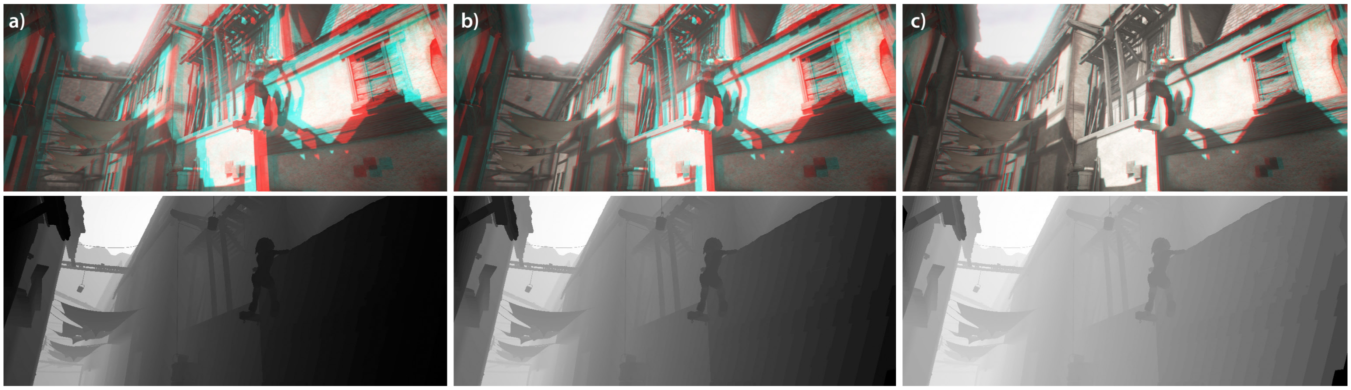

Figure 2: *Remapping. The top row shows the stereo image, the bottom row the disparity. The columns show different amounts of distortion due to remapping.* a) *Original stereo image.* b) *Remapping by a value of $\gamma = 0.8$, leading to equivalence.* c) *Remapping by a value of $\gamma = 0.6$, leading to non-equivalence.*

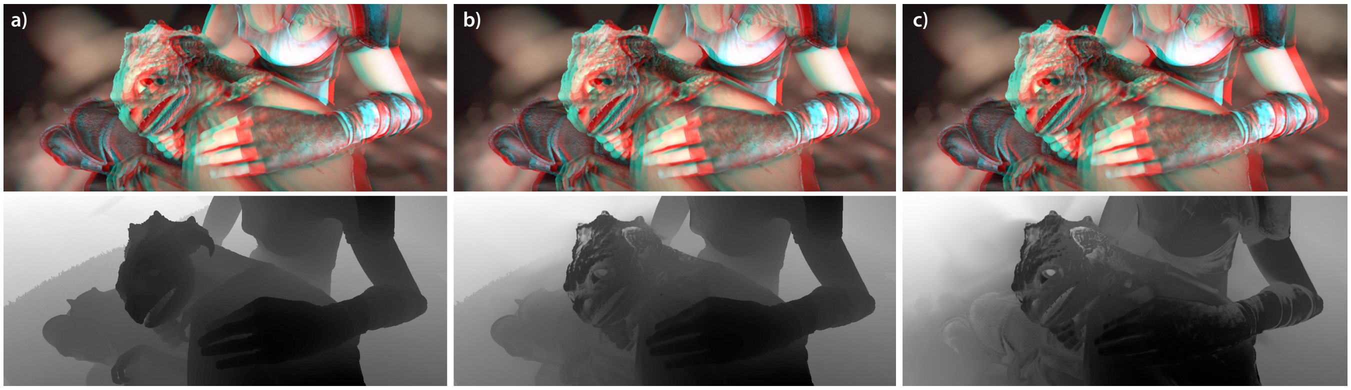

Figure 3: *Object removal. The top row shows the stereo image, the bottom row the disparity. The columns show different amounts of distortion due to object removal.* a) *Original stereo image.* b) *Removal of a circular region of radius $r_2 = 3$ vis. deg. around the character's head, leading to equivalence.* c) *Removal of a region of radius $r_2 = 6$ vis. deg. at the same location, leading to non-equivalence.*

system. It is also inspired by the non-linear operator proposed by Lang et al. [2010]. Example stimuli are shown in Fig. 2.

Removal of entire regions was realized using luminance edge-aware inpainting from surrounding depth that restores structure but not disparity values. The regions removed were randomly positioned circles of radius $r_2 \in \{1, 3, 6, 12\}$ visual degrees. In practice, bilateral filtering with a strongly edge-preserving range radius parameter choice of 0.1 was used. The spatial radius of the filter was set to $0.5 \cdot r_2$ and values in the removed region were weighted by zero in the center of the region and smoothly transitioned to 1 outside the region. This prevented visible discontinuity on the region boundary. The transition was generated by Gaussian blur of the binary mask of the region. Example stimuli are shown in Fig. 3.

Spatial blur was realized using bilateral filtering:

$$D'(\mathbf{x})=\frac{1}{Z(\mathbf{x})}\sum_{x_i\in\Omega}D(\mathbf{x}_i)G(\|\mathbf{x}-\mathbf{x}_i\|,r_{3,1})G(D(\mathbf{x})-D(\mathbf{x}_i),r_{3,2})$$

where $Z(x)$ is the normalizing partition function, Ω the spatial domain of D and $G(d,\sigma)$ is a zero-mean Gaussian distribution function with std. dev. σ chosen as $r_{3,1}\in\{0.25,1,2\}$ visual degrees for spatial range and $r_{3,2}\in\{0.1,0.6,\infty\}$ for the intensity range from 0 to 1. The visual radius of 2 deg corresponds to ca. 80 px in our stimuli. Example stimuli are shown in Fig. 4.

Temporal blurring with a std. dev. of $r_4\in\{0.025,0.25,1\}$ seconds was introduced. The blur was motion-compensated [Shinya 1993], that is, before combining pixels from a different frame, they were moved along their (known) optical flow. This assures, that temporal disparity details are removed for individual objects rather than blending the disparity values of distinct moving objects in a dynamic scene. Example stimuli are shown in Fig. 5.

Subjects 17 participants took part in the experiment (23±4 ys, 8M, 9F). Subjects were naïve with respect to the given task, 4 of them had a background in computer graphics or computer vision. All had corrected or corrected-to-normal vision. None reported any stereo vision deficiency and all were able to identify patterns and digits in test random dot stereograms.

Equipment Stimuli were shown using anaglyph on a DELL U2412M 60 Hz display with spatial resolution of 1920×1200. As the videos are provided at 30 Hz, each frame was displayed twice. The magnitude of crosstalk was not measured. The combination of display, particular glasses and display settings were experimentally chosen so that ghosting was minimal and the same for every experimental condition. We argue that presence of minor ghosting is common in target consumer displays and therefore not violating the purpose of our study aiming to predict user experience from a practical 2D-to-3D stereo conversion system.

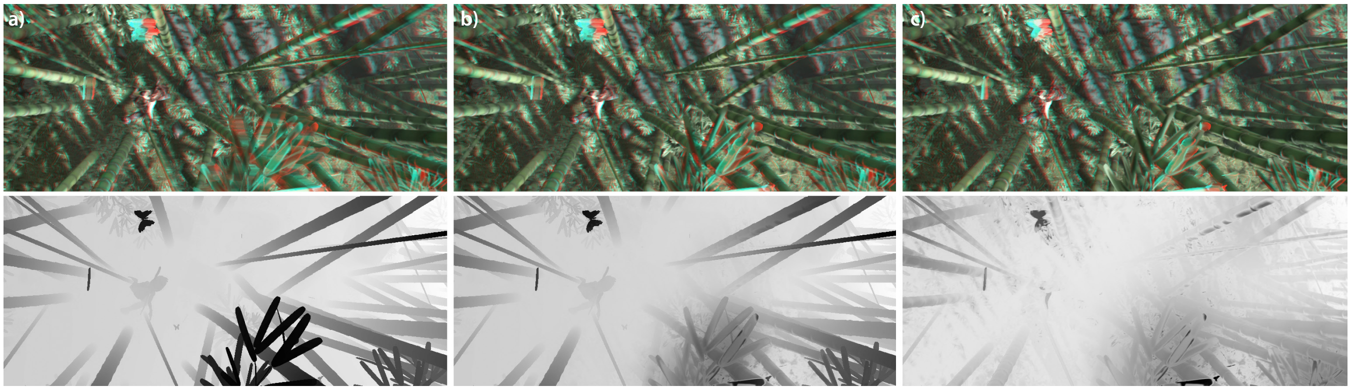

Figure 4: *Spatial blur. The top row shows the stereo image, the bottom row the disparity. The columns show different amounts of distortion due to spatial blur.* a) *Original stereo image.* b) *Blur with a spatial support of* $r_{3,1} = 0.25$ *vis. deg. and a range support of* $r_{3,2} = 0.1$*, leading to equivalence.* c) *Blur with a spatial support of* $r_{3,1} = 2$ *vis. deg. and a range support of* $r_{3,2} = 0.1$*, leading to non-equivalence.*

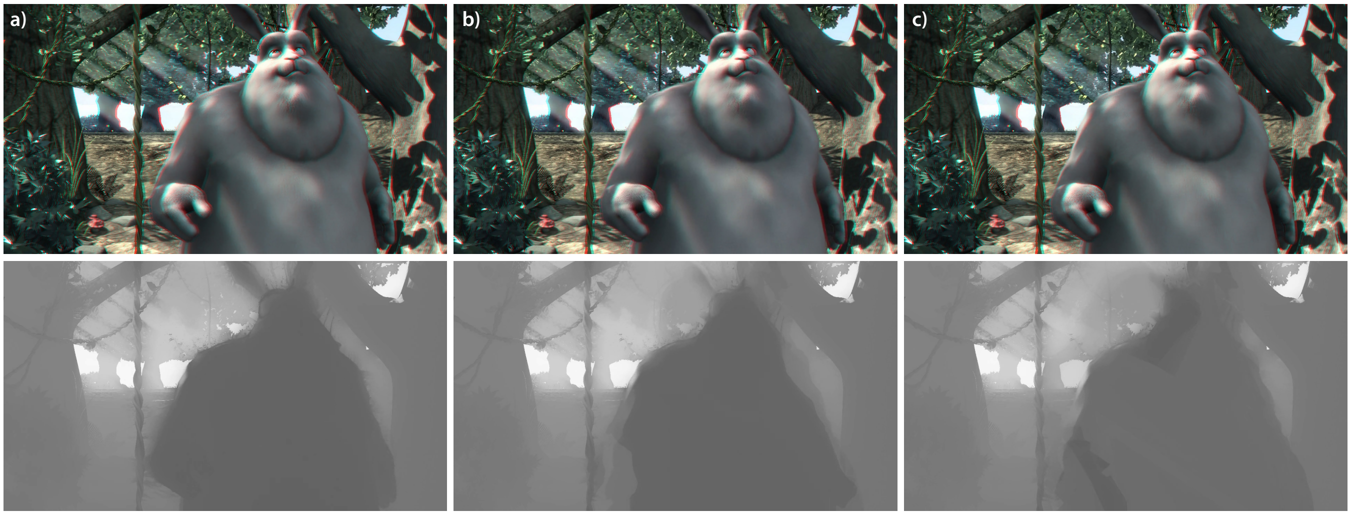

Figure 5: *Temporal blur. The top row shows the stereo image, the bottom row the disparity. The columns show different amounts of distortion due to temporal blur.* a) *Original stereo image.* b) *Blur with a temporal support of* r_4=0.25 *s, leading to equivalence.* c) *Blur with a temporal support of* r_4=1 *s, leading to non-equivalence.*

3.2 Procedure

In each trial, participants were shown the undistorted reference image and a distorted variant in a randomly shuffled vertical arrangement for 3 seconds and were asked to answer the question:

> "Do both images provide equivalent stereo fidelity?"

by pressing one out of two keys on a keyboard. Asking for equivalence instead of preference removes the influence of a subjective bias for a particular disparity distribution which might not even favor the ground-truth in all cases. An example is edge-preserving filtering that typically results in an edge enhancement, which in turn might lead to overall preferred depth appearance by some subjects.

Each trial was followed by a screen with confirmation where the subject could take a rest. A blank screen was displayed for 500 ms immediately before the stimuli was shown. The experiment comprised of 2 repetitions for each of the 4 videos or images being presented with 1 placebo, 5 different remappings, 4 removals, $3{\times}3$ spatial blurs and 3 temporal blurs yielding the total of $2{\times}4{\times}(1{+}5{+}4{+}3{\times}3{+}3){=}64$ trials, and lasted for approx. 40 minutes.

3.3 Analysis

We compute sample means and confidence intervals (binomial test, 95% confidence intervals (CIs), Clopper-Pearson method) for the percentage of trials in which a distorted and an original stimulus are considered equivalent (Fig. 6). The response is aggregated over all four scenes. Equivalence is rejected using two-sample t-testing (all $p{<}0.01$). Additionally, the effect of reduction can be seen from comparing their CIs to the control group, in particular, its lower bound (Fig. 6, dotted line). CIs that do not intersect the placebo CI after the Clopper-Pearson correction indicate the presence of an effect.

4 Discussion

In this section, we discuss the outcome of the above experiment (Sec. 4.1), compare this to observations made for artificial stimuli (Sec. 4.2), compare our equivalence outcome to the prediction of established metrics (Sec. 4.3), recall the scope and limitations of the experiment (Sec. 4.4), and finally propose some recommendations for assessing 2D-to-3D stereo conversion quality (Sec. 4.5).

4.1 Observations

Placebo The control group, which is not distorted at all, is considered equivalent to the reference in $79.0\%\pm 4.0\%$ of the cases (Fig. 6, a). This indicates, that subjects roughly understand the task and do not give random answers. At the same time, it also shows the limits of what to expect from asking for equivalence: a fifth of the subjects reports seeing a difference when images are identical. Note, that this indicates, that a distortion that is equivalent will at best result in observing a score of ca. 80%, not 100%, which is not even achieved when no change at all is present.

Remapping Remapping values for $r_1 \leq 0.6$ (stronger deviation from identity, Fig. 2, c) are significantly nonequivalent (Fig. 6, b), indicating (but not proving) that more subtle remappings might be equivalent (Fig. 2, b). This is in agreement with the general practice of retargeting disparity values using smooth curves [Lang et al. 2010; Didyk et al. 2012] to better account for human perception on limited output devices.

Spatial blur For blurring (Fig. 6, d), not respecting edges ($r_{3,2}=\infty$), or edge-stopping blurring ($r_{3,2}=0.6$ and $r_{3,2}=0.1$) with a spatial Gaussian of std. dev. $r_{3,1}\geq 1$, resp. $r_{3,1}\geq 2$vis. deg. is not equivalent (Fig. 4, c). This indicates that the slightly larger spatial extent and similar range support produce a functionally equivalent result (Fig. 4, b). It also highlights the importance of respecting luminance edges.

Object removal Not reproducing objects as large as $r_2=6$vis. deg. (Fig. 3, c) or larger is significantly nonequivalent (Fig. 6, c), indicating that removal of smaller objects might not be objectionable (Fig. 3, b). As long as such objects are consistently embedded into the environment, which typically happens due to luminance-based edge-aware upsampling, the proper values of depth are not mandatory. This is in agreement with common practice in 2D-to-3D stereo conversion, that does not manually label all objects with depth in a scene exhaustively.

Temporal blur For temporal blurring (Fig. 6, e) all reductions with a temporal Gaussian of std. dev. $r_4\geq 1$ s have been found visually nonequivalent (Fig. 5, c). This indicates that temporal disparity sampling can be surprisingly sparse if it is motion-compensated [Shinya 1993], i. e., only disparity keyframes at ca. 3Hz have to be fully recovered while the intermediate disparity frames can be temporarily interpolated (Fig. 5, b). Temporal upsampling (rotoscoping with keyframes) is a typical component of 2D-to-3D conversion, producing physically incorrect but perceptually valid results.

4.2 Artificial and natural stimuli

Analysis of thresholds, i. e., what can be perceived, has helped to better understand how the human visual system perceives both luminance and stereo [Howard and Rogers 2012, Ch. 18.5]. For finding these, artificial stimuli such as sinusoidal gratings [Didyk et al. 2012] or step edges [Kane et al. 2014] in the absence of other cues are common. However, it is clear, that such thresholds are overly conservative and do not answer the question which two natural stimuli are functionally equivalent. For this reason, visual equivalence has been proposed for specialized natural luminance stimuli involving certain geometry, certain lighting and certain reflectance [Ramanarayanan et al. 2007].

For stereo, the outcome of the above experiment indicates that in natural images, even more edge-aware spatial blurring and temporal filtering is tolerated than what was reported for disparity-only stimuli by Kane et al. [2014]. While the reductions in our experiment (and application) might introduce conflicts between disparity and pictorial cues, the latter seem to play the dominant role in depth perception, and tolerance for disparity reduction is higher. Still, as can be seen in Fig. 6, d, edges at larger depth discontinuities must be preserved, and in the temporal domain (Fig. 6, e) disparity should follow the pixel flow, while the temporal update of specific disparity values can be sparse.

4.3 Comparison to other metrics

To see if common metrics could predict the equivalence found, we compute their prediction of the difference between the reference and all our distorted stimuli and perform both linear ($a+b\cdot x$) and log-linear ($a+b\cdot\log(x+c)$) fits to the equivalence value across all distortions and stimuli. As common metrics we have tested peak signal to noise-ratio on depth and on the image pair [Merkle et al. 2009], a perceptual disparity metric on depth [Didyk et al. 2011] and a structural image similarity metric [Wang et al. 2004] on image pairs.

Table 1: *Linear and log-linear correlation R^2 coefficients of study results with various metrics. Negated values used for PSNR. Measures that explain a certain distortion best are shown in bold face.*

	Depth				Image pair			
	PSNR		Didyk2011		PSNR		DSSIM	
Experiment	Lin.	Log.	Lin.	Log.	Lin.	Log.	Lin.	Log.
Remap.	0.75	0.75	**0.82**	**0.82**	0.53	0.52	0.31	0.35
Removal	0.60	0.55	0.64	0.64	0.65	0.62	0.72	**0.76**
Spat. blur	0.49	0.48	**0.60**	0.56	0.30	0.30	0.15	0.32
Temp. blur	0.42	0.42	0.39	**0.46**	0.31	0.31	0.16	0.17
All	0.43	0.43	0.57	0.57	0.30	0.30	0.17	0.25

Fig. 7, (e) shows a scatterplot relating mean equivalence ratings by subjects to the result of numerical metrics. If any linear fit to a metric would predict equivalence well, its response would need to form a line. Similarly, a log-linear fit would need to form a logarithmic curve. However, we see that all fits predict the actual perceived difference rather poorly. PSNR has a correlation of $R^2=0.44$ for the linear and $R^2=0.44$ for the log-linear fit (all correlation statements in this section are DOF-adjusted R^2 values with $p<.01$ regression significance). As expected, the above-mentioned notions lack perceptual foundation and cannot predict perceived differences. Of greater interest is the finding that the perceptual metrics also do not predict equivalence well. A perceptual model of disparity results in a correlation of $R^2=0.57$ (linear) and $R^2=0.57$ (log-linear). For three out of four distortions, the correlation here was highest. The idea to directly compare the image pair [Merkle et al. 2009] did not result in an improvement, except for the removal distortion.

We see that a linear fit to the perceptual model produces the best result, while providing only a weak correlation. In absence of any better model for equivalence, the fit with $a=0.210$, $b=0.579$ to Didyk et al. [2012] could serve as surrogate. We conclude, that even perceptual metrics cannot capture the task-specific challenge of visual equivalence for 2D-to-3D stereo conversion and that explicit user studies are required until a computational equivalence test is available.

Similarly poor performance of objective metrics was also observed when individual experiment conditions were analyzed separately for object removal (see Fig. 7, b), spatial filtering (see Fig. 7, c) and temporal filtering (see Fig. 7, d). The only exception was found in smooth remapping (see Fig. 7, a) where depth-based metrics performed well and achieved a correlation of up to $R^2=0.82$. It seems

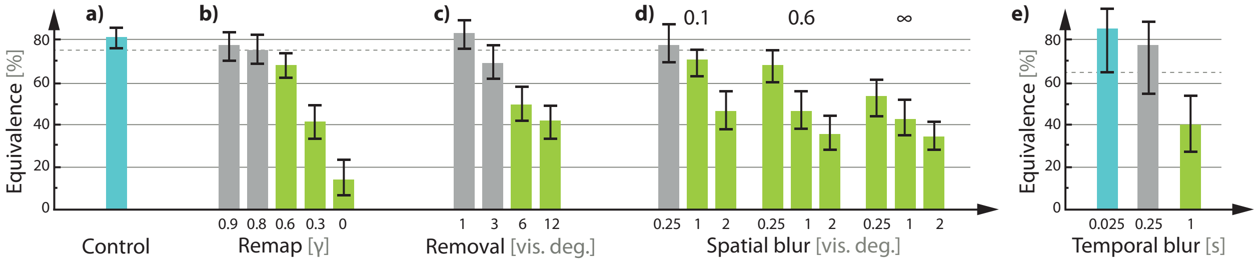

Figure 6: *Perceptual experiment analysis (Sec. 3.3): The horizontal axis shows different bars for different distortions. The vertical axis is equivalence in percentage. A high value means that the distortion is more equivalent to a reference. A green bar has a significantly different equivalence compared to how equivalent the reference is to itself, which is only ca. 80 %, not 100 %. Bars are grouped by distortions. Inside each group the distortion is the same, just more or less strong in one (b,c and e) or two (d) respects. The outcome is discussed in Sec. 4.*

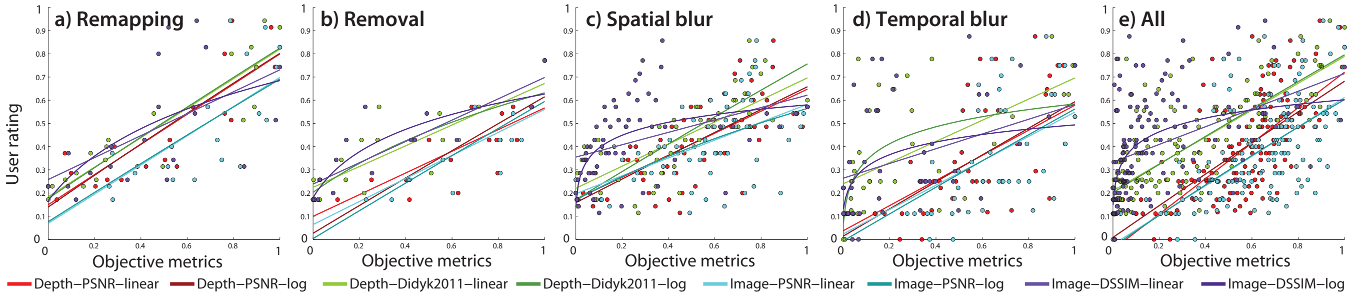

Figure 7: *Linear and log-linear fits of numerical metrics to mean equivalence ratings by subjects. Different metrics are coded as colors. For each point, the response of the common metric defines the horizontal position and the mean equivalence rating of subjects defines the vertical position.*

that existing metrics can deal much better with global monotonic manipulation as introduced by global remapping operators than with spatially varying artifacts that can arise from measurement imperfections or compression. See Table 1 for the complete list of measured correlations.

4.4 Scope and limitations

Working with natural images has inherent difficulties not found for artificial stimuli. Many other cues such as saliency [Borji and Itti 2013; Itti et al. 1998] due to luminance, motion, or stereo itself certainly affect the result. We have addressed this by performing the distortion either globally or in areas that are likely most salient (such as the moving character).

Another limitation of natural images is that we require both ground truth depth and natural images. While this data is easy to acquire for synthetic images, natural video content with ground truth depth maps is hard to come by. Most data sets available could be called to have a substantial “campus and LIDAR”-bias: they result from scanning an open street-level setting with houses and roads using a laser scanner. Practical stereo movie content however is drastically different, involving fractal natural objects, close-ups, human and non-human characters, careful scene arrangement, artistic field-of-view and cues from scene or observer motion.

Even the Sintel dataset does not have certain types of motion that are important in practice, in particular discontinuous motion that makes an important ingredient of vividly moving characters such as in sports broadcasting. This is why we add a selection of four movies with such motion to the set.

Finally, this paper only provides evidence that equivalence is not covered well using common measures. Besides the recommendation for the closest fit with a moderate amount of correlation, this is a partially negative result: It indicates, that user studies are clearly superior over numerical comparison, but does not provide a way to measure functional equivalence computationally. Note, that this however is not yet possible for luminance images and remains future work in computational stereo perception.

4.5 Recommendations

The main conclusion to be drawn from the observations made is, that metrics, be it numerical or perceptual, are poor predictors of perceptual equivalence when assessing 2D-to-3D stereo conversion quality. Instead, results should be compared by explicit user studies, which can reveal a picture entirely different from MSE, PSNR or even from perceptual disparity metrics.

The agreement with luminance edges is of particular importance. When looking at stereoscopic content, nothing is worse than a sharp disparity edge that does not align to a well-visible luminance edge. This leads to the recommendation to actually warp the image and show it to subjects, as only the combination of depth and luminance will allow for any conclusion.

If the objective is stereo video, the comparison has to be made on video, where the tolerance for errors is even higher than it already is for images. In fact it is so high, that for typical natural image footage, blurs with a standard deviation of around an entire second did not show a significant non-equivalence, even after a large number of trials and subjects.

Finally, our results suggest that the procedure of manual 2D-to-3D provides indications of what is important and what is not. Automatic and semi-automatic computational 2D-to-3D stereo conversion should learn from practitioners and adapt measures that reflect their know-how.

5 Conclusion

We have shown how the quality of 2D-to-3D stereo conversion is difficult to quantify using numerical or perceptual metrics, as users have a wide tolerance for several important distortions in comparison to a reference. We have suggested a specific fit to a specific metric that delivers a moderate correlation to user responses. A perceptual experiment has indicated an upper bound with respect to four important distortions: smooth remapping, object removal, spatial blur, and temporal blur. This indicates that quantifying the 3D impression from 2D-to-3D stereo conversion cannot be done using numerical or existing perceptual models of disparity, but requires a metric for stereoscopic visual equivalence, an exciting avenue of further research. Additionally, in future work, we would like to apply our current prediction for computer-generated content with reference 3D information to general content processed by commercial or academic 2D-to-3D systems. This remains challenging, as in general, no reference 3D information is available to quantify the distortion.

References

ASSA, J., AND WOLF, L. 2007. Diorama construction from single images. *Comp. Graph. Forum (Proc. EG) 26*, 3, 599–608.

BORJI, A., AND ITTI, L. 2013. State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 35*, 1, 185–207.

BUTLER, D. J., WULFF, J., STANLEY, G. B., AND BLACK, M. J. 2012. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, 611–625.

CHENG, C.-C., LI, C.-T., AND CHEN, L.-G. 2010. An ultra-low-cost 2D-to-3D video conversion system. *SID 41*, 1, 766–9.

DABALA, Ł., KELLNHOFER, P., RITSCHEL, T., DIDYK, P., TEMPLIN, K., MYSZKOWSKI, K., ROKITA, P., AND SEIDEL, H.-P. 2014. Manipulating refractive and reflective binocular disparity. *Comp. Graph. Forum (Proc. Eurographics 2014) 33*, 2, 53–62.

DIDYK, P., RITSCHEL, T., EISEMANN, E., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2011. A perceptual model for disparity. *ACM Trans. Graph. (Proc. SIGGRAPH) 30*, 96:1–96:10.

DIDYK, P., RITSCHEL, T., EISEMANN, E., MYSZKOWSKI, K., SEIDEL, H.-P., AND MATUSIK, W. 2012. A luminance-contrast-aware disparity model and applications. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 31*, 6.

FEHN, C. 2004. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In *Stereoscopic Displays and Virtual Reality Systems XI*, SPIE, vol. 5291, 93–104.

GUTTMANN, M., WOLF, L., AND COHEN-OR, D. 2009. Semi-automatic stereo extraction from video footage. In *Proc. ICCV*, 136–142.

HOIEM, D., EFROS, A. A., AND HEBERT, M. 2005. Automatic photo pop-up. *ACM Trans. Graph. 24*, 3, 577–584.

HOWARD, I., AND ROGERS, B. 2012. *Perceiving in Depth, Volume 2: Stereoscopic Vision*. Oxford Psychology Series.

HUANG, X., WANG, L., HUANG, J., LI, D., AND ZHANG, M. 2009. A depth extraction method based on motion and geometry for 2D to 3D conversion. In *Proc. IITA*, 294–298.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI 20*, 11, 1254–9.

JONES, G., LEE, D., HOLLIMAN, N., AND EZRA, D. 2001. Controlling perceived depth in stereoscopic images. In *SPIE*, vol. 4297, 42–53.

KANE, D., GUAN, P., AND BANKS, M. 2014. The limits of human stereopsis in space and time. *J Neurosc. 34*, 4, 1397–408.

KARSCH, K., LIU, C., AND KANG, S. B. 2014. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE PAMI 36*, 11, 2144–58.

KONRAD, J., WANG, M., AND ISHWAR, P. 2012. 2D-to-3D image conversion by learning depth from examples. In *CVPR*, 16–22.

KOPF, J., COHEN, M. F., LISCHINSKI, D., AND UYTTENDAELE, M. 2007. Joint bilateral upsampling. *ACM Trans. Graph. (Proc. SIGGRAPH) 26*, 3.

LANG, M., HORNUNG, A., WANG, O., POULAKOS, S., SMOLIC, A., AND GROSS, M. 2010. Nonlinear disparity mapping for stereoscopic 3D. *ACM Trans. Graph. (Proc. SIGGRAPH) 29*, 4.

LANG, M., WANG, O., AYDIN, T., SMOLIC, A., AND GROSS, M. 2012. Practical temporal consistency for image-based graphics applications. *ACM Trans. Graph. (Proc. SIGGRAPH) 31*, 4.

LIU, X., MAO, X., YANG, X., ZHANG, L., AND WONG, T.-T. 2013. Stereoscopizing cel animations. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 32*, 6, 223.

MERKLE, P., MORVAN, Y., SMOLIC, A., FARIN, D., MÜLLER, K., DE WITH, P. H. N., AND WIEGAND, T. 2009. The effects of multiview depth video compression on multiview rendering. *Signal Processing: Image Communcation 24*, 1-2.

MURATA, H., MORI, Y., YAMASHITA, S., MAENAKA, A., OKADA, S., OYAMADA, K., AND KISHIMOTO, S. 1998. A real-time 2-D to 3-D image conversion technique using computed image depth. *SID 29*, 1, 919–23.

PAJAK, D., HERZOG, R., MANTIUK, R., DIDYK, P., EISEMANN, E., MYSZKOWSKI, K., AND PULLI, K. 2014. Perceptual depth compression for stereo applications. *Computer Graphics Forum (Proc. Eurographics) 33*, 2.

RAMANARAYANAN, G., FERWERDA, J., WALTER, B., AND BALA, K. 2007. Visual equivalence: Towards a new standard for image fidelity. *ACM Trans. Graph. (Proc. SIGGRAPH) 26*, 3.

RICHARDT, C., STOLL, C., DODGSON, N., SEIDEL, H.-P., AND THEOBALT, C. 2012. Coherent spatiotemporal filterung, upsampling and rendering of RGBZ videos. *Comp. Graph. Forum 31*, 2.

ROBINSON, A. E., AND MACLEOD, D. I. A. 2013. Depth and luminance edges attract. *Journal of Vision 13*, 11.

SAXENA, A., SUN, M., AND NG, A. Y. 2009. Make3D: Learning 3D scene structure from a single still image. *PAMI 31*, 5, 824–40.

SHINYA, M. 1993. Spatial anti-aliasing for animation sequences with spatio-temporal filtering. In *Proc. SIGGRAPH*, 289–96.

WANDELL, B. A. 1995. *Foundations of vision*. Sinauer Associates.

WANG, Z., BOVIK, A. C., SHEIKH, H. R., AND SIMONCELLI, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing 13*, 4, 600–12.

WARD, B., KANG, S. B., AND BENNETT, E. 2011. Depth director: A system for adding depth to movies. *IEEE Comp. Graph. and App. 31*, 1, 36–48.

YAMADA, K., AND SUZUKI, Y. 2009. Real-time 2D-to-3D conversion at full HD 1080p resolution. In *IEEE ISCE*, 103–106.

YANG, Z., AND PURVES, D. 2003. A statistical explanation of visual space. *Nature Neuroscience 6*, 6, 632–640.

ZHANG, G., HUA, W., QIN, X., WONG, T.-T., AND BAO, H. 2007. Stereoscopic video synthesis from a monocular video. *IEEE TVCG 13*, 4, 686–96.

ZHANG, L., VAZQUEZ, C., AND KNORR, S. 2011. 3D-TV content creation: Automatic 2D-to-3D video conversion. *IEEE Trans. Broadcasting 57*, 2, 372–83.