

PHASE-CODED COMPUTATIONAL IMAGING FOR ACCOMMODATION-INVARIANT NEAR-EYE DISPLAYS

Ugur Akpınar, Erdem Sahin, Atanas Gotchev

Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

ABSTRACT

We present an accommodation-invariant computational near-eye display based on the extended depth of field imaging. The eyepiece of the display consists of a diffractive optical element (DOE) that is used in tandem with a conventional refractive lens. The DOE is co-designed with the pre-processing convolutional neural network, which is analogous to the post-processing deblurring networks in image capture. We demonstrate through simulations that such system achieves accommodation-invariant imaging within 2 Diopters depth range without significantly sacrificing spatial resolution.

Index Terms— Computational near-eye displays, Optics, Neural networks

1. INTRODUCTION

The conventional near-eye displays are based on the simple stereoscopy principle, i.e., they create binocular cues by delivering corresponding view images to each eye. However, such displays suffer from the well-known vergence-accommodation conflict (VAC) [1, 2], due to their failure to recreate the focus cues. The VAC has been reported to be an important factor that causes visual discomfort, such as fatigue, nausea and eye tiredness [1].

Various methods have been proposed in the literature to address the VAC problem mainly by enabling the focus cues. In-depth review of such methods for the near-eye displays can be found in [3, 4], which include varifocal [5] or multifocal approaches [6], light field (LF) displays [7, 8, 9], and holography [10]. Each method has been demonstrated to alleviate the VAC to some extent at the cost of various trade-offs. For instance, the LF displays are able to stimulate all necessary depth cues by delivering the desired LF (the set of rays within the eye pupil with desired 4D spatio-angular distribution); however, such systems often suffer from the so-called spatio-angular trade-off, which decreases the spatial resolution of the displayed content [7, 9].

This paper addresses the VAC problem through extended depth of field (EDoF) imaging. In particular, by designing a

computational display with depth-invariant point spread functions (PSFs) on the retina, we aim to remove the optical defocus blur cue. In such a case, the eye is expected to focus at the intended converged depth, due to disparity-driven accommodation. Such accommodation-invariant (AI) display approach has been previously implemented through various techniques. In the Maxwellian-type displays [11], the image is projected to each eye through a small (pinhole) effective aperture within the pupil, which naturally increases the depth of field (DoF) at the cost of light throughput. In [12], a focus tunable lens is employed as the eyepiece, where a deep scene can be sharply displayed by changing the focal depth of the display faster than the temporal resolution of the eye. Similar idea has been used before in the context of EDoF projectors [13]. In the same use-case, amplitude-coded aperture based EDoF imaging approach [14] is another related work to mention.

Our AI near-eye display method is based on EDoF through phase-coded computational imaging. We co-design the DOE, which is used in tandem with a conventional (refractive) eyepiece, together with the pre-processing deblurring algorithm that is implemented as a convolutional neural network (CNN). In particular, this approach is inspired from and is analogous to our recent work on EDoF image capture, where we propose a hybrid lens computational camera employing a DOE used for wavefront coding and CNN-based post-processing deblurring algorithm that are jointly optimized [15]. The main novelty of our approach in this paper is that it achieves EDoF (hence AI) near-eye visualization in a significantly large depth range (up to 2 Diopters (D)) by using only static optical elements, i.e., a hybrid refractive lens and a DOE. Thus, it has a potential to alleviate several issues related with above-mentioned (active) time-multiplexing based imaging techniques, such as need of high speed display, synchronization, high cost optics, etc.

2. METHOD

Fig. 1 illustrates the proposed algorithm. The inputs to the network are all-in-focus image I , which are desired to be delivered to the viewer, and the depth z , where the eye is assumed to be focused. The network output, I^p , is the simulated perceived image. The network can be divided into two main parts, namely the display and the pre-processing. The

Ugur Akpınar would like to thank for the graduate school funding of Tampere University.

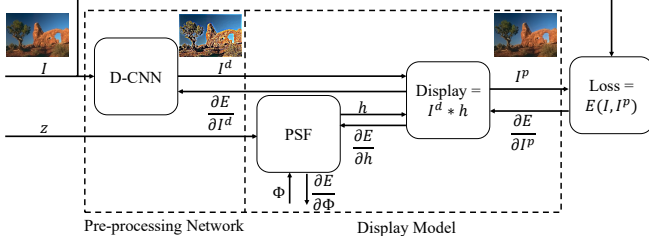


Fig. 1: Overall representation of the proposed method.

display model simulates the perceived image I^p , for a given display input I^d and focused depth z . As analogous to the post-processing in the EDoF cameras [15], we introduce the pre-processing model to compensate the display model in advance, which takes I as input and outputs I^d . Finally, we compute the loss between I and I^p , and co-optimize both the display design (i.e., the DOE) and the pre-processing via gradient descent method. During training, the accommodation distance z is changed randomly at each iteration, optimizing the system to operate in an extended scene depth. In the following, we describe the display model and the pre-processing network in more details.

2.1. Display

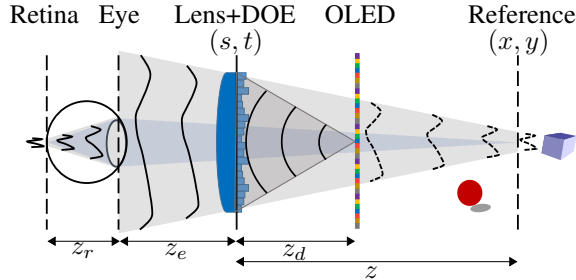


Fig. 2: Proposed near-eye display setup.

Fig. 2 illustrates the image formation model of the proposed computational near-eye display. The display optics at the lens plane consists of an underlying refractive lens and a DOE. We assume an aberration-free eye and in the forward model we simulate the (diffraction-limited) perceived image at the so-called reference plane, at which the eye is assumed to be accommodated. In other words, the reference plane represents the conjugate plane of the retina, which is assumed to be planar. Considering incoherent monochromatic illumination at wavelength λ , the system PSF can be derived via the so-called generalized pupil function [16],

$$Q_{\lambda,z}(s, t) = A(s, t) \exp(j\Phi_{\lambda}(s, t)) \exp\left[j\Psi_{\lambda,z} \frac{s^2 + t^2}{r^2}\right], \quad (1)$$

where $A(s, t)$ is the circular aperture function, $\Phi_{\lambda}(s, t)$ is the

phase delay due to the DOE,

$$\Psi_{\lambda,z} = \frac{k}{2} \left(-\frac{1}{z} + \frac{1}{z_d} - \frac{1}{f_{\lambda}} \right) r^2, \quad (2)$$

is the defocus coefficient, $k = 2\pi/\lambda$ is the wavenumber, f_{λ} is the effective focal length of the underlying refractive lens at the wavelength λ , and r is the radius of the circular lens (and DOE) aperture. The incoherent PSF (for a point source on the display) at the reconstruction plane can then be written as

$$h_{\lambda,z}(x, y) \propto |\mathcal{F}\{Q(s, t)\}|^2, \quad (3)$$

where $\mathcal{F}\{\cdot\}$ is the Fourier transform operator. Finally, the perceived image $I_{z,\lambda}^p(x, y)$, is found as

$$I_{z,\lambda}^p(x, y) = I_{\lambda}^d(x, y) * h_{z,\lambda}(x, y), \quad (4)$$

where $I_{\lambda}^d(x, y)$ is the display image and $*$ is the convolution operator.

During training, the phase delay introduced by the DOE, Φ , is chosen as the optimization parameter of the display model, as illustrated in Fig. 1. Phase modulating elements, such as a free-form refractive lens or a DOE exhibit color dispersion, due to material properties and also diffraction phenomenon in the case of DOE. For such a phase mask with given thickness function (height map) of $d(s, t)$, the introduced phase delay at wavelength λ , Φ_{λ} is given by

$$\Phi_{\lambda}(s, t) = k(n_{\lambda} - 1)d(s, t), \quad (5)$$

where n_{λ} is the wavelength-dependent refractive index. In the proposed implementation, we optimize the DOE using RGB color images in the training, assigning a distinct wavelength to each color channel. The discrete samples of the phase pattern $\Phi_{\lambda_0}(s, t)$ for the specified nominal wavelength of λ_0 are chosen as the optimization parameters, from which the phase delay for another wavelength (color channel) can be derived using Eq. 5 as

$$\Phi_{\lambda}(s, t) = \Phi_{\lambda_0}(s, t) \frac{\lambda_0(n_{\lambda} - 1)}{\lambda(n_{\lambda_0} - 1)}. \quad (6)$$

Similarly, there exists a constraint between the effective focal lengths f_{λ} of the refractive lens for each color channel. Assuming the thin lens model and using Eq. 5, one can derive

$$f_{\lambda} = f_{\lambda_0} \frac{n_{\lambda_0} - 1}{n_{\lambda} - 1}. \quad (7)$$

2.2. Pre-processing

Since the perceived image (at the reconstruction plane) is desired to be as sharp as possible, a pre-processing stage is required before the display stage. Thus, given a sharp desired image I at a given reconstruction plane, the pre-processing stage is to encode this image in such a way that

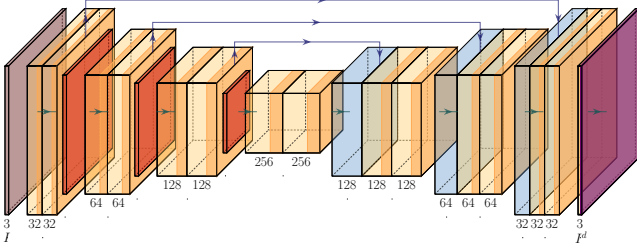


Fig. 3: The pre-processing network (D-CNN) based on U-net architecture [17].

when the display optically decodes the (digitally) encoded image, it appears to be as sharp as possible at the reconstruction plane. This approach is illustrated in Fig. 3, where the pre-processing stage is implemented as a deblurring CNN (D-CNN). We utilize the U-net architecture [17] as the D-CNN, which has been demonstrated to be promising in various image restoration problems. U-net is a multi-level network that includes a contraction (encoding) and an extension (decoding) path, accompanied by skip connections at each level. The encoding part consists of 3×3 convolution kernels with ReLU activation functions, shown as graded yellow in Fig. 3. The feature maps are then downsampled via 2×2 max pooling layer with stride 2 (the red blocks). The channel sizes after each convolution are given under each box, which are doubled after each downsampling. At the decoding step, the feature maps are upsampled with transposed convolution (the blue boxes in the figure) of size 2×2 , with the upsample parameter set as 2. The upsampled feature maps are then concatenated with the output of the encoding layer of the corresponding level, shown as skipped arrows in the figure. Finally, we utilize 1×1 convolution layer to map the network output to the original image size of 3 channels. In order to account for the dynamic range of the physical display, we clamp the network output to be in between 0 and 1, shown as purple. The output of the clamp layer, I^d , is then driven to the display model.

2.3. Loss function

The output of the display model, I^p , is compared to the corresponding all-in-focus image I via a regularized L1-loss. That is, for a given batch of N ground-truth color images (I_1, I_2, \dots, I_N) and outputs ($I_1^p, I_2^p, \dots, I_N^p$), the employed reconstruction loss is

$$E(I, I^p) = \frac{1}{N} \sum_{n=1}^N (\|I_n - I_n^p\|_1 + \alpha \mathcal{R}(I_n, I_n^p)), \quad (8)$$

where $\mathcal{R}(I, I^p)$ is the regularization term and α is the regularization weight. We utilize the dark channel prior as the regularization term [18], which has been demonstrated as a powerful prior for several image restoration problems, such

as image deblurring [19] and dehazing [18]. The dark channel is defined as the minimum color intensity within a patch, i.e.,

$$J(\mathbf{x}) = \min_{\lambda \in \{R, G, B\}} \min_{\mathbf{y} \in \Omega(\mathbf{x})} I_\lambda(\mathbf{y}), \quad (9)$$

where \mathbf{x}, \mathbf{y} are the pixel indices and $\Omega(\mathbf{x})$ is the local patch around \mathbf{x} . We define the regularization term as the weighted L1-norm of the output dark channel, i.e.,

$$\mathcal{R}(I, I^p) = \|\beta \exp(-\gamma J) J^p\|_1, \quad (10)$$

where the weight $\beta \exp(-\gamma J)$ is introduced to decrease the importance of the pixels corresponding to the bright regions of the ground-truth. In our implementation, we set $\beta = 0.005$, $\gamma = 10$, and the local patch size as 17 pixels.

3. SIMULATION RESULTS

We utilize an off-the-shelf plano-convex (fused silica) lens as the underlying refractive lens with the aperture radius of $r = 5mm$ and the effective focal length of $f_{\lambda_s} = 30mm$ at the specification wavelength of $\lambda_s = 587.6nm$. The material of the DOE is also assumed to be fused silica with the refractive indices of $n_{\lambda_R} = 1.458$, $n_{\lambda_G} = 1.461$, $n_{\lambda_B} = 1.466$. The computational display is optimized for color (RGB) images, where the R, G and B channels are assumed to be represented by $\lambda_R = 600nm$, $\lambda_G = 530nm$, $\lambda_B = 450nm$, respectively.

We set the lens-to-display distance as $z_d = 28.9mm$, such that the refractive lens is focused at around 1D from the viewer at the nominal wavelength of $\lambda_0 = \lambda_G$. The assumed display has $8.6\mu m$ pixel pitch, delivering 30 cycles per degree (CPD) spatial resolution. The target depth range is set to between 0D - 2D, which corresponds to the defocus range of $|\Psi| \leq 218$. Following the procedure in [15], the optimum sampling rate of the DOE, Δ_s , is derived based on the scene depth range and found as $\Delta_s = 9\mu m$. Please note that such sampling pitch is only used during training. The test simulations demonstrated below are performed using the (bicubic) upsampled mask at $3\mu m$ sampling, which represents a typical fabrication resolution for the DOE.

The network is trained for 32 epochs using 256×256 image patches from the data set [20]. At each iteration, the reconstruction depth z is picked randomly from uniformly distributed discrete depths of $\{0D, 0.5D, 1D, 1.5D, 2D\}$. We use Adam optimization [21], with the initial learning rate and the weight decay of $1e-3$ and $1e-4$, respectively.

Fig. 4 illustrates the height profile of the optimized phase mask, at upsampled resolution of $3\mu m$. We also provide color PSFs at four different accommodation depths. As one main characteristics of the wavefront coding [22], the PSFs are kept nearly fixed along a wide depth range. In conventional lenses, the PSF spreads quickly as the accommodation distance gets further away from the display focal plane. However, such systems typically provide sharper PSFs at the focal

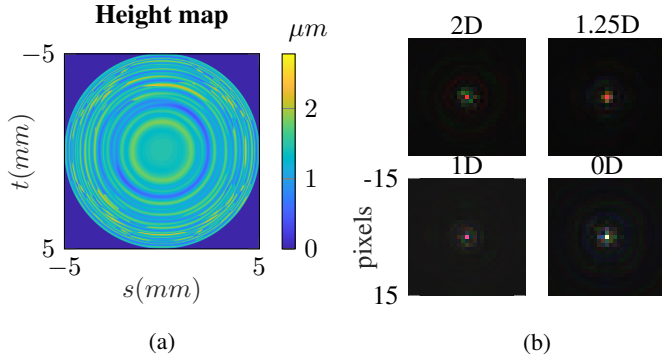


Fig. 4: The optimized height map at assumed fabrication resolution of $3\mu\text{m}$ (a), and color PSFs at different accommodation depths (b).

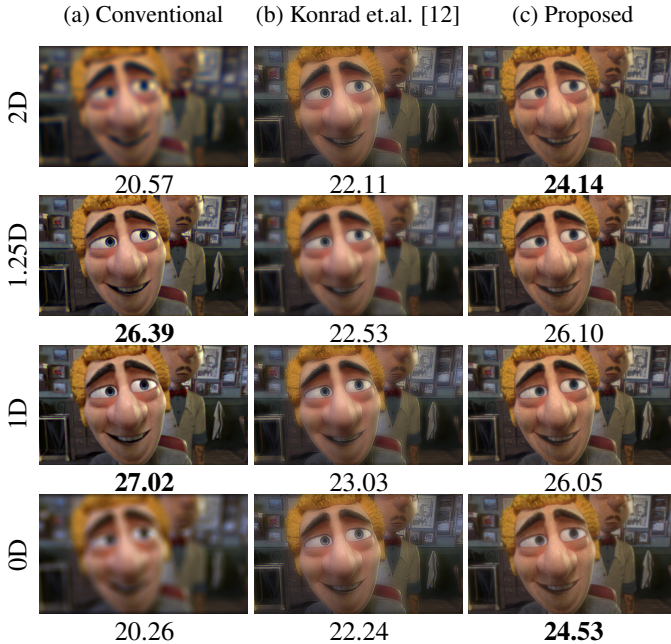


Fig. 5: Comparison of the near-eye displays with three different optics, the conventional stereoscopic display with single refractive lens (a), AI display with focus tunable lens [12] (b), the proposed method with refractive lens and optimized phase mask (c). The PSNR values are given under each image.

plane, therefore higher spatial resolution images. The proposed method can be thought as distributing the spatial image quality equally along the scene depth, while keeping the degradation around the focal plane at acceptable levels. As demonstrated below, this is enabled by the jointly designed pre-processing network and display optics, i.e., DOE.

The proposed algorithm is compared with the conventional stereoscopic display (with refractive lens eyepiece), and one of the state-of-the-art AI displays proposed by Konrad et.al. [12]. The latter takes advantage of a focus tunable

lens to sweep through the desired accommodation range, where the perceived PSF is modeled as the average of the individual PSFs corresponding different focal depths and a Wiener deconvolution-based pre-processing is applied based on this average PSF. We simulate this approach with three different focal depths, namely at 0D, 1D, and 2D. Such configuration is suggested by the authors as a good compromise between the spatial resolution and EDoF range. The methods are tested using a synthetic image from the TAU Agent Data set [23], assuming the accommodation distances of 2D, 1.25D, 1D, and 0D. Fig. 5 illustrates the results. The peak signal-to-noise ratio (PSNR) values are given as objective measures for each approach in the figure, where the best values are given in bold. As can be seen in the figure, the amount of blur significantly changes in the conventional display (a), making the eye accommodate around the focal plane. The focus tunable AI display (b) delivers significantly higher resolution images at 0D and 2D that are away from the focal plane of the conventional display at 1D, at the expense of degradation at and around the focal depth. The proposed method (c), on the other hand, seems to deliver notably higher resolution images at all test depths, which is likely to eliminate the optical defocus blur cue as targeted. In particular, the image quality does not also decrease at the intermediate depth of 1.25D, which was not included during training. This demonstrates that the proposed method provides accommodation invariance in a continuous fashion, i.e., similar quality is achieved both at the training depths and in between.

4. CONCLUSION

This work proposes a novel AI computational near-eye display to address the vergence-accommodation conflict. The underlying EDoF imaging approach with co-designed DOE and pre-processing CNN is demonstrated to advance existing similar techniques in terms of perceived spatial resolution in the entire depth range of interest. The proposed method has also other clear advantages compared to the existing AI near-eye displays in the literature. Most importantly, we employ static optics, which is easily integratable into conventional headsets and compatible with any commercial 2D display units, such as OLEDs.

At this stage we demonstrate AI visualization in 2 Diopters depth range, which might not be sufficient in display of deeper scenes. Nevertheless, we believe that there is still room for improvement, e.g., by employing better optics or pre-processing network. As a future work, we plan to implement a prototype to demonstrate our approach in practice. In particular, we will validate it through subjective accommodation tests.

5. REFERENCES

- [1] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence–accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of vision*, vol. 8, no. 3, pp. 33–33, 2008.
- [2] M. Lambooi, M. Fortuin, I. Heynderickx, and W. IJsselstein, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 30201–1, 2009.
- [3] H. Hua, "Enabling focus cues in head-mounted displays," *Proceedings of the IEEE*, vol. 105, no. 5, pp. 805–824, 2017.
- [4] G. Kramida, "Resolving the vergence-accommodation conflict in head-mounted displays," *IEEE transactions on visualization and computer graphics*, vol. 22, pp. 1912 – 1931, 08 2015.
- [5] N. Padmanaban, R. Konrad, T. Stramer, E. A. Cooper, and G. Wetzstein, "Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays," *Proceedings of the National Academy of Sciences*, vol. 114, no. 9, pp. 2183–2188, 2017.
- [6] K. Akeley, S. J. Watt, A. R. Girshick, and M. S. Banks, "A stereo display prototype with multiple focal distances," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 804–813, 2004.
- [7] H. Hua and B. Javidi, "A 3d integral imaging optical see-through head-mounted display," *Optics express*, vol. 22, no. 11, pp. 13484–13491, 2014.
- [8] F. Huang, K. Chen, and G. Wetzstein, "The Light Field Stereoscope: Immersive Computer Graphics via Factored Near-Eye Light Field Displays with Focus Cues," *ACM Trans. Graph. (SIGGRAPH)*, , no. 4, 2015.
- [9] D. Lanman and D. Luebke, "Near-eye light field displays," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 1–10, 2013.
- [10] A. Maimone, A. Georgiou, and J. S. Kollin, "Holographic near-eye displays for virtual and augmented reality," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–16, 2017.
- [11] T. Ando, K. Yamasaki, M. Okamoto, T. Matsumoto, and E. Shimizu, "Retinal projection display using holographic optical element," in *Practical Holography XIV and Holographic Materials VI*. International Society for Optics and Photonics, 2000, vol. 3956, pp. 211–216.
- [12] R. Konrad, N. Padmanaban, K. Molner, E. A. Cooper, and G. Wetzstein, "Accommodation-invariant computational near-eye displays," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 88, 2017.
- [13] D. Iwai, S. Mihara, and K. Sato, "Extended depth-of-field projector by fast focal sweep projection," *IEEE transactions on visualization and computer graphics*, vol. 21, no. 4, pp. 462–470, 2015.
- [14] M. Grosse, G. Wetzstein, A. Grundhöfer, and O. Bimber, "Coded aperture projection," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 3, pp. 1–12, 2010.
- [15] U. Akpınar, E. Sahin, and A. Gotchev, "Learning optimal phase-coded aperture for depth of field extension," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 4315–4319.
- [16] J. W. Goodman, *Introduction to Fourier Optics*, Roberts and Company Publishers, 2005.
- [17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [18] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.
- [19] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, "Blind image deblurring using dark channel prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1628–1636.
- [20] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European conference on computer vision*. Springer, 2008, pp. 304–317.
- [21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [22] E. R. Dowski and W. T. Cathey, "Extended depth of field through wave-front coding," *Applied optics*, vol. 34, no. 11, pp. 1859–1866, 1995.
- [23] H. Haim, S. Elmaleh, R. Giryes, A. Bronstein, and E. Marom, "Depth Estimation from a Single Image using Deep Learned Phase Coded Mask," *IEEE Transactions on Computational Imaging*, pp. 298 – 310, 2018.