A FRAMEWORK FOR ASSESSING RENDERING TECHNIQUES FOR NEAR-EYE INTEGRAL IMAGING DISPLAYS

Oleksii Doronin, Erdem Sahin, Robert Bregovic, Atanas Gotchev

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

ABSTRACT

We address the problem of 3D scene rendering on near-eye integral imaging displays and evaluation of different rendering methods in terms of human perception. We compare three rendering techniques in terms of perceived spatial resolution at different focused depths, simulating the display in virtual environment and representing the eye through a thin-lens camera model.

***Index Terms*—** Integral imaging, light field, 3d display, sampling, ray tracing.

1. INTRODUCTION

Near-eye displays can provide a highly realistic visual experience for a wide range of VR/AR solutions. While conventional stereoscopic displays offer acceptable spatial resolution in a rather wide field of view, they suffer from to the *vergence-accommodation conflict* (VAC) [1]. Techniques such as varifocal, multi-focal and light field (LF) [2] aim to address the VAC by enabling additional focus cues. The *integral imaging* (InIm) technique, being one of LF ones, is of particular interest due to its relatively simple optical setting [3, 4].

The graphical information for *light field* (LF) displays is usually represented as a set of 2D *display images*. For better immersion, they have to be rendered fast and accurately, which often poses a challenge. Thus, it is critical to have a reliable tool or framework to assess various LF display rendering methods with respect to the *human visual system* (HVS).

In this paper, we briefly discuss the principles of InIm displays and the related near-eye optical setting. Then, we present the framework and assess different rendering techniques in terms of eye-perceived image for a typical near-eye InIm display. For now, it utilizes relatively simple eye model, without considering advanced effects like Stiles-Crawford [5, 6]. This framework is publicly available on GitHub [7].

Related problems of InIm display design have been discussed in [8]. Visual performance of near-eye and close to near-eye displays in terms of wave optics has been addressed in [9] and [10]. More general LF display analysis in terms of LF atoms has been introduced in [11].

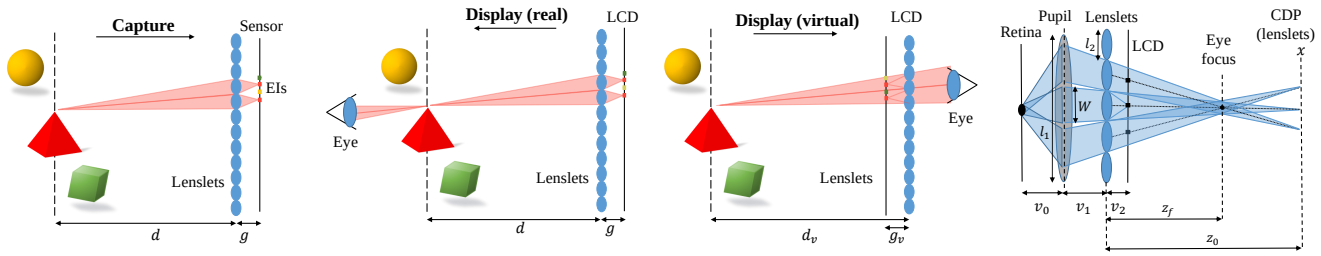
2. INTEGRAL IMAGING

2.1. Overview

InIm provides an efficient way to optically capture and display 3D visual scenes by use of a microlens (lenslets) array. As shown in Figure 1a, during capture, each microlens samples a directional view of the scene, i.e., the angular dimension of the LF defined on some hypothetical LF parametrization plane at the scene (e.g., at d where microlenses are focused). The microlenses sample this directional LF information very densely (in an integral manner), and each *elemental image* (EI) beneath the microlenses record the perspective image at the given location. During display stage, the 3D scene can be reconstructed by mapping the recorded sensor image onto the 2D display, e.g., LCD. Using the direct reconstruction technique as shown in Figure 1b, where $d = d_v$ and the recorded EIs can be directly used on the LCD (assuming that they have the same pixel size), results in pseudoscopic perceived 3D images that are reversed in depth. To obtain orthoscopic images with correct relative depth, one needs to recompute the EIs for the display [12]. In the technique of virtual image reconstruction behind the scene, as shown in Figure 1c, the display elemental images can be found by rotating the captured elemental images 180 degrees around their centers and setting the microlens-to-LCD distance g_v as $g_v = g - 2f^2/(d - f)$, where f is the focal length of the microlenses [8].

2.2. Near-eye InIm display

The InIm display technique has been demonstrated to be an attractive candidate for near-eye displays, particularly due to its capability of delivering focus cues and thus addressing the VAC [3, 4]. The focus cues are basically delivered by providing multiple view images through the eye pupil [13], which is satisfied in InIm by sampling the directional information of LF due to scene by a sufficiently dense set of microlenses. Figure 1d illustrates the near-eye integral setup. Please note that in the near-eye use-case the InIm display is used in the virtual image formation mode as discussed above. As illus-



(a) Scene capture. (b) Real pseudoscopic image. (c) Virtual orthoscopic image. (d) Integral imaging setup.

Fig. 1. Optical capture and display of a 3D scene through integral imaging.

trated in the figure, the display projects multiple rays within the pupil evoking the necessary focus cues for the eye so that the eye is actually focused at the intended object depth, rather than the conjugate image plane of the LCD or central depth plane (CDP) as we call here, to achieve a sharp image on the retina up to the available spatial resolution provided by the InIm display.

As the number of projected rays within the pupil (l_1/W) increases, the focus cues are created more accurately and correspondingly the depth perception is improved. This can be achieved by decreasing the microlens aperture size l_2. However, as a trade-off, the delivered spatial resolution is degraded by up to the same factor (if for the given object depth the magnified elemental images overlap with subpixel shift, this degradation can be less due to superresolution effects). So, it is critical to optimize this trade-off between angular ray density and spatial resolution mainly based on the intended scene depth range. Generally speaking, 3×3 or 4×4 views within the pupil are demonstrated to provide desired accommodation [13]. In this paper we implement the latter case.

2.3. Ray propagation

The simulation of a retinal image basically involves back-projecting a set of rays from each pixel on the retina through the eye pupil. Let us consider a thin-lens eye model (planes P_0 and P_1), and InIm display (planes P_2 and P_3), as in Figure 2a. Given a light ray at point x_0 on the retina plane with a particular direction, in can be traced through the corresponding point x_1 on the plane P_1, point x_2 on P_2 and point x_3 on P_3. Let us denote the optical powers (reciprocal focal length) of the pupil and lenslets by d_1 and d_2 respectively, and consider that the central axis of the optical system goes through the pupil center. Then, the relations between path points are expressed as

$$x_{i+1} = x_i + \omega_i v_i, \quad i = 0, 1, 2, \tag{1a}$$

$$\omega_1 = \omega_0 - x_1 d_1, \quad \omega_2 = \omega_1 - (x_2 - x_c) d_2, \tag{1b}$$

where x_c is the center of the lenslet that intercepts the ray, and ω_i is the angular direction of the ray segment that starts at x_i.

3. DISPLAY IMAGE RENDERING

3.1. Overview

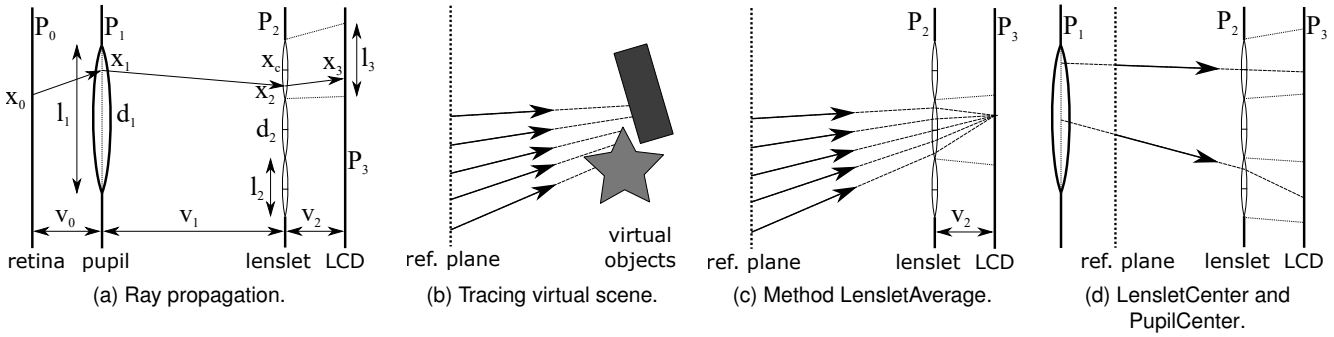
Given a 3D scene to be visualized, one needs to render the corresponding (spatially multiplexed) display image, which, in interaction with the lenslet array, generates the corresponding LF. In general, there are two established techniques to render display image when a synthetic scene is given: rasterization and ray tracing. We favor ray tracing since it is flexible and brings higher realism.

In ray tracing, one image pixel corresponds to its own set of rays, referred to as *camera rays* (or *primary rays*), each of them carrying the light information. These rays are positioned in a virtual world, together with all the objects of the synthetic scene (see Figure 2b). Each ray is traversed until it hits an object, and after this the resulting color of hitpoint is evaluated and stored as the ray result [14]. After all rays for a particular pixel are calculated, their averaged value is stored as a pixel value.

3.2. Camera ray setup

The camera ray setup simulates the physically-based simulation of real-life InIm capturing. Namely, the value of particular display image pixel is determined as the estimated contribution of all light rays that should come to the corresponding location on LCD panel. Numerically this can be evaluated by sampling a discrete number of points on the lenslet surface, and tracing rays that go through these points towards the pixel position (see Figure 2c), including lenslet refraction. After being refracted, rays can be placed in virtual world to calculate the corresponding color contribution (see Figure 2b). Starting position of each ray may be defined with one degree of freedom. Usually, it is defined as the intersection with an imaginary *reference plane* in the virtual world, which can be placed arbitrary. Intuitively, the reference plane can be understood as an analogue of the *near-clip plane*, which is part of camera model in conventional rendering pipeline.

We consider three alternative ways to generate the set of sampling points on the lenslet surface:



(a) Ray propagation. (b) Tracing virtual scene. (c) Method LensletAverage. (d) LensletCenter and PupilCenter.

Fig. 2. Display image rendering.

- LensletAverage (LA) – samples are placed in regular grid pattern (see Figure 2c, we take 5x5 grid);
- LensletCenter (LC) – only one sample at lenslet center (see Figure 2d, the upper lenslet);
- PupilCenter (PC) – only one sample that corresponds to the ray which would hit the pupil center (see Figure 2d, the lower lenslet).

There are two main reasons to apply the LA technique: first, it simulates the real-life InIm capture (see Figure 1a); where many rays are integrated to form a pixel point; second, it models the LF formation and the resulting angular anti-aliasing effect of the lenslet. However, this technique is comparably slow, which may impose burdens when it comes to real-time rendering. Both LC and PC techniques are more efficient, as they assume only on one ray per sample.

The LC technique assumes that during the capturing phase, the lenslet array is replaced by the corresponding pinhole array. Another argument for using this technique is the assumption that the center-most ray that goes through the lenslet center makes the strongest light contribution among the others, and can be considered as a (rough) estimation of the mean value of all contributing rays.

The LC and LA techniques require no knowledge about the expected eye position. If the eye position is known (e.g., by eye-tracking), one can apply the PC technique. It is based on the assumption that the most important ray is the one that goes through the pupil center. In practice, it does not take the pupil size into account assuming it is infinitely small.

4. EXPERIMENTAL RESULTS

4.1. System setup

We test the three candidate rendering techniques in a simulated virtual environment with the following settings: pupil diameter $l_1 = 5$ mm, pupil-to-retina dist. $v_0 = 17$ mm, pupil-to-lenslets dist. $v_1 = 20$ mm, lenslets-to-LCD dist. $v_2 = 6.32$ mm, lenslet size $l_2 = 1.2$ mm, EI size $l_3 = 1.58$ mm. The CPD of such setup is thus placed at $z_0 = 400$ mm, for a point at which there are 4×4 rays are projected within the eye pupil.

The virtual scene is composed of three square-shaped planar objects, placed perpendicular to the eye main direction, at different depths (see Figure 3). Each object contains vector-graphics texture of USAF 1951 resolution test charts. To simulate focusing, we place these objects at three depths, relative to the pupil, as follows: near focus at 300 mm, centre focus at 400 mm, far focus at 600 mm.

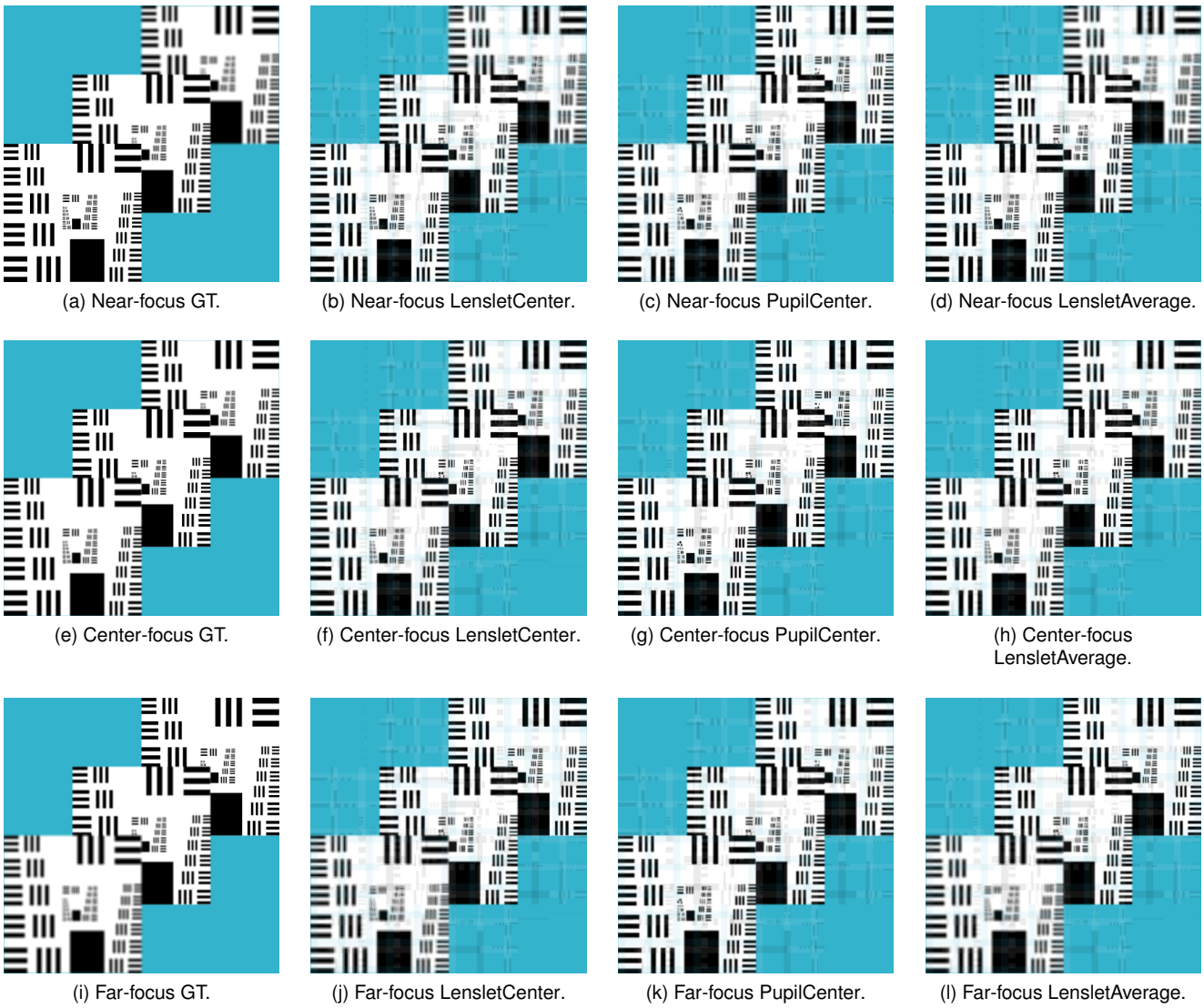
We use two camera models: Pinhole (conventional perspective camera) and Realistic (thin-lens camera simulating the human eye). The ground-truth (GT) images are generated by directly capturing the scene image with the realistic camera model. The ray sampling of the realistic cameras is done through an 8×8 regular grid on the pupil surface, taking into account only those samples that fall into the circle inscribed into 5×5 mm square, which represents the pupil sampling. This is in line with the assumed ray density for evoking focus cues, as discussed in Section 2.2. The pinhole camera represents the whole scene in focus and is used to represent the limit condition of the pupil size, which is of particular importance for the PC technique. The realistic camera is focused at near, centre, and far object during each test case, as shown in Figure 3a, 3e, 3i). For the sake of space we do not show the all-in-focus pinhole images.

4.2. Simulation results

Results of the simulation are shown in Figure 3 and tabulated in Table 1 in terms of PSNR in dB and SSIM [15]. The two metrics are calculated for the corresponding image pairs: a differently focused perceived GT image and a retinal image formed by sampled at the pupil LF generated by a display image, rendered by one of the tested techniques.

	LC	PC	LA	LC	PC	LA
Pin.	21.48	22.31	22.87	0.958	0.961	0.954
Near	24.08	20.85	24.13	0.915	0.902	0.914
Cen.	24.06	22.17	24.36	0.909	0.898	0.908
Far	24.77	21.54	24.90	0.917	0.903	0.917

Table 1. PSNR (left part) and SSIM (right part).



(a) Near-focus GT. (b) Near-focus LensletCenter. (c) Near-focus PupilCenter. (d) Near-focus LensletAverage.

(e) Center-focus GT. (f) Center-focus LensletCenter. (g) Center-focus PupilCenter. (h) Center-focus LensletAverage.

(i) Far-focus GT. (j) Far-focus LensletCenter. (k) Far-focus PupilCenter. (l) Far-focus LensletAverage.

Fig. 3. Rendered images. The GT (ground-truth) designates images obtained directly from the synthetic scene, without display interaction. Near-, Center-, and Far-focus designates the plane at which the simulated eye is focusing.

First of all, in all rendering methods, the focus cues delivered by the near-eye InIm display are observed. That is, the image of the object at the intended depth appears to be sharper when the eye is focused on it. As expected, for the realistic camera models, the LA technique outperforms the two others in all test cases. However, the numbers are pretty close to the ones obtained by LC. For the same realistic camera cases, the PC technique shows inferior performance, which justifies the role of the lenslet in forming the correct LF. For the case of infinitely small pupil, represented by the pinhole camera model, as expected, PC slightly outperforms LC, and even LA in terms of SSIM. Based on the perceived images, a more rigorous analysis can be carried out, e.g., by Fourier domain analysis, to better characterize the capabilities of rendering methods at different spatial frequencies. Furthermore, such analysis would also enable to profile the capabilities of the given display itself.

5. CONCLUSION

The proposed framework [7] is instrumental for assessing and comparing different display designs and LF imaging methods, aimed at visualization on near-eye InIm displays. Our experiments in simulated environment demonstrate that the LC technique is quite competitive with respect to the more elaborated LA one. The performance of the PC technique is comparable to the other two for the case of infinitely small pupil. Further analysis of the perceived images generated by different displays or processing methods by employing Fourier domain methods is part of our future work.

6. REFERENCES

[1] Martin S Banks, David M Hoffman, Joohwan Kim, and Gordon Wetzstein, "3D Displays," *Annual Review of Vision Science*, vol. 2, pp. 397–435, 2016, PMID: 28532351.

[2] Nathan Matsuda, Alexander Fix, and Douglas Lanman, "Focal surface displays," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, July 2017.

[3] Hong Hua and Bahram Javidi, "A 3d integral imaging optical see-through head-mounted display," *Opt. Express*, vol. 22, no. 11, pp. 13484–13491, 2014.

[4] Douglas Lanman and David Luebke, "Near-eye light field displays," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 220:1—-220:10, 2013.

[5] Walter Stanley Stiles and BH Crawford, "The luminous efficiency of rays entering the eye pupil at different points," *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, vol. 112, no. 778, pp. 428–450, 1933.

[6] Gerald Westheimer, "Directional sensitivity of the retina: 75 years of stiles–crawford effect," *Proceedings of the Royal Society B: Biological Sciences*, vol. 275, no. 1653, pp. 2777–2786, 2008.

[7] "LFDisplay project," https://github.com/LeksiDor/LFDisplay, (retrieved May 2020).

[8] Xiao Xiao, Bahram Javidi, Manuel Martinez-Corral, and Adrian Stern, "Advances in three-dimensional integral imaging: sensing, display, and applications," *Applied optics*, vol. 52, no. 4, pp. 546–560, 2013.

[9] Hekun Huang and Hong Hua, "Systematic characterization and optimization of 3d light field displays," *Optics express*, vol. 25, no. 16, pp. 18508–18525, 2017.

[10] Ugur Akpinar, Erdem Sahin, and Atanas Gotchev, "Viewing simulation of integral imaging display based on wave optics," in *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, 2018, pp. 1–4.

[11] Adrian Stern, Yitzhak Yitzhaky, and Bahram Javidi, "Perceivable light fields: Matching the requirements between the human visual system and autostereoscopic 3-d displays," *Proceedings of the IEEE*, vol. 102, no. 10, pp. 1571–1587, 2014.

[12] H. Navarro, R. Martinez-Cuenca, G. Saavedra, M. Martinez-Corral, and B. Javidi, "3d integral imaging display by smart pseudoscopic-to-orthoscopic conversion (spoc)," *Opt. Express*, vol. 18, no. 25, pp. 25573–25583, Dec 2010.

[13] Hekun Huang and Hong Hua, "Systematic characterization and optimization of 3d light field displays," *Opt. Express*, vol. 25, no. 16, pp. 18508–18525, 2017.

[14] Matt Pharr, Wenzel Jakob, and Greg Humphreys, *Physically based rendering: From theory to implementation*, Morgan Kaufmann, 2016.

[15] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.