COMPUTATIONAL COHERENT IMAGING FOR ACCOMMODATION-INVARIANT NEAR-EYE DISPLAYS

Jani Mäkinen, Erdem Sahin, Ugur Akpinar, and Atanas Gotchev

Faculty of Information Technology and Communication Sciences, Tampere University,
Korkeakoulunkatu 1, Tampere, Finland

ABSTRACT

We present a computational accommodation-invariant near-eye display, which relies on imaging with coherent light and utilizes static optics together with convolutional neural network-based preprocessing. The network and the display optics are co-optimized to obtain a depth-invariant display point spread function, and thus relieve the conflict between accommodation and ocular vergence cues that typically exists in conventional near-eye displays. We demonstrate through simulations that the computational near-eye display designed based on the proposed approach can deliver sharp images within a depth range of 3 diopters for an effective aperture (eyepiece) size of 10 mm. Thus, it provides a competitive alternative to the existing accommodation-invariant displays.

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

1. INTRODUCTION

In the pursuit of realistic recreation of 3D content in near-eye displays (NED), depth cues are of vital importance. Conventional NEDs provide the binocular vergence and disparity cues by displaying a separate stereoscopic view for each eye. However, in such displays the monocular accommodation and retinal defocus blur cues are typically missing. As a result, such displays suffer from the well-known phenomenon called vergence-accommodation conflict (VAC), which creates visual discomfort and fatigue [1].

Solving the VAC can be approached by either providing the depth cues accurately (i.e., accommodation enabling displays) or by making the display accommodation-invariant (AI). The former can be achieved, for instance, through varifocal [2], multifocal plane [3] or light field (LF) display [4, 5], all of which usually work in the incoherent imaging regime. In the case of coherent imaging, holographic displays are capable of producing accommodation enabling content with correct defocus blur cues, due to their accurate recreation of the complete wavefield [6, 7]. On the other hand, AI displays provide depth-independent defocus blur cue, and thus the accommodation is mainly (cross-)driven by the disparity signal [8]. By this way, the VAC is avoided as both the vergence and accommodation are consistently driven by the same signal. In a Maxwellian-view display [9], such accommodation response can be achieved by focusing the light rays (or image) onto the entrance pupil of the eye, thus projecting the image to the retina through an aperture effectively smaller than the eye pupil size. The display point spread function (PSF) at the retina can be also designed to be depth-invariant by using varifocal optics, where the display content is projected rapidly at different depths at a rate faster than the human temporal resolution [10]. Here the effective pupil size is equal to the full pupil size, thus avoiding the limitations in Maxwellian-view displays regarding, e.g., light throughput and eyebox.

Alternatively, an attractive approach to AI displays is to employ end-to-end learning to co-design the optics of the display and the preprocessing algorithm, e.g., where the former design (optimization) component is chosen to be a static diffractive optical element (DOE), working in tandem with a refractive lens, and the latter one is implemented via a convolutional neural network (CNN) [11, 12]. Here we propose a computational AI NED based on similar end-to-end learning based design framework, considering the promising results reported in those works despite the simplicity of the optical setup. The main distinction of our approach is that it relies on coherent imaging, employing coherent light sources, which can provide crucial advantages in designing AI displays compared to its incoherent counterparts [11, 12]. Importantly, coherent imaging can provide wider depth of field (DoF) compared to incoherent imaging assuming same imaging optics [13]. Furthermore, coherent illumination ensures operation on distinct wavelengths, whereas incoherently illuminated displays operate in broadband spectrum. As typically three channel (red, green, blue) images are utilized in the modeling and design of display systems, in the former case the wavelength-dependent behaviours of the display system components can be more accurately incorporated in the display design process. Though the coherent illumination for the display imposes further constraints and issues to be addressed, e.g, speckle noise, the inherently larger DoF is demonstrated to ease the end-to-end optimization task.

The paper is partially supported by the Academy of Finland research project "Modeling and Visualization of Perceivable Light Fields", decision number 325530.

2. COMPUTATIONAL NEAR-EYE DISPLAY

Our display design comprises of a coherently illuminated display panel and the display optics, which are a combination of a refractive lens and a DOE (to provide desired phase modulation), as depicted in Fig. 1. The aim of an AI NED is to produce as sharp images as possible within a certain accommodation depth range, thus the goal is to find the optimal set of display optics to achieve such a response. We can extend the display design by incorporating a computational element in the form of a preprocessing algorithm. The algorithm further assists in achieving an AI display response by compensating for the effect of the display optics. This is analogous to the post-processing deblurring step of similar computational camera designs [14]. To find the optimal combination of display optics (i.e. DOE) and preprocessing algorithm (CNN), we propose utilizing a learning based approach; the design (optimization) procedure of the computational NED system is illustrated in Fig. 2. During the design procedure, the optics and the preprocessing CNN are jointly optimized utilizing a set of training images, such that the display response is as sharp as possible across the desired depth range. Such depth range is covered by randomly changing the accommodation distance z during learning.

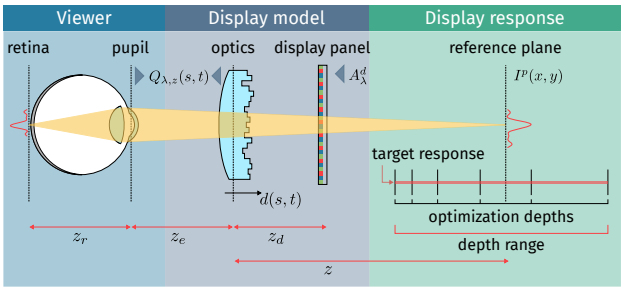

Fig. 1. Display design and imaging model.

As illustrated in Fig. 2, the training starts with feeding the sharp input image I to the CNN, which pre-processes it and outputs the display amplitude A^d. We then simulate, based on the current state of the optics, the image I^p as it would be perceived by a viewer given the distance of accommodation. The perceived image is compared against the ground truth sharp image using the loss function $\mathcal{L}\{I, I^p\}$. The loss function drives the optimization procedure based on its gradient, optimizing the entire display system. The preprocessing network in our proposed method is based on the well-known U-net architecture [15], similar to our previous work [11].

2.1. Computational display model

We simulate the viewing process of the proposed display design, as depicted in Fig. 1, by assuming an aberration-free eye accommodating at distance z. Assuming a planar retina, the image can be found on a conjugate plane, denoted here as

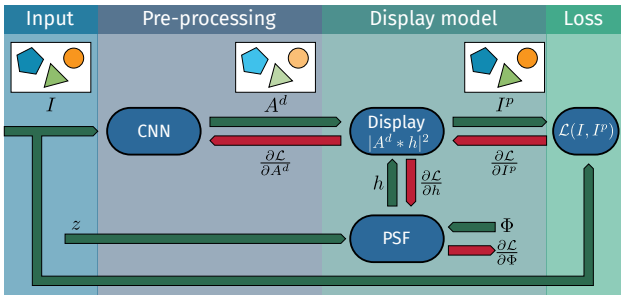

Fig. 2. The proposed learning based computational NED system design approach.

the reference plane (x, y). Assuming monochromatic illumination at wavelength λ, the coherent system PSF $h_{\lambda,z}(x, y)$ on the reference plane is

$$h_{\lambda,z}(x,y) \propto \mathcal{F}\left\{Q(s,t)|_{\left(\frac{x}{\lambda z}, \frac{y}{\lambda z}\right)}\right\}, \tag{1}$$

where $Q_{\lambda,z}(s,t)$ is the generalized pupil function and $\mathcal{F}\{\cdot\}$ is the Fourier transform operator. The pupil function is obtained as a combination of the (circular) aperture function $A(s,t)$, the phase delay of the DOE $\Phi_\lambda(s,t)$ and the refractive lens defocus as

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

where

$$\Psi_{\lambda,z} = \frac{\pi}{\lambda}\left(\frac{1}{z_d} - \frac{1}{z} - \frac{1}{f_\lambda}\right)r^2 \tag{3}$$

is the defocus coefficient, r is the radius of the circular lens and f_λ is the effective focal length of the refractive lens. The DOE phase delay is related to its height $d(s,t)$ through the wavelength-dependent refractive index n_λ of the DOE material as

$$\Phi_\lambda(s,t) = k\left(n_\lambda - 1\right)d(s,t). \tag{4}$$

In the coherent image formation model, the reference plane amplitude is a convolution between the coherent PSF and the source amplitude, i.e., the amplitude from the display panel $A^d_\lambda(x,y) = \sqrt{I^d_\lambda(x,y)}$. Thus, the reference plane image is obtained as

$$I^p_{\lambda,z}(x,y) = \left|A^d_\lambda(x,y) * h_{\lambda,z}(x,y)\right|^2. \tag{5}$$

The coherent image formation poses an issue not present in incoherent imaging, as the display amplitude $A^d_\lambda(x,y)$ contains additional phase values, which are randomly distributed in nature across the display pixels due to randomness in the heights of the pixels at the scale of wavelength [16]. When such values are convolved with the coherent PSF, the neighbouring points (or pixels on the display) interfere such that the resulting intensity pattern contains high frequency and contrast noise known as speckle noise. Though it significantly

degrades the image quality, it can be diminished by, for instance, reducing the temporal coherence of the light source [17]. During training, however, we assume the light to be perfectly coherent and have equal phase across the display, thus discarding the speckle effects, for simplicity.

We optimize, in addition to the preprocessing network, the phase delay of the DOE during the training phase. The optimization is done for a nominal wavelength λ_0, meaning that the delay $\Phi_{\lambda_0}(s,t)$ is optimized, from which the delay for another wavelength λ is calculated as

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

The focal length of the underlying refractive lens f_λ is also wavelength-dependent and can be related to the nominal wavelength focal length f_{λ_0}:

$$f_\lambda = f_{\lambda_0}\frac{n_{\lambda_0} - 1}{n_\lambda - 1}. \tag{7}$$

2.2. Loss function

The loss function drives the learning process by minimizing the difference, or error, between the desired and current output of the system over time. Here we employ a combination of two metrics: the L_1 distance and the structural similarity index measure (SSIM). The aggregated loss function attempts to minimize L_1 and maximize the SSIM as a compromise between the texture detail quality and the perceived change in structural information. The overall loss is calculated as

$$\mathcal{L}(I, I^p) = \mathcal{L}_{L_1}(I, I^p) + \mathcal{L}_{\mathrm{S}}(I, I^p) + \alpha\mathcal{R}(I, I^p) + \gamma\mathcal{R}^d(I^d), \tag{8}$$

where $\mathcal{L}_{L_1}(I, I^p)$ is the L_1 distance, $\mathcal{L}_{\mathrm{S}}(I, I^p)$ is the SSIM-based loss, $\alpha\mathcal{R}(I, I^p)$ is the network output regularization and $\gamma\mathcal{R}^d(I^d)$ is the display image regularization. In order to maximize the SSIM, we define its loss function as

$$\mathcal{L}_{\mathrm{S}}(I, I^p) = 1 - \mathrm{SSIM}(I, I^p). \tag{9}$$

The regularization terms in Equation (8) serve for two purposes: the first one equalizes the intensity levels between the network output and the input, whereas the latter term applies to the dynamic range of the display image. Here the dark channel prior [18] is utilized as $\mathcal{R}(I, I^p)$, weighted by α. As we aim to obtain display images within the range $[0, 1]$, the cut regularization term punishes (i.e., increases the overall loss value) intensity values outside this range. Specifically, the function $\mathcal{R}^d(I^d)$ indicates whether a pixel is within the range or not, which is then weighted by γ to achieve the desired strictness of the regularization.

3. RESULTS

The display design is optimized for color images, where the corresponding wavelengths are chosen as $\lambda_R = 675$ nm for red, $\lambda_G = 540$ nm for green, and $\lambda_B = 450$ nm for blue. The green channel is utilized as the nominal wavelength during training. We assume physically feasible materials for the optics: fused silica for the refractive lens and S1813 photoresist film for the DOE. We set the aperture radius r as 5 mm and the distance between the optics and the display z_d as 28.6 mm. The underlying refractive lens is assumed to have an effective focal length of 30 mm at 587.6 nm wavelength, a central thickness of 2.90 mm and is set up to focus at 1.5 D at the nominal wavelength. We aim to learn AI PSF response within the depth range of 0–3 D, thus limiting the defocus coefficient in Equation (3) by $|\Psi| < 327.3$. As per the discussion in [19], this imposes a minimum display sampling rate; the spectrum of the optimized DOE operates in the same band as the defocus term, therefore band-limiting the PSF spectrum. The corresponding minimum sampling derived from the limited defocus coefficient is 6 µm. The display pixel pitch is set as 12.48 µm, corresponding to 20 cycles per degree in spatial resolution. During training, the accommodation depth is set randomly within the target depth range at 0.5 D discrete steps.

The optimized coherent display properties are presented in Fig. 3. Notably, the PSFs across the target depth range illustrate the relative sharpness when accommodating at different depths. Here the PSFs remain relatively sharp in the middle of the depth range and begin to spread out when approaching 0 and 3 D depths. Though not optimal in terms of an AI response, the proposed display setup is still capable of delivering a large DoF when combined with the optimized preprocessing network, as will be demonstrated in the simulations. Also note that in the coherent imaging regime, the final reconstructed images include not only the linear superposition of such PSFs but also their interference patterns (cross-terms).

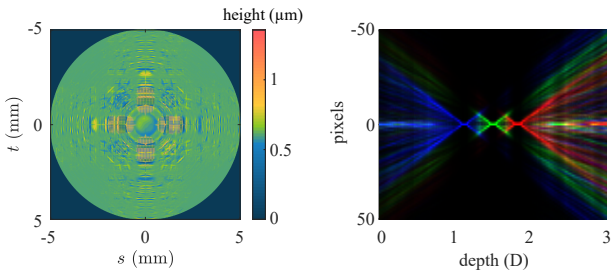

Fig. 3. The optimized DOE height map of the proposed design (left), and the 1D cross-sections of PSF intensities across the target depth range (right).

In the case of coherent light assumption, the resulting reconstructed images are naturally expected to suffer from severe speckle noise. However, in reality the available light sources, even the most coherent ones, have typically finite temporal coherence length, i.e., they are actually partially coherent. In such case, the noise contrast can be observed to be reduced compared to the ideal perfectly coherent case, as the perceived images are effectively integrated (intensity-

wise) in spectrum. To take advantage of this and also to better comply with practical scenarios, in our simulations we consider partially coherent illumination. Based on the method in [16], for each color channel, we take the weighted sum of reconstructed images corresponding to different wavelengths (within the spectrum of the simulated light source) and independently applied random surface height profiles for the display phase components. The random height profiles are assumed to follow a Gaussian distribution, where the standard deviation is equal to the wavelength. We further assume that the partially coherent light source has a Gaussian profile spectral response with a standard deviation of 13 nm, which corresponds to a superluminescent LED [17].

To compare our simulation results against the state of the art approaches in computational AI NEDs, we consider three notable cases: an ideal coherent display, where the phase components of all display pixels are assumed to be zero and thus the reconstructed images have no speckle noise, a partially coherently illuminated display exhibiting speckle noise, and a time-multiplexed partially coherent display, which averages 25 frames of independent noise realizations intensity-wise to reduce the speckle noise. Such time-multiplexed reconstruction can be implemented in practice, for instance, by introducing a moving diffuser in the illumination path [20]. In addition to these, we evaluate the results of an equivalent optimized display using incoherent illumination (similar to [11]), and the AI display proposed by Konrad et. al. [10]. The latter display sweeps through the desired accommodation range using a focus tunable lens, thus resulting in the perceived PSF being modeled as the average of the individual PSFs corresponding to different focal depths. Based on the averaged PSF, the method applies a Wiener deconvolution-based preprocessing step. We simulate this approach with three different focal depths of 0.5 D, 1.5 D, and 2.5 D as a fair compromise between the spatial resolution and extended DoF range.

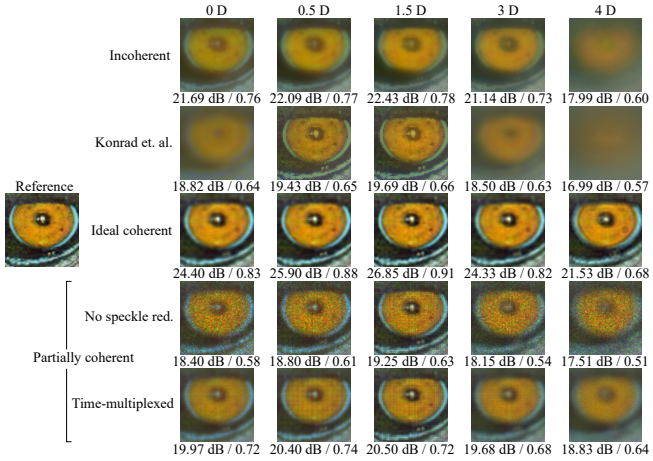

Fig. 4. Zoomed-in simulation results for the Baboon test image at five different accommodation depths, and the corresponding PSNR/SSIM values of the full image.

The zoomed-in simulated reconstructions are presented in Fig. 4 for the well-known Baboon test image, where we compare the reconstructed images with the sharp ground truth image via PSNR and SSIM. In both metrics, the ideal coherent setup achieves the highest scores, followed by the incoherent setup and then the focus tunable lens AI display. The results of the partially coherent display setups are as expected: without any additional speckle reduction methods, the overall scores are among the lowest. However, by introducing even a simplistic speckle reduction approach, the result scores improve noticeably and are relatively competitive with the two rivaling approaches. One can argue that the proposed approach produces visually better (sharper) results in some or most of the cases, despite the PSNR and/or SSIM metrics suggesting the opposite. Although, we keep the more elaborate discussion on perceptual metrics out of scope of this paper, the presented results in Fig. 4 addresses the importance of that topic, especially when comparing (reconstructed) images for displays with different characteristics of introduced distortions. Importantly, when time-multiplexing based speckle reduction method is employed, the proposed display can keep the reconstruction quality also beyond the intended depth range of 0 to 3 D that is assumed during training. Besides these inspiring results, the reconstruction quality of images observed in the ideal coherent case indicate the potential of our approach: if the speckle noise can be further eliminated, via computational or optical means, the resulting display can outperform the state-of-the-art methods.

4. CONCLUSIONS

In this work we present an AI NED that employs computational coherent imaging to address the VAC. By using jointly learned (co-optimized) preprocessing network and hybrid refractive-diffractive optics, the presented display achieves sharp display response within the intended (training) depth range of 0–3 D, which can be further kept beyond the training depth range (e.g., up to 4D) without significant loss in the reconstruction quality. The comparative analysis shows the benefits, and compromises, of replacing an incoherent imaging setup with a coherent one: the reconstructed images are noticeably sharper, however, at the cost of introducing speckle noise. We demonstrate that the reconstruction results can be significantly improved by incorporating speckle reduction methods. Additionally, as the proposed design uses static optical elements, our design provides a simpler construction in comparison, e.g., to varifocal optics based solutions. In the future work, we target including the modeling of partial coherence in the training stage, which can potentially provide better optimized results based on the coherence length of the display illumination source. Furthermore, we plan to incorporate more sophisticated learning-based speckle reduction methods, which is to be also optimized in the initial training stage.

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, 03 2008.

[2] 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.

[3] K. Akeley, S. J. Watt, A. R. Girshick, and M. S. Banks, "A stereo display prototype with multiple focal distances," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 804–813, Aug. 2004.

[4] T. Ueno and Y. Takaki, "Super multi-view near-eye display to solve vergence–accommodation conflict," *Optics Express*, vol. 26, no. 23, pp. 30703, 2018.

[5] H. Hua and B. Javidi, "A 3D integral imaging optical see-through head-mounted display," *Opt. Express*, vol. 22, no. 11, pp. 13484–13491, Jun 2014.

[6] T. Yoneyama, E. Murakami, Y. Oguro, H. Kubo, K. Yamaguchi, and Y. Sakamoto, "Holographic head-mounted display with correct accommodation and vergence stimuli," *Optical Engineering*, vol. 57, no. 6, pp. 1 – 9 – 9, 2018.

[7] A. Maimone, A. Georgiou, and J. S. Kollin, "Holographic near-eye displays for virtual and augmented reality," *ACM Trans. Graph.*, vol. 36, no. 4, July 2017.

[8] C. Schor, "A dynamic model of cross-coupling between accommodation and convergence: simulations of step and frequency responses.," *Optometry and vision science : official publication of the American Academy of Optometry*, vol. 69 4, pp. 258–69, 1992.

[9] M. Von Waldkirch, P. Lukowicz, and G. Tröster, "LCD-based coherent wearable projection display for quasi-accommodation-free imaging," *Optics Communications*, vol. 217, no. 1-6, pp. 133–140, 2003.

[10] R. Konrad, N. Padmanaban, K. Molner, E. A. Cooper, and G. Wetzstein, "Accommodation-invariant computational near-eye displays," *ACM Trans. Graph.*, vol. 36, no. 4, July 2017.

[11] U. Akpinar, E. Sahin, and A. Gotchev, "Phase-coded computational imaging for accommodation-invariant near-eye displays," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3159–3163.

[12] U. Akpinar, E. Sahin, and A. Gotchev, "Computational multifocal near-eye display with hybrid refractive-diffractive optics," in *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2020, pp. 1–6.

[13] J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, 1996.

[14] U. Akpinar, E. Sahin, M. Meem, R. Menon, and A. Gotchev, "Learning wavefront coding for extended depth of field imaging," *IEEE Transactions on Image Processing*, vol. 30, pp. 3307–3320, 2021.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, vol. 9351 of *LNCS*, pp. 234–241.

[16] D. G. Voelz, K. A. Bush, and P. S. Idell, "Illumination coherence effects in laser-speckle imaging: modeling and experimental demonstration," *Applied optics (2004)*, vol. 36, no. 8, pp. 1781–1788, 1997.

[17] Y. Deng and D. Chu, "Coherence properties of different light sources and their effect on the image sharpness and speckle of holographic displays," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[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, 2011.

[19] U. Akpinar, E. Sahin, and A. Gotchev, "Learning Optimal Phase-Coded Aperture for Depth of Field Extension," *Proceedings - International Conference on Image Processing, ICIP*, vol. 2019-September, pp. 4315–4319, 2019.

[20] M. Akram and X. Chen, "Speckle reduction methods in laser-based picture projectors," *Optical Review*, vol. 23, 12 2015.