DOI: 10.1111/j.1467-8659.2012.03006.x
EUROGRAPHICS 2012 / P. Cignoni, T. Ertl
(Guest Editors)

Volume 31 (2012), Number 2

Real-Time Disparity Map-Based Pictorial Depth Cue Enhancement

Christoph Rößing[1] Johannes Hanika[2] Hendrik Lensch[3]

[1]Daimler AG, Ulm, Germany [2]Weta Digital, New Zealand [3]Tübingen University, Germany

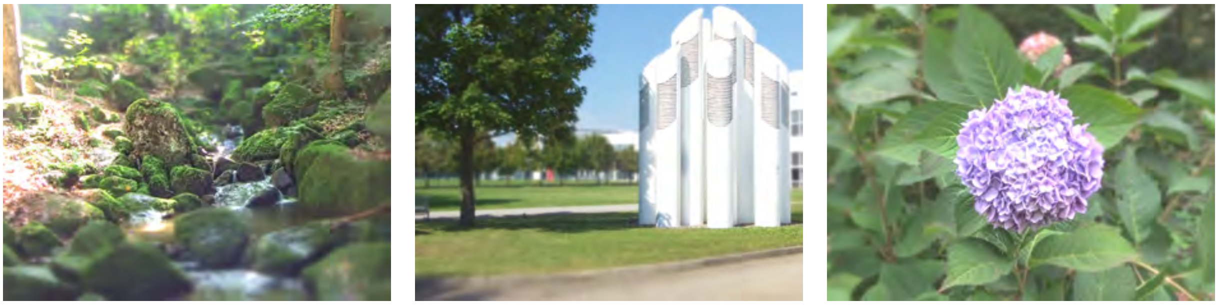


Figure 1: *Monocular depth cue enhancement by adding depth of field and selective local contrast boosting*

Abstract

The availability of stereoscopic image material is increasing rapidly. In contrast to the generation of distance information, displaying it is still a challenging task. To overcome the need for special 3D display hardware, we present a novel real-time video processing framework-based on edge-avoiding à trous wavelets. The framework adds and emphasizes monocular depth cues corresponding to the depth information of a supplemental disparity map. This creates a compelling depth sensation on 2D display devices. The framework enhances multiple depth cues in parallel, such as depth of field, local contrast, ambient occlusion and saturation. At the same time, it improves the disparity map quality. Depth cues control how a human explores an image, since the perception of distance is coupled to visual attention. The presented work demonstrates the effectiveness of the proposed framework in guiding the viewer, without destroying the image content, by evaluating the performance in search-and-find tasks. A user study analyzes the connection between faster response times and the boosting of particular monocular depth cues.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Picture/Three-Dimensional Graphics and Realism—Color, shading, shadowing, and texture

1 Introduction

As the 3D hype has hit the consumer market, the availability of 3D information will increase dramatically in the next years. Portable 3D cameras like the Fujifilm W3, the Rollei Powerflex 3D or the Minox PX3D made stereoscopic image generation very simple. Besides still images, the upcoming camcorder generation is capable of capturing high resolution stereoscopic video streams. Incorporating a CPU optimized stereo algorithm, one can produce supplemental distance information in real-time [GR10].

Unlike the simple process of gathering 3D information, displaying it is still a challenging task [RHFL10]. The most common display systems are 3D eyeglass-based monitors using shutter or polarization techniques [RHM11]. Even though those techniques are capable of creating a compelling 3D sensation, they all have weaknesses: Shutter glasses alternatively darken over one eye. This halves the image brightness and may cause flickering and crosstalk, if the monitor refresh-rate is low or not in sync with the glasses. 3D polarization displays prevent flickering, but they halve the horizontal resolution and reduce the brightness. Lenticular lens systems don't require glasses, but they reduce the resolution even further. In addition, current lenticular systems are limited in viewing angle and distance and therefore limit the maximum number of viewers. A problem inherent to all of these 3D systems is that they create a discrepancy between focal distance and displayed distance, causing visual discomfort [LFIH09].

Any of these limitations may restrict the use of such devices not only in consumer electronics, but also in applications where perceiving the spatial arrangement of the presented scene may be crucial. For instance, rear-view cameras with 2D displays are going to be standard equipment for cars, but in contrast to the rear-view mirror, binocular depth cues will be removed. But even in the absence of binocular depth cues, the Human Visual System (HVS) is still capable of estimating distances by using monocular cues in the presented scene. A stereo camera setup may capture the missing depth information, but conveying this information to the viewer is still challenging.

We propose a novel, simple and powerful edge-avoiding à trous wavelet framework for disparity map refinement and depth-based introduction and enhancement of monocular depth cues. It embeds the captured 3D information by selective boosting of shading and saturation and by adding blur from defocus and ambient occlusion to the output image. This conveys the 3D information in a nondestructive and natural way, to support the viewer in understanding the scene and in the perception of object distances and shapes.

Besides conveying depth information to the viewer, monocular cues may also draw the viewer's attention to specific regions of the presented image. The suggested rendering techniques are capable of amplifying this effect. The viewer could be guided in a potential search-and-find task towards specific image regions and distances, e.g., perceiving an elevated object while driving a car in reverse.

2 Related Work

2.1 Human Monocular Depth Perception

The HVS integrates monocular and binocular depth cues for the sensation of distance. Hence the perceived depth is quite accurate for short distances ($\approx 10m$) where accommodation, vergence and disparity are still present [Nag91]. But even in the total absence of binocular cues, the visual system is capable of integrating several monocular cues to induce an accurate impression of depth [Mat09]. The monocular cues of relative size, occlusion, size of familiar objects and parallel perspective are predefined by the scene itself.

The second category of cues results from sharpness, color, and contrast and can be enhanced for an improved sensation of depth. Held *et al.* [HCOB10] recently investigated the impact of defocus blur as a monocular depth cue and its influence on the human perception of distance and size. Their user study indicates that the HVS is combining defocus blur with other monocular depth cues to estimate distances in pictures. They also showed that physically-correct rendered defocus blur can overrule other depth cues and may even create a miniaturization effect (see Figure 2). Accordingly, adding depth of field to a sharp input image, rendered with the optical parameters of an eye, could support the HVS in estimating depths.

The visual system is capable of perceiving surface structure and object arrangement in 2D images by analyzing highlights and shadows as a result of illumination and shape. Those shading effects are boosted by local contrast enhancement. The field of high dynamic range (HDR) imaging has confirmed that high local contrasts intensify the sensation of depth on LDR and HDR displays [RHM11]. However, enhancing local contrasts and texture in the background has a negative impact on the sensation of depth [KT02]. Local contrast enhancement should therefore be restricted to important objects in the presented scenery.

The combination of color and contrast as depth cues was observed by Treismann [Tre62]. Furthermore, Troscianko *et al.* [TML*91] confirmed that saturation can be used to induce monocular depth sensation. The presented framework is capable of enhancing arbitrary combinations of these monocular depth cues to convey a convincing and natural sensation of depth on 2D displays.

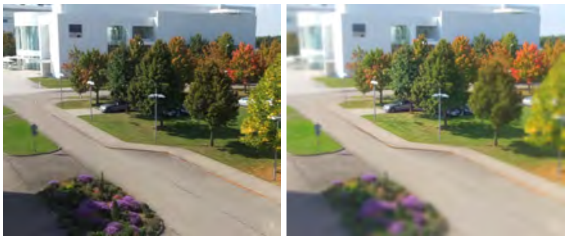


Figure 2: *Left: Original; Right: DOF rendering with a large apertures induces the tilt-shift effect. The scene looks miniaturized*

2.2 Feature Integration Theory and Guided Search

Visual attention and the perception of depth are connected. Treismann and Gelade [TG80] investigated how features in visual search tasks are processed and how they affect the response times in finding a specific target. Treismann divided search tasks into two categories.

The parallel, or preattentive search finds objects which distinguishes themselves from the distractors through a unique feature (*e.g.* a field of red squares with a single green square that has to be found). The target pops out and therefore the response time is nearly independent of the number of distractors. Accordingly, this is called the *pop-out effect*.

The serial or feature conjunction search is performed by the HVS if the object to be found distinguishes itself from its surrounding distractors through multiple features (*e.g.* a field of blue squares and orange triangles with a single orange square that has to be found). The response time for this search increases linearly with the number of distractors. Treisman identified four basic features for the preattentive search: depth, orientation of lines (resulting in shapes), color, and size of objects. In addition, boundaries of homogeneous conjunctions of features are perceived preattentively. Those homogeneous textures may also trigger a pop-out effect (*e.g.* a closed group of green letters in a field of random mixed colored letters).

Wolfe [Wol94] extended the bottom-up theory of Treisman with a top-down component resulting in the theory of *Guided Search*. He discovered that prior knowledge of target features speeds up search-and-find tasks (*e.g.* to find a green T in a field of letters with different colors, our HVS considers only the green objects in its serial search).

Following those theories, visual search is simplified by the presence of monocular depth cues. Focused or sharp areas may serve as top-down prior to a serial search. Additionally, the perceived distance is suitable as a distinguishing feature. Local contrast clarifies texture, shapes and orientations. As a consequence, boosting monocular depth cues speeds up search-and-find tasks. This effect is evaluated in a comparative study in Section 6.

2.3 Unsharp Masking the Depth Buffer

Adding shadows at depth discontinuities supports the HVS in perceiving the spatial arrangement and structure of the presented scene. *Ambient occlusion* rendering [LB00] utilizes this effect to render more realistic images. It is performed through calculating the attenuation of light by casting rays in every direction. If the ray hits the sky, the lightness is increased. A screen space approximation of ambient occlusion was proposed by Luft *et al.* [LCD06]. They introduced depth lighting and darkening based on unsharp masking of the depth buffer. This is performed by Gaussian blurring the depth buffer and subtracting it from the original. The resulting contrast signal is added to the luminance channel of the input image. This introduces artificial depth-based shading and lighting in the original image, helping the HVS to perceive structure and spatial arrangement.

2.4 Dynamic Depth of Field

A real-time system for depth of field rendering was presented by Zhan *et al.* [YTY*11]. They used light field synthesis similar to that presented by Yu *et al.* [YWY10] for performing GPU-accelerated depth of field rendering.

2.5 Edge Avoiding À Trous Wavelets

Depth of field rendering and the unsharp masking of the depth buffer require Gaussian image blurring with multiple kernel sizes. Texture dampening, edge preserving disparity map de-noising and local contrast enhancement can be performed by bilateral image filtering [TM98]. Fortunately, the fast approximations of both algorithms have a close resemblance.

A fast approximation of wide Gaussian filters has been introduced by Burt *et al.* [Bur81]. They show that repeated convolutions with generating kernels of small size converge to the same output as a wide Gaussian filter. Derived from their findings, the undecimated wavelet transform was extended with a similar filter kernel that spread its extent at every level i by a factor 2^i by inserting zeros between the filter coefficients. This is known as the *algorithme à trous* [HKMMT89]. Maintaining a constant number of non-zero coefficients, a non-naive implementation keeps the computational cost for every decomposition level constant.

The fast approximation of the bilateral filter based on *decimated edge avoiding* wavelets was introduced by Fattal [Fat09]. Similar edge crossing functions were also applied to the à trous wavelets [FAR07] extending them to *undecimated edge avoiding* second generation wavelets [Swe10]. This made edge avoiding à trous wavelets (EAAW) also applicable for fast approximation of bilateral [FAR07] and multilateral [DSHL10] image filtering.

In contrast to other accelerations of the bilateral filter (*e.g.* [QTA09]), EAAW decompose the input at multiple scales. Selective boosting and dampening of details on different scales allow noise reduction and local contrast enhancement at the same time [HDL11].

Our work extends the application domain of EAAW to the manipulation and introduction of monocular depth cues (see Section 4) and disparity map refinement (see Section 3.1).

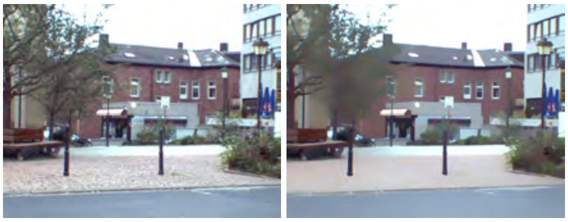


Figure 3: *Left: Input Image; Right: The bilateral filter decomposition removes the details from the image and preserves sharp edges*

3 EAAW Decomposition Pipeline

For real-time manipulation and introduction of multiple monocular depth cues, an edge avoiding à trous wavelet-based parallel rendering pipeline is introduced (see Figure 4, 5 and 7). It decomposes the input image I into a bilateral filtered image B_c, in which the local contrasts in smooth areas are removed (see Figure 3). These removed details are saved separately in B_d (see Section 3.1). In the recomposition stage of the pipeline, the local contrast can then be enhanced or reduced by boosting or dampening the detail coefficients B_d, before they are added back to B_c (see Section 4).

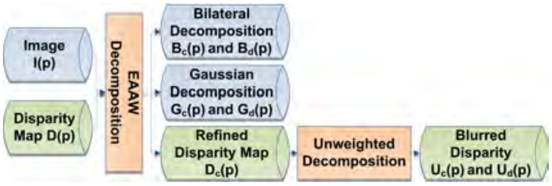


Image I(p)
Disparity Map D(p)
EAAW Decomposition
Bilateral Decomposition $B_c(p)$ and $B_d(p)$
Gaussian Decomposition $G_c(p)$ and $G_d(p)$
Refined Disparity Map $D_c(p)$
Unweighted Decomposition
Blurred Disparity $U_c(p)$ and $U_d(p)$

Figure 4: *The EAAW framework decomposes image and disparity in parallel*

The EAAW are derived from an algorithm for fast approximation of Gaussian blurring (see Section 2.5). Hence it is convenient to perform a Gaussian blurring G_c of the input image within the same wavelet decomposition (see Section 3.1). A depth-dependent full or partial reconstruction of the sharp input image can be performed by adding back the difference to the original image, which is stored in G_d. In the recomposition stage of the pipeline, this is used for DOF rendering (see Section 4).

The quality of the monocular depth cue rendering depends on the coherence between the disparity map and the input image. Disparity maps generated with a real-time SGM algorithm [GR10] are usually noisy and contain artifacts like gaps and border bleeding. Gaps are a result of ambiguous or missing stereo correspondences. The largest gaps are caused by foreground objects, which occlude different scenery regions in both camera images (stereo shadows). Especially within stereo shadows, the smoothness constrain of the SGM algorithm causes border bleeding, resulting in outlines of foreground objects which are too broad (see Figure 6(b)).

The presented pipeline performs a modified EAAW decomposition of the input disparity map for gap interpolation and noise reduction (see Section 3.1). To correct the border bleeding artifacts, the edge information generated in the EAAW decomposition of the corresponding input image and the disparity map are compared. If the edge locations differ considerably, the disparity map outline is corrected (see Section 3.1). These refinement techniques generate a dense, smooth and outline-corrected disparity map D_c.

In the last stage of the pipeline, the unweighted disparity decomposition U_c is calculated by performing Gaussian blurring on D_c. U_c is used to remove blur discontinuities while rendering the DOF (see Section 4). The difference to the input signal D_c is stored in U_d and is used in the recomposition stage as an unsharp mask of the disparity map (see Section 4). To reduce computational cost, the first three decompositions are performed in parallel, which avoids the recalculation of intermediate results of the EAAW image decomposition.

3.1 EAAW Decomposition Module

The image and the disparity map are decomposed by the weighted à trous wavelet transformation as proposed by Dammertz *et al.* [DSHL10]. For N decomposition levels i of the input signal c_0 one has to perform:

$$c_{i+1}(p) = \frac{1}{k} \sum_{q \in \Omega} h_i(q) \cdot w(p,q) \cdot c_i(q) \tag{1}$$

$$k = \sum_{q \in \Omega} h_i(q) \cdot w(p,q) \tag{2}$$

$$w(p,q) = e^{-\frac{\|c_i(p)-c_i(q)\|^2}{\sigma_i}} \tag{3}$$

$$d_i(p) = c_{i+1}(p) - c_i(p) \tag{4}$$

$$c_0 = c_N + \sum_{0}^{i=N-1} d_i \tag{5}$$

Filter h_i is based on a third order B-spline interpolation. At each level, the non-zero coefficients are spread by a factor 2 by filling in 2^{i-1} zeros. To keep the computational cost for every level constant, Ω are only the non-zero elements of the filter h_i. The weighting function $w(p,q)$ avoids pixel interpolation across image edges. The indices p and q are pixel positions and σ_i controls the smoothing in every decomposition step. The quotient $1/k$ normalizes the sum. The decomposition steps are repeated until $i = N$. The resulting images $\{d_0, d_1, ..., d_{N-1}, c_N\}$ are the weighted à trous wavelet transform of c_0. The original signal c_0 is reconstructed by adding back the detail coefficients to the coarsest decomposition c_N.

The EAAW Module calculates two EAAW decompositions in parallel, to decompose the image and the disparity map. A third decomposition identifies and removes boarder-bleeding artifacts in the disparity map (see Figure5).

Image Decomposition Applying the transform to the input image I generates the coarse output $B_c(p)$ and N levels of detail in $B_d^i(p)$:

$$B_c(p) = c_N(p) \qquad |c_0 = I \tag{6}$$

$$B_d^i(p) = d_i(p) \qquad |c_0 = I \tag{7}$$

The number of decomposition levels N define the kernel size, whereas σ_i controls how much detail is moved from the image to $B_d^i(p)$. Increasing σ_0 dampens textures and shading in the bilateral filtered coarse image B_c (see Figure 9). To allow bilateral smoothing across larger image regions, σ_i is doubled at each decomposition level i.

Gaussian blurring of I is calculated by setting $w(p,q) = 1$, to generate G_c.

$$G_c(p) = c_N(p) \qquad |w(p,q) = 1, c_0 = I \tag{8}$$

$$G_d^i(p) = d_i(p) \qquad |w(p,q) = 1, c_0 = I \tag{9}$$

The output image $G_c(p)$ is an approximation of Gaussian blurring with a filter radius $r = 2^N$. The blurring can be undone by adding back the details $G_d^i(p)$. A partial restoration of the details approximates filter radii between $0 < r < 2^N$ (see Section 4).

Disparity Map Decomposition Disparity maps with noise, gaps and border bleeding create artifacts in the output. Three steps are taken to improve the input disparity map D:

Disparity Map Bilateral Decomposition removes noise, but preserves sharp edges. It is carried out similarly to the image decomposition:

$$D_c(p) = c_N(p) \qquad |c_0 = D \tag{10}$$

$$D_d^i(p) = d_i(p) \qquad |c_0 = D \tag{11}$$

The details of this decomposition $D_d^i(p)$ are not needed in the recomposition pipeline, thus they don't have to be calculated or stored. The parameter σ_i controls the noise reduction of the bilateral filtering. For large scale smoothing, σ_i is doubled at each level i. In the first levels, while calculating $D_c(p)$, some of the disparity values might be marked as missing ($c_i(p) =$ missing) by the SGM algorithm.

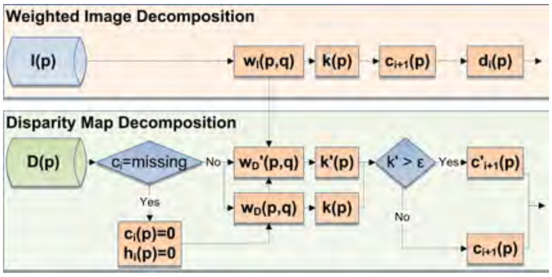


Figure 5: *Stages of the EAAW Module for the weighted decomposition of the image $I(p)$ and the disparity map $D(p)$*

Missing Disparity Interpolation is performed to create a dense disparity map at every location. The largest gaps are caused by occlusion. If a foreground object occludes different background areas in the stereo image pair, correspondence matching fails, which causes gaps in the disparity map (stereo shadows). Therefore, gaps in the disparity map shouldn't be interpolated with foreground disparity values. This is avoided by setting the current disparity value to $c_i'(p) = 0$ and the filter kernel at the location p to $h_i'(p) = 0$:

$$c_i'(p) = 0 \qquad |\text{if } c_i(p) = \text{missing} \qquad (12)$$

$$h_i'(p) = 0 \qquad |\text{if } c_i(p) = \text{missing} \qquad (13)$$

As a result, the current disparity is interpolated by the smallest (closest to 0) neighboring disparities. Setting $h_i'(p) = 0$ removes the influence of the incorrect zero value $c_i'(p) = 0$ to the interpolation. As a result, missing values are interpolated by the most distant disparities in their surrounding. This avoids filling of stereo shadows with incorrect foreground disparities (see Figure 6(c)).

Disparity Map Outline Refinement removes border bleeding artifacts. Following the assertion of Bleyer and Gelautz, it is assumed that regions of smooth color deviation also differ smoothly in disparity [Ble05]. Sharp edges in the disparity map should have a counterpart in the input image at the exact same location. If a disparity edge is located only near an image edge, it is an indication of border bleeding. During decomposing, the edge locations are encoded into the disparity w_D and image w_I weights. These locations are compared by calculating a third EAAW transform with a new weighting function w_D':

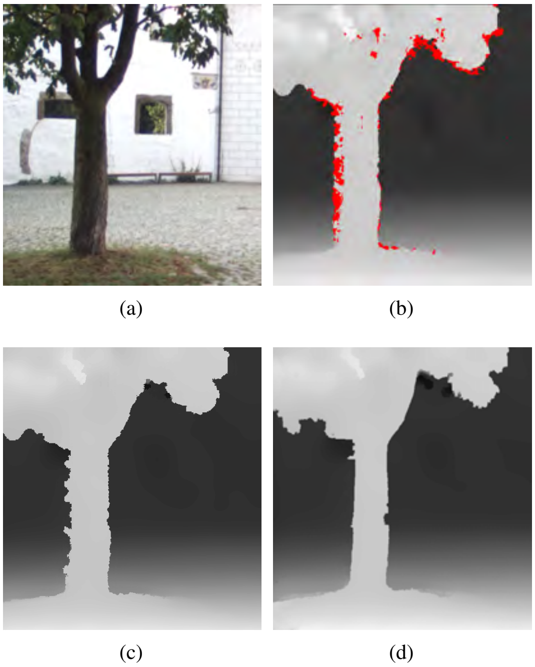
$$w_D'(p,q) = (1 - w_D(p,q)) \cdot w_I(p,q) \qquad (14)$$

$$k' = \sum_{q \in \Omega} h_i(q) \cdot {w_D}'(p,q) \qquad (15)$$

$$c_{i+1}'(p) = \frac{1}{k'} \sum_{q \in \Omega} h_i(q) \cdot w'(p,q) \cdot c_i'(q) \qquad (16)$$

If the neighboring pixels at the location q are similar in their color and disparity, it follows that $w_D(p,q) \approx w_I(p,q)$ and $w_D'(p,q)$ is low (≤ 0.25). On the other hand, if the pixels at the location q are similar in color, but not in disparity, $w_D'(p,q)$ is high. As a consequence, high values for k' indicate a border bleeding artifact at the location p. In this case $c_{i+1}(p)$ has to be calculated differently, to correct the outline in the next level $i+1$. A simple approach would be to replace the weighting function w_D by w_I for the calculation of $c_{i+1}(p)$. But this might take other neighboring border bleeding artifacts into account, causing blurry and incorrect disparity edges. To keep the edges sharp, the new weighting function $w_D'(p,q)$ is used for the decomposition at the current location (see Figure 5). It only generates high values for neighbors with similar color ($w_I \approx 1$) but different disparity ($w_D \approx 0$). As a result, the current disparity value $c_i(p)$ is replaced in the next level with an average of the most different neighboring disparity pixel having a similar color. This replacement of $c_{i+1}(p) = c_{i+1}'(p)$ is only performed if a border bleeding artifact is detected by the indicator $k' > \varepsilon$. The presented decomposition removes border bleeding artifacts for the next level $i+1$, but maintains sharp edges (see Figure 6(d)).

To avoid recalculating $w_D(p,q)$ and $w_I(p,q)$, this operation is performed in parallel to the image and disparity map decomposition.


(a) (b)

(c) (d)

Figure 6: *(a) Input image; (b) Disparity map with gaps (red); (c) Filled disparity map;*
(d) Refined disparity map $D_c(p)$

3.2 Unweighted Disparity Map Decomposition

After the parallel EAAW decomposition, the refined disparity map D_c is Gaussian blurred by performing an unweighted wavelet decomposition. Similar to the calculation in Section 3.1, the Gaussian blurred output U_c and its removed details are obtained U_d by choosing D_c as the input c_0 of our EAAW decomposition and disabling the weighting function ($w(p,q) = 1$):

$$U_c(p) = c_N(p) \qquad |w(p,q) = 1, c_0 = D_c \tag{17}$$

$$U_d^i(p) = d_i(p) \qquad |w(p,q) = 1, c_0 = D_c \tag{18}$$

The Gaussian blurred disparity map $U_c(p)$ is used for removing depth discontinuities while DOF rendering. The sum $\sum U_d^i(p)$ is an unsharp mask of the disparity map which is used for depth lighting and darkening in the composition pipeline.

4 Adding and Boosting Pictorial Depth Cues

After decomposing, the image is re-synthesized with additional and enhanced pictorial depth cues. Those cues are based on the depth information gathered from the refined disparity map. The presented framework is capable of rendering multiple combinations of pictorial depth cues (see Figure 7). Depending on the application, the focal distance z_0 could be set by the user or an automatic object recognition algorithm.

To avoid color artifacts, the pipeline operates in the CIELAB color space.

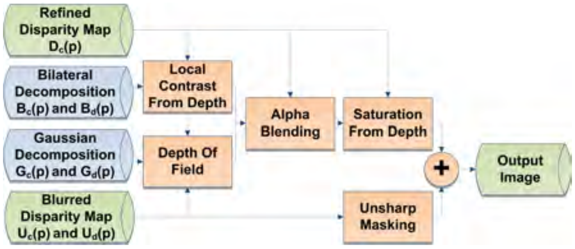

Refined Disparity Map D_c(p)
Bilateral Decomposition B_c(p) and B_d(p)
Gaussian Decomposition G_c(p) and G_d(p)
Blurred Disparity Map U_c(p) and U_d(p)
Local Contrast From Depth
Depth Of Field
Alpha Blending
Saturation From Depth
Unsharp Masking
+
Output Image

Figure 7: *Monocular enhancement pipeline rendering multiple depth cues generated by the same underlying EAAW processing*

Depth of Field Rendering adds an additional monocular depth cue to the re-synthesized output image. Held *et al.* [HCOB10] have shown that blur from defocus is a strong depth cue, capable of overruling even other monocular cues (see Figure 2). A natural depth of field supports the HVS while estimating distances and scene arrangement in 2D images. Performing a DOF rendering with the optical parameters of a human eye adds a natural depth of field to a sharp input image. The rendering is performed while re-synthesizing the output image $O(p)$, based on the depth-dependent blur radius $r(z_1)$. The distance z_1 is calculated from the refined disparity map D_c.

$$O_{DOF}(p) = G_c(p) + \sum_{i=N-1}^{0} G_d^i(p) \cdot \beta_i(z_1) \tag{19}$$

$$\beta_i(z_1) = \begin{cases} 0.0 & \text{for } i < \lfloor \eta(p) \rfloor \\ \eta(p) - \lfloor \eta(p) \rfloor & \text{for } i = \lfloor \eta(p) \rfloor \\ 1.0 & \text{for } i > \lfloor \eta(p) \rfloor \end{cases} \tag{20}$$

$$\eta(p) = \log_2(r(p)) \tag{21}$$

$$r(p) = \left| A \frac{v_0}{z_0} \left(1 - \frac{z_0}{z_1 \cdot D_c(p)} \right) \right| \tag{22}$$

To achieve a depth-dependent defocus as a human observer would see it, the blur radius $r(z_1)$ corresponding to the optical parameters of the human eye has to be calculated: Aperture $A \approx 4,6mm$ and $v_0 \approx 24mm$ [HCOB10]. The blur radius $r(z_1)$ in pixel defines $\beta_i(z_1)$, which controls how many detail levels $G_d^i(p)$ are fully or partially restored in the output image $O_{DOF}(p)$.

Removing Blur Discontinuities should be applied, if the focal distance is not set to foreground objects. Screen space DOF rendering creates unnatural sharp edges at large depth discontinuities. This is most apparent at the crossing of unfocused foreground and focused background objects. To remove these apparent artifacts, the blurred disparity map U_c is used in part for the DOF rendering. If $z_1 \leq z_0$ the renderer uses the blurred $U_c(p)$, otherwise the sharp disparity map $D_c(p)$ for the calculation of $r(p)$. This rendering keeps the edges of the focused objects sharp, but removes discontinuities in the foreground (see Figure 8).

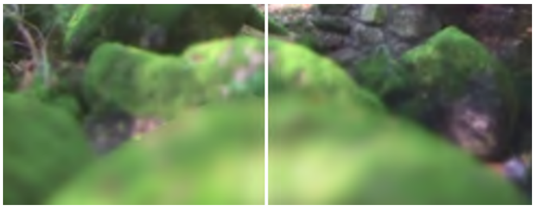


Figure 8: *Left: With blur discontinuities; Right: Removed blur discontinuities*

Depth Dependent Local Contrast Enhancement boosts the shading on objects, which supports the HVS in the perception of 3D texture and shape. Boosting all detail coefficients while reconstructing enhances local contrast, resulting in improved depth sensation [RHM11]. But Hubona *et al.* [HS05] have shown that heavily textured backgrounds disturb depth perception. To account for these findings, the presented framework is capable of depth-dependent texture enhancement/dampening while re-synthesizing the weighted decomposition:

$$O_{BiLat}(p) = B_c(p) + \gamma_i \cdot \beta_i(z_1) \sum_{i=N-1}^{0} B_d(p) \tag{23}$$

The boosting factor γ_i has to be set by user preference. The user can control if small or large scale details are enhanced or removed. *E.g.* Removing the smallest scale $\gamma_0 = 0$ dampens camera noise. The distance-dependent reconstruction factor $\beta_i(z_1)$ is the same as in the DOF rendering. Objects in the plane z_0 are rendered with boosted local contrast, whereas out-of-focus objects appear dull and textureless (see Figure 9). This rendering increases the perceived three-dimensionality of enhanced objects, whereas dampened regions appear to be flat.

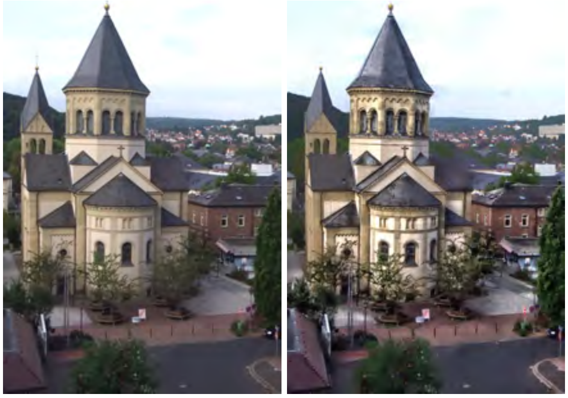


Figure 9: *Left: Original; Right: Depth-dependent local contrast enhancement. The church is boosted whereas the background appears dull*

DOF and Local Contrast Enhancement are combined to create an even more compelling sensation of depth. Whereas DOF supports the HVS for perceiving the coarse scene arrangement and different focal planes, local contrast supports the perception of fine surface structure and discontinuities. The suggested rendering pipeline is able to combine both monocular cues by performing a smooth alpha blending:

$$O(p) = \theta \cdot O_{BiLat}(p) + (1-\theta) \cdot O_{DOF}(p) \tag{24}$$

$$\theta = \frac{1}{2}\left(1 + \tanh\left(5\left(\beta_i(z_1) - \frac{1}{2}\right)\right)\right) \tag{25}$$

The hyperbolic tangent blends the local contrast enhanced image into the focal plane and degrades fast for out-of-focus areas (see Figure 1).

Unsharp Masking The Depth Buffer introduces an additional depth cue to the synthesized output image. The unsharp mask of the disparity map is stored in the detail levels $U_d^i(p)$ of the unweighted disparity map decomposition. The sum over all details is used as a contrast signal for the unsharp masking which is added to the luminance channel of the output image:

$$O_{USM} = O(p) + \sum_{i=N-1}^{0} U_d^i(p) \cdot \rho \tag{26}$$

This rendering adds shadows and highlights at depth discontinuities to the output image. The effect strength is user-controlled by the parameter ρ. As a result, spatial arrangement and intersections are more apparent, supporting the HVS in perceiving scene arrangement (see Figure 10).

Depth-Based Desaturation decreases saturation of background objects. According to the findings of Troscianko *et al*. [TML*91], distance is encoded as desaturation. This supports the HVS in separating foreground from background. The desaturation is performed by multiplying the color channels A and B in the CIELAB color space with the damping factor:

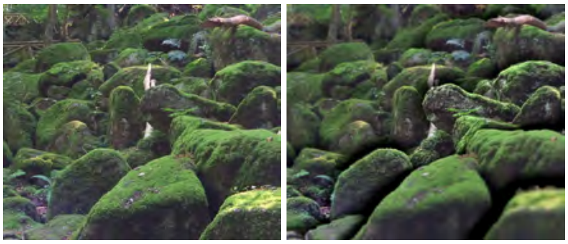


Figure 10: *Left: Original; Right: Unsharp Masking the Depth Buffer with DOF and Local Contrast Enhancement*

$$O'_{AB}(p) = \begin{cases} O(p)_{AB} \cdot \left(\frac{\beta_i(z_1)}{\varsigma} + 1\right)^{-1} & \text{for } z1 > z0 \\ O(p)_{AB} & \text{else} \end{cases} \tag{27}$$

Including $\beta_i(z_1)$ prevents desaturation of focused objects, whereas distant objects in the background are de-saturated and appear even more far away. This supports the perception of scene arrangement and amplifies foreground saliency (see Figure 11).

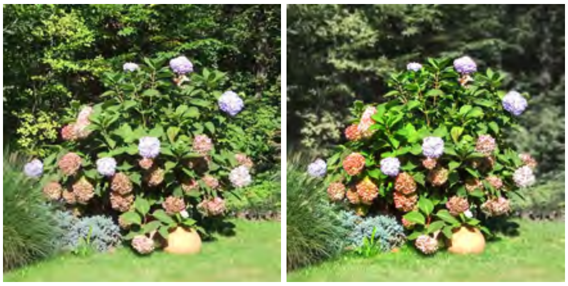


Figure 11: *Left: Original; Right: Depth-Based Desaturation with DOF and local contrast enhancement*

Combination of Multiple Cues The HVS combines all the available cues to estimate depth and scene arrangement in 2D images. Combining multiple cues in one image conveys more of the available depth information to the viewer. The computational costs for multiple renderings increase just slightly. The parallel decomposition of the image and the disparity map in the EAAW module is the most expensive calculation and has to be performed in any case. As a consequence, adding additional effects showed nearly no impact on rendering time (see Section 5).

5 GPU Accelerated Implementation and Benchmark

The per-pixel operations for decomposition and synthesis of the presented pipeline can be performed in parallel. A GPU architecture using CUDA is well suited for such applications. The calculation time depends only on image size, number of wavelet decomposition levels N and the number of enabled monocular depth cue enhancement effects. CIELAB color conversion was performed on the CPU and is excluded from the presented benchmarking results. The benchmark was evaluated on a Geforce GTX 580. As expected, the runtime for the whole CUDA monocular depth cue pipeline is quite low. Even 2MP images can be processed in real-time at 26 frames per second.

Module	0.5 MPix	1 MPix	2 MPix
EAAW Decomp	6.2ms	12.8ms	25.6ms
Unweighted Decomp	1.3ms	2.6ms	5.1ms
DOF	0.2ms	0.4ms	0.9ms
Local Contrast	0.2ms	0.5ms	0.8ms
Alpha Blend	0.8ms	1.7ms	3.0ms
Saturation	0.1ms	0.2ms	0.4ms
Unsharp Masking	0.4ms	1.0ms	1.8ms

Table 1: *Averaged runtimes over 5 runs for single modules of the monocular depth cue enhancing framework for 5 wavelet decomposition levels*

6 Study: Visual Guidance by Monocular Depth Cues

As pointed out in section 2.2 monocular depth cues and visual attention are coupled to each other. Treismann identified depth as a feature for preattentive search. Accordingly, enhanced monocular depth perception speeds up search-and-find tasks. The additional features orientation, texture and shape are boosted by local contrast enhancement. Depth of field rendering blurs textures of objects at out-of-focus distances, whereas objects in focus are rendered sharp and draw the viewer's attention [KT02]. The combination of depth cues can create a pop-out effect in a visual search, lowering the response time independently from the scene complexity. If triggering of the pop-out effect fails, the visual search is performed sequentially. Nevertheless, incorporating the prior knowledge that objects to be found are located in sharp or focused image areas guides the visual attention towards potential target locations. The response time increases, but is still significantly lower than for a sequential search on the whole image.

A comparative study evaluates the impact of monocular depth cue enhancement on visual searches. DOF rendering, Depth-Dependent Local-Contrast Enhancement and their combination are compared with the original image. The user study identifies the impact of supportive monocular depth cue enhancement on human visual search performance. A second experiment evaluates the impact on visual search for objects at depths not targeted by the enhancement.

Study Design A series of response time evaluations of the three suggested renderings and the unchanged input image were conducted. All images were converted to grayscale to avoid the interference of color as feature. Embedding the visual search task in outdoor (street) scenery created a natural context. The depicted search task was to find a ball located within the scenery as fast as possible. The single images were captured from a video stream, presenting the ball rolling or flying within every quadrant of the image.

Study Setup A camera captured stereo video streams with a 2 x 1 MPix rig and computed a disparity map for every frame by incorporating a real-time SGM stereo algorithm [GR10]. Single frames out of the video stream were selected and processed with the presented monocular depth cue enhancing framework. The following parameters were selected for depth cue enhancement: $A = 11,5mm$; $v0 = 60mm$; $\sigma_I = 1$; $\sigma_D = 0.1$; $\gamma = 3$. The focal distance z_0 was set by hand. The resulting images were presented at a viewing distance of $\approx 60cm$ on a Dell 3008WFP monitor with a brightness of $200cd$ at ambient lighting conditions. The images had a resolution of 1288 x 964 pixels and were presented 1:1 on the display, leaving a black frame around the stimulus. The response times were captured by key pressing events on a low latency (Razer Arctosa) keyboard. The image presentation and response time capturing was controlled by the Matlab Psych-Toolbox [Bra97].

For each scene displayed with a ball, the same scene without a ball was displayed. This allows the detection of random guessing within the subject responses, which would reveal itself through high error-rates. Three scenes (dirt road, dwelling zone I, dwelling zone II) were used and the ball moved to multiple positions on or above ground in the range of 10m to 20m. The arbitrary ball positions prevented learning effects, because the connection between scenery and position of the ball was not given. After every image, a white screen with a centered 2 second countdown was shown, to allow time for re-fixation and to prepare for the next image. The participants were told to press the right shift key as fast as possible if they recognized a ball within the scene, and to press the left key if they didn't. The keys were reversed for left-handed participants. All participants had to pass a training phase containing 16 images to learn how to operate the keyboard interface.

Experiment 1 The first experiment was designed to evaluate the impact on search-and-find tasks: if the object to be found was enhanced by the suggested monocular depth cue framework. 13 Participants (11 male and 2 female) between the ages of 22 and 32 with normal or corrected to normal vision took part in the experiment. They were not aware of the goal of the experiment. To trigger a guided top-down search, they received the additional information that the ball was to be found only within sharp image areas. After the training phase, 5 different ball positions enhanced by one of the proposed renderings within the 3 scenes were shown (see Figure 12). The ball was presented at least once in all four image quadrants in every scene. The same number of ball and no-ball images was shown. All participants performed two runs, resulting in 240 captured response times. The images were shuffled randomly and resorted to prevent presenting the same scenery subsequently.

Evaluation The evaluated error rate was between 2.5% and 17.0% (mean 7,9%) indicating, that all subjects tried to find the ball, but some had difficulties detecting it. Response times for wrong answers were disregarded. A Bartlett's test and a Kolmogorow-Smirnow test confirmed normal distribution and homogeneity of variances. Subsequently, all response times for finding a ball in a one-way ANOVA were analyzed. This revealed a statistical significance of the chosen rendering on the resulting response time ($F(3,1344) =$

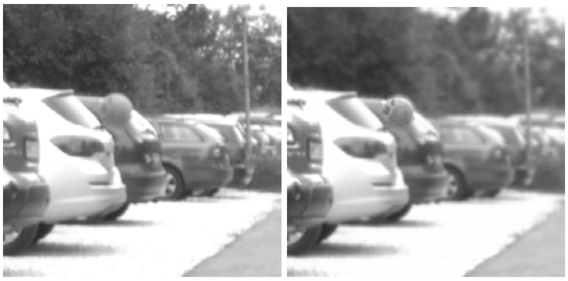


Figure 12: *Enlarged region of Experiment 1, Scene 1 Left: Original; Right: DOF + local contrast enhanced*

$12.01, p < .001$). In addition, the response means of all renderings using Tukey's honestly significant difference criterion ($p < 0.05$) were compared. The mean response time for detecting the target in the unchanged image was $2.46s$ whereas all other renderings showed a statistical significant lower response time (Local Contrast: $2.09s$, depth of field: $1.95s$ and the combination of DOF and Local Contrast: $1.69s$)(see Figure 14). An average speedup of $770ms$ for the last rendering indicates a major impact on detectability of critical objects.

Experiment 2 The second experiment was designed to evaluate the impact on search-and-find tasks if the object to be found was not presented in the enhanced distance of the monocular depth cue framework. As a consequence, only response times of depth of field and the Depth Dependent Local Contrast rendering were evaluated. Five new images rendered with 2 different algorithms out of the same 3 scenes were presented. For this experiment, the focal distance z_0 was set to $100m$, whereas the ball was located at $\approx 10m$. The close position of the ball was chosen to simplify the search task and to avoid wrong answers caused by missed balls (see Figure 13). 13 participants (11 male and 2 female) between the ages of 22 and 32 with normal or corrected to normal vision took part in the second experiment. 7 participants had already taken part in the first experiment. All subjects were unaware of the purpose of the experiment. As an additional instruction, the participants were told to search for the ball in the entire scene, since it might be not within the sharp image area.

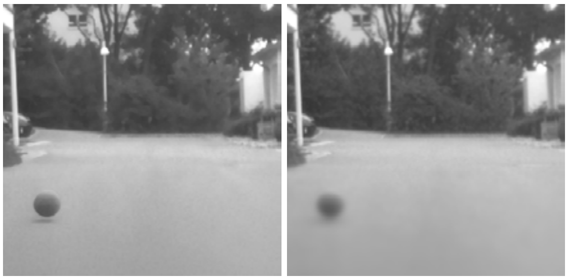


Figure 13: *Enlarged region of Experiment 2, Scene 3 Left: Dampened local contrast; Right DOF*

Evaluation To simplify the second experiment, the ball was presented closer to the subjects, resulting in lower error rates between 1.6% and 8.1% (mean 4,9%). Response times of wrong answers were disregarded. No evidence for irregularities was found in normal distribution and homogeneity of variances. The second ANOVA revealed a significant difference in response times of the Local Contrast Enhanced and DOF rendered images ($F(1,752) = 10.99, p < 0.001$). Tukey's honestly difference criterion indicated significant lower response times ($p < 0.05$) for the Depth-Dependent Local-Contrast rendering. The means of response times were $1.0s$ for the DOF and $0.91s$ for Local Contrast Enhancement (see Figure 14).

Discussion The first experiment indicates that monocular depth cues are capable of simplifying search-and-find tasks. The findings agree with the Feature Integration Theory [TG80] and the Guided Search [Wol94] (see Section 2.2). The staggered response times of the three monocular enhancement renderings suggest that combining multiple cues enhances multiple features for visual search, shortening response times. The mean response time was above $200msec$, indicating that the pop-out effect for the most part was not triggered [TG80]. But the significantly lower response times for rendered images indicate that the serial search was guided towards potential target locations.

The second experiment revealed the impact of renderings with unhelpful focal distances. In this experiment, the target object was much closer, simplifying the task and resulting in shorter response times. As expected, blurring in DOF rendering lowers response times more than the texture dampening in Depth Dependent Local Contrast rendering. For critical applications, one may prefer the second rendering: It still speeds up the search-and-find task, but in the error case a critical object is still detectable in reasonable time.

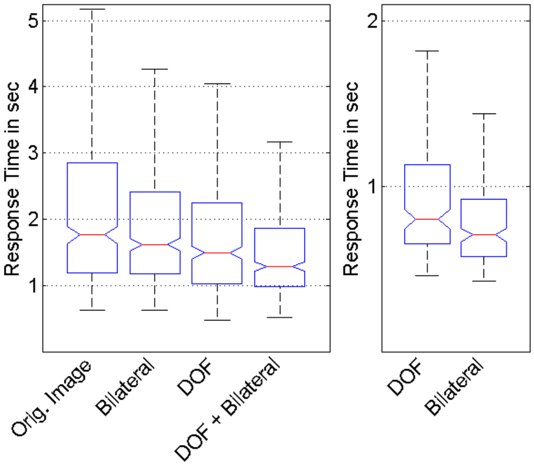


Figure 14: *Left: Box plot of experiment 1 response times; Right: Box plot of experiment 2 response times*

7 Conclusion and Future Work

The presented real-time framework is capable of enhancing and synthesizing multiple monocular depth cues, based on a supplemental disparity map. These cues increase the sensation of depth and support the HVS in the perception of scene arrangement. This is achieved by adding blur from defocus and ambient occlusion as additional depth cues to a sharp input image. along with depth-dependent local contrast enhancement and saturation dampening. Additionally, the disparity map is refined by removing noise, gaps and border bleeding artifacts, which improves the depth cue rendering quality.

Besides conveying disparity map-based distance information, the conducted study reveals the positive impact of monocular depth cues on visual attention and response times in search-and-finds tasks. In particular, the combination of multiple monocular depth cues showed a significantly faster response time in search-and-find tasks.

Future extensions to this framework will be the automatic detection of relevant objects within the presented scene to automatically set the correct focal distance. Ultra-compact cameras with small camera lenses might also benefit from such a framework. A built-in stereo camera would allow a correct simulation of lenses with much bigger apertures.

The next evolutionary step of the framework will be the extension to 3D displays and the coherent combination of binocular and monocular depth cues. Merging both cues might create a highly improved 3D visualization on such devices and have a significant impact on spatial image quality.

References

[Ble05] BLEYER M.: Graph-based surface reconstruction from stereo pairs using image segmentation. *Proceedings of SPIE 5665* (2005), 288–299. 5

[Bra97] BRAINARD D. H.: The Psychophysics Toolbox. *Spatial Vision 10*, 4 (1997), 433–436. 8

[Bur81] BURT P. J.: Fast Filter Transforms for Image Processing. *Computer Graphics and Image Processing 16*, 1 (1981), 20–51. 3

[DSHL10] DAMMERTZ H., SEWTZ D., HANIKA J., LENSCH H.: Edge-avoiding À-Trous wavelet transform for fast global illumination filtering. In *Proceedings of the Conference on High Performance Graphics* (2010), Eurographics Association, pp. 67–75. 3, 4

[FAR07] FATTAL R., AGRAWALA M., RUSINKIEWICZ S.: Multiscale shape and detail enhancement from multi-light image collections. *ACM Transactions on Graphics 26*, 3 (July 2007), 51. 3

[Fat09] FATTAL R.: Edge-avoiding wavelets and their applications. *ACM Transactions on Graphics 28*, 3 (July 2009), 1. 3

[GR10] GEHRIG S. K., RABE C.: Real-Time Semi-Global Matching on the CPU. *Image Rochester NY* (2010), 1–8. 1, 4, 8

[HCOB10] HELD R. T., COOPER E. A., O'BRIEN J. F., BANKS M. S.: Using blur to affect perceived distance and size. *ACM Transactions on Graphics 29*, 2 (Mar. 2010), 1–16. 2, 6

[HDL11] HANIKA J., DAMMERTZ H., LENSCH H.: Edge-Optimized À-TrousWavelets for Local Contrast Enhancement with Robust Denoising. In *Pacific Graphics* (2011). 3

[HKMMT89] HOLSCHNEIDER M., KRONLAND-MARTINET R., MORLET J., TCHAMITCHIAN P.: *A real-time algorithm for signal analysis with the help of the wavelet transform*. Springer-Verlag, 1989, pp. 289–297. 3

[HS05] HUBONA G. S., SHIRAH G. W.: Spatial Cues in 3D Visualization. *Intelligence*, 2 (2005), 104–128. 6

[KT02] KOSARA R., TSCHELIGI M.: Useful Properties of Semantic Depth of Field for Better F + C Visualization. *Image (Rochester, N.Y.)* (2002). 2, 8

[LB00] LANGER M. S., BÜLTHOFF H. H.: Depth discrimination from shading under diffuse lighting. *Perception 29*, 6 (2000), 649–660. 3

[LCD06] LUFT T., COLDITZ C., DEUSSEN O.: Image enhancement by unsharp masking the depth buffer. *ACM Transactions on Graphics 25*, 3 (July 2006), 1206. 3

[LFIH09] LAMBOOIJ M., FORTUIN M., IJSSELSTEIJN W. A., HEYNDERICKX I.: Measuring visual discomfort associated with 3D displays. *Proceedings of SPIE 7237*, 1 (2009), 72370K–72370K–12. 2

[Mat09] MATHER G.: *Foundations of Sensation and Perception*, vol. 2. Psychology Press, 2009. 2

[Nag91] NAGATA S.: How to reinforce perception of depth in single two-dimensional pictures. *Pictorial communication in virtual and real environments* (1991), 527–545. 2

[QTA09] QINGXIONG YANG, TAN K.-H., AHUJA N.: Real-time O(1) bilateral filtering. *IEEE Conference on Computer Vision and Pattern Recognition (2009)*, 1 (2009), 557–564. 3

[RHFL10] REICHELT S., HAUSSLER R., FÜTTERER G., LEISTER N.: Depth cues in human visual perception and their realization in 3D displays. *Most 76900B*, 0 (2010), 76900B–76900B–12. 1

[RHM11] REMPEL A., HEIDRICH W., MANTIUK R.: *The Role of Contrast in the Perceived Depth of Monocular Imagery*. Tech. rep., Tech. Rep. TR-2011-07, The University of British Columbia, 2011. 1, 2, 6

[Swe10] SWELDENS W.: The lifting scheme for wavelet biframes: theory, structure, and algorithm. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 19*, 3 (Mar. 2010), 612–24. 3

[TG80] TREISMAN A., GELADE G.: A feature-integration theory of attention. *Cognitive Psychology 12*, 1 (1980), 97–136. 2, 9

[TM98] TOMASI C., MANDUCHI R.: Bilateral filtering for gray and color images. *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)* (1998), 839–846. 3

[TML*91] TROSCIANKO T., MONTAGNON R., LE CLERC J., MALBERT E., CHANTEAU P. L.: The role of colour as a monocular depth cue. *Vision Research 31*, 11 (1991), 1923–1929. 2, 7

[Tre62] TREISMAN A.: Binocular rivalry and stereoscopic depth perception. *The Quarterly Journal Of Experimental Psychology 14*, 1 (1962), 23–37. 2

[Wol94] WOLFE J. M.: Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review 1*, 2 (1994), 202–238. 3, 9

[YTY*11] YU Z. Y., THORPE C., YU X., GRAUER-GRAY S., LI F., JINGYI: Dynamic Depth of Field on Live Video Streams: A Stereo Solution. In *Computer Graphics International* (2011). 3

[YWY10] YU X., WANG R., YU J.: Real-time Depth of Field Rendering via Dynamic Light Field Generation and Filtering. *eeci.sudel.edu 29*, 7 (2010). 3