1-1-2015

Estimation of signal distortion using effective sampling density for light field-based free viewpoint video

Hooman Shidanshidi
University of Wollongong, hooman@uow.edu.au

Farzad Safaei
University of Wollongong, farzad@uow.edu.au

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

Estimation of signal distortion using effective sampling density for light field-based free viewpoint video

Abstract

In a light field-based free viewpoint video (LF-based FVV) system, effective sampling density (ESD) is defined as the number of rays per unit area of the scene that has been acquired and is selected in the rendering process for reconstructing an unknown ray. This paper extends the concept of ESD and shows that ESD is a tractable metric that quantifies the joint impact of the imperfections of LF acquisition and rendering. By deriving and analyzing ESD for the commonly used LF acquisition and rendering methods, it is shown that ESD is an effective indicator determined by system parameters and can be used to directly estimate output video distortion without access to the ground truth. This claim is verified by extensive numerical simulations and comparison to PSNR. Furthermore, an empirical relationship between the output distortion (in PSNR) and the calculated ESD is established to allow direct assessment of the overall video distortion without an actual implementation of the system. A small scale subjective user study is also conducted which indicates a correlation of 0.91 between ESD and perceived quality.

Keywords

signal, distortion, effective, sampling, estimation, density, video, light, field, free, viewpoint

Disciplines

Engineering | Science and Technology Studies

Publication Details

H. Shidanshidi, F. Safaei & W. Li, "Estimation of signal distortion using effective sampling density for light field-based free viewpoint video," IEEE Transactions on Multimedia, vol. 17, (10) pp. 1677-1693, 2015.

Estimation of Signal Distortion Using Effective Sampling Density for Light Field-based Free Viewpoint Video

Hooman Shidanshidi, *Member, IEEE*, Farzad Safaei, *Senior Member, IEEE*, and Wanqing Li, *Senior Member, IEEE*

***Abstract*— In a light field-based free viewpoint system (LF-based FVV), effective sampling density (ESD) is defined as the number of rays per unit area of the scene that has been acquired and is selected in the rendering process for reconstructing an unknown ray. This paper extends the concept of ESD and shows that ESD is a tractable metric that quantifies the joint impact of the imperfections of LF acquisition and rendering. By deriving and analyzing ESD for the commonly used LF acquisition and rendering methods, it is shown that ESD is an effective indicator determined by system parameters and can be used to directly estimate output video distortion without access to the ground truth. This claim is verified by extensive numerical simulations and comparison to PSNR. Furthermore, an empirical relationship between the output distortion (in PSNR) and the calculated ESD is established to allow direct assessment of the overall video distortion without an actual implementation of the system. A small scale subjective user study is also conducted which indicates a correlation of 0.91 between ESD and perceived quality.**

***Index Terms*—Free Viewpoint Video, Light Field, Rendering Quality Assessment**

I. Introduction

FREE VIEWPOINT VIDEO (FVV) [1, 2] aims to provide users the ability to select arbitrary views of a dynamic scene in real-time. A FVV system consists of three main components: *acquisition* [3-7] that captures the scene using a number of cameras, *rendering* [8-15] that reconstructs the desired view from the acquired information, and *compression/transmission* [1, 2, 16-19] of captured or processed information. The performance, in particular the quality of the output video of a FVV system, depends on the efficacy of these components and their collaboration. While existing research studies individual components independently, this paper presents a study on the joint performance of the acquisition and rendering components. The effect of compression is ignored.

In the past, studies of FVV are mainly based on simplified plenoptic signal [20] representation. In particular, by assuming that the viewer is outside of the scene, the 7D plenoptic signal is reduced to a 4D light field (LF) [21, 22]. LF refers to all the rays reflected from every point of the scene in all directions captured outside of the convex hull of the scene and a 'sample' of LF refers to a discrete ray from the scene captured by a single pixel of cameras. Such LF representation has enabled the studies [3-6, 23] on the minimum sampling density under the assumption that the signal of the scene is band-limited and a perfect rendering is available. Results have shown that a very high camera density is required to acquire a light field, which would be infeasible in practice.

On the other hand, reference-based measurements, such as peak-to-signal noise ratio (PSNR) and subjective tests [24] are usually used to assess the rendering component. These measurements require both the ground truth information as well as the output videos of the system, which may be a significant limitation in practice.

It is evident that both acquisition and rendering will contribute simultaneously to the signal distortion of the output video. This is particularly true for a FVV system that works in the *under-sampled regime* where the number of cameras deployed is not adequate to enable error-free reconstruction. To the best knowledge of the authors, there has not been any reported research on the joint impact of the two components on the output video quality. This paper proposes a method to estimate the signal distortion that accounts for both acquisition and rendering. Specifically, this paper

- extends the concept of effective sampling density (ESD) proposed by the authors in [25, 26] and employs it as an indicator of signal distortion for a LF-based FVV system. Calculation of ESD requires neither a reference/ground truth nor the actual output images/video. It can be derived from the key parameters of acquisition and rendering components,
- presents an analytical form of the ESD for the commonly used regular-grid camera systems and rendering algorithms,
- provides theoretical and empirical verification of ESD as an effective indicator of signal distortion,
- compares ESD with PSNR, establishes an empirical relationship between them, and verifies the correlation between ESD and perceived quality through a subjective test.

The rest of the paper is organized as follows. Section II reviews the related work. Section III analyses the acquisition and rendering components and describes in detail the concept of ESD. Section IV presents the application of ESD to analyze

Manuscript submitted 13 Sep. 2013; approved for publication 30 Apr. 2015.

The authors are with the ICT Research Institute, Faculty of Engineering, University of Wollongong, Australia. E-mails: hooman@uow.edu.au, farzad@uow.edu.au, wanqing@uow.edu.au.

LF systems with commonly used regular-grid cameras and rendering methods. Numerical simulation and validations are presented in Section V. Section VI presents the empirical relationship between the ESD and PSNR. Section VII reports the subjective test and its correlation with ESD. Section VIII concludes the paper with remarks.

II. Related Work

This section provides a review of the existing approaches for evaluating LF acquisition and rendering methods.

A. Evaluation of the Acquisition Component

Light field can be expressed as a simplified four dimensional plenoptic signal [20], first introduced by Levoy and Hanrahan [21] and Gortler et al [22] (as Lumigraph) in mid-1990s. LF acquisition aims to sample the plenoptic signal by using limited number of cameras configured in 3D space. Several parameterization schemes have been proposed to represent the camera configurations and the rays captured by the cameras. For instance, Levoy and Hanrahan [21] employed a regular grid of cameras and represented the rays by using their intersection points with two parallel planes/slabs defined by variables (s, t, u, v) respectively, where (s, t) represents the image plane and (u, v) represents the camera plane. The 4D space is then represented as a set of oriented lines, i.e., *rays* in 3D space. This parallel plane parameterization has been enhanced by more complicated parameterization schemes such as Two-Sphere (2SP) and Sphere-Plane Parameterization (SPP) [27].

Existing approaches for evaluating LF acquisition mainly focus on the minimum required sampling density for error-free signal reconstruction. Two major approaches have been adopted so far. The first one is based on plenoptic signal spectral analysis [3, 23] and, more specifically, the light field spectral and frequency analysis [4, 5]. In this approach the spectral analysis is applied to a surface plenoptic function (SPF) representing the light rays starting from the object surface and the minimum sampling density is estimated based on the sampling theory by computing the Fourier transform of the light field signal. However, the spectrum of a light field is usually not band-limited due to non-Lambertian reflections, depth variations and occlusions. Therefore, approximations such as the first-order approximation [1-2] is often applied to the signal by assuming that the range of depth is limited.

The second approach is based on the view interpolation geometric analysis rather than frequency analysis. This approach is based on blurriness and ghost (shadow)-effect error measurements and elimination in rendered images. In [6] the artifact of "double image" (a geometric counterpart of spectral aliasing) is proposed to measure the ghost effect for a given acquisition configuration. This artifact is geometrically measured by calculating the intensity contribution of rays employed in interpolation. Finally, the minimum sampling density is calculated to avoid this error for all points in the scene. This approach can be used to derive the minimum sampling curve against scene depth information, showing how the adverse effect of depth estimation error can be compensated by increasing the sampling density, i.e., the number of cameras. This method is more flexible, especially for irregular capturing and rendering configurations, and leads to a more accurate and smaller sampling density compared with the first approach.

In addition to these two approaches, optical analysis by considering light field as a virtual optical imaging system is also employed in acquisition analysis [28, 29]. The original light field [21] shows that the distance between two adjacent cameras can be considered as the aperture for ray filtering. This concept is generalized in [13] by introducing a "discrete synthetic aperture", encompassing of several cameras. It is also shown in [13] that the size of this synthetic aperture can change the field of view very similar to an analog aperture. This optical analysis is mostly used to calculate the optimum light field filtering [30].

Due to the assumption of perfect signal reconstruction, all of these approaches result in very high sampling densities, which are hardly achievable in practice. For instance [3] shows that for a typical scenario a camera grid with more than 10,000 cameras is required. They also assume general Whittaker–Shannon interpolation method for signal reconstruction. However, having some geometric information about the scene, such as estimated depth map, could enable more sophisticated interpolation for signal reconstruction and rendering. Consequently, an indicator to measure signal distortion without any reference or ground truth, that works in the *under-sampled regime* is desirable.

B. Evaluation of the Rendering Methods

Along with the acquisition configuration and parameterization schemes, different LF rendering methods have been developed to generate images for arbitrary viewpoints from the captured rays by implicitly or explicitly using geometric information about the scene [31]. These include layered light field [8], surface light field [9] , scam light field [10], pop-up light field [11], all-in-focused light field [12], and dynamic reparameterized light field [13].

Previous works on FVV evaluation and quality assessment with respect to rendering are mainly based on the methods proposed for Image based Rendering (IBR) and are not specifically for LF rendering. Often pixel-wise error metrics such as PSNR with respect to ground-truth images are employed for quality assessment [32]. Ground-truth data is provided by employing a 3D scanner for a real scene or virtual environments such as [33]. In [34], two scenarios are analysed: human performance in a studio environment and sports production in a large-scale environment. A method was introduced for both studio and large-scale environment to quantify error at the point of view synthesis [34]. This method was used as a full-reference metric to measure the fidelity of the rendered images with respect to the ground-truth as well as a no-reference metric to measure the error in rendering. In the no-reference metric, without explicitly having the ground truth, a virtual viewpoint is placed at the mid-point between the two cameras in a camera grid. From this viewpoint, two images are rendered, each using one set of the original cameras. These images are then compared against each other with the same metrics as before.

Quality evaluation has also been carried out with two different categories of metrics, modelling the human visual system (HVS) and employing more direct pixel fidelity

indicators. HVS-based measures of the fidelity of an image include a variety of techniques such as measuring mutual information in the wavelet domain [35], contrast perception modelling [36] and modelling the contrast gain control of the HVS [37]. However, HVS techniques and objective evaluation of a visual system are not able to fully model the human perception as discussed in [38-40]. Pixel-wise fidelity metrics such as MSE and PSNR are simple fidelity indicators but with a low correlation with visual quality [41]. In [42] a full review of pixel-wise fidelity metrics is discussed. Also [43] shows a statistical analysis of pixel metrics and HVS-based metrics.

While the need for analytical quality evaluation of FVV systems is highlighted in several studies such as [44, 45], the current research on LF rendering evaluation and quality assessment focuses mostly on case-based study of applying these metrics. Little development has been reported on an analytical model that can evaluate LF rendering methods. In contrast, the proposed ESD provides an analytical evaluation of the effect of LF rendering as well as LF acquisition on the final video distortion.

III. EFFECTIVE SAMPLING DENSITY (ESD)

Fig. 1 shows a general FVV system that utilizes depth information. The light field is sampled by multiple cameras through the *ray capturing* process, which results in a certain sampling density (SD). SD at a given location is defined as the number of rays acquired per unit area of the convex hull of the surface of the scene in that location. The acquisition can have a variety of configurations, such as regular/irregular 2D or 3D camera grids or even a set of mobile cameras at random positions and orientations.

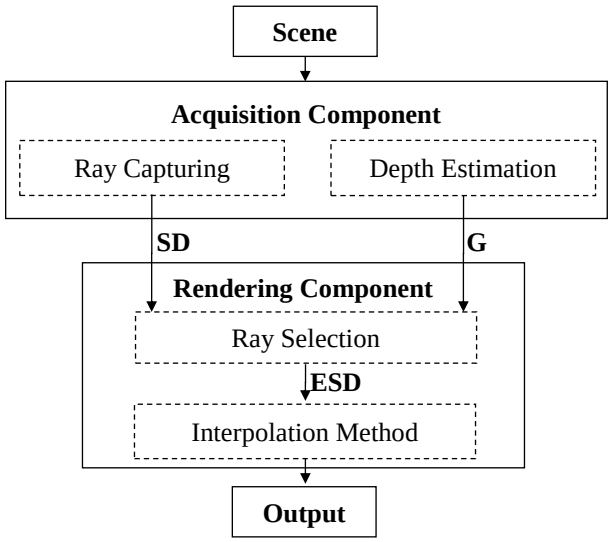

Fig. 1. The schematic diagram of a typical LF-based FVV system that utilizes scene geometric information G

In addition, the *depth estimation* process provides an estimation of depth (e.g. depth map) to improve rendering. This could be obtained by specialized hardware, such as depth cameras, or computed from the images obtained by cameras. In either case, the depth estimation will have some error.

To estimate/reconstruct an unknown ray r from the acquired rays and the depth information, the rendering essentially goes through two processes: (i) the *ray selection* that chooses a subset of acquired rays, purported to be in the vicinity of r, for the purpose of interpolation; and (ii) the *interpolation* that provides an estimate of r from these rays.

The *ray selection process*, in particular, is often prone to error. For example, imperfect knowledge of depth may cause this process to miss some neighboring rays and choose others that are indeed sub-optimal (with respect to proximity to r) for interpolation. Consider the case shown in Fig. 2, where the actual surface is at depth d and the unknown ray r intercepts the object at point p. There are four rays r_1, r_2, r_3, and r_4 captured by the cameras that lie within the interpolation neighbourhood of p, shown as a solid rectangle, and could be used to estimate r. However, since the estimation of depth is in error by Δd, the algorithm would select four other rays, r_1', r_2', r_3', and r_4' as the closest candidates for interpolation. As a result, the sampling density has been effectively reduced from $4/A$ to $4/A'$, where A and A' are the areas of solid and dashed rectangles in the Figure respectively. In addition, the rendering algorithm may not be able to use all available rays for interpolation due to computational constraint.

The output of this process, therefore, represents an *effective sampling density* (ESD) which is *lower* than the SD obtained by the cameras and distortion is inevitably introduced in the reconstructed video. ESD is defined as the number of rays per unit area of the scene that have been captured by *acquisition* component and chosen by *ray selection process* to be employed in the rendering. Clearly, ESD $\leq$ SD with equality holding only when the rendering process has perfect knowledge of depth and sufficient computational resources. Not surprisingly, ESD can be a true indicator of output quality, *not* SD, and its key advantage is that it provides an analytically tractable way for evaluating the influence of the imperfections of *both* acquisition and rendering components.

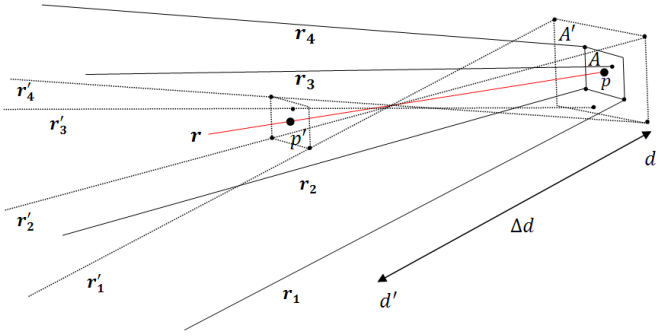

Fig. 2. Selection of rays in a LF rendering and the concept of ESD

Let Θ be the set of all rays captured by the cameras. The *ray selection mechanism* M chooses a subset ω of rays from Θ. Subsequently, an *interpolation function* F is applied to ω to estimate the value of the unknown ray r. A is an imaginary convex hull area around p which intersects with all the rays in ω at depth d. The size of A would depend on the choice of ω, hence, the rendering method. Since each squared pixel in an image sensor integrates light rays coming within a squared-based pyramid extending towards the scene. The cut area (square) of this pyramid at distance d is roughly $ld \times ld$, where l is the size of the pixel determined by camera resolution. Therefore, the minimum length of the sides of A is ld, which is referred to as the system resolution in this paper.

There are usually more rays from Θ passing through A, but are not selected by the ray selection process probably because of limited computing resources or real-time requirement. Let all the captured rays passing through A be denoted by Ω. Clearly:

$$\omega \subseteq \Omega \subseteq \theta \quad (1)$$

Both M and F may or may not use some kind of scene geometric information G such as focusing depth (average depth of the scene computed from automatic focusing algorithms or camera distance sensors) or depth map. Mathematically, the rendering can be formulated as

$$\omega = M(\Theta, G) \quad (2)$$

$$r = F(\omega, G) \quad (3)$$

Different rendering methods differ in their respective M and F functions and their auxiliary information G.

Based on these definitions SD and ESD can be expressed as

$$\mathrm{SD} = \frac{|\Omega|}{A} \quad (4)$$

$$\mathrm{ESD} = \frac{|\omega|}{A} = \frac{|M(\Theta,G)|}{A} \quad (5)$$

where $|\Omega|$ and $|\omega|$ are the number of rays in Ω and ω respectively. A is the area of interpolation convex hull, and can be calculated by deriving the line equations for the boundary rays β_i's and finding the vertexes of convex hull A at depth d. Fig. 3 shows this process for a simple 2D LF acquisition, generated by applying a 2D projection to a 3D light field with 2 planes parameterization, that is, camera plane uv and image plane st over (u, s). Assume that rays in ω are surrounded by the boundary rays β_1 and β_2. The rays in ω are selected by the selection method M and are bounded by $n+1$ cameras in u (u_i to u_{i+n}) and $m+1$ pixels in s (s_j to s_{j+m}). As it can be seen, A is at least a function of k, l , n, m and d, where k is the distance between the cameras, l is the pixel length, n and m are the number of cameras and pixels bounded by boundary rays respectively, and d is the depth of p. The rays intersect with A from these $n+1$ cameras are the rays employed by rendering method, i.e., ω set. However, as it is shown in Fig. 3, there are more than $n+1$ cameras in the grid, (in addition to cameras bounded between u_i to u_{i+n}) that are able to see area A. u_x is shown as an example of these cameras. The rays from these cameras to A, make up the difference between Ω and ω sets.

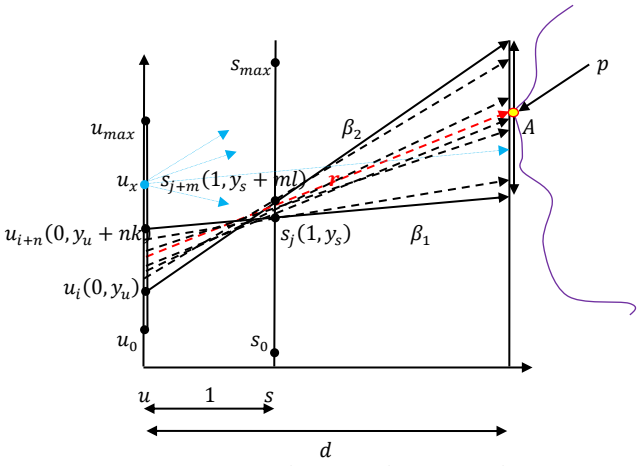

Fig. 3. ESD calculation for a simplified 2D light field system

SD defined in (4) provides the upper bound of ESD. In general, for a given LF acquisition configuration, it is possible to calculate SD on any point over the scene space analytically or numerically. SD is generally not uniform across the field of view, even when a regular camera grid is used in capturing. Fig. 4.a shows the SD contour maps at different depths, $d = 30m$, $60m$, and $90m$, for a regular camera grid of $30x30$ with $k = 2m$, camera field of view of 30°, image resolution of $100x100$ pixels, i.e., $l = 0.53cm$ in image plane st, and ideal area $A = (ld)^2$, i.e., LF system resolution. Fig. 4.b shows a 2D slice where d ranges in $[2m, 100m]$.

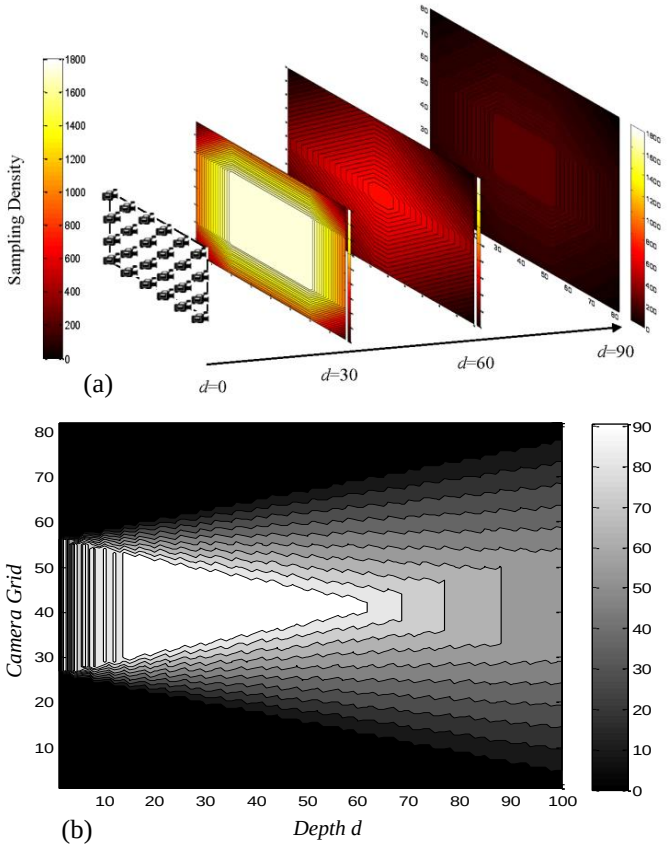

Fig. 4. a) SD contour maps at different depths in 3D; b) SD contour map in 2D

Based on the discussion above, it can be speculated that the output quality of an arbitrary view is determined by three key factors: ESD in each area A, the vicinity of the unknown rays that compose the view, scene complexity in each area A, which could be measured in terms of its spatial frequency components, and the interpolation function F employed for the estimation of the unknown rays.

In particular, for a fixed scene complexity and a given interpolation algorithm, ESD can be used to analytically estimate the signal distortion of a given camera configuration and an adopted rendering algorithm.

IV. ESD Analysis of LF Rendering Methods

Without loss of generality, a simple regular-grid camera system, as shown in Fig.3, is adopted in this section. ESD analysis is presented for different rendering algorithms, specifically, those with and without using depth information. However, the analysis can be extended to other acquisition systems [27]. For a regular-grid camera system, analytical

form of ESD can be obtained for a rendering algorithm with and without using depth information.

A. Rendering Methods without the Depth Information

The LF rendering methods without using depth information, hereafter referred to as *blind* methods, can be categorized into four main groups based on their ray selection mechanism M: Nearest Neighbourhood estimation (NN), 2D interpolation in camera plane (UV), 2D interpolation in image plane (ST) and a full 4D interpolation in both camera and image planes (UVST) [21, 46]. For interpolation function F, bilinear interpolation is often used for the 2D interpolation and a quadrilinear interpolation for the 4D interpolation. However, when $|\omega| > 4$ for UV and ST and when $|\omega| > 16$ for UVST, the convex hull A may not be a grid anymore and other types of 2D and 4D interpolation function F could be employed as discussed in subsection C.

Considering the regular geometry of the cameras shown in Fig.3, analytical form of ESD for these rendering algorithms can be derived. Table I summarizes the ESD derivation for the NN, ST, UV, and UVST methods where $|\omega| = 4$ for UV and ST and $|\omega| = 16$ for UVST. For each one of these rendering methods, the details of selection mechanism M and interpolation function F are given in the second and third columns. The fourth column summarizes the sampling /interpolation length A. Notice that A is a segment in the chosen 2D LF system whereas it is an area in 3D. The fifth column lists the corresponding ESD.

With the analytical ESD forms shown in Table I, it is possible to objectively compare these rendering methods in terms of the signal distortion for the same acquisition. The higher the ESD is, the less distortion is expected. Since when $|\omega|$ is fixed, ESD is a function of the sampling/interpolation area A. The ratio γ of A between two rendering methods is used as a factor for comparison.

Table II summarizes the comparison. The first column shows a pair of rendering methods to be compared, the second column is the ratio γ, the third column gives the relationship between the corresponding ESDs, the fourth column is the minimum value of γ for each pair. Specifically, three particular scenarios are analysed and their corresponding γ are shown in the fifth column of Table II.

Scenario One: $d \to \infty$ and $k \gg l$, which represents a typical low density camera grid and a scene that is very far from the cameras. In this case, the analysis shows that, $4\mathrm{ESD_{NN}} < 4\mathrm{ESD_{UV}} < \mathrm{ESD_{ST}} < \mathrm{ESD_{UVST}}$. In other words, UVST has the highest ESD and is expected to produce the video with least distortion. NN has the lowest ESD and therefore would generate the output with a larger distortion.

Scenario Two: $d \to \infty$ and $k \cong l$, a hypothetical very high density camera grid for a scene that is very far from the grid. The analysis indicates that, $1.7\mathrm{ESD_{NN}} < \mathrm{ESD_{UV}} < \mathrm{ESD_{ST}}$, $4\mathrm{ESD_{NN}} < \mathrm{ESD_{UVST}}$, and $2.2\mathrm{ESD_{UV}} < 2.2\mathrm{ESD_{ST}} < \mathrm{ESD_{UVST}}$. This shows the same order as first scenario, but both NN and UV methods work much better in comparison with ST, though UVST still has the best performance.

Scenario Three: $d \cong 1$, a hypothetical scene very close to the image plane. The analysis indicates that $4\mathrm{ESD_{NN}} < 4\mathrm{ESD_{ST}} < \mathrm{ESD_{UV}} < \mathrm{ESD_{UVST}}$. This shows that UV outperforms ST in such a scenario with ESD more than four times higher than ST. Hence, for a scene close to the grid, UV is a better choice for rendering method compared with ST, which is intuitively appealing.

Similar analysis can be applied to other scenarios, which can offer a choice of rendering algorithms for a given acquisition system.

Table I: ESD for the LF rendering methods without using depth information [25]

Rendering method	*Selection Mechanism M*	*Interpolation Function F*	*Sampling/Interpolation length A in 2D LF*	ESD *for symmetric 3D light field*
NN	Select the nearest ray in 4D space, $\|\omega\| = 1$	No interpolation, neighbourhood estimation	$A_{NN} = (\frac{l+k}{2})d - \frac{k}{2}$	$ESD_{NN} = \frac{1}{{A_{NN}}^2}$
ST	Select 4 or more rays from the neighbourhood pixels in st plane to the nearest camera in uv plane, $\|\omega\| \geq 4$	Any type of 2D interpolation, e.g., bilinear interpolation for 2D grid selection of rays	$A_{ST} = (l + \frac{k}{2})d - \frac{k}{2}$	$ESD_{ST} = \frac{4}{{A_{ST}}^2}$
UV	Select 4 or more rays from the neighbourhood cameras in uv plane to the nearest pixel in the st plane, $\|\omega\| \geq 4$	Any type of 2D interpolation, e.g., bilinear interpolation for 2D grid selection of rays	$A_{UV} = (k + \frac{l}{2})d - k$	$ESD_{UV} = \frac{4}{{A_{UV}}^2}$
UVST	Select 16 or more rays from four neighbourhood cameras in uv to four neighbourhood pixels in st, $\|\omega\| \geq 16$	Any type of 4D interpolation, e.g., quadrilinear interpolation for grid selection of rays	$A_{UVST} = (l + k)d - k$	$ESD_{UVST} = \frac{16}{{A_{UVST}}^2}$

Table II: Comparison of ESD of the LF rendering methods without using depth information [25]

Methods	*Sampling length comparison*	ESD *comparison*	γ *(the ratio of* ESD*'s)*	γ *Analysis*
NN vs. ST	$A_{NN}.\gamma > A_{ST}$	$ESD_{NN}.\frac{4}{\gamma^2} < ESD_{ST}$	$\gamma > 1 + \frac{ld}{(l+k)d-k}$	$d \to \infty$ *and* $k \gg l \Rightarrow \gamma = 1$ $d \to \infty$ *and* $k \cong l \Rightarrow \gamma = 1.5$ $d \cong 1 \Rightarrow \gamma = 2$
NN vs. UV	$A_{NN}.\gamma > A_{UV}$	$ESD_{NN}.\frac{4}{\gamma^2} < ESD_{UV}$	$\gamma > 1 + \frac{kd-k}{(l+k)d-k}$	$d \to \infty$ *and* $k \gg l \Rightarrow \gamma = 2$ $d \to \infty$ *and* $k \cong l \Rightarrow \gamma = 1.5$ $d \cong 1 \Rightarrow \gamma = 1$
NN vs. UVST	$A_{NN}.\gamma > A_{UVST}$	$ESD_{NN}.\frac{16}{\gamma^2} < ESD_{UVST}$	$\gamma > 2$	$\gamma > 2$

ST vs. UVST	$A_{ST}.\gamma > A_{UVST}$	$ESD_{ST}.\frac{4}{\gamma^2} < ESD_{UVST}$	$\gamma > 1 + \frac{d-1}{(\frac{2l}{k}+1)d-1}$	$d \to \infty \text{ and } k \gg l \Rightarrow \gamma = 2$ $d \to \infty \text{ and } k \cong l \Rightarrow \gamma = 1.33$ $d \cong 1 \Rightarrow \gamma = 1$
UV vs. UVST	$A_{UV}.\gamma > A_{UVST}$	$ESD_{UV}.\frac{4}{\gamma^2} < ESD_{UVST}$	$\gamma > 1 + \frac{ld}{(l+2k)d-2k}$	$d \to \infty \text{ and } k \gg l \Rightarrow \gamma = 1$ $d \to \infty \text{ and } k \cong l \Rightarrow \gamma = 1.33$ $d \cong 1 \Rightarrow \gamma = 2$
ST vs. UV	$A_{UV} > \gamma.A_{ST}$	$ESD_{UV}.\gamma^2 < ESD_{ST}$	$\gamma < 1 + \frac{(k-l)d-k}{(2l+k)d-k}$	$d \to \infty \text{ and } k \gg l \Rightarrow \gamma = 2$ $d \to \infty \text{ and } k \cong l \Rightarrow \gamma = 1$ $d \cong 1 \Rightarrow \gamma = 0.5$

B. Rendering Methods with the Depth Information

Utilization of depth information G in rendering can compensate to some extent for insufficient number of samples acquired in an *under-sampling* situation [47]. It can make the ray selection mechanism M more effective compared with blind rendering methods. The amount of depth information G could vary from a crude estimate, such as the focusing depth, to the full depth map or even full 3D geometric model of the scene. A mechanism M in this case may choose a number of rays intersecting the scene in the vicinity of point p at depth d. A rendering method whose interpolation function F is a 2D interpolation over uv plane and utilizes only the focusing depth is referred to as UV-D (**UV**+**D**epth) and the one with a full depth map is referred to as UV-DM (**UV**+**D**epth **M**ap). By extending the selection mechanism M and interpolation function F to a full 4D interpolation over both uv and st planes, the rendering methods are referred to as UVST-D (**UVST**+**D**epth) and UVST-DM (**UVST**+**D**epth **M**ap) respectively, the former using focusing depth only. Many LF rendering methods with depth information can be mathematically expressed in the form of one of these 4 groups. These include layered light field [8], surface light field [9], scam light field [10], pop-up light field [11], all-in-focused light field [12], and dynamic reparameterized light field [13].

Again, without loss of generality, we study the cases where $|\omega| = 4$ and bilinear interpolation as F for UV-D and UV-DM and $|\omega| = 16$ and quadrilinear interpolation as F for UVST-D and UVST-DM.

Fig. 5 illustrates the rendering methods with depth information. If the exact depth d at point p, the intersection of unknown ray r with the scene, is known, applying a back projection can find a subset of known rays Ω intersecting the scene at the vicinity of p. Subsequently, an adequate subset ω of these rays can be selected by mechanism M to be employed in interpolation F.

However, in practice, the estimated depth of p has an error Δd. This makes the rays intersect in an imaginary point p' in the space and going through the vicinity of area A on the scene instead of intersecting with the exact point p on the scene surface. Subsequently, this estimation error Δd would result in reduction of ESD and increase the distortion. To compute Ω in this case, back projection should be applied to the vertexes of A and not p to find all the rays passing through A.

The size of area A depends on Δd and as Δd gets larger, it also increases. Usually only the upper bound of the error is known and therefore in this paper, the worst-case scenario, i.e., largest A is computed in the LF analysis which corresponds to the lower bound of ESD.

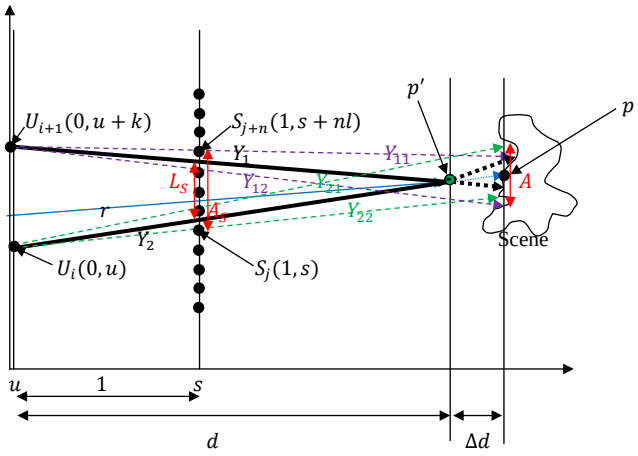

Fig. 5. Light field rendering methods using depth information (UV-D, UVST-D, UV-DM /UVST-DM) with Δd error in depth estimation

Considering scenario in Fig. 5, Y_1 and Y_2 are two immediate neighbour rays, intersecting with the desired ray r at depth d on object surface. If these two rays don't pass through the known s values in image plane, Y_1 from Y_{11} and Y_{12} and Y_2 from Y_{21} and Y_{22} can be estimated. Finally, a bilinear interpolation in uv plane (or a linear interpolation over u in this 2D example) is applied to estimate r from Y_1 and Y_2.

Here, ω includes only two samples for UV-D/UV-DM and four samples for UVST-D/UVST-DM though all acquired rays that intersect the object surface at point p in vicinity A at depth d can be employed in the rendering ($\omega = \Omega$) to reduce distortion. Y_{12} and Y_{21} are boundary rays used for interpolation. If the depth estimation has no error, i.e., $\Delta d = 0$, then, $A_S = L_S + \frac{l}{2} + \frac{l}{2} = \frac{k(d-1)+ld}{d}$, $A_{UVD/UVDM} = ld$ and $A_{UVSTD/UVSTDM} = 2ld$. In a case that $\Delta d > 0$, p is somewhere in the range of $d \pm \Delta d$, and the sampling area A would be increased to:

$$A = max[|Y_{11}(d+\Delta d) - Y_{22}(d+\Delta d)|, |Y_{12}(d+\Delta d) - Y_{21}(d+\Delta d)|] = l(d+\Delta d) + \frac{\Delta d.k}{d} \quad (6)$$

Using this approach, it can be shown that the difference between the rendering methods with focusing depth (UV-D/UVST-D) and the rendering methods with full depth map (UV-DM/UVST-DM) is in the scale of Δd. For focusing depth, a fixed depth is used for all points of the scene. This makes the depth estimation error, $\Delta d = \frac{object\ length}{2} + focusing\ depth\ estimation\ error$. When the full depth map of the scene is used as G, the depth of each point p of the scene possibly with some estimation error Δd is known. Δd is much less than the focusing depth error, which makes the UV-DM/UVST-DM rendering less distorted than UV-D/UVST-D.

C. General Case of Rendering Methods with Depth Maps

Fig. 6 demonstrates a LF rendering method with 2 plane parameterization using a depth map as the auxiliary information G. Again ray r is the unknown ray that needs to be estimated for an arbitrary viewpoint reconstruction. r is assumed to intersect the scene on point p at depth d.

In Fig. 6, seven rays from all rays intersecting imaginary p are selected by M, i.e., $|\omega| = 7$, assuming these rays pass through known pixel values or if neighbourhood estimation is used. In the case of bilinear interpolation in st plane, 28 rays are chosen by M to estimate these 7 rays. The chosen cameras in uv plane are bounded by a convex hull A'. It is easy to show that interpolation convex hull A is proportional to A'.

Finally a 2D interpolation F over convex hull A' on uv plane can be applied to estimate unknown ray r from the rays in ω. This rendering method with depth information is a generalization of UV-DM described in subsection B but with arbitrary number of rays for interpolation when 2D interpolation is performed over neighbouring cameras in the uv plane and neighbourhood estimation, i.e., choosing the closest pixel in the st plane. Again the generalization of UVST-DM is in the case of 2D interpolation over neighbouring cameras in the uv plane and bilinear interpolation over neighbouring pixels in the st plane.

In a simple form of UV-DM and UVST-DM, the rays in ω are selected in a way that A' becomes rectangular, i.e., 2D grid selection and therefore 2D interpolation over A' can be converted into a familiar bilinear interpolation.

The ESD for the UV-DM and UVST-DM demonstrated in Fig. 6 can be derived as:

$$\mathrm{ESD}_{\mathrm{UVDM}} = \frac{|\omega|}{A} = \frac{|\omega|}{\frac{\Delta d}{d}A' + \mu(l(d+\Delta d), A')} \tag{7}$$

$$\mathrm{ESD}_{\mathrm{UVSTDM}} = \frac{|\omega|}{A} = \frac{|\omega|}{\frac{\Delta d}{d}A' + \mu(2l(d+\Delta d), A')} \tag{8}$$

where μ is a function to calculate the effect of pixel interpolation over st plane on the area A. A is mainly determined by A', but the pixel interpolation μ which is added to (7) and (8) also has small effect on A. The pixel interpolation over st even when $\Delta d = 0$ makes $A = (ld)^2$.

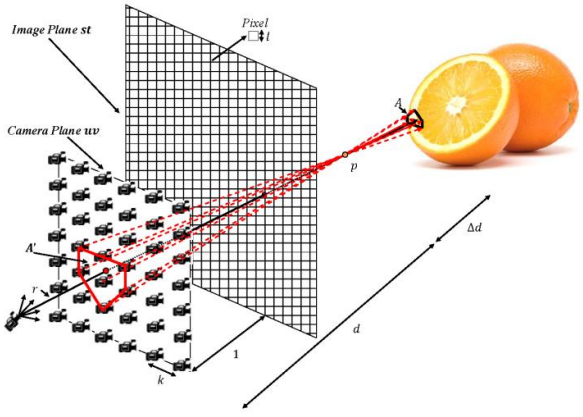

Fig. 6. General light field rendering method using depth information (UV-DM /UVST-DM) with Δd error in depth estimation

Simple forms of UV-DM and UVST-DM described in subsection B can be formulated for a regular camera grid and 2D grid selection of rays, i.e., A' as a rectangular area with 4 and 16 samples in $|\omega|$ respectively, then (7) and (8) become:

$$\mathrm{ESD}_{\mathrm{UVDM}} = \frac{4}{\left(\frac{\Delta d.k}{d} + l(d+\Delta d)\right)^2} \tag{9}$$

$$\mathrm{ESD}_{\mathrm{UVSTDM}} = \frac{16}{\left(\frac{\Delta d.k}{d} + 2l(d+\Delta d)\right)^2} \tag{10}$$

where k is the distance between the two neighbouring cameras in the cameras grid and l is the length of the pixel in the image plane as illustrated in Fig. 6. Note that the edge of A' rectangular is equal to k and that is how (9) and (10) are derived from (7) and (8).

Mathematically, a general representation of simplified UV-DM rendering method with arbitrary number of rays for interpolation is $r = \mathrm{UVDM}(d, \Delta d, k, l, |\omega|)$. By extending (9) and considering the edge of A' rectangular to be equal to $(\sqrt{|\omega|} - 1)k$, the ESD could be calculated for $\mathrm{UVDM}(d, \Delta d, k, l, |\omega|)$ as follows:

$$\mathrm{ESD}_{\mathrm{UVDM}(d,\Delta d,k,l,|\omega|)} = \frac{|\omega|}{\left(l(d+\Delta d) + \frac{\Delta d.k}{d}(\sqrt{|\omega|}-1)\right)^2} \tag{11}$$

Equation (11) assumes that the rays are chosen for interpolation symmetrically around the vertical and horizontal axes, such as $4x4$. In this case, $\sqrt{|\omega|}$ would be an integer.

ESD for the rendering methods using either focusing depth or depth maps can be analytically derived based on the geometry of the regular grid camera system as described in Fig. 5 and Fig. 6 and (6) to (11). Table III summarizes derivation. The first column shows the rendering methods: UV-D and UVST-D methods that use focusing depth and UV-DM and UVST-DM that use depth maps, with $|\omega| = 4\ or\ 16$ and $|\omega| > 4\ or\ 16$. The second and third columns describe the selection mechanism M and interpolation function F respectively. The fourth and fifth column give the sampling/interpolation length A and ESD respectively.

Table IV summarizes comparison of the ESD among UVST, UV-D, and UVST-D. It is clear from Table III that (UV-DM and UV-D) and (UVST-DM and UVST-D) have the same ESD, the difference between them being the scale of Δd, thus UV-DM and UVST-DM are omitted in Table IV. Similar to the analysis of the blind methods, ratio γ is used and two scenarios, one with $d \to \infty, k \cong l\ and\ \Delta d \ll d$ and the other with $d \to \infty, k \gg l\ and\ \Delta d \ll d$ are analysed. The second scenario corresponds to a typical FVV system where the scene is far from the grid, depth estimation error is small compared with the depth and there are a finite number of cameras.

The γ values allow us to compare the rendering methods with and without using depth information. Table II and Table IV have shown that: $4\mathrm{ESD}_{\mathrm{NN}} < 4\mathrm{ESD}_{\mathrm{UV}} < \mathrm{ESD}_{\mathrm{ST}} < \mathrm{ESD}_{\mathrm{UVST}} \ll \mathrm{ESD}_{\mathrm{UVD/UVDM}} < \mathrm{ESD}_{\mathrm{UVSTD/UVSTDM}}$,i.e., for a given acquisition, the NN rendering method has the lowest ESD and hence results in the highest video distortion following by UV, ST, UVST, UV-D/UV-DM, and UVST-D/UVST-DM respectively. The experimental validation in next section will not only confirm this, but also show that ESD is highly correlated with PSNR.

Equations shown in Table III and Table IV can be used in LF system analysis and design. In addition to LF system evaluation and comparison, by knowing the upper bound of the depth estimation error, optimum system parameters such as camera density k, cameras resolution in terms of l , and

rendering complexity in terms of number of rays employed in interpolation $|\omega|$ can be theoretically calculated. For example, in [26], the authors have used the above relationships to obtain the minimum camera density for capturing a scene. We will show in future publications how ESD can be used to optimize the acquisition and rendering parameters of a LF system individually and jointly for a target output video quality.

Table III: ESD for the LF rendering methods with depth information [25]

Rendering method category	*Selection Mechanism M*	*Interpolation Function F*	*Sampling/Interpolation length A in 2D LF*	ESD *for symmetric 3D light field*
UV-D $\|\omega\| = 4$	Select 4 rays sourcing from neighbourhood cameras in uv and intersecting with expected p	Neighbourhood estimation in st and 2D interpolation over uv	$A_{UVD} = l(d + \Delta d) + \frac{\Delta d.k}{d}$	$ESD_{UVD} = \frac{4}{A_{UVD}{}^2}$
UVST-D $\|\omega\| = 16$	Select 16 rays sourcing from neighbourhood cameras in uv, through known pixels in st and intersecting with expected p	4D interpolation over st and uv planes, e.g., quadlinear interpolation	$A_{UVSTD} = 2l(d + \Delta d) + \frac{\Delta d.k}{d}$	$ESD_{UVSTD} = \frac{4}{A_{UVSTD}{}^2}$
UV-DM $\|\omega\| = 4$	The same as UV-D but with more accurate depth estimation of p employing depth maps.	The same as UV-D	$A_{UVDM} = l(d + \Delta d) + \frac{\Delta d.k}{d}$	$ESD_{UVDM} = \frac{4}{A_{UVDM}{}^2}$
UVST-DM $\|\omega\| = 16$	The same as UVST-D but with more accurate depth estimation of p employing depth maps.	The same as UVST-D	$A_{UVSTDM} = 2l(d + \Delta d) + \frac{\Delta d.k}{d}$	$ESD_{UVSTDM} = \frac{16}{A_{UVSTDM}{}^2}$
UV-DM $\|\omega\| > 4$	Select $\|\omega\|$ rays sourcing from neighbourhood cameras in uv and intersecting with expected p	2D interpolation over chosen rays in ω and estimate each ray from closest known pixel in st	$A_{UVDM(d,\Delta d,k,l,\|\omega\|)} = l(d + \Delta d) + \frac{\Delta d.k}{d}(\sqrt{\|\omega\|} - 1)^*$	$ESD_{UVDM(d,\Delta d,k,l,\|\omega\|)} = \frac{\|\omega\|}{A_{UVDM(d,\Delta d,k,l,\|\omega\|)}{}^2}$
UVST-DM $\|\omega\| > 16$	Select $\|\omega\|$ rays sourcing from neighbourhood cameras in uv, through known pixels in st and intersecting with expected p	4D interpolation over chosen rays in ω in both uv and st planes	$A_{UVSTDM(d,\Delta d,k,l,\|\omega\|)} = 2l(d + \Delta d) + \frac{\Delta d.k}{d}(\sqrt{\|\omega\|} - 1)^*$	$ESD_{UVSTDM(d,\Delta d,k,l,\|\omega\|)} = \frac{\|\omega\|}{A_{UVSTDM(d,\Delta d,k,l,\|\omega\|)}{}^2}$

**This is calculated by assuming that chosen rays are form a rectangular grid in uv plane for simplification*

Table IV: Comparison of the UVST, UV-D/UV-DM and UVST-D/UVST-DM methods [25]

Methods	*Sampling length comparison*	ESD *comparison*	*γ Ratio*	*γ Analysis*
UVST vs. UV-D	$A_{UVST} > \gamma . A_{UVD}$	$ESD_{UVST}\frac{\gamma^2}{4} < ESD_{UVD}$	$\gamma < \frac{(k+l)d^2 - kd}{ld^2 + l\Delta\mathrm{dd} + \mathrm{k}\Delta\mathrm{d}}$	$d \to \infty, k \cong l \text{ and } \Delta d \ll d \Rightarrow \gamma = 2$ $d \to \infty, k \gg l \text{ and } \Delta d \ll d \Rightarrow \gamma = \infty$
UVST vs. UVST-D	$A_{UVST} > \gamma . A_{UVSTD}$	$ESD_{UVST}\gamma^2 < ESD_{UVSTD}$	$\gamma < \frac{(k+l)d^2 - kd}{2ld^2 + 2l\Delta\mathrm{dd} + \mathrm{k}\Delta\mathrm{d}}$	$d \to \infty, k \cong l \text{ and } \Delta d \ll d \Rightarrow \gamma = 1$ $d \to \infty, k \gg l \text{ and } \Delta d \ll d \Rightarrow \gamma = \infty$
UV-D vs. UVST-D	$A_{UVD} > \gamma . A_{UVSTD}$	$ESD_{UVD}4\gamma^2 < ESD_{UVSTD}$	$\gamma < 1 - \frac{ld^2 + l\Delta\mathrm{dd}}{2ld^2 + 2l\Delta\mathrm{dd} + \mathrm{k}\Delta\mathrm{d}}$	$d \to \infty \Rightarrow \gamma = \frac{1}{2}$

V. Theoretical and Simulation Results

To verify the effectiveness of ESD as an indicator to estimate the distortion introduced by the acquisition and rendering components in a LF-based FVV system, a computer simulation system employing a 3D engine has been developed to generate the ground truth data [48]. The system takes a 3D model of a scene and simulates a multiple camera system to capture the scene. For any virtual views to be reconstructed, the system generates its ground truth image as a reference for comparison. Fig. 7 illustrates a simulated regular-camera grid for acquisition. Virtual views were randomly generated as the ground truth and used to evaluate the performance of ESD as a distortion indicator.

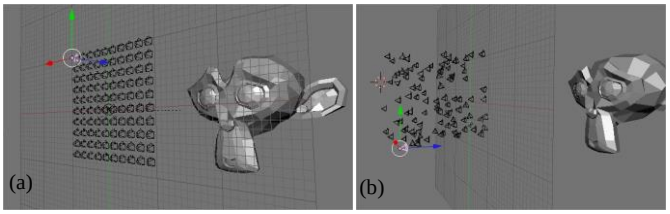

Fig. 7. a) A simulated regular camera grid; b) Random virtual viewpoints

In addition, since 3D models were used to represent the scene, a full precise depth map was available for rendering. Error is simulated and added to the depth map in order to evaluate ESD when inaccurate depth is employed in the rendering. In the following, details on the depth error model and experimental settings are presented.

A. Depth Error Model

There are two commonly used approaches to obtain depth information for FVV systems [49]: triangularization-based through either stereoscopic vision or structure light, and time-of-flight (ToF) based. When depth is estimated using the former approach, the error Δd is normally distributed whose standard deviation is proportional to the square of distance d^2, i.e. $\Delta d \approx \tau * d^2$, where τ depends on the system parameters [50]. For ToF, the error tends to be approximated coarsely as $\Delta d \approx \tau * d$ [51]. The linear model is adopted for the experimental validation in this paper. In the experiments, the ground truth depth map is known from the simulator. Based on the prescribed depth estimation error, for each pixel of the exact depth map, a random error with normal distribution and standard deviation of $\Delta d = \tau * d$ is introduced to create a noisy depth map with average of $\tau\%$ error.

B. ESD of Scenes

The ESD equations summarized in Table I and Table III are all for a small vicinity of scene around a given point p.

Clearly, ESD varies over the scene, depending on the depth. On the other hand, the overall distortion of output in addition to ESD is also scene dependent. Estimation of overall distortion for a given scene requires integration of ESD over the entire scene and at each point considering the scene texture complexity. In this paper, an approximation is adopted by using the average depth of the scene. This allows analysing acquisition configurations or rendering methods based on ESD independently of the scene complexity. To compare acquisition configurations and rendering methods an $\overline{\mathrm{ESD}}$ for each configuration/method is calculated for comparison using an average depth of the scene $\bar{d}$ with an average $\overline{\Delta d}$ of absolute depth error.

C. Simulation Settings

For the experiments reported in this paper, the LF engine is customized for the eight LF rendering methods: NN, UV, ST, UVST, UV-D, UVST-D, UV-DM and UVST-DM with $|\omega|$ = 1, 4, 4, 16, 4, 16, 4 and 16 respectively with default rectangular grid ray selection for M and bilinear and quadrilinear interpolations for F.

To assess the effect of scene complexity on output distortion, four 3D models, a "*room*", a "*chess board*", "*blender monkey*", and "*Stanford bunny*", as shown in Fig. 8, were selected, where the complexity decreases in this order. In the simulation, the centre of the 3D model was placed at $d = 10m$ by default, if depth is not given in the experiment. A $16x16$ regular camera grid were placed for acquisition and the image resolution was originally set to $1024x768$ pixels, i.e., $l = 0.05$. However, for experiments reported in Fig. 12 , to evaluate the effect of the 3D model depth in output PSNR, $\bar{d}$ is changed between $[10m, 50m]$, in Fig. 18 to evaluate the effect of the camera grid density in output PSNR, k is changed between $[0.1m, 0.9m]$, and in Fig. 19 to evaluate the effect of the reference cameras resolution on output PSNR, l is changed between [0.02cm,0.1cm], to analyse the effects of these factors on the output distortion. Please note that the term pixel size in the following experiments refers to l, the projected pixel size on image plane st at depth $d = 1$. Hence, $l = 0.02cm$ on st plane corresponds to a real pixel size equal to $4.8x10^{-4}cm$ for a typical $1/2^{"}$ camera sensor or capturing resolution of $2560x1920$. With the same assumptions, $l = 0.5cm$ corresponds to capturing resolution of $1024x768$ and $l = 0.1cm$ to resolution of $512x384$.

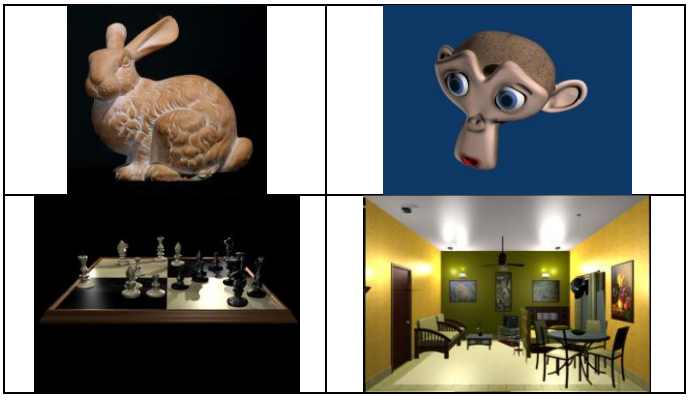

Fig. 8. Four 3D scenes chosen for experimental validation

For each 3D model, 1000 random virtual cameras at different distances from the scene were generated and average PSNR between the rendering images and the ground truth was calculated. In the following, the theoretical expectations in terms of calculated $\overline{\mathrm{ESD}}$ and the actual measurement of output video distortion in PSNR are reported and compared for different rendering methods and acquisition configurations.

D. Results on Rendering Methods

1) Theoretical expectation

Fig. 9 shows the ESD for the above-mentioned LF rendering methods in addition to the ideal rendering ($\Delta d = 0$) where $k = 0.4m$, $l = 0.05cm$, $d \in [10m, 50m]$, the object length is $5m$ and $\Delta d = 0.1d$ i.e., ten percent error in depth estimation. The ideal case is when there is no error in the depth map and refers to the maximum value for ESD at depth d. The vertical axis is logarithmic. For UV-D and UVST-D the actual error is $\frac{object\ length}{2} + \Delta\mathrm{d}$, which in this example is equal to $2.5m + 0.1d$.

It can be seen from Fig. 9 that, for all depths, the expected relative relationship of ESD among the eight LF rendering methods is maintained. A quadrilinear interpolation over UVST makes UVST-D and UVST-DM perform slightly better than their corresponding UV-D and UV-DM, especially for small d. For large depths, UV-D/UVST-D performance approaches that of UV-DM/UVST-DM, because the object length is small compared to depth error in this case.

Fig. 10 demonstrates a bar chart of theoretical ESD values for different rendering methods for $k = 0.4m$, $l = 0.05cm$, for a point p with $d = 10m$ and $\Delta d = 1m$.

Fig. 11 shows the effect of depth map error on ESD for UV-DM for $l = 0.01cm$, $|\omega| = 4$, $\bar{d} = 100$, $\frac{\Delta d}{d}$ between 0% to 20%, for $k = 5$, 10, 20 and 50. As it can be seen, higher errors in depth estimation result in less ESD when k is fixed. However, small k could increase the ESD.

2) Simulation results

Fig. 12 shows the simulated results, where the object depth d is changed from $10m$ to $50m$ with steps of $5m$ to analyze the effect of d on rendering output distortion in PSNR for different rendering methods. The acquisition parameters are: $k = 0.4m$ and $l = 0.05cm$ (i.e., camera resolution of $1024x768$). Notice that all the parameters for camera configuration and rendering algorithm were set the same as those used to obtain the theoretical results shown in Fig. 9. 10% depth error was added in the experiments. Fig. 12 shows the average results calculated from 288,000 experiments for 9 depths, 8 rendering methods, four 3D models and 1000 virtual viewpoints for each experiment. As it can be seen, rendering methods with full depth information UVST-DM and then UV-DM performed the best with the least distortion (in PSNR) followed by rendering methods with focusing depth information UVST-D and then UV-D. Not surprisingly, the blind rendering methods with no depth information had the highest distortion with UVST performing the best among blind methods followed by ST, UV and NN. The distance of the scene to the camera grid had a direct effect on output distortion, where further distance caused higher distortion for all methods, more significantly for methods with depth

information and less pronounced for blind methods. More importantly, the results show the same trends with the theoretical ESD values shown in Fig. 9.

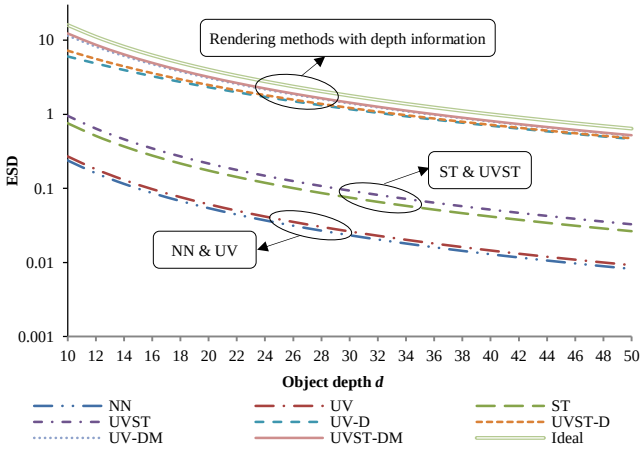

Fig. 9. Theoretical $\overline{ESD}$ for different LF rendering methods based on object depth $\bar{d}$ for $k = 0.4m$ and $l = 0.05cm$ (i.e., camera resolution of $1024x768$)

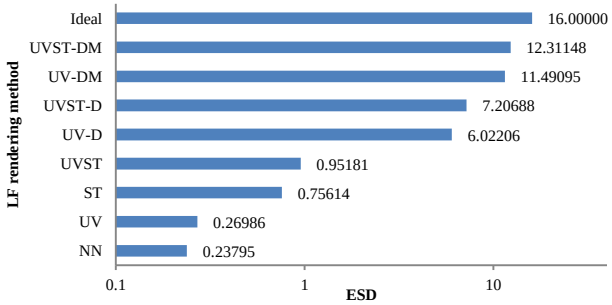

Fig. 10. Theoretical $\overline{ESD}$ for different rendering methods for $k = 0.4m$, $l = 0.05cm$, $\bar{d} = 10m$, and $\overline{\Delta d} = 1m$

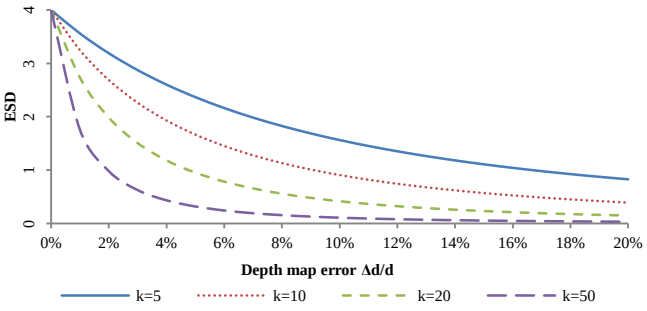

Fig. 11. Theoretical ESD for UV-DM for $\bar{\boldsymbol{d}} = \mathbf{100}$, $\overline{\boldsymbol{\Delta d}}$ in the range of $[\mathbf{0}\%, \mathbf{20}\%]$, $\boldsymbol{l} = \mathbf{0.01}$, $|\boldsymbol{\omega}| = \mathbf{4}$, for $\boldsymbol{k} = \mathbf{5}, \mathbf{10}, \mathbf{20}$ and $\mathbf{50}$

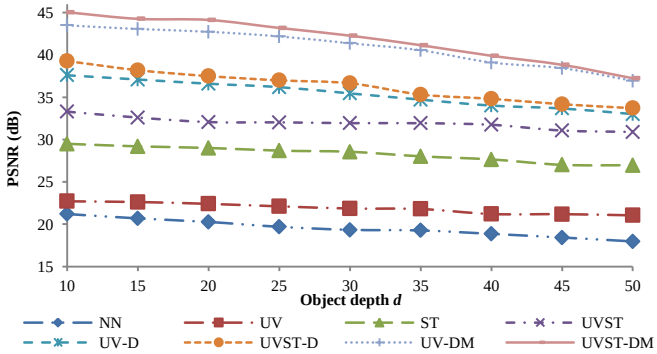

Fig. 12. Experimental rendering quality in PSNR for different LF rendering methods vs. object depth $\bar{d}$

Fig. 13 shows the average PSNR values over 32,000 simulations at $d = 10m$. NN interpolation performs the worst; UVST-DM is the best while UVST is the best blind rendering method. This order is consistent with the theoretically calculated ESD shown in Fig. 10.

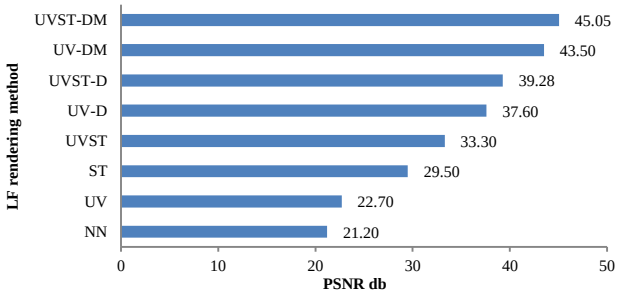

Fig. 13. Experimental rendering quality in PSNR for different LR methods

Fig. 14 shows the mean PSNR from 144,000 experiments for different rendering methods, categorized based on the complexity of the scene. As can be seen, more complex scenes result in reduced rendering quality. This can be explained due to fixed ESD for different scenes with different complexities in term of higher spatial frequency components. Nevertheless, ESD provides the right ranking on the performance amongst the various methods. Fig. 15 shows the rendering distortion from 144,000 experiments based on the distance of the virtual camera to the scene. As it is shown, far navigation results in higher rendering quality compared with closer observations. Again, this can be explained as a consequence of reduction in the required high frequency components to be sampled. Note that this experiment is different from experiments demonstrated in Fig. 12 and that is why the results are different. In this experiment, the light field system was fixed and the depth of virtual cameras was changed. In the previous experiment, the object depth is changed and the PSNR is calculated as the mean of 1000 random virtual cameras.

E. Results on Acquisition Configurations

By changing l and k respectively, various LF acquisition configurations were simulated.

1) Theoretical expectations

Fig. 16 demonstrates the theoretical relationship between k, the distance between the cameras in the camera grid, and ESD. As expected, for all methods, dense camera grid (small k) results in high ESD and therefore high rendering quality. In this Figure, $d = 50m$, $l = 0.05cm$ (camera resolution of $1024x768$), and $k \in [0.1m, 0.9m]$ with the same assumption for depth error as the case shown in Fig. 9.

As it can be seen, changing the value of k has limited effects on UV-D/UVST-D and UV-DM/UVST-DM, though at large k, UV-D and UV-DM performance gets worse compared to UVST-D and UVST-DM respectively. Also ESD of the ideal case (when there is no error in depth) is independent of k as demonstrated before. However, for blind methods, k has a significant effect on ESD values. NN, UV, ST and UVST all perform poorly especially for a large k. This confirms the view that by utilizing depth information, the cost of acquisition system can be significantly reduced.

Fig. 17 presents the theoretical relationship between l, the pixel size and ESD. It is clear that for all methods, high

resolution (small l) results in high ESD and therefore high rendering quality. In this Figure, $d = 50m$, $k = 0.4m$ and $l \in [0.02cm, 0.1cm]$, i.e., camera resolution of $2560x1920$ to $512x384$ respectively, with the same assumption for depth error as the case shown in Fig. 9.

As it can be seen, changing l has a direct effect on all methods. This effect is much more significant for UV-D, UVST-D, UV-DM, UVST-DM and the ideal case and less significant for blind methods. NN/UV and also ST/UVST performed similarly especially for a small l (high resolution).

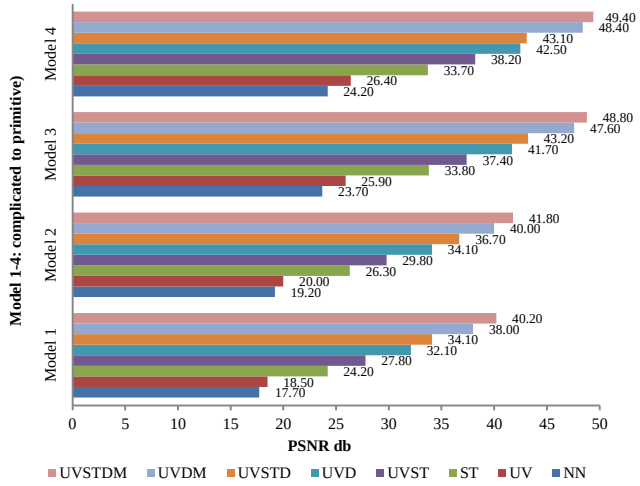

Fig. 14. Rendering quality and scene complexity

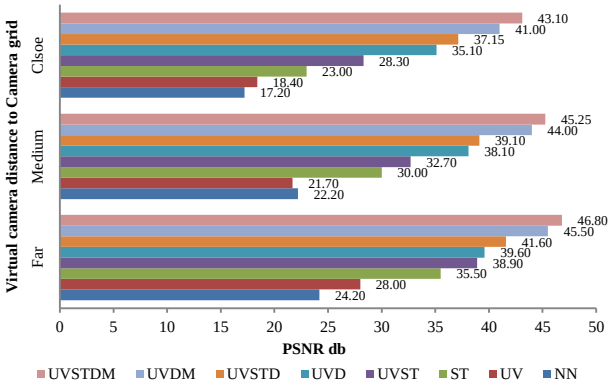

Fig. 15. Rendering quality and observation distance

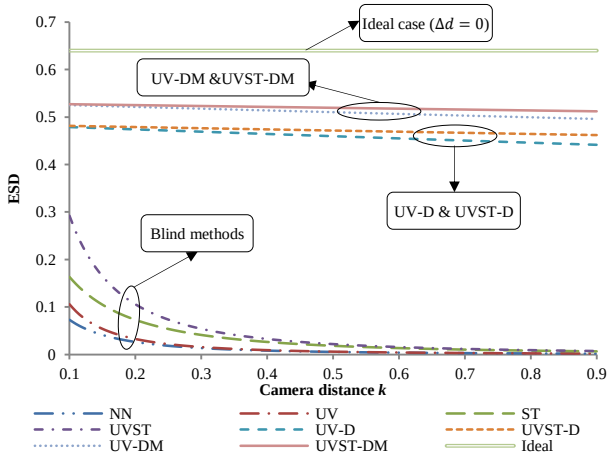

Fig. 16. Theoretical $\overline{ESD}$ for different LF rendering methods based on camera distance k between $0.1m$ to $0.9m$ for $l = 0.05cm$

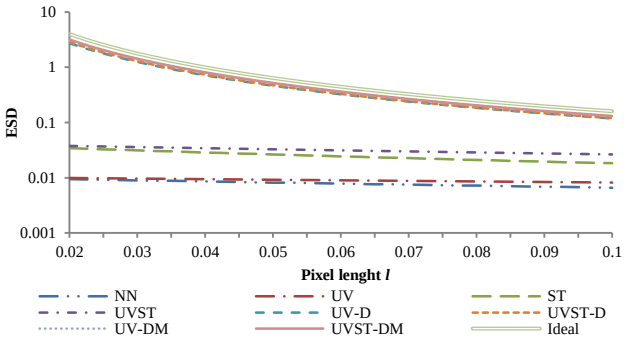

Fig. 17. Theoretical $\overline{ESD}$ for different LF rendering methods based on pixel length l between $0.02cm$ (camera resolution of $2560x1920$) to $0.1cm$ (camera resolution of $512x384$)

2) Simulation results

Experiments were carried out to see the effect of k in rendering distortion in term of PSNR so as to make a comparison to the theoretical ESD values. In first experiment, $d = 50m$, object length $= 5m$, $l = 0.05cm$ and $k \in [0.1m, 0.9m]$ and 10% depth error was added. Fig. 18 shows the results calculated from random 288,000 trials. As it can be seen, large separation between the cameras decreases the rendering PSNR as expected. However, the impact of increasing k is less significant for UV-D, UVST-D, UV-DM and UVST-DM compared to the blind methods.

The second experiment shows the relationship between the resolution of cameras (in term of pixel length l) and the rendering distortion in term of PSNR. In this experiment $d = 50m$, object length $= 5m$, $k = 0.4m$, $l \in [0.02cm, 0.1cm]$, i.e., resolution of $2560x1920$ to $512x384$ respectively, and 10% depth error. Fig. 19 illustrates the results calculated from 288,000 trials. As it can be seen, high resolution (smaller value of l) increases the rendering PSNR as expected. However, l has less impact on the blind rendering methods and more on UV-D, UVST-D, UV-DM and UVST-DM.

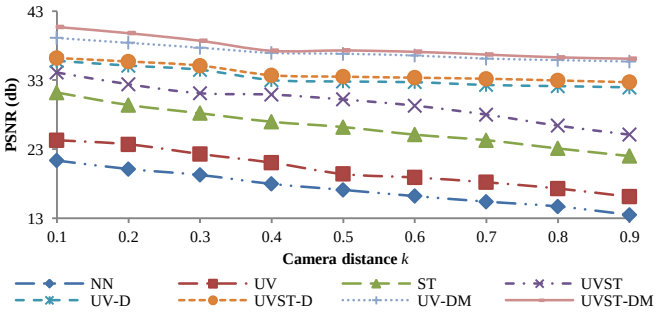

Fig. 18. Experimental rendering quality in PSNR for different LF rendering methods vs. camera distance k

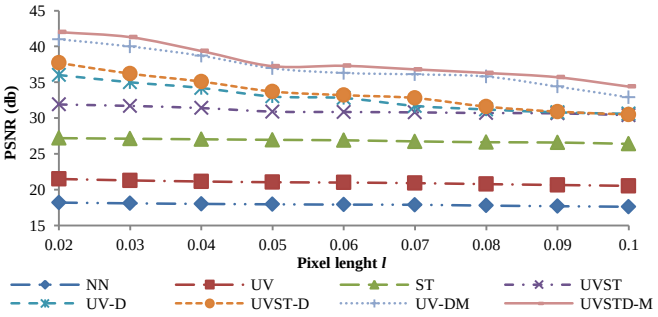

Fig. 19. Experimental rendering quality in PSNR for different LF rendering methods vs. pixel length l

Therefore, the theoretical expectations based on ESD analysis are confirmed by the empirical results. This can be seen clearly by comparing Fig. 16 with 18 and Fig. 17 with Fig. 19. Notice that the theoretical expectation is shown in ESD while the simulation results are shown in PSNR, and their relationship will be examined in the next section.

F. Discussions

Figures 9 to 19 present the theoretical expectations in term of ESD and experimental results in term of PSNR for different scenarios. To verify whether ESD is a good distortion indicator, an analysis was conducted of ESD vs. its counterpart PSNR, i.e., pairs of Figures (9, 12), (16, 18) and (17, 19). Fig. 20 shows the average experimental PSNR from Fig. 12 vs. theoretical ESD from Fig. 9, both obtained by changing the object depth $\bar{d}$. The trendline, covariance, and correlation of PSNR vs. ESD are also shown in Fig. 20.

Similarly, Fig. 21 demonstrates the observed PSNR from Fig. 18 vs. calculated ESD from Fig. 16, both obtained by changing the camera density. Again, the trendline, covariance, and correlation of PSNR vs. ESD are shown. Fig. 22 shows the observed PSNR from Fig. 19 vs. calculated ESD from Fig. 17, both obtained by changing the camera resolution.

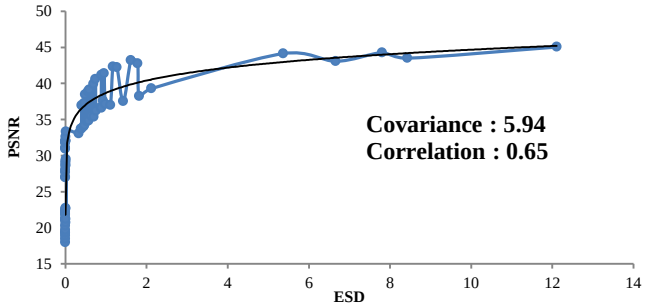

Fig. 20. Theoretical calculated ESD from Fig. 9 vs. experimental PSNR from Fig. 12, both obtained by changing the object depth ($\bar{d}$ from $10m$ to $50m$)

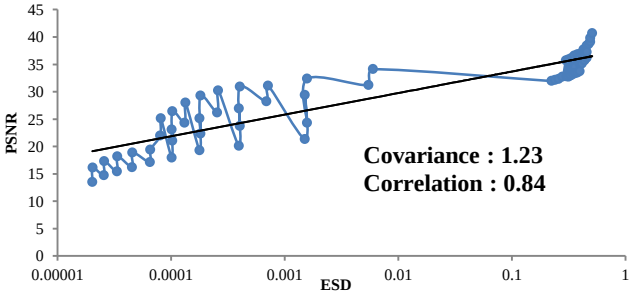

Fig. 21. Theoretical calculated ESD from Fig. 16 vs. experimental PSNR from Fig. 18, both obtained by changing the camera density (k from $1m$ to $9m$)

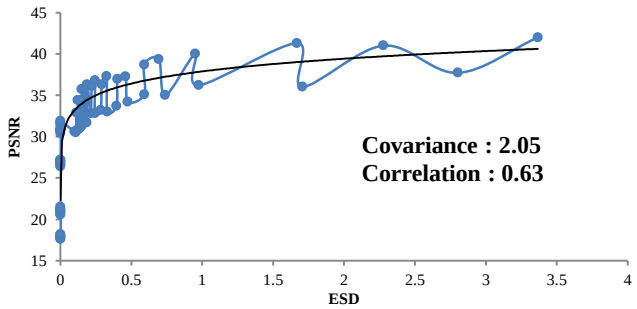

Fig. 22. Theoretical calculated ESD from Fig. 17 vs. experimental PSNR from Fig. 19, both obtained by changing the resolution (l from $0.02cm$ to $0.1cm$)

Fig. 20, Fig. 21, and Fig. 22 show a high correlation between theoretically calculated ESD and observed PSNR. In addition, as the trendlines demonstrate, there is an empirical relationship that can be explored to estimate output distortion in PSNR directly from calculated ESD without experiments. This will be explored in the next section.

VI. Empirical Relationship between ESD and PSNR

The experiments have shown that there is a relationship between ESD and PSNR. Since PSNR is a function of MSE (Mean Squared Error), it is expected that that MSE is a function of $\overline{\text{ESD}}$ for each given LF rendering method, denoted by ESD_{method}, and for a given fixed scene, i.e., $\text{MSE} = f(\text{ESD}_{method})$. In general, empirical f can be formulated as,

$$f(ESD_{method}) = Q * ESD_{method}{}^{P} \quad (12)$$

To find f, a subset of existing data is chosen as training set for curve fitting and the rest of the data as a validation set to test the accuracy of the empirical model f. To generate the curve fitting data, a map between observed PSNR and expected MSE is calculated as follows:

$$f(\text{ESD}_{method}) = \text{Expected MSE} = \frac{255^2}{10^{\left(\frac{\text{Observed PSNR}}{10}\right)}} \quad (13)$$

The data presented in Figures 9 and 12 (theoretical and experimental results based on changing the object depth) is used as the training set and data demonstrated in Figures (16, 18) and (17, 19) for validation. Fig. 23 demonstrates the overall curve fitting. This curve fitting is done on all the data and without clustering the data based on the rendering methods. Fig. 24 shows the curve fitting for each LF rendering method separately (method-dependent). The optimum value for $f(ESD_{method})$ for best estimation is when it is equal to expected MSE.

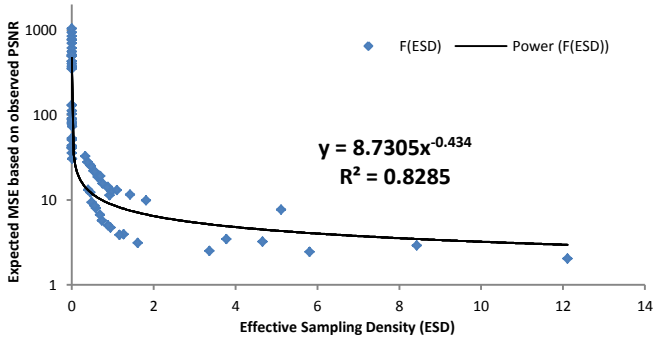

Fig. 23. A general curve fitting for $f(\text{ESD})$ estimation based on calculated $\overline{\text{ESD}}$ vs. expected MSE

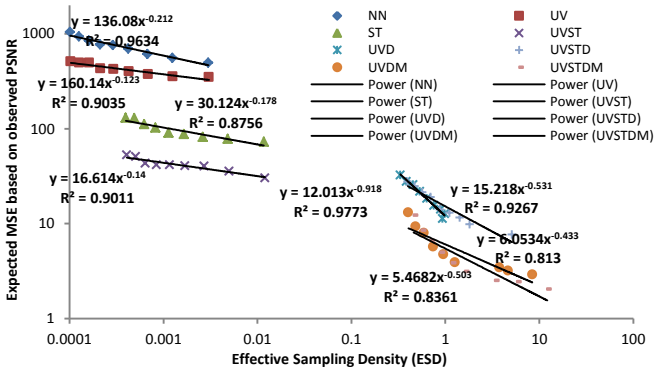

Fig. 24. Method-dependent curve fittings for $f(\mathrm{ESD}_{method})$

Fig. 25 shows a summary of curve fitting and validation errors of PSNR estimation for all LF rendering methods. As it can be seen from Fig. 25, the method-dependent estimation error for validation tests is less than 3%. If the method-dependent equations are not available, the estimation error for the overall equation is less than 12%. This shows that empirical equations for $f(\mathrm{ESD}_{method})$ are accurate to indicate the rendering distortion in term of PSNR. These equations offer a way to directly estimate the overall rendering distortion of a LF-based FVV system from the calculated ESD without implementation and experiments.

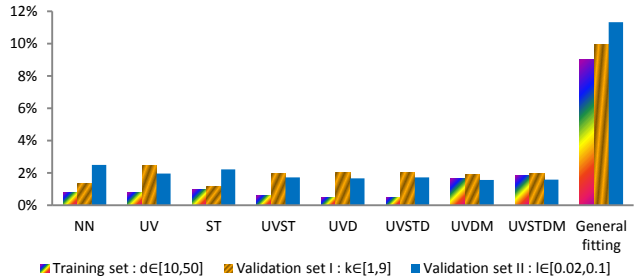

Fig. 25. Summary of curve fitting training and validation errors of PSNR estimation

By applying the analytical ESD equations to the proposed empirical equations, a direct model to estimate the rendering quality in PSNR from LF system parameters can be formulated. This helps the system designers to optimize the LF acquisition and LF rendering components without exhaustive experimental implementation of each configuration. For instance, for a general $\mathrm{UVDM}(d, \Delta d, k, l, |\omega|)$ method, by applying the ESD from (11), the rendering distortion can be directly calculated as:

$$\mathrm{PSNR}_{\mathrm{UVDM}(d,\Delta d,k,l,|\omega|)} \cong 20\, log_{10} \frac{255}{\sqrt{3.4545\left(\frac{|\omega|}{[l(d+\Delta d)+\frac{\Delta d.k}{d}(\sqrt{|\omega|}-1)]^2}\right)^{-0.256}}} \quad (14)$$

Table V summarizes the empirical boundaries of Q and P for different LF rendering methods, estimated for different scenes and acquisitions.

Table V: Empirical boundaries of P and Q

LF rendering method type	*LF rendering method*	*Q*	*P*
LF rendering methods with no depth information	NN	$50 < Q_{NN} < 300$	$-0.3 < P_{NN} < -0.2$
	ST	$20 < Q_{ST} < 200$	$-0.2 < P_{ST} < -0.1$
	UV	$20 < Q_{UV} < 250$	$-0.25 < P_{UV} < -0.1$
$10 < Q < 300$ $-0.3 < P < -0.1$	UVST	$10 < Q_{UVST} < 200$	$-0.2 < P_{UVST} < -0.1$
LF rendering methods with focusing depth information $10 < Q < 40$ $-1.0 < P < -0.15$	UVD	$10 < Q_{UVD} < 40$	$-1.0 < P_{UVD} < -0.15$
	UVSTD	$10 < Q_{UVSTD} < 40$	$-1.0 < P_{UVSTD} < -0.15$
LF rendering methods with full depth information $1 < Q < 15$ $-0.9 < P < -0.2$	UVDM	$1 < Q_{UVDM} < 15$	$-0.9 < P_{UVDM} < -0.2$
	UVSTDM	$1 < Q_{UVSTDM} < 15$	$-0.9 < P_{UVSTDM} < -0.2$
	General Method	$1 < Q < 10$	$-1.4 < P < -0.2$

The differences in $f(\mathrm{ESD}_{method})$ equations can be directly explained due to differences in the scene complexities and interpolation methods. Despite these differences, the general model offers a good indication on what the overall distortion in terms of PSNR should be expected by a given $\overline{\mathrm{ESD}}$.

VII. SUBJECTIVE ASSESSMENT

While previous section discussed the correlation between ESD and output video distortion in term of PSNR, this section demonstrates that ESD is also highly correlated with subjective assessment of the perceived video quality. A subjective quality assessment based on ITU-T standardization and guidelines on "subjective video quality assessment methods for multimedia applications" [24] and using degradation category rating (DCR) method was carried out. The test procedure is based on recommendations proposed in VQEG reports [52, 53]. Three rendering methods, UVST as a candidate of rendering methods with no depth information, UV-D with focusing depth and UV-DM with full depth information were selected for subjective test. The ground truth from the simulator and Stanford light field archive [54] was used as reference images. The original Stanford camera grid to capture real scenes is $17x17$, i.e., 289 reference images. To provide the ground truth for real scenes with real depth values, a subset of these reference images as a sparse $8x8$ camera grid was selected for acquisition component and a subset of other cameras were used as ground truth. 18 subjects participated in the test. For each of three candidate rendering methods, eight rendering outputs from different viewpoints for four different scenes, "*chess board*" and "*room*" from simulator and "*eucalyptus flowers*" and " *Lego knights*" from Stanford real data were generated. These 96 test sequences as a pair of reference and rendering output were presented to each subject with the recommended time pattern and experiment conditions as proposed in [24, 55]. The subjects were asked to rate the impairment of the second stimulus in relation to the reference into one of the five-level scales: 5-Imperceptible, 4-Perceptible but not annoying, 3-Slightly annoying, 2-Annoying, and 1-Very annoying.

The ESD is also calculated for each pair of scene and rendering method using the equations presented in Table I and III. There are totally 12 values for ESD (4 scenes and 3 rendering methods). Each value of ESD is corresponded to 8 different views.

Fig. 26 shows samples of the test sequences, presented to the subject panel. Note that Fig. 26 shows twelve different pairs out of 96 test sequences which were presented to each

subject. Fig. 27 illustrates the results of the subjective test for each rendering method. The average and variance of the impairment for each rendering method was calculated from 576 collected scores (32 test sequences among 18 subjects).

To validate the relationship between ESD and subjective DCR rating, the procedure for specifying accuracy and cross-calibration of video quality metrics proposed in VQEG reports [52, 53] were employed. Fig. 28 shows the scatter plot for the ESD-DCR couples for all 96 test sequences. Please note that for each 8 test sequences for different views, there is only one calculated ESD. To obtain the empirical relationship between DCR impairment rating and ESD, a polynomial curve fitting, as one of the candidates in VQEG reports, is applied over the data. The *Pearson correlation coefficient* is calculated as 0.91 which demonstrates a high relationship among ESD and DCR. The curve fitting has a *root mean square error* of 0.34 which shows around 10% error to predict DCR from ESD which is technically satisfactory. Fig. 29 shows an outdoor scene rendered with the proposed FVV system for subjective comparison of ground truth with the rendered output.

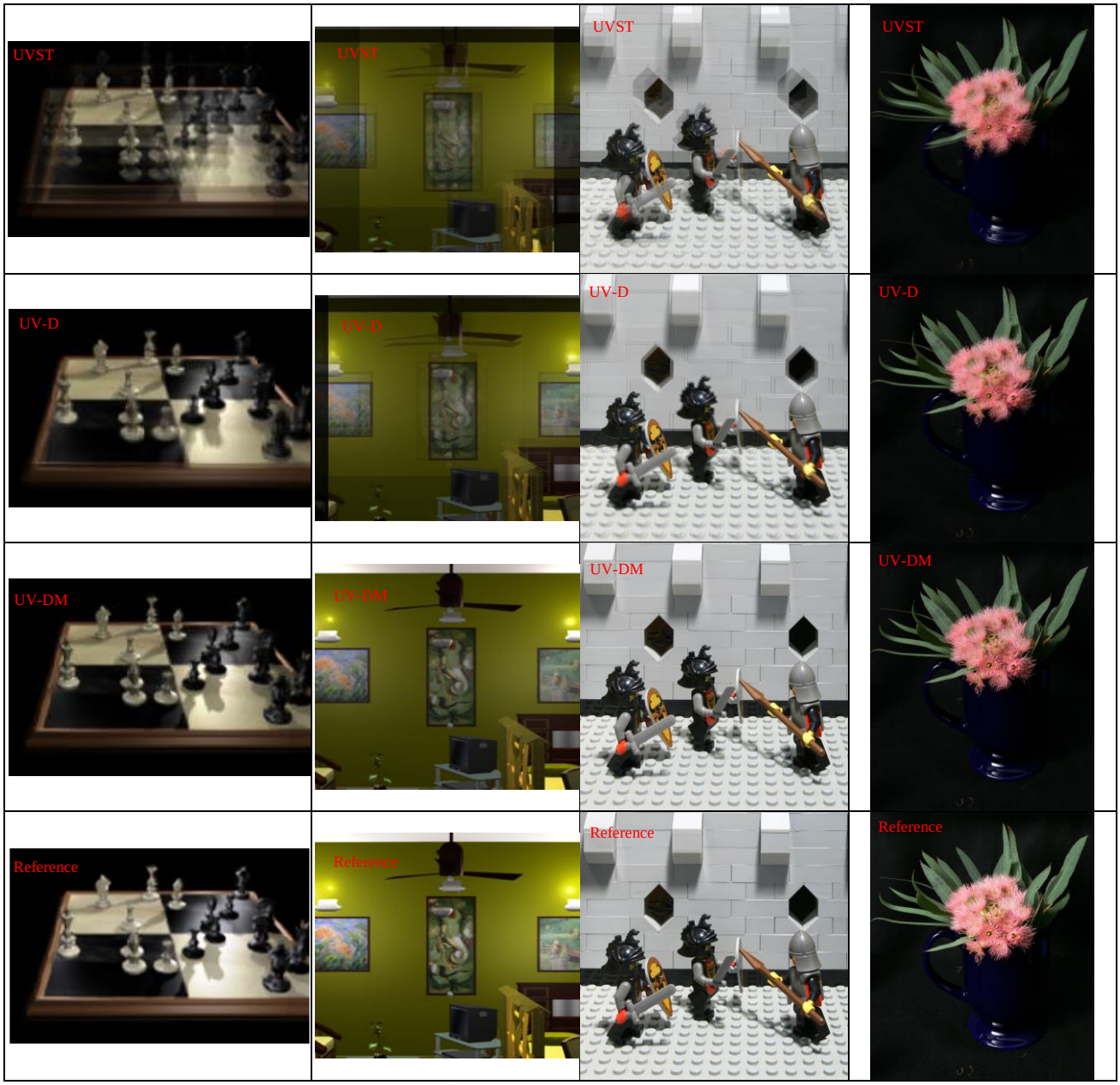

Fig. 26. Samples of test sequences used in the subjective assessment.

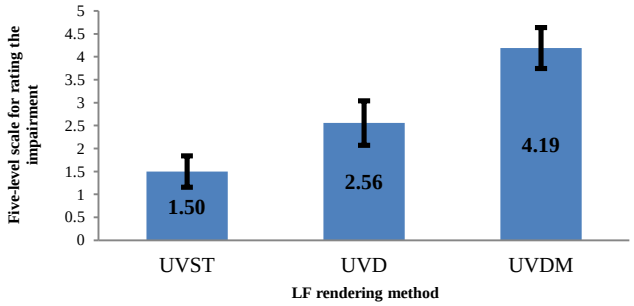

Fig. 27. Subjective assessment of three LF rendering methods by using degradation category rating (DCR), showing the Mean and Variance of rating from 576 collected scores for each method (32 test sequences among 18 subjects) with a five-level scale for rating the impairment

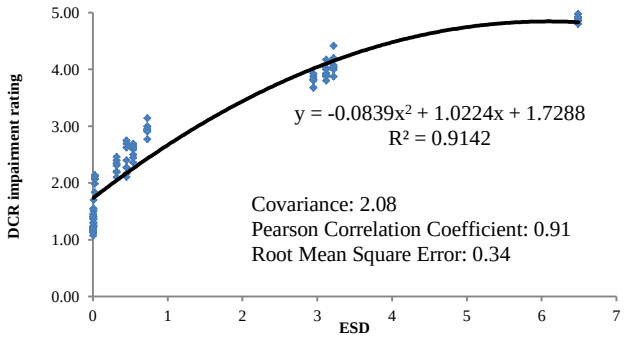

Fig. 28. DCR impairment rating for subjective assessment vs. theoretical ESD and the empirical relationship between these two parameters

Fig. 29. An outdoor scene, ground truth and the rendered output for subjective comparison

VIII. Conclusion

This paper has extended the concept of ESD. Using ESD different LF rendering methods and LF acquisition configurations can be theoretically evaluated and compared. Eight well-known rendering methods with different acquisition configurations have been analyzed through ESD and simulation. The results have shown that ESD is an effective indicator of distortion that can be obtained directly form system parameters and takes into consideration both acquisition and rendering. In addition, an empirical relationship between the theoretical ESD and achievable PSNR has been established. Furthermore, a subjective assessment has confirmed that ESD is highly correlated with the perceived output quality. Although this paper focuses on the overall distortion of a LF-based FVV system, the concept is readily extended to measure the rendering quality at a specific location or part of the scene. A further study on the impact of depth estimation errors on ESD and optimization of ESD with respect to the *camera density* and *ray selection complexity* for a given output quality will be our future work.

References

[1] M. Tanimoto, *et al.*, "Free-Viewpoint TV," *IEEE Signal Processing Magazine*, vol. 28, pp. 67-76, 2011.
[2] M. Tanimoto, "FTV: Free-viewpoint Television," *Signal Processing: Image Communication*, vol. 27, pp. 555-570, 2012.
[3] J.X. Chai, *et al.*, "Plenoptic sampling," *Proc. SIGGRAPH (ACM Trans. Graphics)*, pp. 307-318, Jul. 2000.
[4] C. Zhang and T. Chen, "Spectral analysis for sampling image-based rendering data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 1038-1050, 2003.
[5] C. Zhang and T. Chen, "Light field sampling," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, pp. 1-102, 2006.
[6] L. Zhouchen and S. Heung-Yeung, "A Geometric Analysis of Light Field Rendering," *Int. J. Comput. Vision*, vol. 58, pp. 121-138, 2004.
[7] N. King-To, *et al.*, "A Multi-Camera Approach to Image-Based Rendering and 3-D/Multiview Display of Ancient Chinese Artifacts," *Multimedia, IEEE Transactions on*, vol. 14, pp. 1631-1641, 2012.
[8] K. Takahashi and T. Naemura, "Layered light-field rendering with focus measurement," *Signal Processing: Image Communication*, vol. 21, pp. 519-530, 2006.
[9] N. W. Daniel, *et al.*, "Surface light fields for 3D photography," presented at the Proceedings of the 27th annual conference on Computer graphics and interactive techniques, 2000.
[10] Y. Jingyi, *et al.*, "Scam light field rendering," in *Computer Graphics and Applications, 2002. Proceedings. 10th Pacific Conference on*, 2002, pp. 137-144.
[11] H. Y. Shum, *et al.*, "Pop-up light field: An interactive image-based modeling and rendering system," *ACM Trans. Graphics*, vol. 23, pp. 143-162, Apr. 2004.
[12] W. Wen, *et al.*, "An efficient method for all-in-focused light field rendering," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, pp. 399-404.
[13] I. Aaron, *et al.*, "Dynamically reparameterized light fields," presented at the Proceedings of the 27th annual conference on Computer graphics and interactive techniques, 2000.

[14] K. Hansung, *et al.*, "Outdoor Dynamic 3-D Scene Reconstruction," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, pp. 1611-1622, 2012.
[15] S. X. Liu, *et al.*, "High quality virtual view synthesis based on corrected surface mapping and image fusion," *Electronics Letters*, vol. 45, pp. 30-32, 2009.
[16] E. Ekmekcioglu, *et al.*, "Content Adaptive Enhancement of Multi-View Depth Maps for Free Viewpoint Video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, pp. 352-361, 2011.
[17] T. Scandarolli, *et al.*, "Attention-Weighted Rate Allocation in Free-Viewpoint Television," *Signal Processing Letters, IEEE*, vol. 20, pp. 359-362, 2013.
[18] W. Qifei, *et al.*, "Free Viewpoint Video Coding With Rate-Distortion Analysis," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, pp. 875-889, 2012.
[19] H. Zhun and D. Qionghai, "A New Scalable Free Viewpoint Video Streaming System Over IP Network," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. II-773-II-776.
[20] E. H. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," *Computational Models of Visual Processing*, pp. 3-20, 1991.
[21] M. Levoy and P. Hanrahan, "Light field rendering," *Proc. SIGGRAPH (ACM Trans. Graphics)*, pp. 31-42, Aug. 1996.
[22] S. J. Gortler, *et al.*, "The lumigraph," *Proc. SIGGRAPH (ACM Trans. Graphics)*, pp. 43-54, Aug. 1996.
[23] M. N. Do, *et al.*, "On the bandwidth of the plenoptic function," *Image Processing, IEEE Transactions on*, vol. 21, pp. 708-717, 2012.
[24] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.
[25] H. Shidanshidi, *et al.*, "Objective evaluation of light field rendering methods using effective sampling density," in *MMSP*, 2011, pp. 1-6.
[26] H. Shidanshidi, *et al.*, "A Method for Calculating the Minimum Number of Cameras in a Light Field Based Free Viewpoint Video System," presented at the ICME, 2013.
[27] E. Camahort, *et al.*, "Uniformly sampled light fields," *Rendering Techniques*, vol. 98, pp. 117-130, 1998.
[28] T. Feng and H. Y. Shum, "An optical analysis of light field rendering," in *Proceedings of Fifth Asian Conference on Computer Vision*, 2000, pp. 394-399.
[29] A. Lumsdaine and T. Georgiev, "Full resolution lightfield rendering," *Indiana University and Adobe Systems, Tech. Rep*, 2008.
[30] J. Stewart, *et al.*, "A new reconstruction filter for undersampled light fields," presented at the Proceedings of the 14th Eurographics workshop on Rendering, Leuven, Belgium, 2003.
[31] L. Wenfeng, *et al.*, "Virtual View Specification and Synthesis for Free Viewpoint Television," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, pp. 533-546, 2009.
[32] C. L. Zitnick, *et al.*, "High-quality video view interpolation using a layered representation," *Proc. SIGGRAPH (ACM Trans. Graphics)*, pp. 600-609, Aug 2004.
[33] S. M. Seitz, *et al.*, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *CVPR*, 2006, pp. 519-528.
[34] J. Kilner, *et al.*, "Objective quality assessment in free-viewpoint video production," *Image Commun.*, vol. 24, pp. 3-16, 2009.
[35] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, pp. 430-444, 2006.
[36] A. Pons, *et al.*, "Image quality metric based on multidimensional contrast perception models," *Displays*, vol. 20, pp. 93-110, 1999.
[37] S. Winkler, "A perceptual distortion metric for digital color images," in *ICIP*, 1998, pp. 399-403 vol. 3.
[38] T. Brandão and P. Queluz, "Towards objective metrics for blind assessment of images quality," in *IEEE International Conference on Image Processing (ICIP)* 2006, pp. 2933-2936.
[39] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," 2007, pp. I-869-I-872.
[40] S. Winkler, "Video quality and beyond," in *Proc. European Signal Processing Conference*, 2007, pp. 3-7.
[41] Z. Wang, *et al.*, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600-612, 2004.
[42] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on Communications*, vol. 43, pp. 2959-2965, 1995.
[43] İ. Avcıbaş, *et al.*, "Statistical evaluation of image quality measures," *Journal of Electronic imaging*, vol. 11, p. 206, 2002.
[44] E. Bosc, *et al.*, "Towards a New Quality Metric for 3-D Synthesized View Assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, pp. 1332-1343, 2011.
[45] E. Bosc, *et al.*, "Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols?," in *18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2597-2600.
[46] R. Raskar and A. K. Agrawal, "4D light field cameras," ed: Google Patents, 2010.
[47] K. Takahashi, "Theoretical Analysis of View Interpolation With Inaccurate Depth Information," *Image Processing, IEEE Transactions on*, vol. 21, pp. 718-732, 2012.
[48] H. Shidanshidi, *et al.*, "A quantitative approach for comparison and evaluation of light field rendering techniques," in *ICME*, 2011, pp. 1-4.
[49] S. Schwarz, *et al.*, "Depth Sensing for 3DTV: A Survey," *MultiMedia, IEEE*, vol. 20, pp. 10-17, 2013.
[50] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, pp. 1437-1454, 2012.
[51] T. Pattinson, "Quantification and description of distance measurement errors of a time-of-flight camera," M. Sc. Thesis, University of Stuttgart, Stuttgart, Germany, 2010.
[52] P. b. T. A. W. G. o. M. C. C. a. Performance", "Methodological Framework for Specifying Accuracy and Cross-Calibration of Video Quality Metrics," 2001.
[53] M. H. Brill, *et al.*, "Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1A1," *Signal Processing: Image Communication*, vol. 19, pp. 101-107, 2004.
[54] S. U. Computer Graphics Laboratory. *The (New) Stanford Light Field Archive*. Available: http://lightfield.stanford.edu/lfs.html
[55] R. K. Mantiuk, *et al.*, "Comparison of four subjective methods for image quality assessment," in *Computer Graphics Forum*, 2012, pp. 2478-2491.

Hooman Shidanshidi graduated from the Bahá'í Institute for Higher Education (BIHE) University, Iran with the degree of Bachelor of Software Engineering and received his Master of Research and PhD in Computer Engineering from the University of Wollongong, Australia. He has been a Lecturer and Faculty Member at Bahá'í Institute for Higher Education (BIHE) University since 1998 and a Postdoctoral Research Fellow in ICT Research Institute at the University of Wollongong since 2013. Before joining the University of Wollongong, he was also the Senior Project Manager in several software development companies. His research areas include

computer vision, multimedia signal processing, free viewpoint video, computational intelligence, and simulation optimization.

Farzad Safaei graduated from the University of Western Australia with the degree of Bachelor of Engineering (Electronics) and obtained his PhD in Telecommunications Engineering from Monash University, Australia. Currently, he is the Professor of Telecommunications Engineering and Managing Director of ICT Research Institute at the University of Wollongong. Before joining the University of Wollongong, he was the Manager of Internetworking Architecture and Services Section in Telstra Research Laboratories. His research interests include immersive multimedia communications and free viewpoint TV.

Wanqing Li received his PhD in electronic engineering from The University of Western Australia. He joined Motorola Lab in Sydney (98-03) as a Senior Researcher and later a Principal Researcher and was a visiting researcher at Microsoft Research US in 2008, 2010 and 2013. He is currently an Associate Professor and Co-Director of Advanced Multimedia Research Lab (AMRL) of University of Wollongong, Australia. His research areas are 3D computer vision and 3D multimedia signal processing, including 3D reconstruction, human motion analysis, detection of objects and events, and free-viewpoint video. Dr. Li is currently a co-chair of the 3D Rendering, Processing and Communications interest group, Multimedia Technical Committee of IEEE Communication Society. He is the guest editor of the special issue on Human activity understanding from 2D and 3D data (2015), International Journal of Computer Vision, and the special issue on Visual Understanding and Applications with RGB-D Cameras (2013), Journal of Visual Communication and Image Representation. He served as a Co-organizer of many IEEE international conferences and workshops.