Perceptual real-time 2D-to-3D conversion using cue fusion

Thomas Leimkühler* MPI Informatik
Petr Kellnhofer MPI Informatik
Tobias Ritschel University College London
Karol Myszkowski MPI Informatik
Hans-Peter Seidel MPI Informatik

Figure 1: Stereo images produced from mono images by our automatic real-time 2D-to-3D conversion.

ABSTRACT

We propose a system to infer binocular disparity from a monocular video stream in real-time. Different from classic reconstruction of physical depth in computer vision, we compute perceptually plausible disparity, that is numerically inaccurate, but results in a very similar overall depth impression with plausible overall layout, sharp edges, fine details and agreement between luminance and disparity. We use several simple monocular cues to estimate disparity maps and confidence maps of low spatial and temporal resolution in real-time. These are complemented by spatially-varying, appearance-dependent and class-specific disparity prior maps, learned from example stereo images. Scene classification selects this prior at runtime. Fusion of prior and cues is done by means of robust MAP inference on a dense spatio-temporal conditional random field with high spatial and temporal resolution. Using normal distributions allows this in constant-time, parallel per-pixel work. We compare our approach to previous 2D-to-3D conversion systems in terms of different metrics, as well as a user study.

Index Terms: Computer Graphics [I.3.3]: Picture/Image Generation—Viewing algorithms

1 INTRODUCTION

The majority of images and videos available is 2D and automatic conversion to 3D is a long-standing challenge [39]. For applications such as view synthesis, for surveillance, autonomous driving, human body tracking, relighting or fabrication, accurate physical depth is mandatory, and obviously binocular disparity can be computed from such data, resulting in a perfect stereo image pair. For 2D-to-3D stereo conversion, such physical depth however is not required. Instead, we seek to compute perceptually plausible disparity in this work. It differs from physical depth by three properties. First, the absolute scale of disparity is not relevant, and any reasonable smooth remapping [5, 18] is perceived equally plausible and may even be preferred in terms of viewing comfort and realism. Second, the natural statistics of depth and luminance indicate that depth is typically spatially smooth, except at luminance discontinuities [23, 37]. Therefore, not reproducing disparity details can be acceptable and is often not even perceived, except at luminance edges [10]. Third, the temporal perception of disparity allows for a temporally coarse solution, as fine temporal variations of disparity are not perceivable [8, 10]. Consequently, as long as the error is 2D-motion compensated, depth from one point in time can be used to replace depth at a different, nearby point in time.

Our method is modular (Sec. 3) and based on priors learned in a pre-process (Sec. 3.1) combined with stereo cues extracted from 2D images or videos at runtime (Sec. 3.2). Both priors and cues are represented as normal distributions allowing to fuse a plausible disparity map with high spatial and temporal resolution in real-time (Sec. 3.3). Image-based rendering produces a stereo video stream from this map (Sec. 3.4). The results shown in Sec. 4 are computed at ca. 35 Hz for HD video and compare favorable to off-line methods in terms of different error metrics as well as user ratings.

*{tleimkueh, pkellnho, karol, hps}@mpi-inf.mpg.de, t.ritschel@ucl.ac.uk

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.

Manual conversion produces high-quality results but requires human intervention, which can result in substantial cost. It is based on painting depth annotations [7] with special user interfaces [35] and propagation in space and time [19]. The semi-supervised method of Assa and Wolf [1] combines cues extracted from an image with user intervention to create depth parallax. User intervention can be included in our approach as an additional depth cue.

Automatic conversion does not induce manual effort, but results in long computation times to produce results of medium quality. Make3D [28] is based on learning appearance features to infer depth. This approach shows good results for static street-level scenes with super-pixel resolution but requires substantial computation. Non-parametric approaches rely on a large collection of 3D images [15] or 3D videos [11] that have to contain an exemplar similar to a 2D input. Conceptually, such an approach aligns all 3D images or 3D videos in a large collection (hundreds of exemplars) with a monocular query input image or video and transfers their depth to the query. Aligning to a large collection of images or videos of hundreds of elements contradicts our real-time requirements. We include prior disparity knowledge learned from exemplars into our inference by means of per-category disparity and confidence maps conditioned by image location and appearance. For cel animation with outlines, T-junctions have been shown to provide sufficient information to add approximate depth [20]. Our approach includes T-junctions in combination with other cues.

Real-time methods to produce disparity from 2D input videos usually come at low visual quality. Individual cues such as color [4], motion [9] or templates [36] are combined in an ad-hoc fashion. For rigid motions in animations, structure-from-motion (SfM) can directly be used to produce depth maps [38]. Classical SfM makes strong assumptions about the scene content such as a rigid scene with

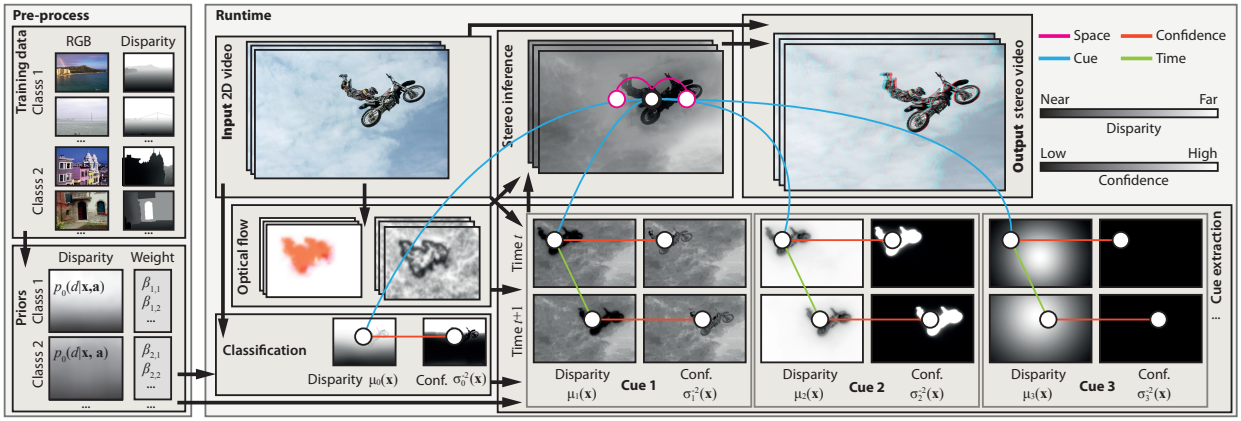

Figure 2: Overview of our approach *(from left to right)* as described in Sec. 3. The grey coding used is annotated in the top right.

camera motion. More recent work relaxes these assumptions [33], but comes along with high computational costs. In the case of no or very unstructured motion, no stereo is provided by SfM alone, whereas in our fusion-based approach, motion is just one of many cues used when available. Commercial 2D-to-3D solutions [39] 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 [11]. We subsume all such approaches in a principled framework that combines an arbitrary selection of cues in a common disparity-plus-confidence representation that can be effectively computed. Finally, most approaches produce low spatial resolution, and lack agreement between depth and luminance edges, as discussed next.

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 [16] and the spatio-temporal [26] 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 [37]. 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 [23]. This in turn leads to a significantly better depth-image-based rendering (DIBR) quality than is possible at the same bandwidth of MPEG-style compressed depth, which preserves more depth features at the expense of blurring depth edges. In this work, we follow all these guidelines while reconstructing depth maps, as we also use DIBR to secure a high-quality 3D experience.

In previous work, perception was taken into account for stereography when disparity is given [5], but it was routinely ignored when inferring disparity from monocular input. In this work, we employ depth perception models to guide the 2D-to-3D reconstruction. Inference of depth from monocular images is based on depth *cues*. In this work, we use monocular cues (Sec. 3.2) to infer the missing binocular cue. A discussion of individual cues is beyond the scope of this article and can be found in Howard and Rogers [8]. The combination of cues into a perception of depth is called *fusion*. If multiple cues are extracted, their computational fusion is considered difficult, and left to the user as in the system of Assa and Wolf [1]. Two main opposing paradigms of fusion exist: the *weak* and the *strong* model [17]. In the weak model, cues act in isolation to produce an estimate of depth which is directly combined in a fixed linear weighting. In a strong model, cues interact in an unspecified and arbitrarily complex way. Our work is based on *modified weak* fusion [17], in which cues are independent, but their combination is not a linear mixture with fixed weights, as it adapts to the confidence of each cue. Bayesian fusion [13] using normal distributions is a formal way to achieve modified weak fusion. Here, cues are weighted by their confidence before they are combined. Besides using only the cues of the present stimulus, one strength of Bayesian inference is that it can account for prior experience [13]. We acquire disparity distribution priors for different scene classes using range scanners [37] or by manual annotation. While Bayesian fusion has been considered in perception literature [8, Ch. 30] for weighting specific cues according to their confidence to explain certain observations, we show for the first time a computational model to fuse multiple cues and a prior in order to solve a real-world task such as 2D-to-3D stereo conversion in real-time.

The spatial disparity sensitivity function determines the minimum disparity magnitude required to detect sinusoidal depth corrugations of various spatial frequencies [8]. 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 [34]. Similar investigations in the temporal domain indicate that the highest sinusoidal disparity modulation that can be resolved is about 6–8 Hz [8], which is significantly lower than the 70 Hz measured for luminance [34]. As analyzed by Kane et al. [10], 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 the 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. In this work, we follow this strategy by maintaining high precision in reconstructing sharp depth discontinuities, while otherwise allowing for substantial disparity blurring. Kane et al. 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. 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 [27]. In this work, we explicitly align imprecisely reconstructed and excessively blurred depth edges with detailed luminance edges. Interestingly, depth discontinuities that are not accompanied by color edges of sufficient contrast poorly contribute to the depth perception and do not require precise reconstruction in stereo 3D rendering [5].

3 Our Approach

An overview of our approach is shown in Fig. 2. It has two main parts: a pre-process (Sec. 3.1) to extract disparity priors (Fig. 2, left) and a runtime component (Fig. 2, right). While the pre-process uses

many example images and requires considerable time, the runtime components execute in real time.

At runtime, first disparity and disparity confidence maps are extracted from monocular images (Sec. 3.2). This is the most computationally intensive part of our pipeline and implemented as parallel algorithms to require only a few milliseconds each. We support a flexible combination of both static cues (aerial perspective, defocus, vanishing points and occlusions) and dynamic cues (depth-from-motion). Each cue alone often has a low confidence in many areas and might contradict other cues. The cue evidence is then fused into plausible disparity maps (Sec. 3.3) using a robust maximum a posteriori (MAP) estimate [13]. This fusion happens again in real-time, producing results that are smooth in time and space, except at luminance edges. Finally, the monocular input image is converted into a stereo image pair (Sec. 3.4).

3.1 Pre-processing

In a pre-process, we learn prior information about disparity for certain classes of images and how to detect those classes.

3.1.1 Disparity priors

Priors model what is known about disparity in general without considering any specific image. This information is acquired from example depth images, validated and calibrated, and finally fit to a conditional distribution.

A disparity prior is the probability distribution of disparity $p_0(d)$. For efficient storage and computation, the probability distribution $p_0(d) = \mathcal{N}(d|\mu_0, \sigma_0)$ is modeled as a normal distribution $\mathcal{N}$ of a certain mean μ_0, standard deviation σ_0, and variance σ_0^2 in this work. Furthermore, our priors $p_0(d|c, \mathbf{x}, \mathbf{a})$ are *conditioned* on three parameters: the scene class c (the depth distribution in streets is different from open countries), the location $\mathbf{x} \in \mathbb{R}^2$ inside the image (the upper areas are more likely to be distant) and the appearance (RGB color) $\mathbf{a} \in \mathbb{R}^3$ (blue in the top of a forest image is more likely distant than green). For final cue fusion, scene class, image location and appearance are known and *unconditioned* priors will be used. Formally, the conditioned prior is defined as two 6D maps containing mean disparity $\bar{\mu}_0(c, \mathbf{x}, \mathbf{a})$ and the confidence of disparity $\bar{\sigma}_0^{-2}(c, \mathbf{x}, \mathbf{a})$. A high-variance value is found for a wide and unreliable distribution, while a high-confidence value σ_0^{-2} indicates a reliable estimate.

Disparity maps were acquired both by sensors and by human annotation. Sensor-acquired classes are streets and indoor. For all other classes, depth maps were painted manually. Annotation was done in parts by 2D-to-3D conversion professionals, and experienced users of image manipulation software. Images have a resolution of ca. 100 k pixels. We used 10 classes consisting of about 40 example images each. We provide the annotated database of our hand-painted depth maps and the resulting priors in our supplemental materials.

Priors are extracted from example data independently for each class (see examples in Fig. 3). Each prior is represented as a 5D regular grid where the spatial dimension is discretized into 62×38 and the color dimension into 3×3×3 bins. Normalized image coordinates between 0 and 1 are used for the spatial component and YCrCb color coordinates for the color component. Consequently, our prior contains $n_\mathrm{b} = 63\,612$ bins, with coordinates denoted as $\mathbf{b}_i \in \mathbb{R}^5$. The 2D positions and 3D colors of the n_s input pixels from all input images from that class are concatenated into a set of 5D samples $\mathbf{s}_i \in \mathbb{R}^5$, where each sample is labeled with its disparity d_i. Note that the number of bins is much smaller than the number of samples, $n_\mathrm{b} \ll n_\mathrm{s}$. Prior mean and confidence are each computed independently for all grid cells in two consecutive passes. In the first pass, the prior mean is computed as

$$\bar{\mu}_{0,i} = \sum_{j=1}^{n_\mathrm{s}} w_{ij} d_j \Big/ \sum_{j=1}^{n_\mathrm{s}} w_{ij} \quad \text{where} \quad w_{ij} = \alpha_i \psi(\mathbf{s}_j, \mathbf{b}_i),$$

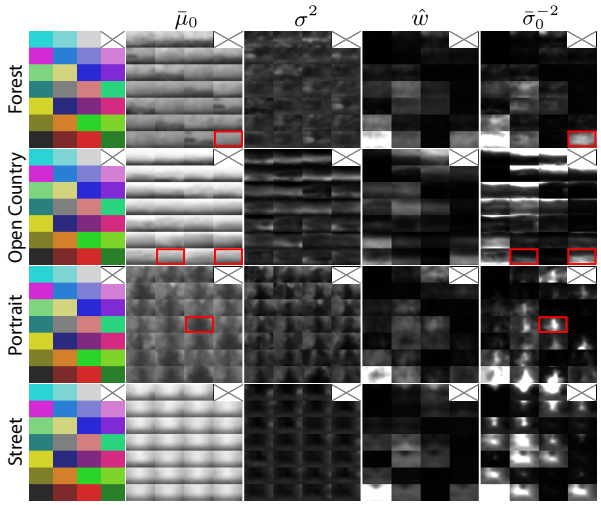

Figure 3: Mean, variance, weight and confidence *(columns)* at different colors *(tiles)* for priors of different classes *(rows)*. For forests, green central pixels have a medium depth. For countries, brown and green lower pixels have a nearby depth. For portraits, skin-colored central pixels are more nearby.

and $\psi(\mathbf{s}, \mathbf{b}) = \exp(-(\mathbf{s}-\mathbf{b})^\mathsf{T} \mathsf{A} (\mathbf{s}-\mathbf{b}))$ is a Gaussian kernel with a diagonal precision matrix A. For all results in this paper, the empirically chosen matrix entries are $\mathsf{A}_{11} = \mathsf{A}_{22} = 75$ for the spatial and $\mathsf{A}_{33} = \mathsf{A}_{44} = \mathsf{A}_{55} = 40$ for the appearance term. The normalization α_i for bin i is required because the 5D population can be highly non-uniform, and we use Gaussian filters of infinite support instead of compact kernels. At the same time, our number of bins introduces a boundary bias for bins closer to the surface of the space-appearance cube which would receive a lower total weight compared to other pixels. To compensate for this effect, we normalize each bin by its total weight. In the next pass, prior per-bin variance and weight

$$\sigma_i^2 = \frac{\sum_{j=1}^{n_\mathrm{s}} w_{ij} (\bar{\mu}_{0,i} - d_j)^2}{\sum_{j=1}^{n_\mathrm{s}} w_{ij} - \frac{\sum_{j=1}^{n_\mathrm{s}} w_{ij}^2}{\sum_{j=1}^{n_\mathrm{s}} w_{ij}}} \quad \text{and} \quad \hat{w}_i = \frac{\sum_{j=1}^{n_\mathrm{s}} w_{ij}}{\sum_{i=1}^{n_\mathrm{b}} \sum_{j=1}^{n_\mathrm{s}} w_{ij}}$$

are computed. The final prior confidence is $\bar{\sigma}_{0,i}^{-2} = \hat{w}_i / \sigma_i^2$.

3.1.2 Scene classification

Priors depend on the scene class c which is found from the monocular input RGB image. To this end, an image classifier is trained from example images that were manually labeled by their scene class. To meet our real-time requirements at test time and following ideas from Torralba [32], the image downscaled to 8×8 pixels is used as a feature vector. A linear Support Vector Machine is trained using gradient descent to separate each class from the other classes (one-vs.-one). At test time, we count the number of wins for each class over the other classes and pick the class c with the largest number of wins.

3.2 Depth cues

We model the i-th depth cue as a conditional probability distribution $p_i(d|\mathbf{x})$ of disparity d given a position $\mathbf{x}$. This distribution is described by a spatially-varying map of normal distributions in our approach. We store and process maps of mean disparity $\mu_i(\mathbf{x})$ and their confidence $\beta_{i,c}\sigma_i^{-2}(\mathbf{x})$ at position $\mathbf{x}$. The factor $\beta_{i,c}$ is a global per-cue i and per-category c weight that gives higher weights to cues that have shown to work better for certain scene categories. Actual values were determined empirically and are provided in the supplemental. We now briefly explain the $n_\mathrm{c} = 6$ cues we use. While the input sequence might have an arbitrary spatio-temporal resolution,

the typical resolution to store each cue p_i is 300×170 pixels at 3 Hz, which will later be upsampled in space and time by the pairwise fusion. We refer to frames of the image sequence holding depth cues as *keyframes*. Their position is not essential to our approach, and we refer to the supplemental material for details and example responses of cues to different input images and videos.

3.2.1 Aerial perspective

Distant objects in images showing a landscape-scale range of depth undergo changes in appearance due to atmospheric scattering. Consequently, this cue is most effective in scenes showing landscapes. Disparity is extracted following Tam et al. [30] in constant time, parallel for all pixels. As atmospheric scattering reduces contrast, pixels with little local contrast in their vicinity (low variance) have higher confidence.

3.2.2 Defocus

Scenes imaged with a finite-size aperture are increasingly blurry at image locations with distances different from the distance of the focal plane. Notably, the defocus only indicates a difference of distance to the focal plane, but not the sign. For the cue to be effective, the image has to contain this depth-of-field, which mostly occurs in images taken with a larger aperture for nearby objects. Depth-from-defocus is computed by measuring the local frequency content around a pixel [25], computed using a Laplacian pyramid in multiple passes but constant amortized time per pixel. Areas with only low-frequency content are considered out of focus. Such areas are found by first soft thresholding of contrast (Laplacian pixels) up to 0.02 using a sigmoid and second blurring this map with a kernel of size 7×7 on each level. The thresholding is required to avoid interpreting high-contrast features as being more in-focus. Out-of-focus regions are assumed to be behind the in-focus regions. Consequently, sharp regions map to a disparity of 0 and sufficiently blurred regions to a value of 1. Confidence for defocus is inversely proportional to disparity, which means that we are sure that high-frequency regions are close to the focal plane, but low-frequency regions might be or might not. Additionally, the overall confidence of this cue is reduced if in-focus features dominate the image.

3.2.3 Vanishing points

Perspective projections of parallel 3D lines cross in a 2D vanishing point. If dominant lines are visible in an image, their point of convergence is a strong depth cue we would like to exploit as well. We use an approach based on edge extraction and line accumulation [2]. First, edge orientation is found at multiple image scales and edge strength is measured by counting the number of scales at which the edge is present. Next, all pixels along a line elongating the orientation of every edge pixel are incremented by splatting a line primitive with additive blending. The value of the line increases linearly with the distance to the pixel creating this line. This gradient is required, as vanishing points are more stable if they result in agreement with other lines at an image position far away from the respective pixel causing them. Finally, the pixel in the accumulated line-image that has the highest response to a Harris corner detector is considered the vanishing point pixel. This pixel is found using a parallel reduction. The vanishing point itself is additionally low-pass filtered in time using a temporal cut-off of 0.5 Hz. Disparity is created according to this vanishing point using a radial gradient that is 1 at the vanishing point and 0 at the pixel farthest away from this point. Confidence is computed by the curvature of the accumulated value: If all lines concentrate on a single pixel, the confidence is high and the vanishing point is reliable. If multiple or only a diffuse vanishing point is found, the cue is considered less confident. While images can contain multiple vanishing points, we found it more stable in practice to only pick the dominant one.

3.2.4 Occlusion

Occlusion is a strong depth cue that works on all scales of depth: If an object A occludes object B, A is closer. However, occlusion is only a relative cue and furthermore cannot be measured directly, only inferred. Occlusions are found by detecting T-junctions of edges and lines. This is done by convolving the image with a bank of twelve separable filters, tuned to the detection of incident edges and lines at different scales. We implement filters of increasing size by executing same-sized (15-tap) oriented 1D filters on an image pyramid. The approach of Michaelis and Sommer [24] is used to detect T configurations based on these responses. As occlusion only indicates ordering, not absolute disparity, it cannot directly produce disparity and confidence, but produces sparse spatial disparity gradients with high confidence. More precisely, if a T-junction is found at position $\mathbf{x}$ with a vertical bar in direction $\mathbf{d}$ at scale s, a line orthogonal to $\mathbf{d}$ with length $10s$ is drawn with high confidence and a positive gradient at $\mathbf{x} + s\mathbf{d}$ and with a negative gradient at $\mathbf{x} - s\mathbf{d}$.

3.2.5 Motion

Several different depth cues are related to motion. Particular observer motions result in typical depth patterns and typical motions in the scene allow predictions about the relative depth of objects. In this work we use the computationally most simple cue that works based on optical flow alone. First, optical flow is computed between consecutive frames using a GPU implementation of Lucas-Kanade [21] registration. Although the output of the stereo cues is at low temporal resolution, the flow is computed at the full temporal, but reduced spatial resolution of the input image sequence, as we found flow between consecutive frames to work more reliably than registration of stronger deformations. Flow is augmented by a confidence map, computed from the luminance variance: Flow in featureless regions is considered unreliable. This flow and its confidence is later also used for temporal upsampling and propagation. Next, the confidence-weighted flow average is removed from the flow, leading to a motion residual. Finally, residual motion magnitude is mapped to disparity, such that fast moving objects are closer. Confidence is computed based on the average residual motion magnitude. In cases where this value is high, motion parallax is present and the cue is considered confident at image locations, where the flow is confident.

3.2.6 User input

Optionally, user input can be included as another depth cue to augment traditional manual stereo painting with automatic inference in the propagation. A user simply paints a disparity and confidence map and the system includes this additional cue into the inference. No results in this paper were produced using any manual intervention. The supplemental materials demonstrate the stereo improvement achieved by adding a few sparse strokes to the automatic solution.

3.3 Cue fusion

Cue fusion combines evidence from cues over space and time with the scene-specific prior (Fig. 4). Here, we will first explain the use of MLE to fuse evidence from multiple cues in a single pixel. Second, we extend the idea to include priors (MAP estimate). Next, we describe an iteratively reweighted variant of the estimate to make it robust to outliers and contradicting cues. Finally, we include interactions over time and space, and compute them using efficient edge-aware filtering.

3.3.1 Unary estimate

The unary estimate predicts the most likely value, given multiple observations with different levels of confidence. For a pixel $\mathbf{x}$, the MLE estimate of disparity $\mu_{\mathrm{MLE}}(\mathbf{x})$ is the confidence-weighted average of

Figure 4: Cue fusion *(left to right)*. Here, unary fusion combines confident occlusion, aerial perspective and defocus. The prior overrides values in the sky. Inconclusive evidence between prior and other cues is resolved by iterated re-weighting. The pairwise step propagates confident estimates to other locations, preserving space-time luminance discontinuities.

disparity means

$$\mu_{\mathrm{MLE}}(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{i=1}^{n_c} \mu_i(\mathbf{x}) \beta_{i,c} \sigma_i^{-2}(\mathbf{x}),$$

where Z is the normalizing partition function. Furthermore, the MLE of the confidence simply is

$$\sigma_{\mathrm{MLE}}^{-2}(\mathbf{x}) = \sum_{i=1}^{n_c} \beta_{i,c} \sigma_i^{-2}(\mathbf{x}). \tag{1}$$

This approach was taken in computer vision for measurements in the presence of sensor uncertainty [29] but not for 2D-to-3D conversion.

3.3.2 Prior

Priors are included in the fusion using Bayesian inference, which states that the probability distribution $p(h|e)$ of the hypothesis h given the evidence e is $p(h|e) = p(e|h)p(h)p(e)^{-1}$, where $p(e|h)$ is the probability distribution that the evidence e would be observed when the hypothesis is h, $p(h)$ is the probability distribution of the hypothesis h and $p(e)$ is the probability distribution of the evidence e [13]. A prior is included as an additional observation $\{\mu_0, \sigma_0^{-2}\}$, producing the MAP estimate of disparity

$$\mu_{\mathrm{MAP}}(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \left(\mu_0(\mathbf{x}) \beta_{0,c} \sigma_0^{-2}(\mathbf{x}) + \sum_{i=1}^{n_c} \mu_i(\mathbf{x}) \beta_{i,c} \sigma_i^{-2}(\mathbf{x}) \right).$$

The MAP estimate of variance $\sigma_{\mathrm{MAP}}^{-2}(\mathbf{x})$ is computed by extending the sum of the MLE confidence (Eq. 1).

In practice, the prior extracted in the pre-process (Sec. 3.1) that expresses information for all possible appearances at a location (conditioned prior) is used for an image with a specific appearance at a specific location (unconditioned prior). Let $L(\mathbf{x}) \in \mathbb{R}^2 \to \mathbb{R}^3$ be this appearance, a simple RGB image. We denote the final unconditioned priors mean and variance as $\mu_0(\mathbf{x}) = \mathtt{fetch}(\bar{\mu}_0, (\mathbf{x}|L(\mathbf{x}))$ and $\sigma_0^{-2}(\mathbf{x}) = \mathtt{fetch}(\bar{\sigma}_0^{-2}, (\mathbf{x}|L(\mathbf{x}))$. The function $\mathtt{fetch}(X, \mathbf{y}) \in \mathbb{R}^5 \to \mathbb{R}$ is the 5D linear filtering of a grid X at position $\mathbf{y}$ that is implemented efficiently leveraging common GPU texture filtering.

3.3.3 Robust estimate

If multiple high-confidence cues (including the prior) indicate different disparities, not all can be correct and at least one of them has to be considered an outlier. As MLE and MAP estimates for Gaussian noise models are generalized least-squares fits, they do not perform well in such conditions [6], as a single outlier quadratically skews the entire solution. Consider an example of two cues (e. g., focus and aerial perspective) and the prior that indicate a blurry blue pixel in the top to be far away, and a single cue (e. g., motion) to indicate it is close, all with the same confidence. A least squares-fit would indicate a medium disparity value. A more robust fit would result in a distant disparity and ignore the other cue as an outlier. This can be achieved by an iteratively reweighted MAP estimation. In each step (3 in our implementation) a weighted MAP is computed. In the first iteration, the weight is 1 for all evidence. In later iterations, the weight of evidence not supporting the MAP estimate of the previous iteration is decreased. Evidence does not support the estimate, if it is very different from it. The Cauchy weight function [6] is used to control the reweighting.

3.3.4 Pairwise estimate

The disparity at one space-time location $\mathbf{x}$ also depends on evidence from other pixels at nearby space-time positions $\mathbf{y}$. This serves both as an additional regularization constraint and as an opportunity to share information between less confident and more confident space-time locations. This dependency is modeled by the *domain* weight (disparity of nearby pixels should be similar) and the *range* weight (pixels with similar luminance values should have similar disparity),

$$v(\mathbf{x}, \mathbf{y}) = \mathcal{N}(\|\mathbf{x} - \mathbf{y}\|, \sigma_{\mathrm{d}}) \mathcal{N}(I(\mathbf{x}) - I(\mathbf{y}), \sigma_{\mathrm{r}}),$$

where I is the monocular image intensity [16, 26, 31]. Here, we assume the images have been motion-compensated, i. e., $\|\mathbf{x} - \mathbf{y}\|$ is the spatial distance of $\mathbf{x}$ and $\mathbf{y}$ moved to the time coordinate of $\mathbf{x}$ along the optical flow, or infinite if they are not related by optical flow. Then the final inference that combines spatially-varying cues and priors with confidence maps and interactions of pixels in space and time is

$$\mu(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \int_{\Omega} v(\mathbf{x}, \mathbf{y}) \sigma_{\mathrm{MAP}}^{-2}(\mathbf{y}) \mu_{\mathrm{MAP}}(\mathbf{y}) \mathrm{d}\mathbf{y} \tag{2}$$

with confidence

$$\sigma^{-2}(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \int_{\Omega} v(\mathbf{x}, \mathbf{y}) \sigma_{\mathrm{MAP}}^{-4}(\mathbf{y}) \mathrm{d}\mathbf{y}, \tag{3}$$

where Ω is the entire space-time domain. This inference is realized in three steps: i) Pixel-wise pre-multiplication of the mean disparity map μ_{MAP} by its confidence map $\sigma_{\mathrm{MAP}}^{-2}$; ii) edge-aware blurring of both the pre-multiplied mean disparity and confidence maps in time and space; iii) per-pixel division of the propagated mean disparity by its confidence [14].

Steps i) and iii) are trivially parallel and equivalent to compositing using pre-multiplied alpha. For propagation in time, the two nearby keyframes are first motion-compensated and then blended. Recall that we compute the flow in full temporal resolution in the depth-from-motion cue component. For motion compensation, we forward-concatenate the flow from the past keyframe and backward-concatenate the flow from the future keyframe and use this flow to warp depth from the respective keyframes into the current frame. Warping disocclusions are filled using push-pull from a Gaussian MIP map. The backward flow is approximated using the negated forward flow, assuming motion is linear on small time scales. The result is then linearly blended using the temporal distance to the future and past keyframe as weights. The output of this step is at full temporal, but still at low spatial resolution. For propagation in space, a two-channel bilateral grid [3] with 8 layers and the full spatial resolution is used. Confidence-weighted disparity and confidence values are inserted into the layers of that grid using the final image intensity I as a guide with a standard deviation of $\sigma_{\mathrm{r}} = 0.1$. This grid is then blurred using a standard deviation of $\sigma_{\mathrm{d}} = 0.5$ deg using a Gaussian

Figure 5: Our 2D-to-3D video result with motion-compensated filtering provides temporally stable stereo with fine details.

MIP map. Next, the bilateral grid is upsampled to the desired high resolution, using the high-resolution luminance as a guide. After this step, the filtered, high-resolution disparity-confidence product is finally divided by the filtered confidence component.

3.4 Stereo image generation

The final step converts the acquired disparity maps into a stereo image pair. This step is a standard 2D-to-3D procedure for which many alternatives exist. We use grid-based image deformation [22] with a grid size of one pixel.

4 Evaluation

Example results of our real-time system are shown in Fig. 1 and Fig. 7. All are produced at 35 fps on a Geforce GTX 780 with an Intel Xeon E5-1620 CPU. Please see the supplemental material for a timing breakdown. Results for video are seen in Fig. 5. An example comparison between our cue-guided manual 2D-to-3D conversion and a conventional scribble interface is seen in Fig. 6 and eleven similar results are provided in the supplemental materials. We found that our system works well over a range of scenes, while other approaches are more specific to a certain class, e. g., static street-level outdoor images. While other approaches are specialized to a specific cue (like vanishing points), certain motion (like rigid), a certain shape (like ground plane), or the image is similar to a previous image, our technique relies on a greater variety of pictorial depth cues combined with priors based on scene types. Finding a balance between prior information and individual cues is an important component of our system (Fig. 8). To use a prior, the scene needs to be classified, and if classification fails, disparity quality degrades as seen in Fig. 9. Failure cases are discussed in Fig. 10.

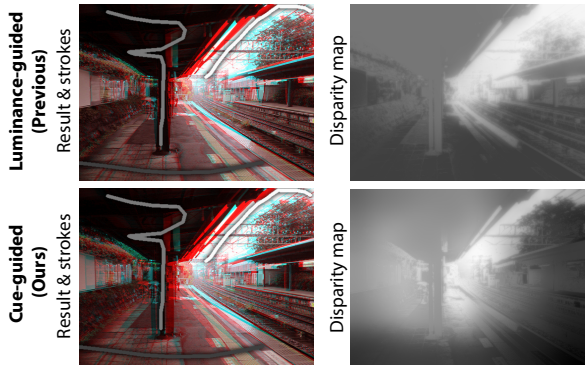

Figure 6: Manual 2D-to-3D stereo conversion without *(top)* and with *(bottom)* using our cue fusion. Our approach results in a better disparity layout and keeps details, such as the wires, for which the vanishing point was confident.

4.1 Perceptual study

We would like to know if the results produced in real time by our method are preferred over other approaches. Overall, 77 image pairs produced by our method and a previous method were shown to 41 participants wearing anaglyph glasses. The image pairs were presented in a random horizontal arrangement and participants were

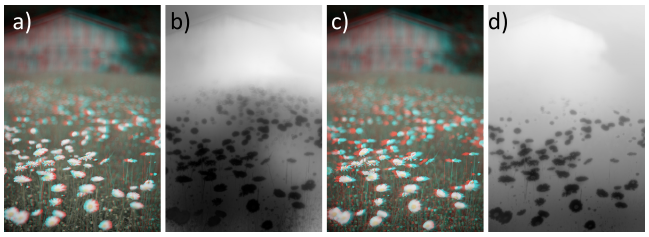

Figure 8: Result *(a)* and depth map produced by cue fusion without priors *(b)*, and including the prior for open country *(c and d)*. The cues have identified the sharpness gradient from the defocus complemented by the prior.

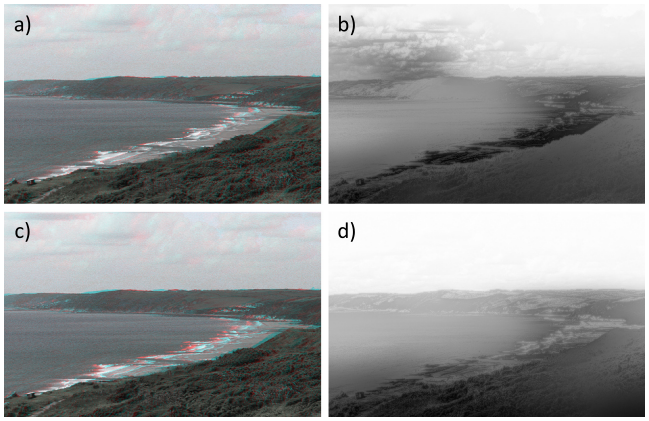

Figure 9: An image *(a)* was classified to show mountains, resulting in a disparity map *(b)* that is more vertical as seen from the low vertical contrast and the light-grey beach is mapped to near depth values. With correct classification as a coast *(c)*, the beach will be placed at medium disparity *(d)*.

asked which image provides a better 3D impression. The images have been produced using methods proposed by Saxena et al. [28], Cheng et al. [4], and Karsch et al. [11]. In our study, we include results on our images for the method of Cheng et al., images and depths provided by the original publication for the method of Saxena et al. and a mixture of both for the method of Karsh et al. To produce results for our images the method of Karsh et al. was trained using 400 outdoor images from the Make3D dataset [28] as done in the original paper. Our main goal in this study was to maximize the participants' performance in seeing differences between the methods. Therefore we chose to use static images instead of videos, since human disparity sensitivity decreases with motion [10] and participants were less likely to overlook artifacts. Our method is preferred over the method of Cheng et al. in $58.7\% \pm 1.7\%$ (0.95 confidence intervals, binomial) of the cases, over the one of Saxena et al. in $55.8\% \pm 3.5\%$ and over the approach of Karsch et al. in $50.9\% \pm 2.1\%$ of the case. The first two comparisons are significant ($p < 0.001$). Comparing our result and the method of Karsch et al. on a subset containing their images, leads to a significant preference for their results ($41.3\% \pm 3.3\%$ prefer ours) while comparing on a subset only containing our images provides significant preference for our method ($58.5\% \pm 2.7\%$). This can be attributed to a non-optimal

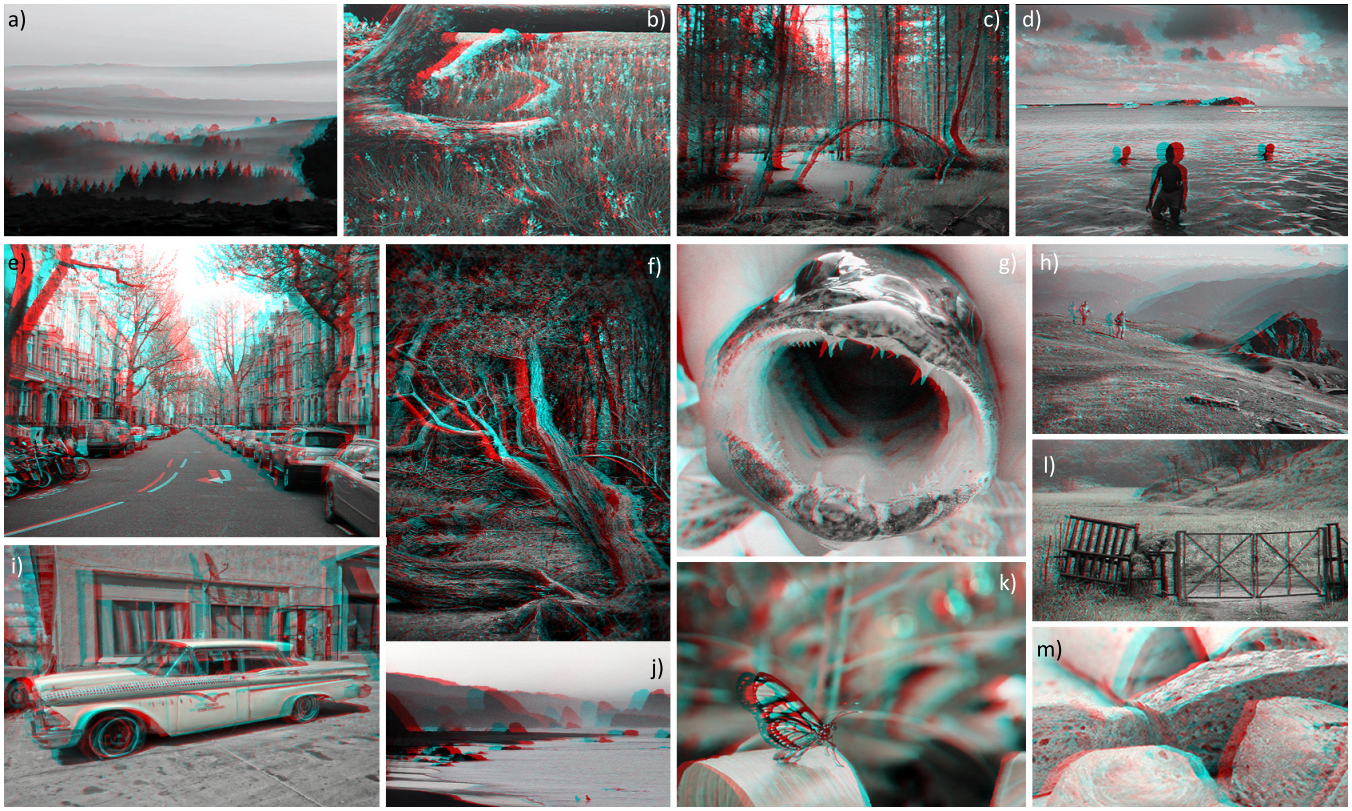

Figure 7: Results for static images with a multitude of different cues. Disparity mean and confidence maps, the response of all cues and the prior used for more than 60 images and more than 30 videos are found in our supplemental materials.

training set for certain images used in the study. We conclude that we can outperform state-of-the-art real-time 2D-to-3D conversion of Cheng et al. as well as a classic method by Saxena et al. while differences to the offline method of Karsch et al. remain insignificant and no conclusion can be made, even after a substantial number of participants and small confidence intervals.

4.2 Quantitative evaluation

The final quality of a stereo image is due to the complex interaction of monocular and binocular stereo cues, for which no computational model is available. The perceived error of a 2D-to-3D stereo conversion consequently correlates only very little with the predictions of classic image quality metrics such as the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index when they are applied to the disparity maps [12]. Merkle et al. [23] show that more meaningful quality predictions can be obtained when the reconstructed disparity is actually applied to generate stereo-image pairs and those are compared to the ground-truth images. Tbl. 1 nonetheless lists the numerical error with respect to the ground truth NYU (Kinect sensor; well-aligned key luminance and depth edges) and Make3D (laser scanning; low-resolution depth maps) data sets for the approaches of Cheng et al. [4], Karsch et al. [11] as well as ours and a baseline that uses low-frequency fractal noise as a disparity map. We see that according to the PSNR (which is poor in detecting localized disparity distortions and rather assumes their spatially uniform distribution), the approach of Karsch performs best and that most approaches perform better than random, but not on all datasets and according to all metrics. Overall, in terms of SSIM, the margin starts to get smaller. Finally, when using the most recommended metric by Merkle et al. [23], the difference between all three methods is marginalized. We conclude, that we can achieve similar quality in terms of error numbers as the competitors that either take much longer to compute and / or have a lower user preference, where the latter is clearly the most reliable quality measure. Interestingly, although the visual quality of the baseline stereo-image pairs is clearly not acceptable, the metric predictions (Tbl. 1) do not show them as clear outliers in all cases.

Table 1: Numerical comparison (larger is better).

	NYU Range				Make3D			
	Disparity		Image pair		Disparity		Image pair	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Cheng et al. [4]	9.96	0.72	21.84	0.80	10.30	0.56	16.91	0.42
Karsch et al. [11]	10.77	0.76	21.77	0.80	11.60	0.77	18.40	0.49
Ours	10.18	0.74	21.03	0.78	10.03	0.66	17.02	0.43
Baseline	10.11	0.75	18.20	0.68	8.69	0.69	16.29	0.37

5 Conclusion

We proposed a system to infer perceptually plausible binocular disparity from a monocular video stream in real time. Several monocular cues estimate disparity and confidence maps of low spatial and temporal resolution. These are complemented by spatially varying, class-specific disparity priors. Robust MAP fusion produces stereo image streams with high spatial and temporal resolution. Perceptual experiments favorably compared our approach to existing techniques. Our method reconstructs perceptually plausible disparity and not physical depth. Instead, we rather draw inspiration from how humans proceed when manually annotating disparity in 2D-to-3D conversion. If physical accuracy is required, e. g., for viewpoint changes larger than inter-ocular distance or for refocusing, it is not advised to use our method. We found our method to produce images that might have physically incorrect depth, yet, they almost always provide a 3D look due to the agreement to high-frequency luminance

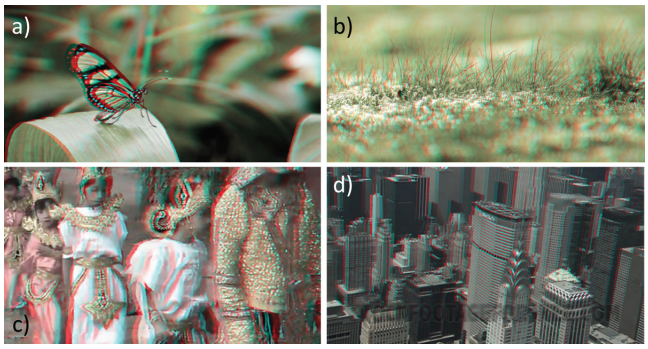

Figure 10: Failure cases: *a)* High-contrast textures can cause problems in the cue extraction as well as the cue fusion phase. Here, the occlusion module detected several T-junctions in the butterfly wing and therefore hallucinated depth gradients. This misinterpretation cannot be compensated by the pairwise fusion, since it does not distribute the available depth information across the whole object, but rather stops at the luminance edges. This leads to false-positive depth edges in the final disparity map. *b)* If an assumption made in a cue extraction module is violated, the module may produce wrong disparity values with high confidence. If only a small number of cues is present in the input video, there is not much reliable information to compensate for that. In this case, the assumption of the defocus cue, that blurred regions are distant, is violated. Since there is no other strong cue present, this leads to a large disparity in the foreground. *c)* The motion cue fails, because the walking subjects cover the image in large part. This leads to residual motion, whose magnitude is low for the subjects and high for the background, hence turning the latter into foreground. *d)* For a camera rotating around an object, both close and far points with a high velocity get classified as close.

features and overall plausible layout. Our approach seems to be less sensitive to the variety of scenes and works on priors created by painting. Depending on the problem at hand, working with sensor data can be more or less efficient than our pragmatic approach.

In future work we would like to integrate more sophisticated cues into our method. Structure-from-motion could be introduced into our system as a cue itself. More elaborate priors conditioned on texture and flow could add to the inference without imposing additional complexity and compute cost. We also would like to model cue fusion with the goal of improving the quality of stereoscopic experience when binocular disparity is given, instead of producing it from monocular images. Finally, our fusion is not limited to inference of depth, but could include other modalities such as observer motion, multiple images or real-time sensor data.

References

[1] J. Assa and L. Wolf. Diorama construction from single images. *Comp. Graph. Forum (Proc. EG)*, 26(3):599–608, 2007.

[2] S. T. Barnard. Interpreting perspective images. *Artificial Intelligence*, 21(4):435–462, 1983.

[3] J. Chen, S. Paris, and F. Durand. Real-time edge-aware image processing with the bilateral grid. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 26(3):103, 2007.

[4] C.-C. Cheng, C.-T. Li, and L.-G. Chen. An ultra-low-cost 2D-to-3D video conversion system. *SID*, 41(1):766–9, 2010.

[5] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, H.-P. Seidel, and W. Matusik. A luminance-contrast-aware disparity model and applications. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 31(6), 2012.

[6] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J Royal Stat. Soc. B*, pages 149–92, 1984.

[7] M. Guttmann, L. Wolf, and D. Cohen-Or. Semi-automatic stereo extraction from video footage. In *Proc. ICCV*, Sept 2009.

[8] I. Howard and B. Rogers. *Perceiving in Depth*. Oxford Psychology Series, 2012.

[9] X. Huang, L. Wang, J. Huang, D. Li, and M. Zhang. A depth extraction method based on motion and geometry for 2D to 3D conversion. In *Proc. IITA*, pages 294–298, 2009.

[10] D. Kane, P. Guan, and M. Banks. The limits of human stereopsis in space and time. *J Neurosc.*, 34(4):1397–408, 2014.

[11] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE PAMI*, 36(11), 2014.

[12] P. Kellnhofer, T. Leimkühler, T. Ritschel, K. Myszkowsi, and H.-P. Seidel. What makes 2D-to-3D conversion perceptually plausible? In *Proc. Symp. Applied Perception*, 2015.

[13] D. C. Knill and W. Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

[14] H. Knutsson and C.-F. Westin. Normalized and differential convolution. In *CVPR*, pages 515–23, 1993.

[15] J. Konrad, M. Wang, and P. Ishwar. 2D-to-3D image conversion by learning depth from examples. In *CVPR*, pages 16–22, 2012.

[16] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 26(3), 2007.

[17] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young. Measurement and modeling of depth cue combination: In defense of weak fusion. *Vis. Res.*, 35(3):389–412, 1995.

[18] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. Nonlinear disparity mapping for stereoscopic 3D. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 29(4), 2010.

[19] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross. Practical temporal consistency for image-based graphics applications. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4), 2012.

[20] X. Liu, X. Mao, X. Yang, L. Zhang, and T.-T. Wong. Stereoscopizing cel animations. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 32(6):223, 2013.

[21] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *IJCAI*, 81:74–679, 1981.

[22] W. R. Mark, L. McMillan, and G. Bishop. Post-rendering 3D warping. In *Proc. I3D*, pages 7–16, 1997.

[23] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P. H. N. de With, and T. Wiegand. The effects of multiview depth video compression on multiview rendering. *Signal Processing: Im. Commun.*, 24(1-2), 2009.

[24] M. Michaelis and G. Sommer. Junction classification by multiple orientation detection. In *Proc. ECCV*, pages 101–8. 1994.

[25] A. P. Pentland. A new sense for depth of field. *IEEE PAMI*, (4), 1987.

[26] C. Richardt, C. Stoll, N. Dodgson, H.-P. Seidel, and C. Theobalt. Coherent spatiotemporal filterung, upsampling and rendering of RGBZ videos. *Comp. Graph. Forum*, 31(2), 2012.

[27] A. E. Robinson and D. I. A. MacLeod. Depth and luminance edges attract. *Journal of Vision*, 13(11), 2013.

[28] A. Saxena, M. Sun, and A. Y. Ng. Make3D: Learning 3D scene structure from a single still image. *PAMI*, 31(5):824–40, 2009.

[29] R. Szeliski. Bayesian modeling of uncertainty in low-level vision. *IJCV*, 5(3):271–301, 1990.

[30] W. J. Tam, C. Vázquez, and F. Speranza. Three-dimensional TV: A novel method for generating surrogate depth maps using colour information. In *Proc. SPIE*, 2009.

[31] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Int Conf. Comp. Vis.*, pages 839–846, 1998.

[32] A. Torralba. How many pixels make an image? *Visual Neuroscience*, 26:123–131, 1 2009.

[33] C. Vogel, K. Schindler, and S. Roth. 3d scene flow estimation with a piecewise rigid scene model. *Int. J Comp. Vis.*, 115(1):1–28, 2015.

[34] B. A. Wandell. *Foundations of vision*. Sinauer Associates, 1995.

[35] B. Ward, S. B. Kang, and E. Bennett. Depth director: A system for adding depth to movies. *IEEE Comp. Graph. and App.*, 31(1), 2011.

[36] K. Yamada and Y. Suzuki. Real-time 2D-to-3D conversion at full HD 1080p resolution. In *IEEE ISCE*, pages 103–6, 2009.

[37] Z. Yang and D. Purves. A statistical explanation of visual space. *Nature Neuroscience*, 6(6):632–640, 2003.

[38] G. Zhang, W. Hua, X. Qin, T.-T. Wong, and H. Bao. Stereoscopic video synthesis from a monocular video. *IEEE TVCG*, 13(4), 2007.

[39] L. Zhang, C. Vazquez, and S. Knorr. 3D-TV content creation: Automatic 2D-to-3D video conversion. *IEEE Trans. Broadcasting*, 57(2):372–83, 2011.