Fast Gaze-Contingent Optimal Decompositions for Multifocal Displays

OLIVIER MERCIER, Université de Montréal and Oculus Research
YUSUFU SULAI, Oculus Research
KEVIN MACKENZIE, Oculus Research
MARINA ZANNOLI, Oculus Research
JAMES HILLIS, Oculus Research
DEREK NOWROUZEZAHRAI, McGill University
DOUGLAS LANMAN, Oculus Research

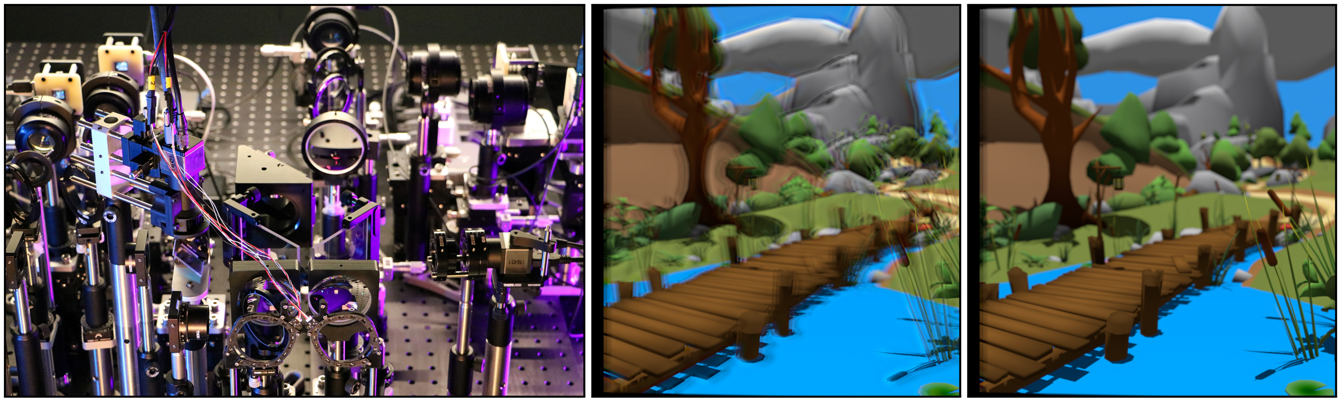

(a) Multifocal Testbed with Eye and Accommodation Tracking (b) Eye Movement without Correction (c) Eye Movement with Correction

Fig. 1. Multifocal displays require a decomposition of the scene onto the display planes, which often assumes perfect alignment of the viewer with the system. Otherwise, parallax introduced by eye rotation and head offsets relative to the display may result in misregistration between the images, creating halos and increasing blurriness. (a) In this paper, we present the first multifocal display with eye tracking and accommodation measurement. (b-c) We also introduce the first computationally efficient optimal decomposition algorithm, enabling interactive content that utilizes eye tracking to directly maintain image alignment. Both our hardware and algorithmic contributions are necessary steps towards better understanding the practical requirements of multifocal displays.

As head-mounted displays (HMDs) commonly present a single, fixed-focus display plane, a conflict can be created between the vergence and accommodation responses of the viewer. Multifocal HMDs have long been investigated as a potential solution in which multiple image planes span the viewer's accommodation range. Such displays require a scene decomposition algorithm to distribute the depiction of objects across image planes, and previous work has shown that simple decompositions can be achieved in real-time. However, recent optimal decompositions further improve image quality, particularly with complex content. Such decompositions are more computationally involved and likely require better alignment of the image planes with the viewer's eyes, which are potential barriers to practical applications.

Our goal is to enable interactive optimal decomposition algorithms capable of driving a vergence- and accommodation-tracked multifocal testbed. Ultimately, such a testbed is necessary to establish the requirements for the practical use of multifocal displays, in terms of computational demand and hardware accuracy. To this end, we present an efficient algorithm for optimal decompositions, incorporating insights from vision science. Our method is amenable to GPU implementations and achieves a three-orders-of-magnitude speedup over previous work. We further show that eye tracking can be used for adequate plane alignment with efficient image-based deformations, adjusting for both eye rotation and head movement relative to the display. We also build the first binocular multifocal testbed with integrated eye tracking and accommodation measurement, paving the way to establish practical eye tracking and rendering requirements for this promising class of display. Finally, we report preliminary results from a pilot user study utilizing our testbed, investigating the accommodation response of users to dynamic stimuli presented under optimal decomposition.

CCS Concepts: • **Hardware → Displays and imagers**; • **Computing methodologies → Rendering**;

Additional Key Words and Phrases: computational displays, multifocal displays, multiview rendering, vergence-accommodation conflict

ACM Reference format:
Olivier Mercier, Yusufu Sulai, Kevin MacKenzie, Marina Zannoli, James Hillis, Derek Nowrouzezahrai, and Douglas Lanman. 2017. Fast Gaze-Contingent Optimal Decompositions for Multifocal Displays. *ACM Trans. Graph.* 36, 6, Article 237 (November 2017), 15 pages.
https://doi.org/10.1145/3130800.3130846

0730-0301/2017/11-ART237 $15.00
https://doi.org/10.1145/3130800.3130846

1 INTRODUCTION

More than a century of research into stereoscopic and multiscopic displays has worked toward an accurate reproduction of the three-dimensional world [Urey et al. 2011]. Today's binocular head-mounted displays (HMDs) offer an accessible means to resolve persistent deficiencies of 3D displays, achieving accurate reproduction of motion parallax, as well as depicting 360-degree imagery enveloping the viewer. However, modern HMDs do not correctly reproduce all natural depth cues available to the human visual system. In particular, due to the fixed optical focus of current HMDs, the retinal blur created by out-of-focus scene components is synthesized inaccurately. Correspondingly, the use of HMDs may lead to vergence-accommodation conflict (VAC), which biases perceived depth [Watt et al. 2005], and has been linked to visual fatigue and visual discomfort [Hoffman et al. 2008; Shibata et al. 2011].

Volumetric displays are one solution to alleviate the issues associated with VAC. This widely studied class of glasses-free 3D display can depict accurate retinal defocus blur by synthesizing an additive volume of modulated light sources [Blundell and Schwartz 1999]. Rolland et al. [2000] were among the first to propose a *multifocal* volumetric HMD, capable of generating multiple virtual image planes spanning a range of depths. By incorporating an eyepiece, Rolland et al. demonstrated that a compact multifocal HMD can reproduce a volume of light sources extending throughout a viewer's accommodative range.

As first described by Akeley et al. [2004], a *scene decomposition* must be performed to distribute virtual objects across the various image planes to produce near-correct retinal defocus blur. Specifically, they introduced a *linear blending* algorithm to divide the depiction of objects across the nearest enclosing image planes. In this paper, we significantly expand upon the capabilities and practicality of the more recent *optimized blending* algorithm of Narain et al. [2015], which is better suited for depicting occlusions and reflections, as well as accurate retinal defocus blur.

Despite nearly two decades of investigation, multifocal displays remain potentially unsuitable for practical applications, primarily due to two unresolved issues. First, computing high-quality scene decompositions is inefficient, as evidenced by the minutes-long run times reported by Narain et al. and other more complex decomposition approaches [Matsuda et al. 2017]. Second, all existing multifocal display decompositions assume a single, fixed viewpoint. As shown in Figure 1, this can cause the projections of the image planes to be misregistered on the retina if the position and direction of the viewer's eye are not exactly the same as those assumed during the scene decomposition, which can significantly reduce image quality.

In this paper, we present solutions to these long-standing challenges. First, we show that high-quality scene decompositions can be computed at interactive frame rates, leveraging insights from numerical methods and perceptual science. Second, we demonstrate how eye tracking measurements can be efficiently used to correct for eye movements. We apply these methods to drive the first multifocal testbed with integrated eye tracking and accommodation measurement, demonstrating the feasibility of gaze tracking within a multifocal display. Our hardware and algorithmic contributions enable the use and study of multifocal displays with dynamic content, and open the way to a better understanding of practical requirements for multifocal displays.

1.1 Contributions

- We achieve a three-orders-of-magnitude improvement in computation time relative to state-of-the-art optimal scene decompositions, reaching interactive performance through a different numerical method that is provably stable and amenable to GPU implementations;
- From prior perceptual studies, we derive a modified decomposition algorithm to optimize the retinal defocus blur gradient, as well as the retinal defocus blur itself, further accelerating computations;
- We develop an efficient algorithm to correct for eye movements detected after scene decomposition, showing eye tracking can be used to solve the misalignments due to eye rotation and head movements relative to the display;
- We develop the first adaptive multifocal system with integrated vergence and accommodation eye tracking, supporting three adjustable-focus displays per eye. This system is the first to support dynamic content, leveraging our efficient decomposition method and eye movement correction;
- We report preliminary results from empirically measured accommodation responses that show, for the first time, that optimal decomposition correctly drives accommodation.

2 RELATED WORK

2.1 Driving Accommodation with HMDs

Volumetric displays, through their evolution into multifocal HMDs, are not the only means to address the vergence-accommodation conflict. As reviewed by Kramida [2016], there exists a broad spectrum of such designs, spanning comparatively modest modifications (e.g., varifocal HMDs) to nearly complete overhauls (e.g., near-eye light field displays). In this context, multifocal displays present a moderate, but technically challenging, progression, adding display elements and computational complexity in exchange for extending the supported accommodation range.

With any HMD, the viewer's pupil must remain within the designed eye box. Correspondingly, to mitigate VAC, the stimulus to accommodation should be depicted correctly over this limited region. A direct solution is offered by near-eye light field displays, faithfully reproducing wavefronts of natural scenes for perspectives within the eye box. Lanman and Luebke [2013], Hua and Javidi [2014], and Song et al. [2014] demonstrate microlens-based architectures for this purpose, whereas Huang et al. [2015] apply multilayer LCDs; however, in all these examples, resolution remains limited with current display technologies. Similarly, Konrad et al. [2017] recently showed that accommodation-invariant displays can be used to alleviate the VAC problem, but again at the cost of a tradeoff in resolution.

Another approach to mitigate VAC is offered by varifocal HMDs in which the virtual image distance is varied to match the vergence distance reported by an eye tracking system. This concept has been explored using electronically-tunable lenses, in part, by Liu et al. [2010], Johnson et al. [2016], Konrad et al. [2016] and Padmanaban et al. [2017]. Relative to near-eye light field displays, varifocal HMDs can offer higher resolutions and larger field of views [Dunn et al. 2017], but require tunable optics that must rely on accurate eye tracking. In addition, retinal defocus blur can only be rendered synthetically. Although eye tracking can improve the rendered blur [Kellnhofer et al. 2016], it cannot be properly reproduced optically as the viewer accommodates.

In contrast to light field displays, volumetric displays utilize an additive superposition of display elements located at different depths. This construction raises natural questions regarding the density of planes required for accurate depictions. Early research by Rolland et al. [2000] suggests as many as 14 layers would be required to support one arcminute resolution (i.e., 20/20 visual acuity) over an accommodation range of two diopters (e.g., from 50 cm to optical infinity). More recently, MacKenzie et al. [2012; 2010] established that a coarser separation between layers, as wide as 0.6 to 1.0 diopters, is sufficient to correctly drive accommodation, requiring only four planes for a two-diopter accommodation range. With this reduced requirement on the number of planes, recent research has focused on identifying practical hardware to support a limited number of planes. For example, Love et al. [2009] apply fast-switching birefringent optics, Hu et al. [2014] investigate deformable membrane mirrors, and Llull et al. [2015] incorporate electronically-tunable lenses. Matsuda et al. [2017] use a spatial light modulator to create non-planar focal surfaces, which more closely adapts the few layers to the scene content but increases processing times. In contrast to these works, our efforts are focused on unresolved, yet fundamental, questions of maintaining image quality under natural eye movements and reducing algorithmic complexity; as a result, our system is optimized as a perceptual testbed, with these prior works showing potential paths toward compact form factors.

2.2 Multifocal Displays, Blur, and Accommodation

Similar to binocular disparity, the magnitude of retinal defocus blur varies monotonically with the separation between an object and the point of focus. Therefore, this retinal blur provides another cue to perceived depth [Held et al. 2012]. Based on statistics of natural scenes and the properties of the human visual system, Burge and Geisler [2011] found that reliable estimates of depth could be obtained from retinal defocus blur alone. This result is consistent with a growing body of psychophysical work showing the importance of retinal defocus blur for depth perception. The tilt-shift illusion provides a convincing example: artificial blur, as added to a photograph or a computer-generated image, can dramatically affect perceived scale [Held et al. 2010; Vishwanath and Blaser 2010]. Moreover, recent studies have employed multifocal displays to show that retinal defocus blur, in isolation, is sufficient to recover depth ordering. Critically, this finding was supported only when retinal defocus blur was created by the optics of the eye, as opposed to synthetically rendered on a conventional display [Zannoli et al. 2016].

As discussed above, multiple HMD architectures have been proposed to depict retinal defocus blur. We emphasize that those relying on rendered defocus blur alone, such as varifocal displays, may not appear correctly or respond quickly enough to changes in the viewer's accommodative state due to unmodeled aspects of the eye or system latency, respectively. Multifocal displays avoid these concerns by creating retinal defocus blur through optical means (i.e, resulting from physiological changes within the eye). MacKenzie et al. [2010] confirmed that the linear blending algorithm of Akeley et al. [2004] approximates retinal defocus blur. Specifically, the accommodative state producing maximum retinal contrast occurs when focusing at the correct depth.

Others have investigated alternative decomposition algorithms. Wu et al. [2016] use a saliency map to optimize the display plane locations for linear blending. Liu and Hua [2010] advocate a nonlinear weighting to maximize the modulation transfer function (MTF) for objects within the display volume. Subsequent analysis by Ravikumar et al. [2011] reported a preference for linear blending over nonlinear weighting when considering biologically plausible metrics of image quality and properties of natural scenes. In later work, Narain et al. [2015] demonstrated that, despite subtle differences between these methods, no prior scene decomposition suppresses salient artifacts at occlusion boundaries and reflections. This deficiency motivated the development of their optimized decomposition, directly using the reconstructed focal stack to compute the displayed images. We build on their work: to be practical, multifocal displays must support artifact-free viewing, but must also demonstrate real-time frame rates with unconstrained eye movements.

As described above, current evaluations of retinal defocus blur depictions have focused on the maximization of retinal contrast. Yet, as with rendered blur, it is not enough to correctly replicate this retinal blur itself, but also its variation as the eye accommodates (i.e., the retinal defocus blur gradient). Current evidence suggests that the accommodative system may exploit the temporal change of contrast that is induced through accommodative microfluctuations. This signal may be applied to resolve the direction of an accommodative stimulus (i.e., whether it is closer or further than the plane of focus) [MacKenzie et al. 2010; Metlapally et al. 2014]. Similarly, others have identified the retinal defocus blur gradient as a critical feedback signal to the accommodative response [Alpern 1958; Kotulak and Schor 1986; Owens 1980]. To our knowledge, we are the first to directly use the retinal defocus blur gradient produced by multifocal displays within the optimization formulation of the scene decomposition.

3 INTERACTIVE SCENE DECOMPOSITION

Any practical application of a multifocal display requires decomposing a virtual scene onto the layers of the display. In order to do so both accurately and efficiently, we formulate the scene decomposition as an optimization problem with an efficient numerical solution. We begin with a simplified formulation of the problem in the theoretical case of a fixed eye, and only later generalize the formulation to support eye movements (Section 4.1). We assume monochromatic (i.e., grayscale) images, but the same formulation can be independently applied to any number of color channels.

We write scalars in math italic (e.g. x, D), n-d points/vectors in boldface italic (e.g. $\boldsymbol{b}$), and matrices/sub-matrices in sans serif (e.g. K). Depending on context, vector and matrix subscripts may either refer to individual scalar entries or to contained sub-vector/sub-matrices.

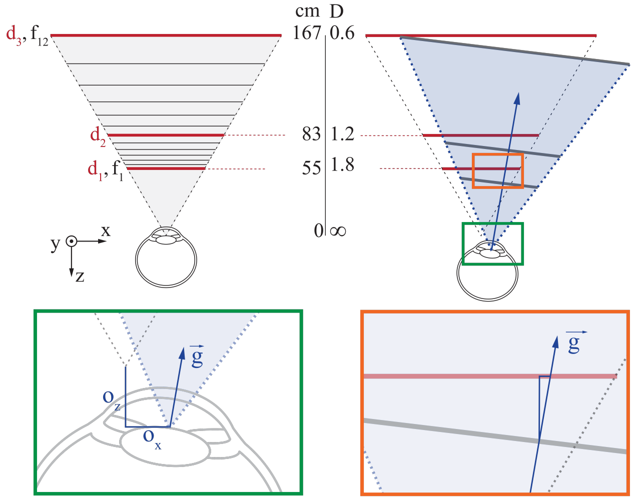


Fig. 2. (*Top left*) Multifocal display diagram, showing the idealized configuration of an eye perfectly aligned with the display view frustum, for $D = 3$ displays and $F = 12$ focal slices. Distances d_i and f_i are in diopters. (*Top right and insets*) In practice, eye rotations and head movements relative to the display cause the entrance pupil, at position $\mathbf{o}$ and with unit gaze direction $\mathbf{g}$, to become misaligned with the displays, distorting the perceived images.

3.1 Optimal Decomposition

Figure 2 provides a schematic of a three-plane multifocal display: a stack of optically additive display planes are positioned at distances d_i from the eye, for $i \in \{1, \dots, D\}$. The optical axis of the system is defined so as to intersect the displays orthogonally at their centers. We additionally assume, for simplicity and without loss of generality, that each display has a square resolution of $N \times N$ pixels. We compute *focal slices* to form a *focal stack*, modeling the retinal defocus blur when a viewer is accommodated at focal distances f_i, for $i \in \{1, \dots, F\}$. All distances are in diopters, and we refer to the display planes and focal slices by their distance from the eye.

When the viewer accommodates at f_i, the superposition of the display images should be as close as possible to the corresponding focal slice. This requirement for multifocal display image formation can be cast as a minimization problem [Narain et al. 2015]. Motivated by this formulation, we propose a novel solution that differs significantly from previous work by its interpretation, efficiency, and stability.

We formalize the optimal decomposition of a scene onto D display planes as the solution of the following constrained block-matrix system:

$$\underset{\mathbf{x}}{\operatorname{argmin}} \left\| \overbrace{\begin{pmatrix} \mathsf{K}_{11} & \dots & \mathsf{K}_{1D} \\ \vdots & \ddots & \vdots \\ \mathsf{K}_{F1} & \dots & \mathsf{K}_{FD} \end{pmatrix}}^{\mathsf{K}} \overbrace{\begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_D \end{pmatrix}}^{\mathbf{x}} - \overbrace{\begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_F \end{pmatrix}}^{\mathbf{b}} \right\| \tag{1}$$

$$\text{such that } 0 \le \mathbf{x}_i \le 1, \ \forall i. \tag{2}$$

Here, $\|\cdot\|$ is the Euclidean norm and we define:

- $\mathsf{K}_{fd} \in \mathbb{R}^{N^2 \times N^2}$ as the *kernel sub-matrix* for focal slice f and display d (see below),
- $\mathbf{x}_d \in \mathbb{R}^{N^2}$ as the unknown optimal pixel intensities for display d, and
- $\mathbf{b}_f \in \mathbb{R}^{N^2}$ as the known pixel values of focal slice f.

The pixels of every display and focal slice are linearized to form vectors $\mathbf{x}$ and $\mathbf{b}$. Each column of a kernel sub-matrix K_{fd} corresponds to the discretized point spread function (PSF) of a given pixel on display d viewed while focusing at distance f, which we refer to as the *kernel* of this pixel. Each column of the entire kernel matrix K therefore comprises the kernels of a displayed pixel as focus spans that of the whole focal stack. The constraints in Equation 2 are necessary to model the finite, nonnegative range of display intensities.

The system in Equation 1 can be solved using the normal equations

$$(\mathsf{K}^\top \mathsf{K})\, \mathbf{x} = \mathsf{K}^\top \mathbf{b}\,. \tag{3}$$

Solving directly for $\mathbf{x}$ in Equation 3 will not generally give a solution that satisfies the constraints, but it provides a way of approaching the constrained solution. We thus study the unconstrained normal equations here, and will discuss constraints further in Section 3.3.

It is useful to expand the left-hand side of Equation 3 as

$$\begin{pmatrix} \sum \mathsf{K}_{i1}^\top \mathsf{K}_{i1} & \dots & \sum \mathsf{K}_{i1}^\top \mathsf{K}_{iD} \\ \vdots & \ddots & \vdots \\ \sum \mathsf{K}_{iD}^\top \mathsf{K}_{i1} & \dots & \sum \mathsf{K}_{iD}^\top \mathsf{K}_{iD} \end{pmatrix} \equiv \begin{pmatrix} \mathsf{C}_{11} & \dots & \mathsf{C}_{1D} \\ \vdots & \ddots & \vdots \\ \mathsf{C}_{D1} & \dots & \mathsf{C}_{DD} \end{pmatrix}.$$

We can similarly expand the right-hand side as:

$$\mathsf{K}^\top \mathbf{b} = \begin{pmatrix} \sum_{i=1}^{F} \mathsf{K}_{i1}^\top \mathbf{b}_i \\ \vdots \\ \sum_{i=1}^{F} \mathsf{K}_{iD}^\top \mathbf{b}_i \end{pmatrix} \equiv \begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_D \end{pmatrix}.$$

This allows us to re-write Equation 3 more concisely as

$$\mathsf{C}\,\mathbf{x} = \mathbf{r} \tag{4}$$

with $\mathsf{C} \in \mathbb{R}^{DN^2 \times DN^2}$ and $\mathbf{r} \in \mathbb{R}^{DN^2}$.

Recalling that columns of K_{ij} are pixel kernels for a single focal slice, the $(a, b)^{\text{th}}$ element of C_{ij} corresponds to the sum of correlations of pixel a's kernel in display i and pixel b's kernel in display j. Similarly, the a^{th} element of $\mathbf{r}_i$ is the sum of the correlations of pixel a's kernel in display i with each focal slice.

Rewritten this way, we see that $\mathsf{C}\,\mathbf{x}$ is a discrete convolution of the displayed images with summed cross-correlated kernels, reducing the optimal decomposition problem to that of discrete deconvolution.

We can compute the kernel for any displayed pixel directly and accurately with a virtual scene consisting of a plane positioned at the same distance as the physical display of that pixel. If we discretize the plane geometry with an $N \times N$ grid, "activate" (i.e., render with unit intensity) the grid element aligned with the pixel of interest, and then render the resulting focal stack, we can compute the pixel's kernels across the focal stack.

If we represented the eye with a physically accurate model, the kernel of each pixel would need to be computed independently since the PSF of a human eye depends non-linearly on position and accommodation distance. All the matrix and vector elements in C and $\mathbf{r}$ would therefore need to be computed independently, requiring significant storage and processing: for example, for 8-bit monochromatic images with $N = 1024$ and $D = 3$, matrix C would require 9 TB of memory. As such, we adopt several carefully chosen approximations to afford a practical, yet accurate, formulation.

3.2 A Thin Lens Approximation of Defocus Blur

To simplify computations, we approximate the optics of the human eye as an ideal thin lens system. This approximation is common in graphics [Potmesil and Chakravarty 1981] and has also been adopted in the vision science community, particularly for multifocal displays [Narain et al. 2015]. Our derivation applies an additional small angle (paraxial) approximation, which is valid since display and focal distances $1/d_i$ and $1/f_i$ are large compared to the pupil diameter ϕ (given in meters). With these approximations, kernels obtained from the thin lens model are spatially invariant and constant over a circular support. We can also express the image formation in terms of tangent angles, i.e., a point in focus on focal slice f positioned $\mathbf{p}$ meters away from the optical axis maps to $\mathbf{p}/(1/f)$ in image space.

For focal slice f, the kernel $k(\mathbf{p})$ of a pixel at position $\mathbf{p}_0$ on display d is

$$k(\mathbf{p}) = \operatorname{circ}\left(\frac{(\mathbf{p}-\mathbf{p}_0)}{(\phi/2)\,|d-f|}\right), \tag{5}$$

where $\mathbf{p}$ and $\mathbf{p}_0$ are given in tangent angles, and $\operatorname{circ}(x)$ is 1 inside a unit disk and 0 elsewhere. After rasterization, kernels are normalized to have unit area.

To avoid complications at image boundaries, we add a band of black pixels around the image so that pixels near the boundary can use the same kernels as inner pixels. The necessary width of this band is easily evaluated from the maximum kernel radius. The value of these black pixels is never changed, and they are removed after optimization. Note that this approach is not applicable to the method of Narain et al., since it changes the frequency information of the images. In our implementation of their method, which we require later for comparison, we use a band of replicated edge pixels around the images with a smooth falloff, as described in their paper.

These approximations drastically simplify the computation of our matrix system and allow us to recast Equation (3) in terms of simple image operations. Columns of each sub-matrix K_{fd} now all have the same structure, differing only by a translation. As such, each sub-matrix K_{fd} can be replaced by a single image $\overline{\mathsf{K}}_{fd} \in \mathbb{R}^{N\times N}$ of the kernel for the display's central pixel. All subsequent matrix operations can also be written in terms of kernel images, instead of using large, impractical kernel matrices. We arrive at

$$\overline{\mathsf{C}}_{ij} = \sum_{f=1}^{F} \overline{\mathsf{K}}_{fi} * \overline{\mathsf{K}}_{fj} \text{ and } \overline{\mathbf{r}}_i = \left(\sum_{f=1}^{F} \overline{\mathsf{K}}_{fi}\right) * \mathbf{b}_i\ , \tag{6}$$

and we need to solve the system $\overline{\mathsf{C}} \star \overline{\mathbf{x}} = \overline{\mathbf{r}}$, or explicitly

$$\begin{pmatrix} \overline{\mathsf{C}}_{11} & \dots & \overline{\mathsf{C}}_{1D} \\ \vdots & \ddots & \vdots \\ \overline{\mathsf{C}}_{D1} & \dots & \overline{\mathsf{C}}_{DD} \end{pmatrix} \star \begin{pmatrix} \overline{\mathbf{x}}_1 \\ \vdots \\ \overline{\mathbf{x}}_D \end{pmatrix} = \begin{pmatrix} \overline{\mathbf{r}}_1 \\ \vdots \\ \overline{\mathbf{r}}_D \end{pmatrix}, \tag{7}$$

where the correlation ($*$) and the convolution ($\star$) of a matrix of images with a vector of images is defined as a regular scalar matrix multiplication, replacing multiplications of scalars by the corresponding image operation. The addition of images is computed pixelwise. Note that since the circular kernels of Equation 5 are symmetric, the correlations are convolutions.

Although conceptually similar to the scalar matrix formulation in Equation 4, the image formulation we obtain in Equation 7 is significantly more compact. The terms $\overline{\mathsf{C}}_{ij}$ and $\sum_{f=1}^{F} \overline{\mathsf{K}}_{fi}$ can be precomputed once as images and easily fit in memory; the matrix of images $\overline{\mathsf{C}}$ now only requires 9 MB of memory.

3.3 Solving the Constrained Minimization

Even if unusable in practice, the full scalar matrix formulation in Equation 4 allows us to reason about using numerical linear algebra to solve the system more efficiently than in previous work. We detail our optimal decomposition solver, relying on over-relaxed Jacobi iterations [Burden and Faires 2011].

Let λ_d be the scalar value of the central pixel of image $\overline{\mathsf{C}}_{dd}$, and let $\boldsymbol{\lambda}^{-1} = (1/\lambda_1, \dots, 1/\lambda_D)^\top$. Given the approximate solution vector $\overline{\mathbf{x}}^{(k)}$ obtained during the k^{th} Jacobi iteration, we can write the next Jacobi iteration $\overline{\mathbf{x}}^{(k+1)}$ of the image matrix system (Equation 7) in the compact image matrix notation of Section 3.2 as

$$\overline{\mathbf{x}}^{(k+1)} = (1-\alpha)\,\overline{\mathbf{x}}^{(k)} + \alpha\,\boldsymbol{\lambda}^{-1}\left(\overline{\mathbf{r}} + \boldsymbol{\lambda}\,\overline{\mathbf{x}}^{(k)} - \overline{\mathsf{C}}\,\overline{\mathbf{x}}^{(k)}\right)\ , \tag{8}$$

where α is a positive scalar, and the product of scalars $\boldsymbol{\lambda}^{-1}$ with a vector of images simply scales each image by the corresponding scalar entry. We leverage the fact that kernels are non-negative to prove (see Appendix A) that this iterative process is guaranteed to converge when

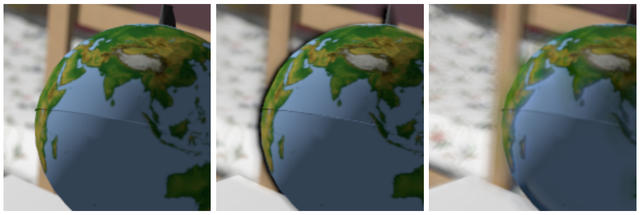
$$0 < \alpha < \underbrace{1\Bigg/\left(\sum_{d=1}^{D}\left(\frac{F}{\lambda_d}\right)\right)}_{\hat{\alpha}}\ . \tag{9}$$

Empirically, $\alpha = 0.75\,\hat{\alpha}$ yields good results in all of our tests, but a more comprehensive analysis of the system could lead to more insights on optimal α settings.

The only remaining step is to deal with constraints (Equation 2). To do so, we simply clamp the pixels of images $\overline{\mathbf{x}}$ to 0 and 1 after each Jacobi iteration. Although simplistic, this projection step does not impact the convergence guarantee (Appendix A), is easy to implement, and yields consistently good results in practice (see Section 6).

4 PRACTICAL CONSIDERATIONS

The Jacobi iterations of Section 3 provide an efficient way of computing the scene decomposition, but do not fully solve the two main issues of current HMDs discussed at the beginning of this paper. This section discusses the correction of errors caused by eye movements, the modification of the objective function to further improve convergence speed, and GPU implementation details. The final algorithm including these modifications is summarized in Algorithm 1.



(a) Ground Truth (b) Linear Blending (c) Optimal Decomposition

Fig. 3. Small eye offsets have significant effects for all multifocal decomposition methods, most noticeably near large depth discontinuities. For linear blending (b), the errors often appear as salient dark bands across edges. For optimal decomposition (c), the errors show more subtly as additional blur at the edges. The supplementary video also shows the effects of misalignments.

4.1 Eye Tracker Deformation

The assumption that the pupil of the user is always at the exact position assumed by the scene decomposition is not easily maintained in practice. This may create plane misalignments that can significantly impair the quality of the perceived images, as shown in Figure 3 and in the accompanying video. This problem is still present in very recent multifocal systems [Matsuda et al. 2017], and diminishes the impact of new advances in multifocal display technology on practical applications.

Even when trying to constrain the position of a user in a display system, for instance using a bite bar, eye rotations move the position of the pupil relative to the displays because the center of rotation of the eye is located behind the pupil. This type of misalignment is predictable (using eye dimensions from an average viewer) and can be alleviated geometrically without eye tracking. In effect, plane misalignment errors are most noticeable near the region fixated by the user. As proposed by Akeley [2004], we can maintain a localized plane alignment by rendering each pixel on given lines of sight from the assumed viewpoint of a rotated pupil. However, this approach requires rendering the scene from multiple viewpoints (one per line of sight), which breaks the assumptions of Section 3.2, prevents us from leveraging modern single-viewpoint oriented hardware, and ultimately hinders an efficient implementation of our decomposition. Note that an approximation of this behavior can be obtained by using the center of rotation of the eye as the center of projection of the camera [Akeley 2004], but we do not use this approximation.

More importantly, a local alignment strategy that only corrects for eye rotations still assumes the head of the viewer is perfectly static within the display device. This requirement may be too constraining for practical applications, since users constantly move their head slightly when looking into benchtop systems, and HMDs cannot be perfectly fixed to a user's head. Therefore, because of the higher dimensionality and possibly larger amplitude of viewpoint displacements in practical applications, misalignments are not easily predictable.

We show here how eye tracking can be used to correct for such eye movements. The eye tracker gives the position of the pupil and gaze direction of the user relative to the origin and direction assumed by the decomposition. First, we offset the virtual camera of the renderer to match the eye-tracked position and gaze. The scene decomposition is then carried out normally, but since the displays are now tilted and shifted with respect to the new frame of reference of the virtual camera (Figure 2), we cannot simply show the decomposed images on the displays.

The display misalignments can be corrected with a simple image-space deformation of the image computed by the decomposition. This transformation is obtained by directly computing the mapping between the physical pixels of the displays and the pixels of the virtual images used for the decomposition (respectively red and blue planes in Figure 2). Let $\mathbf{n} = (n_x, n_y)^\top \in [-1, 1]^2$ be the normalized coordinate of a pixel, $\mathbf{g} = (g_x, g_y, g_z)^\top$ and $\mathbf{o} = (o_x, o_y, o_z)^\top$ be respectively the measured gaze direction and eye offset, and let $\mathbf{t} = (t_x, t_y) = (\tan(\text{fov}_x/2), \tan(\text{fov}_y/2))$. The mapping from a physical pixel to a decomposed image is given explicitly by

$$\mathbf{n} \mapsto \frac{\mathsf{M}_1\,\mathbf{n} + \mathbf{v}_1}{\mathsf{M}_2\,\mathbf{n} + \mathbf{v}_2} \tag{10}$$

where $\mathsf{M}_1, \mathsf{M}_2 \in \mathbb{R}^{2x2}$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2$ are defined as

$$\begin{aligned}
\mathsf{M}_1 &= \begin{pmatrix} d\,g_z\,t_x & 0 \\ -d\,g_x\,g_y\,t_x & d\,t_y\,(1-g_y^2) \end{pmatrix} \\
\mathsf{M}_2 &= \begin{pmatrix} -t_x^2\sqrt{1-g_y^2}\,d\,g_x & -t_x\,t_y\,\sqrt{1-g_y^2}\,d\,g_y \\ -t_y\,t_x\,\sqrt{1-g_y^2}\,d\,g_x & -t_y^2\,\sqrt{1-g_y^2}\,d\,g_y \end{pmatrix} \\
\mathbf{v}_1 &= \begin{pmatrix} (d+o_z)\,g_x - g_z\,o_x \\ g_x\,g_y\,o_x - o_y\,(1-g_y^2)) + (d+o_z)\,g_y\,g_z \end{pmatrix} \\
\mathbf{v}_2 &= \begin{pmatrix} t_x\,\sqrt{1-g_y^2}\,(d\,g_z + o\cdot g) \\ t_y\,\sqrt{1-g_y^2}\,(d\,g_z + o\cdot g) \end{pmatrix}.
\end{aligned} \tag{11}$$

This mapping strategy is easy to implement as operations on images, and its exactness is only limited by the precision of the eye tracker. Furthermore, it is completely decoupled from the decomposition strategy, so it can be applied directly to any other decomposition method, including linear blending.

As shown in Figure 1(c), displaying the decomposed images deformed by Equation 10 exactly solves the display misalignment problem. This is also demonstrated in the accompanying video. Notice that black bands appear at the edges of the displays since the offset virtual view frustum is not entirely contained in the original view frustum formed by the displays. This can be solved by artificially reducing the field of view of the renderer, so the offset virtual view frustum remains within the display frustum for reasonable eye rotations and translations. In practice, we have not found the outside edge artifacts to be disturbing, and we prefer to ignore them in order to maximize the field of view of the system.

4.2 Blur Gradient Heuristic

Our optimal decomposition solver can be further improved by investigating the behavior of our Jacobi iterations. As seen in Figure 5, the solution of the decomposition after a large number of iterations features ring structures around depth discontinuities. These structures appear in the optimal decomposition of most scenes, and therefore seem to be important for the accurate reconstruction of the focal stack, but our algorithm requires many iterations before these patterns emerge.

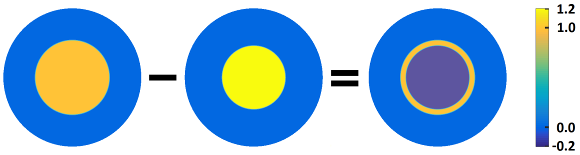
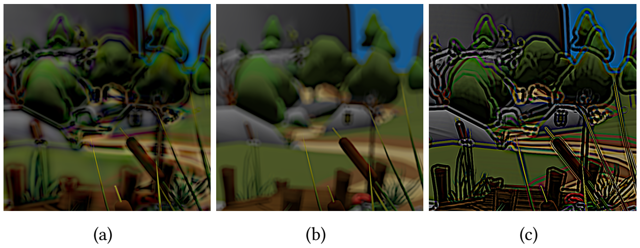


Fig. 4. The blur gradient kernels are obtained by subtracting the kernels of adjacent focal slices, creating desirable ring structures.



(a) (b) (c)

Fig. 5. Front plane of the optimal scene decomposition. (a) The converged solution features ring structures around occlusion boundaries. (b) The ring features take a large number of iterations to appear when using our method without the blur gradient modification, and are not visible after only 10 regular Jacobi iterations. (c) Using our blur gradient modification pushes the solution toward the optimal image more aggressively, and the ring structures already start to appear after a single iteration. As verified in Section 6.3, this consequently improves the convergence speed of our method.

As mentioned in Section 2.2, current perceptual science research suggests the gradient of the blur with respect to changes in depth is key to driving accommodation. To try to force the ring features to appear more quickly, we modify the minimization formulation of Equation 1 in order to explicitly include the blur gradient. Since the scene is densely sampled in depth by the focal stack, a gradient in depth can be approximated as finite differences by subtracting adjacent focal slices. We can thus include the gradient term in the minimization as

$$\underset{\mathbf{x}}{\operatorname{argmin}} \; \|\mathsf{K}\,\mathbf{x}-\mathbf{b}\|^2 + \beta \left\|\mathsf{K}'\,\mathbf{x}-\mathbf{b}'\right\|^2 \tag{12}$$

where β weights the contributions of the reconstructed images and their gradient, and

$$\mathsf{K}' := \begin{pmatrix} \mathsf{K}'_{11} & \cdots & \mathsf{K}'_{1D} \\ \vdots & \ddots & \vdots \\ \mathsf{K}'_{F1} & \cdots & \mathsf{K}'_{FD} \end{pmatrix}, \quad \mathsf{K}'_{f,d} := \mathsf{K}_{f+1,d} - \mathsf{K}_{f,d} \tag{13}$$

$$\mathbf{b}' := \begin{pmatrix} \mathbf{b}'_1 \\ \vdots \\ \mathbf{b}'_F \end{pmatrix}, \quad \mathbf{b}'_f := \mathbf{b}_{f+1} - \mathbf{b}_f\,. \tag{14}$$

Note that the new terms in Equation 12 are obtained by reusing the already-computed focal stack and blur kernels, so this modification does not incur any significant additional cost. The new minimization can be solved through the normal equations

$$(\mathsf{K}^\top\mathsf{K} + \beta\mathsf{K}'^\top\mathsf{K}')\,x = (\mathsf{K}^\top\mathbf{b} + \beta\mathsf{K}'^\top\mathbf{b}'). \tag{15}$$

This system is processed similarly to the original normal equations to obtain an efficient formulation in the image matrix notation of Section 3.2, and likewise reduces to simple operations on the blur gradient kernel images $\overline{\mathsf{K}'}_{fd}$ and blur kernel images $\overline{\mathsf{K}}_{fd}$.

The blur gradient formulation, despite its structural similarity to the original minimization, cannot be solved directly with our Jacobi iterations. Since the blur gradient kernels are composed of differences of the original kernels, they are not non-negative kernels, and the convergence criterion of Equation 9 does not hold. This results in divergent instabilities in the decomposition after a large number of iterations. Still, we use the blur gradient formulation for the first few iterations of the decomposition, and then revert back to the original formulation to maintain stability. We have found using a single blur-gradient-augmented Jacobi step with $\beta \approx 250$ to be sufficient in our experiments to increase convergence speed, but the optimal values for the weights and the number of blur-gradient-augmented steps remain to be investigated.

Some intuition on the effects of the blur gradient term can be gained by looking at the structure of the blur gradient kernels $\overline{\mathsf{K}'}_{fd}$, shown in Figure 4. The new kernels possess ring structures akin to the structures we observe in the converged optimal decomposition in Figure 5, which might explain why they improve the convergence of the decomposition. The computational benefits of using the blur gradient are verified in Section 6.3.

4.3 GPU Implementation

The Jacobi iterations of Equation 8 are mostly composed of per-pixel operations, which are implemented in a pixel shader on the GPU. Only the term $\overline{\mathsf{C}}\,\overline{\mathbf{x}}^{(k)}$ requires more attention, as it corresponds to a convolution. However, for the parameters used throughout this paper, the cross-correlated kernels are small and only convolutions over a few pixels are required. Even if standard GPU convolution techniques can be used [Podlozhnyuk 2007], a naive implementation summing over neighboring pixels in a pixel shader outperformed all other methods we have tested. Approximate downsampled approaches could also be used, but they would introduce errors in the decomposition and would require further analysis.

Generating the focal stack $\overline{\mathbf{r}}$ in Equation 8 is also a challenging part of the decomposition, as it requires accurately rendering depth of field blur for each focal slice. We compute the focal stack by accumulating images over 64 samples on the virtual pupil. As depicted in Figure 6, this would usually require 64 × F renders using a standard pinhole rendering pipeline. We instead approximate this process by using a single view frustum per pupil sample which envelops all focal slices, reducing the number of required renders to 64. The enveloping images are rendered at higher resolution, and the image for each focal slice is extracted by cropping the enveloping images. This greatly improves the efficiency of focal stack generation, and we have found the resulting sampling errors to be negligible. More approximate but faster focal stack rendering methods might give superior results, but we prefer to use a slower but accurate focal stack generation method in this paper, so that no additional error is introduced in our analysis.

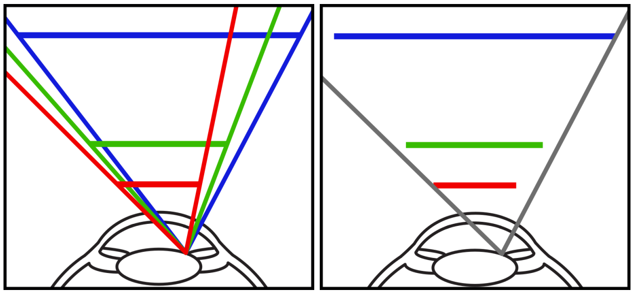


Fig. 6. To generate the focal stack, we accumulate samples on the pupil of the virtual camera. (*left*) This would usually require one render per sample per focal slice, as shown here for a single pupil sample and $F = 3$ focal slices. (*right*) As explained in Section 4.3, we instead use a single render per eye sample which envelops the rendering frustums required by all focal slices, greatly reducing the cost of focal stack generation.

Algorithm 1: Scene decomposition (computed each frame)

```
for each eye do
    - update virtual camera to match measured eye position
    - render focal stack $\bar{\mathbf{r}}$ (Section 4.3)
    - initialize $\bar{\mathbf{x}}$ to zero
    - do one Jacobi step with blur gradient (Section 4.2)
    - do $S$ regular Jacobi steps (Equation 8)
    - apply eye tracker deformation (Equation 10)
end
```

5 EYE-TRACKED MULTIFOCAL DISPLAY TESTBED

We build a multifocal display testbed driven by the methods of Sections 3 and 4, leveraging a combination of off-the-shelf and custom components. Our testbed is purpose-built to explore open questions regarding the accommodation response to, and visual perception of, multifocal displays, as enabled by our efficient gaze compensation and scene decomposition algorithms. As such, three subsystems are necessary: a binocular multifocal display with a reasonably large field of view to reproduce natural viewing conditions; an eye tracker to account for parallax caused by eye rotations and head movements, following Section 4.1; and a way of measuring accommodation to experimentally verify the reaction of subjects under various viewing conditions. This section describes the components selected for our testbed, their use, and their calibration. Figure 7 details the construction, and a detailed view of the system is also presented in the accompanying video.

5.1 Displays

The testbed employs six full-color organic light-emitting diode (OLED) display panels (MicroOLED MDP02), each supporting 1280× 1024 resolution at a 60 Hz refresh rate. The panels are mounted on motorized translation stages to create three variable-focus virtual image planes per eye. Light from the displays is combined using pellicle beamsplitters and relayed to the eye through a pupil-forming optical system. This pupil-forming system affords a 20-degree field of view, a 10-mm-diameter eye box, and is designed to be telecentric in the virtual image space, so as to maintain image resolution of one arcminute per pixel (i.e, 20/20 visual acuity). Each display panel is independently actuated to address a 17-diopter depth of focus (DOF), and these ranges are staggered to address a total DOF spanning from −5 to +12 diopters. This extended range allows for the correction of the spherical component of the viewer's prescription, eliminating the need to use corrective eyewear when viewing the testbed, thereby assisting eye tracking and accommodation measurements. Viewers are positioned, relative to the viewing optics, using a bite bar. Note that the bite bar helps stabilize the user's head, but does not eliminate head movements, so the corrections of Section 4.1 are still required in this system. A manual translation stage controls the interaxial distance (IAD) by altering the separation between the right-eye display subsystem and the remainder of the testbed to adjust to the user's interpupillary distance, if necessary.

Because a different display synthesizes each virtual image plane, the system requires accurate radiometric and color calibrations. These calibrations are obtained from measurements of the gamma curves and primary spectra, as recorded with a Photo Research PR-745 SpectraScan Spectroradiometer. A look-up table converts target sRGB image values to color-corrected, display-specific RGB values, following the method of Brainard [1989]. The focus of each display was measured using a SID4 wavefront sensor from Phasics Corp. Optical distortions and alignment between the virtual images are measured using a method akin to Gilson et al. [2011]. Similar to Watson and Hodges [1995], distortions and alignment are corrected by pre-warping imagery on the GPU.

5.2 Eye Tracker

The use of eye tracking for multifocal displays has been discussed before, for instance in the early design of Rolland et al. [2000]. However, to our knowledge, our testbed is the first to incorporate such eye tracking. We employ a conventional model-based eye tracking algorithm, as surveyed by Hansen and Ji [2010], wherein the position and pose of the eyes are estimated by tracking the boundary of the viewer's pupil and the bright reflections of point light sources from the anterior surface of the cornea. The point light sources consist of an array of near-infrared light-emitting diodes (LEDs) placed into a structure in front of the designed eye box. A pair of infrared-sensitive cameras record images focused over a 25-mm-diameter region centered on this eye box at a sampling rate of 250 Hz. Dichroic "hot" mirrors combine the eye tracking and display paths. We emphasize that our development of this eye tracking system closely follows prior constructions, with extended implementation details provided for a similar design by Stengel et al. [2015].

5.3 Accommodation Measurement

The accommodative state of the viewer's left eye is measured at 67 Hz using the well documented Shack-Hartmann wavefront sensing technique [Liang et al. 1994]. Our system employs near-infrared light created with a Thorlabs SLD830S-A10 superluminescent diode (SLD) that is coupled to the eye using a weakly reflecting beamsplitter. Light passing through the viewer's eye and reflecting from the retina is separated from the display path using another "hot" mirror and relayed to an Imagine Optic HASO wavefront sensing camera with a 34×34 microlens array achieving a 294 μm pitch at the system entrance pupil.

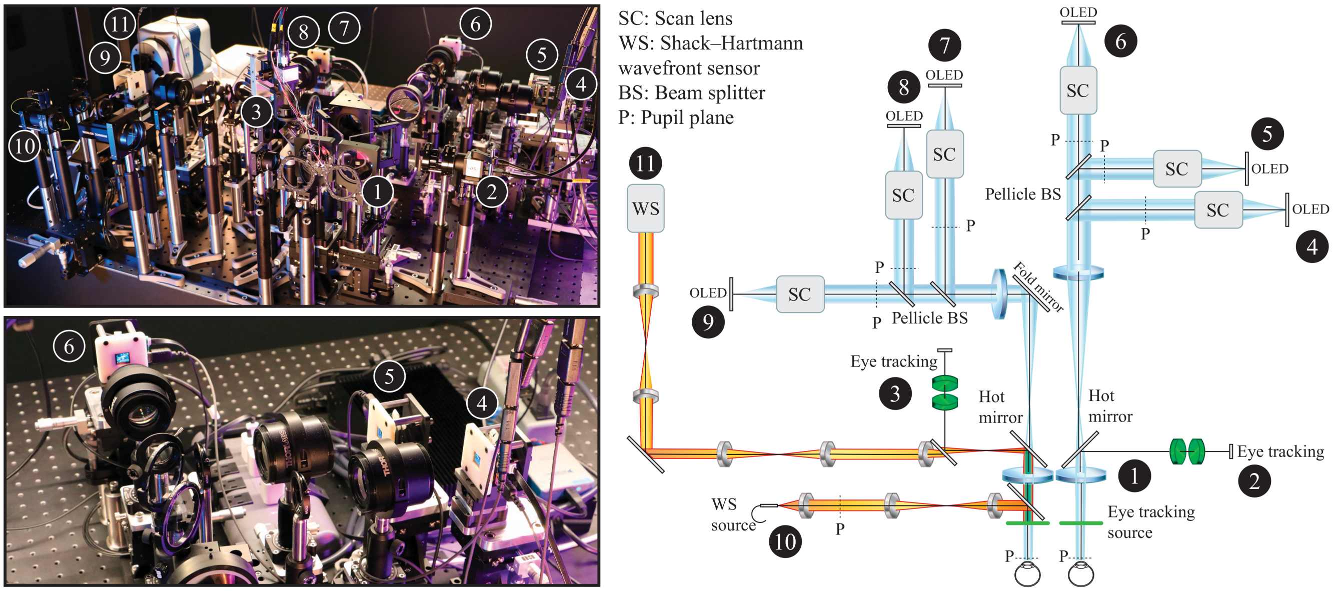


Fig. 7. Our multifocal testbed includes three primary subsystems, as denoted in the photographs (left) and the optical diagram (right). First, the display subsystem uses three OLED panels per eye to achieve a 17-diopter accommodation range over a 20-degree field of view. Lenses in this subsystem are shaded blue, and blue optical rays are traced from display pixels to the eye box. Second, the eye tracking subsystem comprises a pair of 250 Hz cameras and a set of near-infrared LEDs. Lenses in this subsystem are shaded green. Third, the accommodation measurement subsystem uses a Shack-Hartmann wavefront sensor. Lenses in this subsystem are shaded gray and red optical rays are traced from the illumination source, to the eye, and back to the wavefront sensor. See the supplementary video for additional details.

Method	Time 100 iterations	Precomp.	Per Iteration
Narain CPU	-	-	1.8
Narain GPU	20	1.5	0.185
Ours	2.38	0.5	0.025

Table 1. Time comparison (in seconds) of the original CPU implementation of Narain et al. [2015], our GPU implementation of their work, and our method. The total and precomputation times for Narain et al. were not reported. We report the time required to compute 100 iterations for image resolution 512×512, and break down timings into precomputations and the iterations themselves.

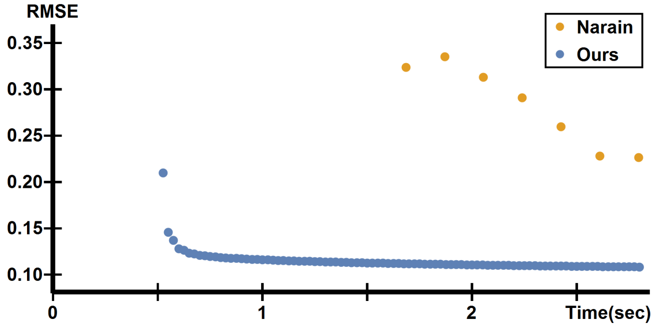


Fig. 8. Comparing the residual mean square error (RMSE) over time of our GPU implementation of Narain et al. [2015] against our method. We can compute 47 Jacobi iterations, which yields a solution close to optimal, before a single iteration of Narain et al. is computed.

6 RESULTS AND DISCUSSION

In this section, we show that our method is an efficient way of solving the optimal decomposition formulation of Equation 1, outperforming previous work, and that it unlocks the interactive use of high-quality decompositions for practical multifocal display applications.

6.1 Efficiency Versus Previous Work

In their original paper, Narain et al. [2015] describe a CPU implementation of their method. They report a computation time of 180 seconds for 100 iterations, or 1.8 seconds per iteration. Note that this time does not appear to include the computation of the ground truth focal stack and other various images (e.g., the Fourier transforms of the PSFs required by their method). To obtain a fair comparison between our method and theirs, we first implement their method more efficiently on the GPU. Doing so, we report a computation time of about 0.185 seconds per iteration for similar conditions, which is an order of magnitude faster. Times are reported in Table 1. All computations are done on a 12-core 3.5 GHz processor and an NVidia TitanX Pascal graphics card. Note that because of possible small differences between our benchmark test and that originally used by Narain et al., the improvement of our GPU implementation could be slightly lower than 10x, but this uncertainty is taken into account later in this section.

The implementation of our Jacobi iterations is also done on the GPU, following Algorithm 1. We report a time of 0.025 seconds per frame, which is an order of magnitude faster than our GPU implementation of Narain et al. The reason for this significant speed up in our implementation is due to fact that the method of Narain et al. solves the deconvolution problem in Fourier space, but applies the constraints projection in the primal domain, which requires two Fourier transforms per iteration. Convolutions in Fourier space become pointwise multiplications, which is efficient for very large kernels. However, in our case, the radii of the kernels are fairly small, especially for a plane spacing of 0.6 diopters. We found it much faster to compute the convolutions directly in the primal domain, as described in Section 4.3.

Figure 8 compares our Jacobi method with our GPU implementation of Narain et al. We use scene A (shown in Figure 10), and image resolution 512 × 512. We use the residual mean square error (RMSE) to compare each reconstructed focal slice to a ground truth image rendered with correct defocus blur. The errors are averaged over the focal range at twice the frequency used by the decomposition, i.e., we average the errors over 23 focal slices whereas the decomposition uses $F = 12$ for all results in this paper. Doing so verifies that no large error appears between the focal slices used by the decomposition.

Our method converges to the optimal solution faster than previous work, both in terms of number of iterations and computation time. We can use at least 10 times fewer iterations with our method compared to Narain et al. to reach the same image quality. As such, the computational time can be further divided by 10, which gives a total of **three orders of magnitude** improvement in computational time for our method compared to the original implementation of Narain et al. Note that, according to Figure 8, the improvement in number of iterations for our Jacobi method over our GPU implementation of Narain et al. is actually significantly more than 10 fold. We report this conservative value to account for the uncertainty, described earlier in this section, related to the comparison between the CPU and GPU implementations of Narain et al.

With our current implementation, we can thus run the optimal decomposition of scenes at 5 frames per second (FPS) for a 512 × 512 image resolution. As indicated by the steep slope of the error curve in Figure 8 for a low number of iterations, we begin to notice errors if we reduce this number of iterations further. Note that this timing does not include precomputation times, which are mostly comprised of the focal stack generation. As seen in Table 1, we clock our focal stack generation at roughly 2 FPS, but we emphasize that this is highly dependent on the renderer, scene complexity, and focal stack generation method. We use 64 pupil samples and 12 focal slices throughout this paper, which generates a high-quality focal stack and allows us to avoid the effects of focal stack errors in our analysis. However, it is very likely that much more efficient focal stack generation methods can be employed. For instance, a simple reduction in the number of pupil samples would directly reduce precomputation times. Furthermore, using well-known approximations, such as a reverse-mapped z-buffer, would trivially bring focal stack generation to real-time rates. Determining whether such fast focal stack generation methods are perceptually sufficient is an interesting research avenue that our system enables.

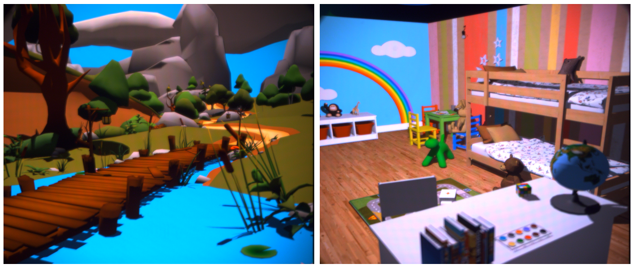


Fig. 9. Captures from our testbed with a camera focused at 0.6 diopters. The accompanying video also shows captures of focal stacks for dynamic content.

Metric	Method	Scene				
		A	B	C	D	E
Q	Converged	72.46	67.70	73.59	74.93	74.84
	Linear	66.85	61.70	60.78	66.52	57.92
	Narain	63.66	62.91	64.09	65.28	62.99
	Ours	71.80	66.52	72.92	74.24	72.25
RMSE	Converged	0.0974	0.0794	0.0441	0.0913	0.0413
	Linear	0.1324	0.1760	0.1259	0.1297	0.1506
	Narain	0.3237	0.3910	0.2965	0.4275	0.1792
	Ours	0.1210	0.1170	0.0700	0.1295	0.0515

Table 2. Quantification of the error for the decomposition methods and scenes of Figure 10. We use an equal-time comparison at 5 frames per second, which corresponds to 1 iteration of Narain et al. and 8 iterations of our method. The HDR-VDP-2 Q metric (higher is better) and the RMSE (lower is better) are averaged over the entire focal stack. In all scenes, our method beats both linear blending and Narain et al. in both metrics.

This performance allows us to compute the optimal decomposition of dynamic content with good quality at interactive frame rates. Figure 9 shows images captured within our system, and the accompanying video shows a sequence with dynamic content captured in real-time in our testbed. However, faster framerates are desirable, and the resolution of the images (stretched to fill the displays vertically) is only half the maximal resolution of our displays. The performance of our method is highly sensitive to differences in equipment and implementation details, and we believe that the significant improvements we report in comparisons with previous work, both in equal time and equal number of iterations, confirm the fundamental benefits of our method. By decreasing optimal decomposition times from the order of minutes to milliseconds, we believe the path to true real-time performances becomes a manageable problem of hardware and implementation efficiency.

6.2 Equal Time Comparison

We use our Jacobi iterations with the blur gradient modification to solve the optimal decomposition for a variety of different scenes, shown in Figure 10. We test our method for a display spacing of 0.6 diopters, as recommended by current research [MacKenzie et al. 2012], but also for larger display distances of up to 2 diopters to test the possibility of covering larger accommodation ranges. We use $F = 12$ focal slices, and image resolution 512 × 512. Since the goal

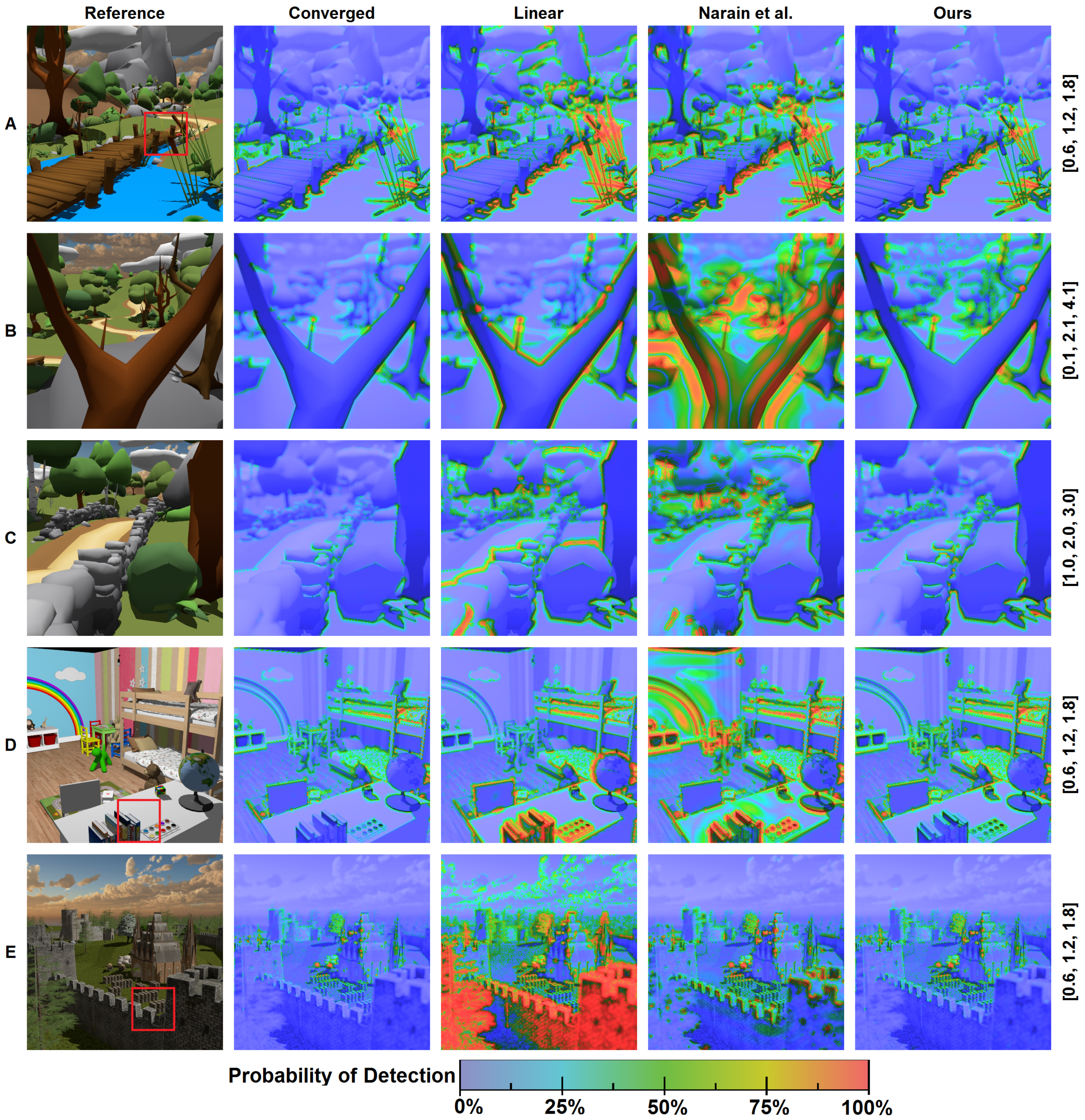


Fig. 10. Comparison of the HDR-VDP-2 metric for different decomposition methods applied to various scenes, compared to the ground truth focal stack. We average the metric over the depth range spanned by the displays, so the colors can be interpreted as the probability of detection of differences between the reconstructed and original focal stacks. The reference images (column 1) show the scenes (identified *A* to *E*) viewed from a pinhole camera, without any defocus blur. The insets in the reference images for scenes A, D and E are used in Figure 11. To compare Narain et al. (column 4) with our method (column 5), we use an equal-time comparison at 5 frames per second, ignoring precomputation time. In this time, we can afford 1 iteration of Narain et al. and 8 iterations of our Jacobi method. We also compare both methods to linear blending (column 3) and to the converged solution of the optimal decomposition (column 2), which we compute using 10,000 iterations of our method. The numbers on the right give the display plane positions used for each scene. For all scenes, our method gives better results than both Narain et al. and linear blending at this interactive frame rate. These results are also quantified in Table 2.

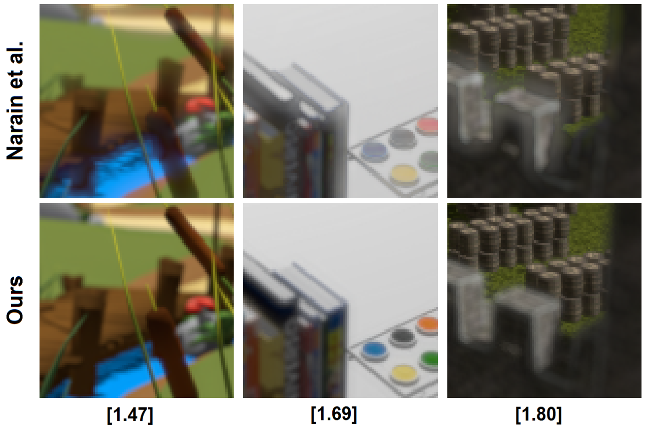


Fig. 11. Insets of scenes A, D and E from Figure 10, using an equal-time comparison of Narain et al. and our method at 5 frames per second. Images are taken at a single focal depth indicated in square brackets. This shows how the errors detected by the HDR-VDP-2 metric of Figure 10 translate to visible artifacts in the reconstructed images. Note how Narain et al. washes out the colors and makes parts of the scene bleed into each other, for instance on the left where the river is visible through the cattail, while our method gives a sharper reconstruction.

of scene decomposition is to reproduce the focal stack as closely as possible, Figure 10 uses the popular perception-based HDR-VDP-2 metric [Mantiuk et al. 2011] to compare reconstructed focal stacks with ground truth images rendered with correct defocus blur. The HDR-VDP-2 metric gives the probability for an average user to detect differences between two images, which we use to compare reconstructed and reference focal slices. The probability is then averaged over the whole focal range, sampled at twice the frequency used by the decomposition (similarly to Section 6.1). Furthermore, as discussed in Section 3.2, the different decomposition methods treat boundary pixels differently. We therefore remove an additional small band of pixels around the images before comparing them to reduce the possible effect of boundary treatment on the image quality metrics.

Since our Jacobi approach and that of Narain et al. are based on the same objective function, they will ultimately converge to similar solutions after a large number of iterations. Focusing on interactive applications, we use an equal time comparison at 5 FPS, without counting precomputation time. In this period, we can compute one iteration of our GPU implementation of Narain et al., and eight iterations of our Jacobi method. Both methods are also compared to linear blending [Akeley et al. 2004], and to the converged solution of the optimal decomposition, computed using 10,000 iterations of our Jacobi method. Note that 5 FPS is the fastest frame rate we can use for the equal-time comparison since we need to compute at least one step of Narain et al. Slower frame rates could be used, but this would only improve the advantage of our method compared to linear blending, and would reduce the gap between our method and Narain et al., making the analysis less clear.

Table 2 also compares the methods and scenes of Figure 10 quantitatively. HDR-VDP-2 provides a global image quality metric Q, which we use to compare reconstructed and reference focal slices, again averaging over the whole focal stack. We also give the same comparison using the RMSE metric, which is proportional to the quantity minimized by the optimal decomposition formulation of Equation 1.

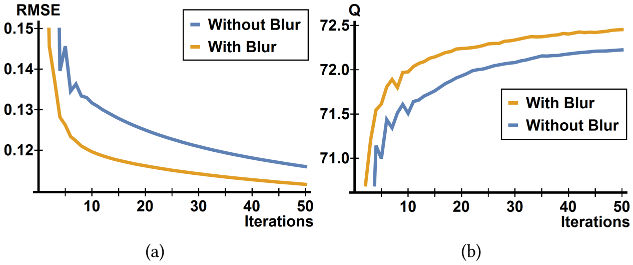


Fig. 12. Comparison of our method with and without the blur gradient modification of Section 4.2. Both in the RMSE (lower is better) and HDR-VDP-2 Q (higher is better) metrics, our method reaches a given error roughly 2 to 4 times faster when using the blur gradient.

For all five test scenes and all metrics, our method performs better than both Narain et al. and linear blending, reaching decomposition results that are perceptually close to the converged solution, at interactive frame rates. Note that even though the RMSE metric sometimes gives similar values for our method and linear blending (e.g. in Scene D), the Q metric and the images in Figure 10 distinctly highlight the advantages of our method. Note also that using the RMSE metric in Figure 10, or using the maximum error over the focal range instead of the average, gave similar results in all cases.

Figure 11 shows insets for three of the scenes presented in Figure 10, as indicated by red squares in the reference images of that figure. These insets compare our method to Narain et al. for a given focal slice and show that the differences identified by the HDR-VDP-2 metric do indeed correspond to perceivable differences in the reconstructed images. In general, for a low number of iterations, the method of Narain et al. tends to create halos around objects, and generates blurrier images with colors from different objects bleeding into one another. This is also visible in the equal time comparison present in the accompanying video.

6.3 Blur Gradient Evaluation

All results presented in Sections 6.1 and 6.2 use the blur gradient modification described in Section 4.2. Figure 12 compares the errors obtained with and without this blur modification, using both the HDR-VDP-2 Q and RMSE metrics described in Section 6.2. As done in previous sections, the metrics are computed by averaging over the whole focal stack, sampled at twice the depth frequency used by the decomposition. From this figure, we see that the blur modification does indeed improve the performance of our method, reducing the number of iterations (and therefore the computation time) required to reach a given error by roughly 2 to 4 times.

6.4 Accommodation of Human Subjects

We tested the capabilities of our system in a pilot user study where we compared the accommodation responses of users looking at dynamic content decomposed using linear blending and our optimal decomposition method. We collected accommodation responses from four observers to a target oscillating sinusoidally in depth between 0.6 and 1.8 diopters at a rate of 0.1 hertz. Observers were asked to maintain fixation on the target, and the image deformation of Section 4.1 was used to adjust to the user's pupil location. The target consisted of a Snellen eye chart embedded into scene C of Figure 10. The target size was held constant (2.4 degrees wide, letter size 0.2-0.7 degrees) in order to remove looming as a potential cue to accommodation. Three repetitions of the movement were collected over 30 seconds for each of the four observers. Observers viewed the scene monocularly to avoid the influence of binocular cues (e.g., vergence distance), and to ensure the changes in accommodation were driven by retinal blur alone.

The results of this study are shown in Figure 13. Individual observer responses were shifted relative to the stimulus position in order to align responses while accounting for subtle shifts in the accommodation response unique to each observer's optics. This was done by computing the average accommodative position through the captured sequence, and then shifting the responses by an amount equal to the difference between that average and the average stimulus position (1.2 diopters).

The results shown in Figure 13 indicate that both decomposition methods provide a stimulus that drives the accommodation response. For linear blending, we replicate the findings expected from literature [MacKenzie et al. 2010], with an accommodative gain of about 0.61. Our Jacobi algorithm also drove changes in the accommodation response, but with a significantly lower gain than linear blending, as confirmed with a repeated measures t-test (0.28, t(3) = 5.08, p=0.015).

Measuring the modulation transfer function (MTF) under both decompositions can help explain the results of the user study. Figure 14(a) shows the MTF measured with a camera looking at a point stimulus in the system, decomposed with either linear blending or optimal decomposition. In all cases, the stimulus is placed between two display planes at 1.5 diopters, and the camera is focused at this same depth. We see that linear blending has a higher MTF, notably in the 4-8 cycles per degree range which maximizes the signal to accommodation [MacKenzie et al. 2012]. Figure 14(b) shows the MTF for the same stimulus, but captured virtually in software. Since the stimulus is a point, and the virtual system is not diffraction limited, the ground truth MTF is constant. Again, we see that linear blending gives a better MTF than optimal decomposition. The method itself is therefore responsible for at least part of the drop in relative contrast observed in the real MTF of Figure 14(a).

Many factors could explain the lower MTF and accommodation gain of optimal decomposition. For instance, optimal decomposition reduces high spatial frequencies at the display planes [Narain et al. 2015], which can thus reduce the strength of the accommodative signal and the MTF, even with a virtual camera. This is particularly important for the scene we used because of the large depth discontinuity at the edges of the eye chart. Furthermore, the display alignment appears to be more critical for optimal decomposition since it distributes light across all three planes, while linear blending distributes light to the two nearest ones. Small calibration errors could therefore increase image blur for optimal decomposition, further reducing the strength of the accommodative signal and the captured MTFs of Figure 14.

These results and their explanation require a more in depth investigation, but the study illustrates that optimal decomposition does provide a stimulus that can drive accommodation. We reiterate that this user study is preliminary and only serves to demonstrate the capabilities of our system. This simple user study already raises many questions and possible research avenues, such as the possibility of modifying the objective function to optimize for the MTF directly, which shows the potential and usefulness of our testbed.

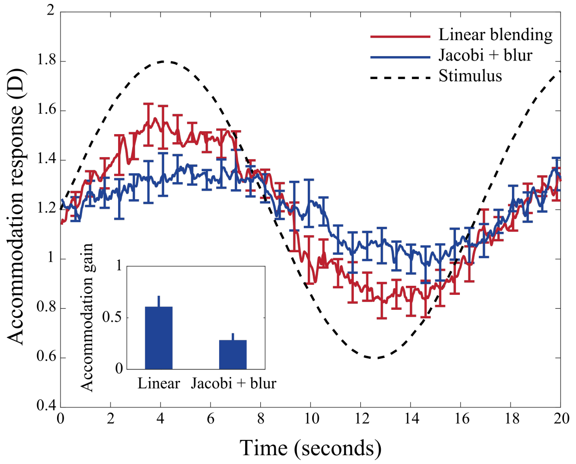


Fig. 13. Results of accommodation measurement to both linear blending (red), and our optimal decomposition method (blue). The black dashed curve shows the stimulus profile. Error bars represent +/-1 standard error of the mean. Accommodative gains (inset) were obtained by computing the difference between the maximum and minimum responses during the stimulus movement, and scaling the difference by the amplitude of the stimulus movement. These results show, for the first time, that optimal decomposition does indeed drive accommodation, albeit with a lower gain than linear blending.

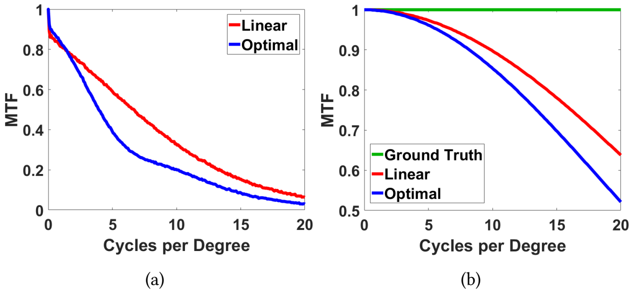


Fig. 14. MTF measurements for a point stimulus placed at 1.5 diopters and decomposed using either linear blending or optimal decomposition. The camera is also focused at 1.5 diopters. (a) MTF using a camera looking into our system. (b) MTF computed in software using a virtual camera with a 5mm aperture.

7 DISCUSSION AND CONCLUSION

We have presented significant, necessary improvements over the current state-of-the-art in multifocal displays. Our efficient scene decomposition method unlocks the use of optimal decomposition for high-quality interactive applications. We have also demonstrated how eye tracking can be used to efficiently maintain plane alignment in multifocal displays.

The way current display technologies drive accommodation is still under active investigation [Koulieris et al. 2017], but accommodation in multifocal displays has so far been difficult to study due to impractical decomposition times, misalignment issues, and the difficulty of integrating measurement paths to a multifocal system. By combining eye tracking and accommodation measurement with our interactive decomposition algorithm, our multifocal testbed is the first to fully enable the investigation of many open questions regarding multifocal displays and the human visual system.

Many of these open questions are intricately coupled with the design of our system. For instance, we hope to investigate the required precision and latency of eye tracking, the effects of our blur gradient heuristic and better optimization functions on accommodation, and the relation between the error metrics and the perceived realism, quality and comfort in multifocal displays. By its significant form-factor, our testbed is however limited to investigating the accommodation of static users, and cannot be used to study such open questions relating to the interactions of more depth cues as a user moves freely in a virtual environment. We hope that a better understanding of these questions will allow us to improve our testbed, and in turn guide the design of future multifocal displays.

A CONVERGENCE PROOF

We prove the convergence criterion of Equation 9. Because the submatrices of $\mathsf{K}^\top \mathsf{K}$ are definite-positive, C only has positive eigenvalues, and the convergence criterion for over-relaxed Jacobi iterations [Burden and Faires 2011] is

$$\alpha < \hat{\alpha} := \frac{2}{\rho(\Lambda^{-1}\,\mathsf{C})} \tag{16}$$

where Λ is the diagonal matrix of C and ρ denotes the spectral radius. The proof thus reduces to computing the largest eigenvalue of $\Lambda^{-1}\,\mathsf{C}$.

For a single display d and a single focal slice f, the action of $\Lambda^{-1}\,\mathsf{C}$ is a convolution by the kernel image $\frac{1}{\lambda_d}\overline{\mathsf{K}}_{fd}^\top * \overline{\mathsf{K}}_{fd}$. The largest eigenvalue for this case is obtained by finding the eigenvector image $\overline{\Omega}$ with largest norm after convolution.

By Parseval's identity, the problem can be solved equivalently in Fourier space. Denoting the Fourier transform by $\mathcal{F}$, we can decompose the convolution into a pixelwise multiplication as

$$\mathcal{F}\left(\left(\frac{1}{\lambda_d}\overline{\mathsf{K}}_{fd}^\top * \overline{\mathsf{K}}_{fd}\right) \star \overline{\Omega}\right) = \mathcal{F}\left(\frac{1}{\lambda_d}\overline{\mathsf{K}}_{fd}^\top * \overline{\mathsf{K}}_{fd}\right) \cdot \mathcal{F}(\overline{\Omega})\,. \tag{17}$$

The largest norm after convolution is thus obtained by using the image $\overline{\Omega}$ which only contains the frequency with the largest amplitude in $\mathcal{F}(\overline{\mathsf{K}}_{fd}^\top * \overline{\mathsf{K}}_{fd})$. Because our kernels are positive, the largest amplitude is located at frequency $(0, 0)$. The image with the largest norm after convolution is thus a constant image, which is also trivially an eigenvector image under convolution.

Because the kernels are normalized, the convolution of a constant image $\overline{\Omega}_i$ on display i with kernel $\frac{1}{\lambda_d}\overline{\mathsf{K}}_{fd}^\top * \overline{\mathsf{K}}_{fd}$ results in the uniform image $\frac{1}{\lambda_d}\overline{\Omega}_i \quad \forall f$. Combining all displays and all focal slices then yields the eigensystem

$$\sum_{f=1}^{F}\sum_{i=1}^{D}\frac{1}{\lambda_d}\overline{\Omega}_i = \frac{F}{\lambda_d}\sum_{i=1}^{D}\overline{\Omega}_i = \gamma\overline{\Omega}_d \quad d \in \{1, \ldots, D\} \tag{18}$$

for eigenvalue γ, whose solution is

$$\gamma = \sum_{i=1}^{D}\frac{F}{\lambda_i}\,. \tag{19}$$

Combining this result with Equation 16 gives the convergence criterion of Equation 9.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their useful feedback, as well as Alexander Fix, Kavitha Ratnam, and Mike Monroe for their help. This work uses the following assets from the Unity Asset Store (www.assetstore.unity3d.com): *Low Poly Series: Landscape* (69650) from Stoolfeather Games (Figures 1, 5, 9-11 and video), *Kids Room* (48596) from 3D Everything (Figures 3 and 9-11), and *Medieval Special Buildings* (75754) from Bretwalda Games (Figures 10 and 11).

REFERENCES

Kurt Akeley. 2004. *Achieving near-correct focus cues using multiple image planes.* Ph.D. Dissertation. Stanford University.

Kurt Akeley, Simon J. Watt, Ahna Reza Girshick, and Martin S. Banks. 2004. A Stereo Display Prototype with Multiple Focal Distances. *ACM Trans. Graph.* 23, 3 (2004), 804–813.

Mathew Alpern. 1958. Variability of accommodation during steady fixation at various levels of illuminance. *Journal of the Optical Society of America* 48 (1958), 193–197.

Barry Blundell and Adam Schwartz. 1999. *Volumetric Three-Dimensional Display Systems.* Wiley-IEEE Press.

David H. Brainard. 1989. Calibration of a computer controlled color monitor. *Color Research and Application* 14, 1 (1989), 23–34.

Richard L. Burden and J. Douglas Faires. 2011. Numerical Analysis, 9th International Edition. *Brooks/Cole, Cencag Learning* (2011).

Johannes Burge and Wilson S. Geisler. 2011. Optimal defocus estimation in individual natural images. *PNAS* 108 (2011), 16849–16854.

David Dunn, Cary Tippets, Kent Torell, Petr Kellnhofer, Kaan Akşit, Piotr Didyk, Karol Myszkowski, David Luebke, and Henry Fuchs. 2017. Wide Field Of View Varifocal Near-Eye Display Using See-Through Deformable Membrane Mirrors. *IEEE Transactions on Visualization and Computer Graphics* 23, 4 (2017), 1322–1331.

Stuart J. Gilson, Andrew W. Fitzgibbon, and Andrew Glennerster. 2011. An automated calibration method for non-see-through head mounted displays. *Journal of Neuroscience Methods* 199, 2 (2011), 328–335.

Dan W. Hansen and Qiang Ji. 2010. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32, 3 (2010), 478–500.

Robert T. Held, Emily A. Cooper, and Martin S. Banks. 2012. Blur and Disparity Are Complementary Cues to Depth. *Current Biology* 22 (2012). Issue 5.

Robert T. Held, Emily A. Cooper, James F. O'Brien, and Martin S. Banks. 2010. Using Blur to Affect Perceived Distance and Size. *ACM Trans. Graph.* (2010).

David M. Hoffman, Ahna R. Girshick, Kurt Akeley, and Martin S. Banks. 2008. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision* 8, 3 (2008), 33.

Xinda Hu and Hong Hua. 2014. High-resolution optical see-through multi-focal-plane head-mounted display using freeform optics. *Optics Express* 22, 11 (2014).

Hong Hua and Bahram Javidi. 2014. A 3D integral imaging optical see-through head-mounted display. *Optics Express* 22, 11 (2014).

Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. 2015. The Light Field Stereoscope: Immersive Computer Graphics via Factored Near-eye Light Field Displays with Focus Cues. *ACM Trans. Graph.* 34, 4, Article 60 (2015), 12 pages.

Paul V. Johnson, Jared A.Q. Parnell, Joohwan Kim, Christopher D. Saunter, Gordon D. Love, and Martin S. Banks. 2016. Dynamic lens and monovision 3D displays to improve viewer comfort. *Optics Express* 24, 11 (2016).

Petr Kellnhofer, Piotr Didyk, Karol Myszkowski, Mohamed M Hefeeda, Hans-Peter Seidel, and Wojciech Matusik. 2016. GazeStereo3D: seamless disparity manipulations. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 68.

Robert Konrad, Emily A. Cooper, and Gordon Wetzstein. 2016. Novel Optical Configurations for Virtual Reality: Evaluating User Preference and Performance with Focus-tunable and Monovision Near-eye Displays. *ACM Conference on Human Factors in Computing Systems (CHI)* (2016), 1211–1220.

Robert Konrad, Nitish Padmanaban, Keenan Molner, Emily A. Cooper, and Gordon Wetzstein. 2017. Accommodation-invariant Computational Near-eye Displays. *ACM Trans. Graph. (SIGGRAPH)* 4 (2017). Issue 36.

Jonh C. Kotulak and Clifton M. Schor. 1986. A computational model of the error detector of human visual accommodation. *Biological Cybernetics* 54 (1986), 189–194.

George-Alex Koulieris, Bee Bui, Martin S. Banks, and George Drettakis. 2017. Accommodation and Comfort in Head-Mounted Displays. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)* 36, 4 (July 2017), 11.

Gregory Kramida. 2016. Resolving the Vergence-Accommodation Conflict in Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics* 22, 7 (2016), 1912–1931.

Douglas Lanman and David Luebke. 2013. Near-eye Light Field Displays. *ACM Trans. Graph.* 32, 6, Article 220 (2013), 10 pages.

Junzhong Liang, Bernhard Grimm, Stefan Goelz, and Josef F. Bille. 1994. Objective measurement of wave aberrations of the human eye with the use of a Hartmann-Shack wave-front sensor. *Journal of the Optical Society of America A* 11, 7 (1994), 1949–1957.

Sheng Liu and Hong Hua. 2010. A systematic method for designing depth-fused multifocal plane three-dimensional displays. *Optics express* 18, 11 (2010), 11562–11573.

Sheng Liu, Hong Hua, and Dewen Cheng. 2010. A Novel Prototype for an Optical See-Through Head-Mounted Display with Addressable Focus Cues. *IEEE Transactions on Visualization and Computer Graphics* 16, 3 (2010), 381–393.

Patrick Llull, Noah Bedard, Wanmin Wu, Ivana Tošić, Kathrin Berkner, and Nikhil Balram. 2015. Design and optimization of a near-eye multifocal display system for augmented reality, In Imaging and Applied Optics. *Imaging and Applied Optics*.

Gordon D. Love, David M. Hoffman, Philip J.W. Hands, James Gao, Andrew K. Kirby, and Martin S. Banks. 2009. High-speed switchable lens enables the development of a volumetric stereoscopic display. *Optics Express* 17, 18 (2009).

Kevin J. MacKenzie, Ruth A. Dickson, and Simon J. Watt. 2012. Vergence and accommodation to multiple-image-plane stereoscopic displays: “real world” responses with practical image-plane separations? *Journal of Electronic Imaging* 21, 1 (2012).

Kevin J MacKenzie, David M Hoffman, and Simon J Watt. 2010. Accommodation to multiple-focal-plane displays: Implications for improving stereoscopic displays and for accommodation control. *Journal of Vision* 10, 8 (2010), 22–22.

Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. 2011. HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In *ACM Transactions on Graphics (TOG)*, Vol. 30. ACM, 40.

Nathan Matsuda, Alexander Fix, and Douglas Lanman. 2017. Focal Surface Displays. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)* 36, 4 (July 2017), 14.

Sangeetha Metlapally, Jianliang L Tong, Humza J Tahir, and Clifton M Schor. 2014. The impact of higher-order aberrations on the strength of directional signals produced by accommodative microfluctuations. *Journal of vision* 14, 12 (2014), 25–25.

Rahul Narain, Rachel A. Albert, Abdullah Bulbul, Gregory J. Ward, Martin S. Banks, and James F. O’Brien. 2015. Optimal Presentation of Imagery with Focus Cues on Multi-plane Displays. *ACM Trans. Graph.* 34, 4, Article 59 (2015), 12 pages.

D. A. Owens. 1980. A comparison of accommodation responses and contrast sensitivity for sinusoidal gratings. *Vision Research* 29 (1980), 159–167.

Nitish Padmanaban, Robert Konrad, Tal Stramer, Emily A Cooper, and Gordon Wetzstein. 2017. Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. *Proceedings of the National Academy of Sciences* (2017), 201617251.

Victor Podlozhnyuk. 2007. Image convolution with CUDA. *NVIDIA Corporation white paper, June* 2097, 3 (2007).

Michael Potmesil and Indranil Chakravarty. 1981. A Lens and Aperture Camera Model for Synthetic Image Generation. *Computer Graphics* 15, 3 (1981), 297–305.

Sowmya Ravikumar, Kurt Akeley, and Martin S. Banks. 2011. Creating effective focus cues in multi-plane 3D displays. *Optics Express* 19, 21 (2011).

Jannick P. Rolland, Myron W. Krueger, and Alexei Goon. 2000. Multifocal Planes Head-Mounted Displays. *Applied Optics* 39 (2000), 3209–3215.

Takashi Shibata, Joohwan Kim, David M. Hoffman, and Martin S. Banks. 2011. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision* 11, 8 (2011), 11.

Weitao Song, Yongtian Wang, Dewen Cheng, and Yue Liu. 2014. Light field head-mounted display with correct focus cue using micro structure array. *Chinese Optics Letters* 12, 6 (2014), 060010.

Michael Stengel, Steve Grogorick, Martin Eisemann, Elmar Eisemann, and Marcus Magnor. 2015. An Affordable Solution for Binocular Eye Tracking and Calibration in Head-mounted Displays. In *Proc. ACM Multimedia.* 15–24.

Hakan Urey, Kishore V Chellappan, Erdem Erden, and Phil Surman. 2011. State of the art in stereoscopic and autostereoscopic displays. *Proc. IEEE* 99, 4 (2011), 540–555.

Dhanraj Vishwanath and Erik Blaser. 2010. Retinal blur and the perception of egocentric distance. *Journal of Vision* 10 (2010).

Benjamin A. Watson and Larry F. Hodges. 1995. Using texture maps to correct for optical distortion in head-mounted displays. In *Virtual Reality Annual International Symposium.* 172–178.

Simon J. Watt, Kevin J. MacKenzie, and Louise Ryan. 2005. Real-world stereoscopic performance in multiple-focal-plane displays: How far apart should the image planes be?. In *SPIE Stereoscopic Displays And Applications*, Vol. 8288.

W. Wu, P. Llull, I. Tošić, N. Bedard, K. Berkner, and N. Balram. 2016. Content-adaptive focus configuration for near-eye multi-focal displays. In *IEEE Multimedia and Expo.*

Marina Zannoli, Gordon D. Love, Rahul Narain, and Martin S. Banks. 2016. Blur and the perception of depth at occlusions. *Journal of Vision* 16, 6 (2016), 17.