Chapter 2
Quality Assessment, Evaluation, and Optimization of Free Viewpoint Video Systems by Using Effective Sampling Density

Hooman Shidanshidi, Farzad Safaei, and Wanqing Li

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. The concept of ESD has been developed in last 7 years by the authors. It is shown 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 from system parameters and can be used to directly estimate output video quality 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. ESD also has been applied to several problems for evaluation and optimization of FVV acquisition and rendering subsystems. This chapter summarizes an overview of the ESD and its application in evaluation and optimization of FVV systems.

2.1 Introduction

Free viewpoint video (FVV) [1, 2] aims to provide users the ability to select arbitrary views of a dynamic scene in real time. An FVV system consists of three main components: *acquisition* [3–8] that captures the scene using a number of cameras, *rendering* [9–16] that reconstructs the desired view from the acquired information, and *compression/transmission* [1, 2, 17–20] of captured or processed information. The performance, in particular the quality of the output video of an FVV system,

H. Shidanshidi (✉) • F. Safaei • W. Li
ICT Research Institute, Faculty of Engineering and Information Sciences,
University of Wollongong, Wollongong, NSW, Australia
e-mail: hooman@uow.edu.au; farzad@uow.edu.au; wanqing@uow.edu.au

A. Kondoz, T. Dagiuklas (eds.), *Connected Media in the Future Internet Era*,
DOI 10.1007/978-1-4939-4026-4_2

depends on the efficacy of these components and their collaboration. While existing research studies individual components independently, this chapter 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 [21] 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) [22, 23]. 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, 24] on the minimum sampling density under the assumption that the signal of the scene is band limited and a perfect rendering procedure 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 [25] are usually used to assess the rendering component. These measurements require both the ground truth information and 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 and hence the quality of the output video. This is particularly true for an 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, before proposing ESD [26–28] there had not been any reported research on the joint impact of the two components on the output video quality. This chapter discusses this problem and reviews the theory of ESD and its application to estimate the signal distortion that accounts for both acquisition and rendering. Specifically, this chapter

- Covers the concept of effective sampling density (ESD) proposed by the authors in [26, 29] and employs it as an indicator of signal distortion for an 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 the 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 extensive 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.
- Demonstrates that how ESD can be employed for the evaluation and optimisation of FVV acquisition and rendering subsystems. Several research problems are discussed and it is shown that how ESD can be applied to these problems. The same framework can be used to similar evaluation and optimisation problems.

2.1.1 An Overview on the ESD Theory and Its Applications

The theory of ESD was first introduced by the authors in [26] followed by its application in evaluation and optimization of FVV acquisition and rendering subsystems in [29–32]. A comprehensive description of ESD and a framework for analytical derivation of ESD for different rendering methods can be found in [27, 28]. It is also shown that how theoretically calculated ESD can be used to empirically predict the output video quality in terms of objective signal distortion in PSNR as well as high correlation between ESD and perceived quality. Other applications of ESD include calculation of the minimum number of cameras for a regular camera grid [29], non-uniform light-field acquisition based on the scene complexity variations [30], and optimisation of acquisition and rendering subsystems [33].

One of the main problems in any FVV system analysis and design is acquisition and rendering evaluation and comparison. For any given acquisition configuration and rendering method, the ESD can be analytically calculated. To evaluate an acquisition component or a rendering method, it was shown in [27, 28] that the configuration or method with higher ESD has a better output video quality. Hence, ESD can be used as an unbiased tractable indicator to directly compare acquisition configurations and rendering methods.

Another important problem is acquisition and rendering optimization. To optimize the parameters of an acquisition system, e.g. camera density for a regular camera grid or the parameters of a rendering method, e.g. number of rays for interpolation, the optimization problem can be derived using the concept of ESD and solved numerically or analytically.

Another related problem is output video prediction and estimation from system parameters without the need for implementation and experiments. In [27, 28] it was shown that there is a high correlation between ESD and output video quality both in terms of objective signal distortion in PSNR and subjective quality perceived by users. In addition, an empirical method was proposed to map calculated ESD directly to rendering quality in PSNR. This allows predicting output video quality directly from FVV system parameters.

The mathematical framework to calculate ESD for a given FVV system and to solve problems of evaluation and optimization is fully addressed in [27, 28]. In this chapter a summary of some of these problems is given to show the applications of ESD.

The rest of the chapter is organized as follows. Section 2.2 reviews the related work. Section 2.3 analyses the acquisition and rendering components and describes in detail the concept of ESD. Section 2.4 presents the application of ESD to analyze LF systems with commonly used regular-grid cameras and rendering methods. Numerical simulation and validations are presented in Sect. 2.5. Section 2.6 presents the empirical relationship between the ESD and PSNR. Section 2.7 reports the subjective test and its correlation with ESD. In Sect. 2.8, several FVV research problems are discussed and it is shown that how ESD has been used or can be extended to address these problems. Section 2.9 concludes the chapter with remarks.

2.2 Related Work

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

2.2.1 Evaluation of the Acquisition Component

Light field can be expressed as a simplified four-dimensional plenoptic signal [21], first introduced by Levoy and Hanrahan [22] and Gortler et al. [23] (as Lumigraph) in mid-1990s. LF acquisition aims to sample the plenoptic signal by using limited number of cameras configured in 3D space. Several parameterisation schemes have been proposed to represent the camera configurations and the rays captured by the cameras. For instance, Levoy and Hanrahan [22] 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 parameterisation has been enhanced by more complicated parameterisation schemes such as two-sphere (2SP) and sphere-plane parameterisation (SPP) [34].

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, 24] 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 [3, 24] are 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 [35, 36]. The original light field [22] shows that the distance between two adjacent cameras can be considered as the aperture for ray filtering. This concept is generalised in [14] by introducing a "discrete synthetic aperture", encompassing of several cameras. It is also shown in [14] 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 [37].

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.

2.2.2 Evaluation of the Rendering Methods

Along with the acquisition configuration and parameterisation 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 [38]. These include layered light field [9], surface light field [10], scam light field [11], pop-up light field [12], all-in-focused light field [13], and dynamic reparameterised light field [14].

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 [39]. Ground-truth data is provided by employing a 3D scanner for a real scene or virtual environments such as [40]. In [41], 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 [41]. 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 [42], contrast perception modelling [43], and modelling the contrast gain control of the HVS [44]. However, HVS techniques and objective evaluation of a visual system are not able to fully model the human perception as discussed in [45–47]. Pixel-wise fidelity metrics such as MSE and PSNR are simple fidelity indicators but with a low correlation with visual quality [48]. In [49] a full review of pixel-wise fidelity metrics is discussed. Also [50] 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 [51, 52], 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.

2.3 Effective Sampling Density

Figure 2.1 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

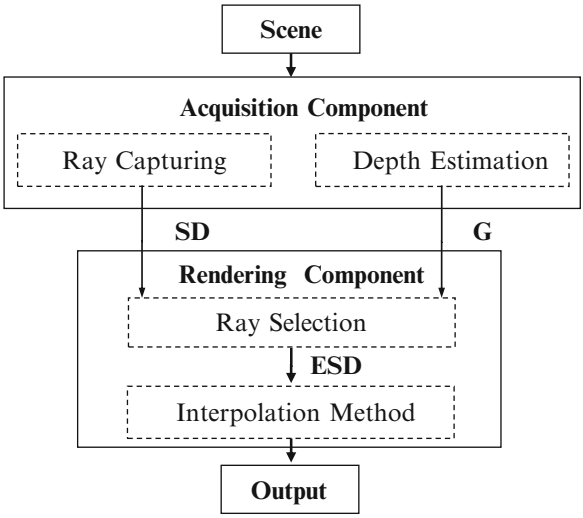

Fig. 2.1 The schematic diagram of a typical LF-based FVV system that utilises scene geometric information *G*

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. In addition, the *depth estimation* process provides an estimation of depth (e.g. depth map) to improve rendering. This could be obtained by specialised hardware, such as depth cameras, or computed from the images obtained by multiple 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: (1) the *ray selection* that chooses a subset of acquired rays, purported to be in the vicinity of r, for the purpose of interpolation, and (2) the *interpolation* that provides an estimate of r from these selected 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 neighbouring rays and choose others that are indeed sub-optimal (with respect to proximity to r) for interpolation. Consider the case shown in Fig. 2.2, 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

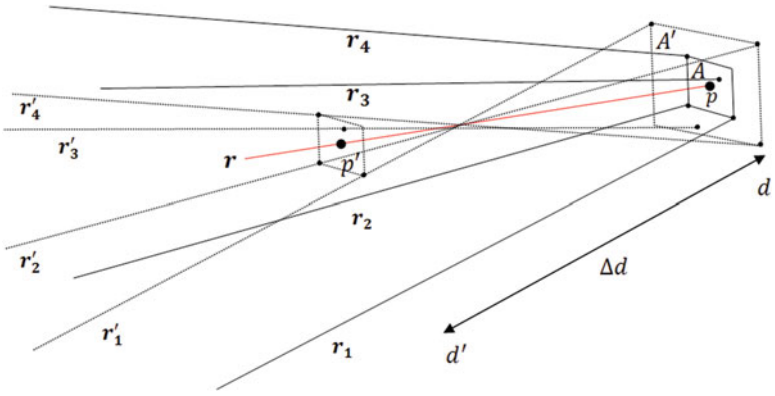

Fig. 2.2 Selection of rays in an LF rendering and the concept of ESD

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.

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. Note that 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 chapter.

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:

$$\omega \subseteq \Omega \subseteq \theta \tag{2.1}$$

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

$$\omega = M(\theta, G) \tag{2.2}$$

$$r = F(\omega, G) \tag{2.3}$$

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

$$\mathrm{SD} = \frac{|\Omega|}{A} \tag{2.4}$$

$$\mathrm{ESD} = \frac{|\omega|}{A} = \frac{|M(\theta, G)|}{A} \tag{2.5}$$

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. Figure 2.3 shows this process for a simple 2D LF acquisition, generated by applying

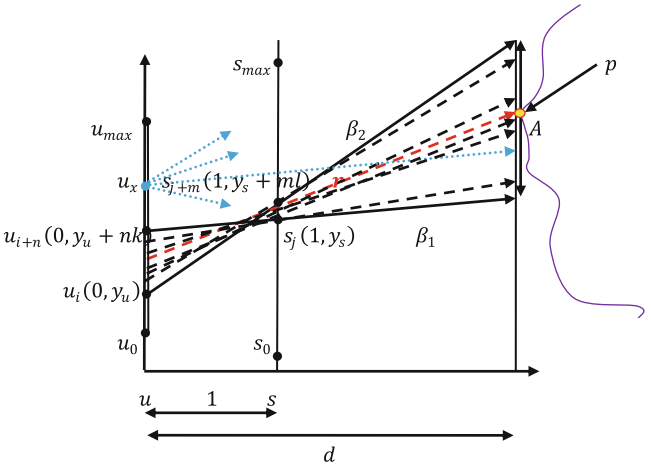

Fig. 2.3 ESD calculation for a simplified 2D light-field system

a 2D projection to a 3D light field with two-plane parameterisation, 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 that 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. 2.3, there are more than $n + 1$ cameras in the grid (in addition to cameras bounded between u_i and 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.

SD defined in Eq. (2.4) 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. Figure 2.4a shows the SD contour maps at different depths, $d = 30$, 60, and 90 m, for a regular camera grid of 30×30 with $k = 2$ m, camera field of view of 30°, image resolution of 100×100 pixels, i.e. $l = 0.53$ cm in image plane st, and ideal area $A = (ld)^2$, i.e. LF system resolution. Figure 2.4b shows a 2D slice where d ranges in [2 m, 100 m].

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.

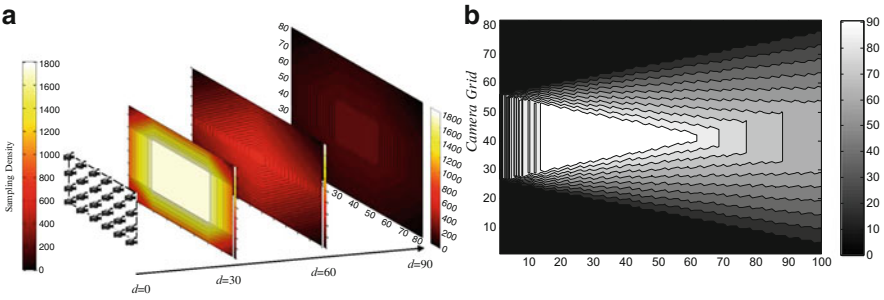

Fig. 2.4 (**a**) SD contour maps at different depths in 3D; (**b**) SD contour map in 2D

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.

2.4 ESD Analysis of LF Rendering Methods

Without loss of generality, a simple regular-grid camera system, as shown in Fig. 2.3, 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 [34]. For a regular-grid camera system, analytical form of ESD can be obtained for a rendering algorithm with and without using depth information.

2.4.1 Rendering Methods Without the Depth Information

The LF rendering methods without using depth information, hereafter referred to as *blind* methods, can be categorised 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) [22, 53]. 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. This will be discussed later in subsection 2.4.3.

Considering the regular geometry of the cameras shown in Fig. 2.3, analytical form of ESD for these rendering algorithms can be derived. Table 2.1 summarises the ESD derivation for the NN, ST, UV, and UVST methods where $|\omega| = 4$ for UV

Table 2.1 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	$A_{\mathrm{NN}} = \left(\frac{l+k}{2}\right) d - \frac{k}{2}$	$\mathrm{ESD}_{\mathrm{NN}} = \frac{1}{\mathrm{A_{NN}}^2}$
ST	Select four 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	$A_{\mathrm{ST}} = \left(l + \frac{k}{2}\right) d - \frac{k}{2}$	$\mathrm{ESD}_{\mathrm{ST}} = \frac{4}{\mathrm{A_{ST}}^2}$
UV	Select four 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	$A_{\mathrm{UV}} = \left(k + \frac{l}{2}\right) d - k$	$\mathrm{ESD}_{\mathrm{UVST}} = \frac{4}{\mathrm{A_{UVST}}^2}$
UVST	Select 16 or more rays from four neighbourhood cameras in *uv* to 4 neighbourhood pixels in *st*, $\|\omega\| \geq 16$	Any type of 4D interpolation, e.g. quadrilinear interpolation for grid selection of rays	$A_{\mathrm{UVST}} = (l + k) d - k$	$\mathrm{ESD}_{\mathrm{UVST}} = \frac{16}{\mathrm{A_{UVST}}^2}$

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 summarises 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 2.1, 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 2.2 summarises 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, and 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 2.2.

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 $4\text{ESD}_{\text{NN}} < 4\text{ESD}_{\text{UV}} < \text{ESD}_{\text{ST}} < \text{ESD}_{\text{UVST}}$. 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 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 $1.7\text{ESD}_{\text{NN}} < \text{ESD}_{\text{UV}} < \text{ESD}_{\text{ST}}$, $4\text{ESD}_{\text{NN}} < \text{ESD}_{\text{UVST}}$, and $2.2\text{ESD}_{\text{UV}} < 2.2\text{ESD}_{\text{ST}} < \text{ESD}_{\text{UVST}}$. 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 $4\text{ESD}_{\text{NN}} < 4\text{ESD}_{\text{ST}} < \text{ESD}_{\text{UV}} < \text{ESD}_{\text{UVST}}$. 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.

2.4.2 Rendering Methods with the Depth Information

Utilisation of depth information G in rendering can compensate to some extent for insufficient number of samples acquired in an *under-sampling* situation [54]. 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 utilises only the focusing depth

Table 2.2 Comparison of ESD of the LF rendering methods without using depth information [25]

Methods	Sampling length comparison	ESD comparison	γ (The ratio of ESDs)	γ Analysis
NN vs. ST	$A_{\mathrm{NN}}\gamma > A_{\mathrm{ST}}$	$\mathrm{ESD}_{\mathrm{NN}}\frac{4}{\gamma^2} < \mathrm{ESD}_{\mathrm{ST}}$	$\gamma > 1 + \frac{ld}{(l+k)d-k}$	$d \to \infty$ 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	$A_{\mathrm{NN}}\gamma > A_{\mathrm{UV}}$	$\mathrm{ESD}_{\mathrm{NN}}\frac{4}{\gamma^2} < \mathrm{ESD}_{\mathrm{UV}}$	$\gamma > 1 + \frac{kd-k}{(l+k)d-k}$	$d \to \infty$ 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	$A_{\mathrm{NN}}\gamma > A_{\mathrm{UVST}}$	$\mathrm{ESD}_{\mathrm{NN}}\frac{16}{\gamma^2} < \mathrm{ESD}_{\mathrm{UVST}}$	$\gamma > 2$	$\gamma > 2$
ST vs. UVST	$A_{\mathrm{ST}}\gamma > A_{\mathrm{UVST}}$	$\mathrm{ESD}_{\mathrm{ST}}\frac{4}{\gamma^2} < \mathrm{ESD}_{\mathrm{UVST}}$	$\gamma > 1 + \frac{d-1}{(\frac{2l}{k}+1)d-1}$	$d \to \infty$ and $k \gg l \Rightarrow \gamma = 2$ $d \to \infty$ and $k \cong l \Rightarrow \gamma = 1.33$ $d \cong 1 \Rightarrow \gamma = 1$
UV vs. UVST	$A_{\mathrm{UV}}\gamma > A_{\mathrm{UVST}}$	$\mathrm{ESD}_{\mathrm{UV}}\frac{4}{\gamma^2} < \mathrm{ESD}_{\mathrm{UVST}}$	$\gamma > 1 + \frac{ld}{(l+2k)d-2k}$	$d \to \infty$ and $k \gg l \Rightarrow \gamma = 1$ $d \to \infty$ and $k \cong l \Rightarrow \gamma = 1.33$ $d \cong 1 \Rightarrow \gamma = 2$
ST vs. UV	$A_{\mathrm{UV}} > \gamma A_{\mathrm{ST}}$	$\mathrm{ESD}_{\mathrm{UV}}\gamma^2 < \mathrm{ESD}_{\mathrm{ST}}$	$\gamma < 1 + \frac{(k-l)d-k}{(2l+k)d-k}$	$d \to \infty$ and $k \gg l \Rightarrow \gamma = 2$ $d \to \infty$ and $k \cong l \Rightarrow \gamma = 1$ $d \cong 1 \Rightarrow \gamma = 0.5$

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 four groups. These include layered light field [9], surface light field [10], scam light field [11], pop-up light field [12], all-in-focused light field [13], and dynamic reparameterised light field [14].

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.

Figure 2.5 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 chapter the worst-

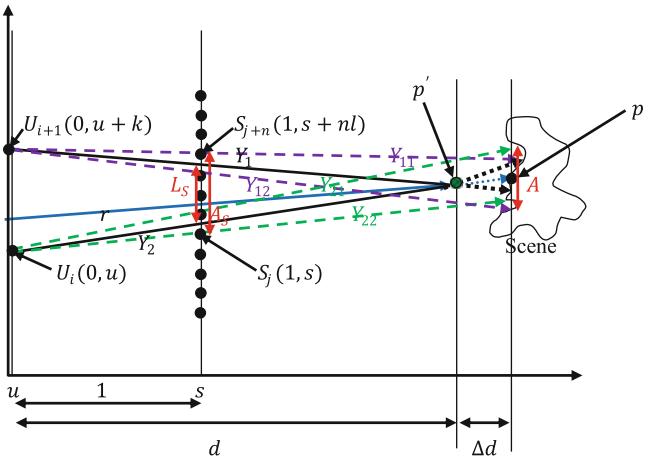

Fig. 2.5 Light-field rendering methods using depth information (UV-D, UVST-D, UV-DM/ UVST-DM) with Δd error in depth estimation

case scenario, i.e. largest A, is computed in the LF analysis which corresponds to the lower bound of ESD.

Considering scenario in Fig. 2.5, 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 $A_S = L_S + \frac{l}{2} + \frac{l}{2} = \frac{k(d-1)+ld}{d}$, $A_{\text{UVD/UVDM}} = ld$, and $A_{\text{UVSTD/UVSTDM}} = 2ld$. 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

$$\begin{aligned} A &= \max\left[\left|Y_{11}\left(d+\Delta d\right) - Y_{22}\left(d+\Delta d\right)\right|, \left|Y_{12}\left(d+\Delta d\right) - Y_{21}\left(d+\Delta d\right)\right|\right] \\ &= l\left(d+\Delta d\right) + \frac{\Delta d \times k}{d} \end{aligned} \tag{2.6}$$

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 $\Delta d = \frac{\textit{object length}}{2} + \textit{focusing depth estimation error}$. 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 usually much less than the focusing depth error, which makes the UV-DM/UVST-DM rendering less distorted than UV-D/UVST-D.

2.4.3 General Case of Rendering Methods with Depth Maps

Figure 2.6 demonstrates an LF rendering method with two-plane parameterisation 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. 2.6, seven rays from all rays intersecting imaginary p are selected by M, i.e. $|\omega| = 7$, assuming that 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 generalisation of UV-DM described in subsection 2.4.2 but with

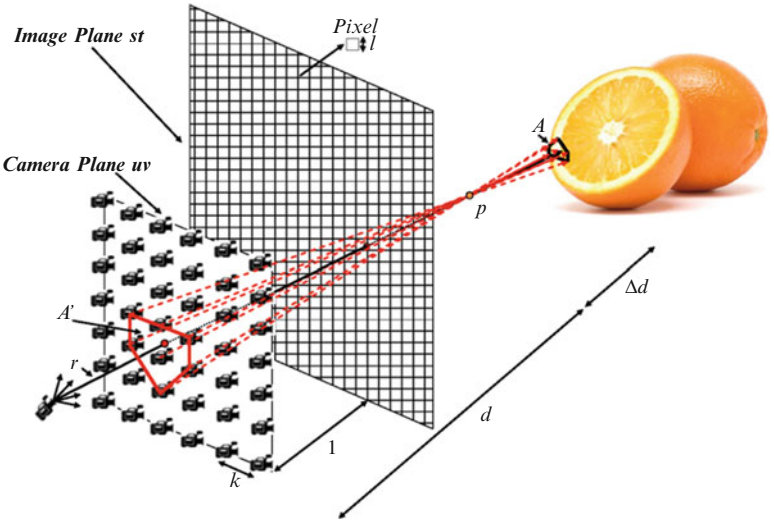

Fig. 2.6 General light-field rendering method using depth information (UV-DM/UVST-DM) with Δd error in depth estimation

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 generalisation 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. 2.6 can be derived as

$$\mathrm{ESD}_{\mathrm{UVDM}} = \frac{|\omega|}{A} = \frac{|\omega|}{\frac{\Delta d}{d}A' + \mu\left(l\left(d + \Delta d\right), A'\right)} \tag{2.7}$$

$$\mathrm{ESD}_{\mathrm{UVSTDM}} = \frac{|\omega|}{A} = \frac{|\omega|}{\frac{\Delta d}{d}A' + \mu\left(2l\left(d + \Delta d\right), A'\right)} \tag{2.8}$$

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 Eqs. (2.7) and (2.8) also has a small effect on A. The pixel interpolation over *st* even when $\Delta d = 0$ makes $A = (ld)^2$.

Simple forms of UV-DM and UVST-DM described in subsection 2.4.2 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; subsequently Eqs. (2.7) and (2.8) become

$$\mathrm{ESD}_{\mathrm{UVDM}} = \frac{4}{\left(\frac{\Delta d.k}{d} + l\,(d + \Delta d)\right)^2} \tag{2.9}$$

$$\mathrm{ESD}_{\mathrm{UVSTDM}} = \frac{16}{\left(\frac{\Delta d \times k}{d} + 2l\,(d + \Delta d)\right)^2} \tag{2.10}$$

where k is the distance between the two neighbouring cameras in the camera grid and l is the length of the pixel in the image plane as illustrated in Fig. 2.6. Note that the edge of rectangular A' is equal to k and that is how Eqs. (2.9) and (2.10) are derived from Eqs. (2.7) and (2.8).

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 Eq. (2.9) and considering the edge of rectangular A' to be equal to $\left(\sqrt{|\omega|} - 1\right)k$, the ESD could be calculated for UVDM$(d, \Delta d, k, l, |\omega|)$ as follows:

$$\mathrm{ESD}_{\mathrm{UVDM}(d,\Delta d,k,l,|\omega|)} = \frac{|\omega|}{\left(l\,(d + \Delta d) + \frac{\Delta d \times k}{d}\left(\sqrt{|\omega|} - 1\right)\right)^2} \tag{2.11}$$

Equation (2.11) assumes that the rays are chosen for interpolation symmetrically around the vertical and horizontal axes, such as 4×4 samples. 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 Figs. 2.5 and 2.6, Eqs. (2.7), (2.8), (2.9), (2.10), and (2.11). Table 2.3 summarises 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 columns give the sampling/interpolation length A and ESD, respectively.

Table 2.4 summarises comparison of the ESD among UVST, UV-D, and UVST-D. It is clear from Table 2.3 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 2.4. 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 camera 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. Tables 2.2 and 2.4 have shown that $4\mathrm{ESD}_{\mathrm{NN}} < 4\mathrm{ESD}_{\mathrm{UV}} <$

Table 2.3 ESD for the LF rendering methods with depth information

Rendering method category	Selection mechanism M	Interpolation function F	Sampling/interpolation length A in 2D LF	ESD for symmetric 3D light field
UV-D $\lvert\omega\rvert = 4$	Select four rays sourcing from neighbourhood cameras in uv and intersecting with expected p	Neighbourhood estimation in st and 2D interpolation over uv	$A_{\mathrm{UVD}} = l\,(d + \Delta d) + \frac{\Delta dk}{d}$	$\mathrm{ESD}_{\mathrm{UVD}} = \frac{4}{{A_{\mathrm{UVD}}}^2}$
UVST-D $\lvert\omega\rvert = 16$	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. quadrilinear interpolation	$A_{\mathrm{UVSTD}} = 2l\,(d + \Delta d) + \frac{\Delta dk}{d}$	$\mathrm{ESD}_{\mathrm{UVSTD}} = \frac{4}{{A_{\mathrm{UVSTD}}}^2}$
UV-DM $\lvert\omega\rvert = 4$	The same as UV-D but with more accurate depth estimation of p employing depth maps.	The same as UV-D	$A_{\mathrm{UVDM}} = l\,(d + \Delta d) + \frac{\Delta dk}{d}$	$\mathrm{ESD}_{\mathrm{UVDM}} = \frac{4}{{A_{\mathrm{UVDM}}}^2}$
UVST-DM $\lvert\omega\rvert = 16$	The same as UVST-D but with more accurate depth estimation of p employing depth maps	The same as UVST-D	$A_{\mathrm{UVSTDM}} = 2l\,(d + \Delta d) + \frac{\Delta dk}{d}$	$\mathrm{ESD}_{\mathrm{UVSTDM}} = \frac{16}{{A_{\mathrm{UVSTDM}}}^2}$
UV-DM $\lvert\omega\rvert > 4$	Select $\lvert\omega\rvert$ 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	$A_{\mathrm{UVDM}(d,\Delta d,k,l,\lvert\omega\rvert)} = l\,(d + \Delta d) + \frac{\Delta dk}{d}\left(\sqrt{\lvert\omega\rvert} - 1\right)$[a]	$\mathrm{ESD}_{\mathrm{UVDM}(d,\Delta d,k,l,\lvert\omega\rvert)} = \frac{\lvert\omega\rvert}{{A_{\mathrm{UVDM}(d,\Delta d,k,l,\lvert\omega\rvert)}}^2}$
UVST-DM $\lvert\omega\rvert > 16$	Select $\lvert\omega\rvert$ 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	$A_{\mathrm{UVSTDM}(d,\Delta d,k,l,\lvert\omega\rvert)} = 2l\,(d + \Delta d) + \frac{\Delta dk}{d}\left(\sqrt{\lvert\omega\rvert} - 1\right)$[a]	$\mathrm{ESD}_{\mathrm{UVSTDM}(d,\Delta d,k,l,\lvert\omega\rvert)} = \frac{\lvert\omega\rvert}{{A_{\mathrm{UVSTDM}(d,\Delta d,k,l,\lvert\omega\rvert)}}^2}$

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

Table 2.4 Comparison of the UVST, UV-D/UV-DM, and UVST-D/UVST-DM methods

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

$ESD_{ST} < ESD_{UVST} \ll ESD_{UVD/UVDM} < ESD_{UVSTD/UVSTDM}$, i.e. for a given acquisition, the NN rendering method has the lowest ESD and hence results in the highest video distortion followed 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 Tables 2.3 and 2.4 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, camera 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 [29], 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 optimise the acquisition and rendering parameters of an LF system individually and jointly for a target output video quality.

2.5 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 an LF-based FVV system, a computer simulation system employing a 3D engine has been developed to generate the ground-truth data [55]. 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. Figure 2.7 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.

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.

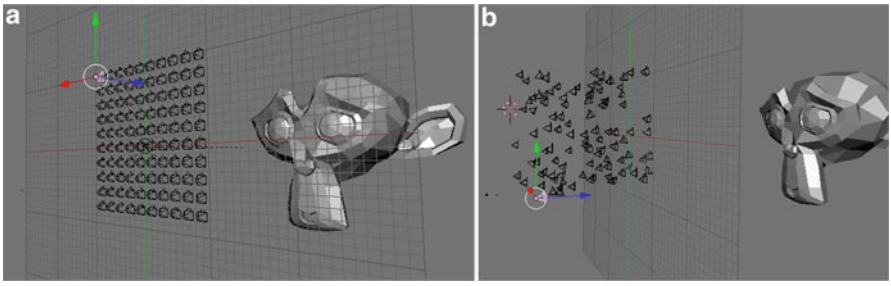

Fig. 2.7 (**a**) A simulated regular camera grid; (**b**) random virtual viewpoints

2.5.1 Depth Error Model

There are two commonly used approaches to obtain depth information for FVV systems [56]: triangularisation 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. $\Delta d \approx \tau \times d^2$, where τ depends on the system parameters [57]. For ToF, the error tends to be approximated coarsely as $\Delta d \approx \tau \times d$ [58]. The linear model is adopted for the experimental validation in this chapter. 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 $\Delta d = \tau \times d$ is introduced to create a noisy depth map with average of τ % error.

2.5.2 ESD of Scenes

The ESD equations summarised in Tables 2.1 and 2.3 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 chapter, 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 $\overline{\text{ESD}}$ for each configuration/method is calculated for comparison using an average depth of the scene $\overline{d}$ with an average $\overline{\Delta d}$ of absolute depth error.

2.5.3 Simulation Settings

For the experiments reported in this chapter, the LF engine is customised for the eight LF rendering methods: NN, UV, ST, UVST, UV-D, UVST-D, UV-DM, and UVST-DM with $|\omega| = 1,\ 4,\ 4, 16, 4, 16, 4, 16$, 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. 2.8, were selected, where the complexity decreases in this order. In the simulation, the centre of the 3D model was placed at $d = 10$ m by default, if depth is not given in the experiment. A 16×16 regular camera grid was placed for

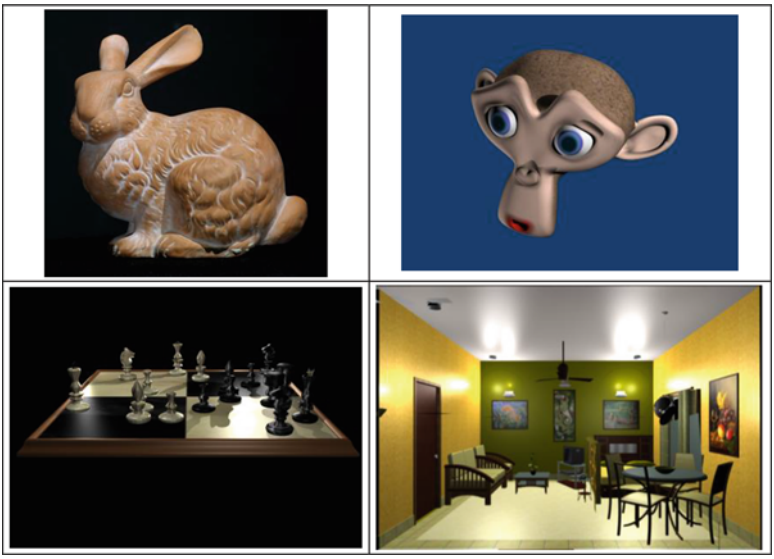

Fig. 2.8 Four 3D scenes chosen for experimental validation

acquisition and the image resolution was originally set to 1024×768 pixels, i.e. $l = 0.05$. However, for experiments reported in Fig. 2.10, to evaluate the effect of the 3D model depth in output PSNR, $\overline{d}$ is changed between [10 m, 50 m], in Fig. 2.17 to evaluate the effect of the camera grid density in output PSNR, k is changed between [0.1 m, 0.9 m], and in Fig. 2.19 to evaluate the effect of the reference camera resolution on output PSNR, l is changed between [0.02 cm,0.1 cm], 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, $l = 0.02$cm on st plane corresponds to a real pixel size equal to 4.8×10^{-4}cm for a typical $1/2''$ camera sensor or capturing resolution of 2560×1920. With the same assumptions, $l = 0.05$ cm corresponds to capturing resolution of 1024×768 and $l = 0.1$ cm to resolution of 512×384.

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 for comparison. In the following, the theoretical expectations in terms of calculated $\overline{\text{ESD}}$ and the actual measurement of output video distortion in PSNR are reported and compared for different rendering methods and different acquisition configurations.

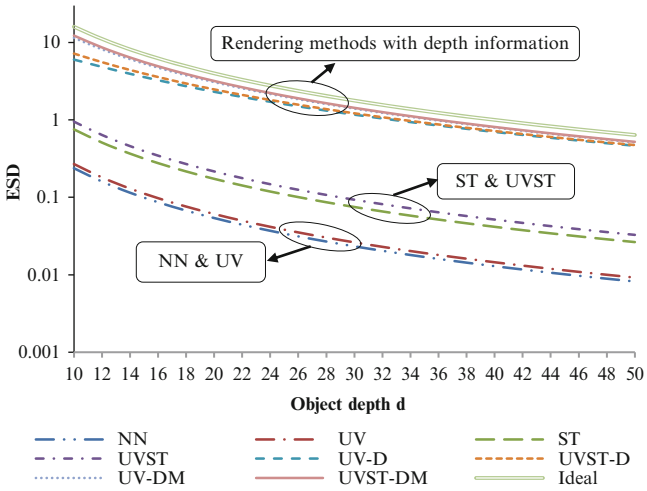

Fig. 2.9 Theoretical $\overline{\text{ESD}}$ for different LF rendering methods based on object depth $\overline{d}$ for $k = 0.4$m and $l = 0.05$cm (i.e. camera resolution of 1024×768)

2.5.4 Results on Rendering Methods

2.5.4.1 Theoretical Expectation

Figure 2.9 shows the ESD for the above-mentioned LF rendering methods in addition to the ideal rendering ($\Delta d = 0$) where $k = 0.4$ m, $l = 0.05$ cm, $d \in [10\,\text{m}, 50\,\text{m}]$, the object length is 5 m, and $\Delta d = 0.1d$, i.e. 10 % 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 $\frac{\text{object length}}{2} + \Delta d$, which in this example is equal to $2.5\text{m} + 0.1d$.

It can be seen from Fig. 2.9 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.

Figure 2.11 demonstrates a bar chart of theoretical ESD values for different rendering methods for $k = 0.4$ m and $l = 0.05$ cm, for a point p with $d = 10$ m and $\Delta d = 1$ m.

Figure 2.13 shows the effect of depth map error on ESD for UV-DM for $l = 0.01$ cm, $|\omega| = 4$, $\overline{d} = 100$, and $\frac{\Delta d}{d}$ between 0 % and 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.

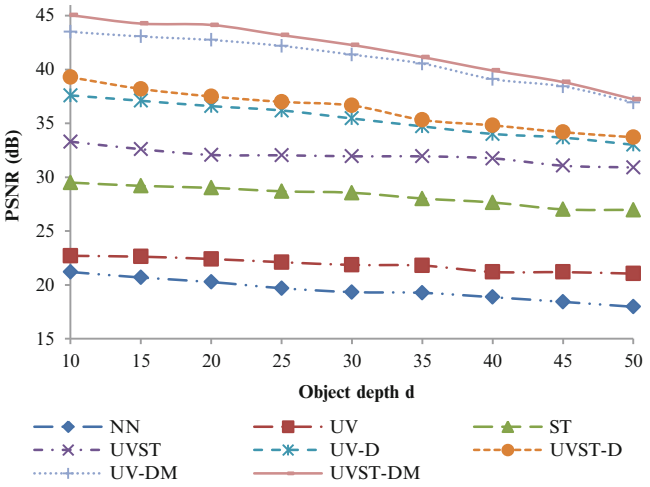

Fig. 2.10 Experimental rendering quality in PSNR for different LF rendering methods vs. object depth $\overline{d}$

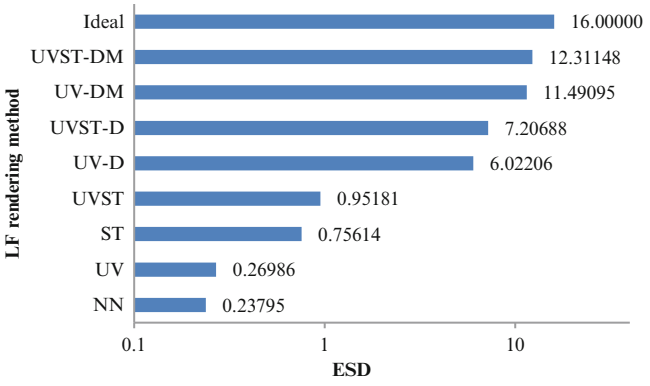

Fig. 2.11 Theoretical $\overline{\text{ESD}}$ for different rendering methods for $k = 0.4$m, $l = 0.05$cm, $\overline{d} = 10$m, and $\overline{\Delta d} = 1$m

2.5.4.2 Simulation Results

Figure 2.10 shows the simulated results, where the object depth d is changed from 10 m to 50 m with steps of 5 m to analyse the effect of d on rendering output distortion in PSNR for different rendering methods. The acquisition parameters are $k = 0.4$ m and $l = 0.05$ cm (i.e. camera resolution of 1024×768). 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. 2.9. 10 % depth error was added in the experiments. Figure 2.10 shows the average results calculated from 288,000 experiments for 9 depths, 8 rendering methods, 4 3D models, and 1000 virtual viewpoints for each experiment. As it can be seen, rendering

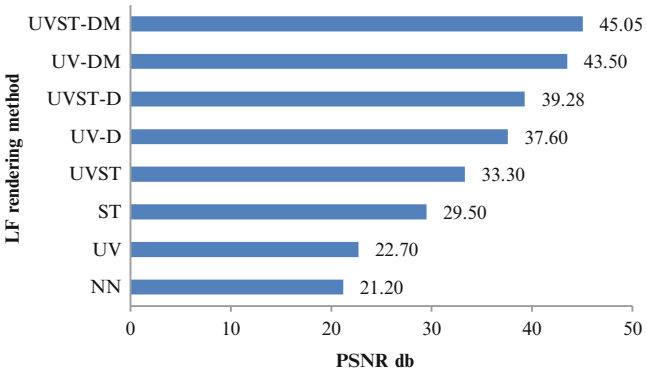

Fig. 2.12 Experimental rendering quality in PSNR for different LR methods

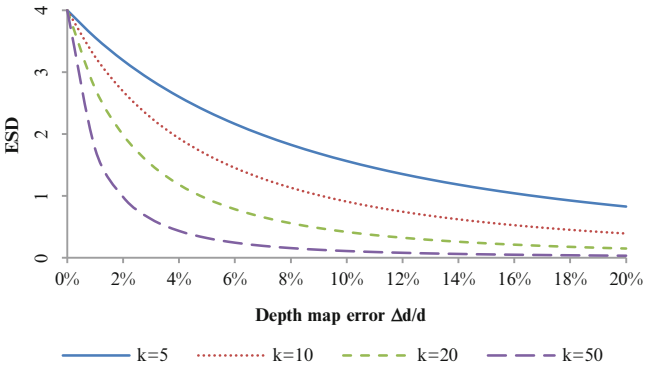

Fig. 2.13 Theoretical ESD for UV-DM for $\overline{d} = 100$, $\overline{\Delta d}$ in the range of $[0\,\%, 20\,\%]$, $l = 0.01$, $|\omega| = 4$ for $k = 5, 10, 20, 50$

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 of 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. 2.9.

Figure 2.12 shows the average PSNR values over 32,000 simulations at $d = 10$ m. 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. 2.11.

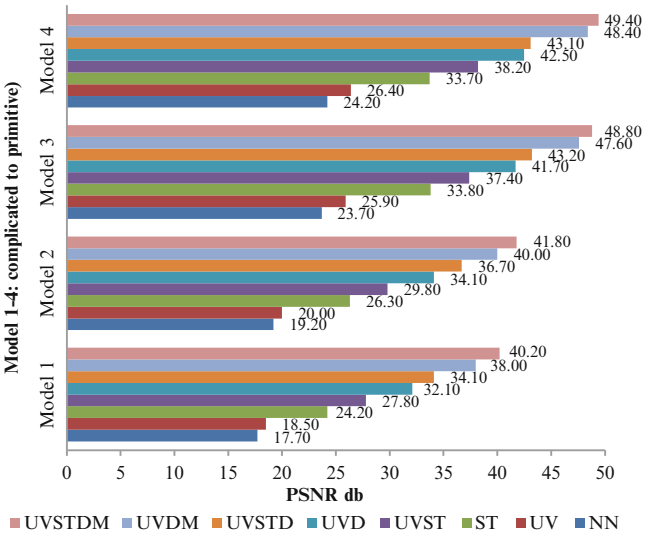

Fig. 2.14 Rendering quality and scene complexity

Figure 2.14 shows the mean PSNR from 144,000 experiments for different rendering methods, categorised 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 terms of higher spatial frequency components. Nevertheless, ESD provides the right ranking on the performance amongst the various methods.

Figure 2.15 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. 2.10 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.

2.5.5 Results on Acquisition Configurations

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

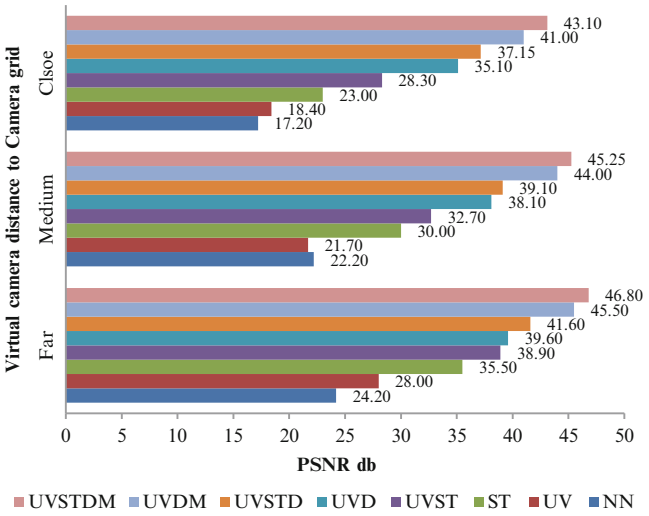

Fig. 2.15 Rendering quality and observation distance

2.5.5.1 Theoretical Expectation

Figure 2.16 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, $d = 50$ m, $l = 0.05$ cm (camera resolution of 1024×768), and $k \in [0.1\text{ m}, 0.9\text{ m}]$ with the same assumption for depth error as the case shown in Fig. 2.9.

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 utilising depth information, the cost of acquisition system can be significantly reduced.

Figure 2.18 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, $d = 50$ m, $k = 0.4$ m, and $l \in [0.02\text{ cm}, 0.1\text{ cm}]$, i.e. camera resolution of 2560×1920 to 512×384, respectively, with the same assumption for depth error as the case shown in Fig. 2.9.

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).

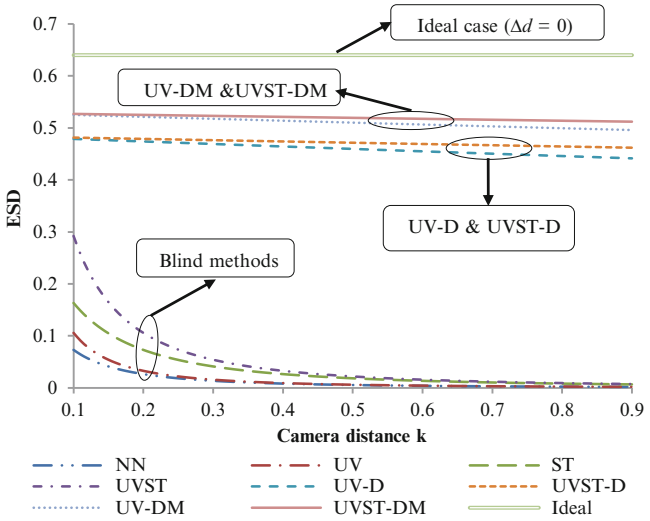

Fig. 2.16 Theoretical $\overline{\text{ESD}}$ for different LF rendering methods based on camera distance k between 0.1m and 0.9m for $l = 0.05$cm

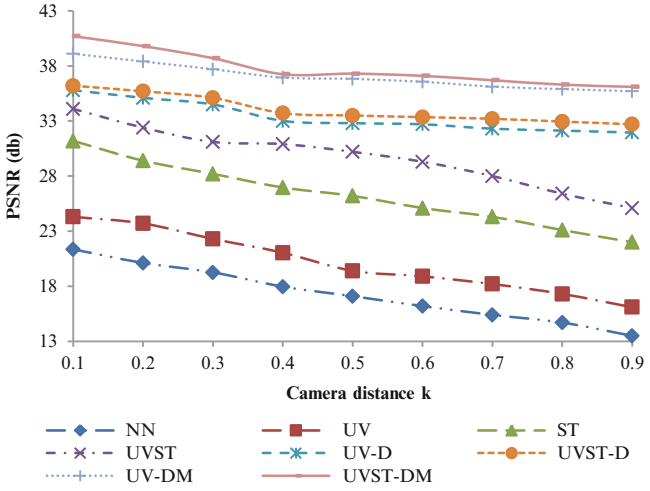

Fig. 2.17 Experimental rendering quality in PSNR for different LF rendering methods vs. camera distance k

2.5.5.2 Simulation Results

Experiments were carried out to see the effect of k in rendering distortion in terms of PSNR so as to make a comparison to the theoretical ESD values. In the first experiment, $d = 50$ m, object length $= 5$ m, $l = 0.05$ cm, and $k \in [0.1\text{ m}, 0.9\text{ m}]$ and 10 % depth error was added. Figure 2.17 shows the results calculated from

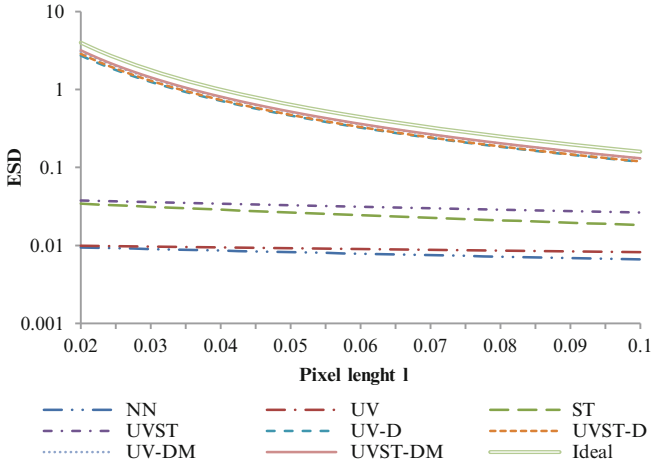

Fig. 2.18 Theoretical $\overline{\text{ESD}}$ for different LF rendering methods based on pixel length l between 0.02cm (camera resolution of 2560 × 1920) and 0.1cm (camera resolution of 512 × 384)

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 terms of pixel length l) and the rendering distortion in terms of PSNR. In this experiment $d = 50$ m, object length $= 5$ m, $k = 0.4$ m, and $l \in [0.02\text{ cm}, \ 0.1\text{ cm}]$, i.e. resolution of 2560 × 1920 to 512 × 384, respectively, and 10 % depth error. Figure 2.19 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.

Therefore, the theoretical expectations based on ESD analysis are confirmed by the empirical results. This can be seen clearly by comparing Fig. 2.16 with Fig. 2.17 and Fig. 2.18 with Fig. 2.19. 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.

2.5.6 Discussions

Figures 2.9, 2.10, 2.11, 2.12, 2.13, 2.14, 2.15, 2.16, 2.17, 2.18, and 2.19 present the theoretical expectations in terms of ESD and experimental results in terms of PSNR for different scenarios. To verify whether ESD is a good distortion indicator,

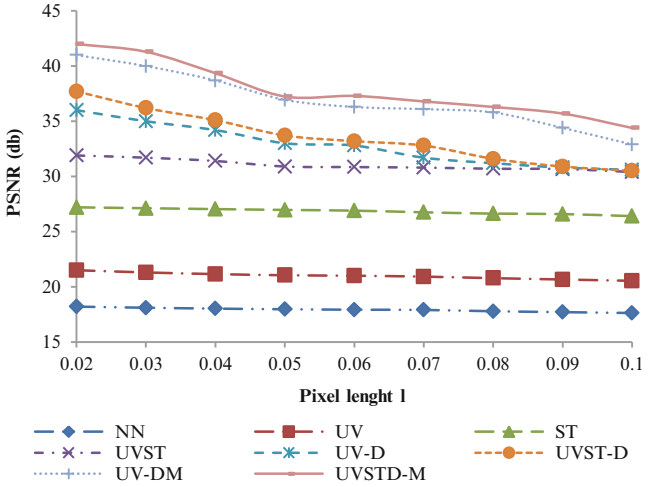

Fig. 2.19 Experimental rendering quality in PSNR for different LF rendering methods vs. pixel length l

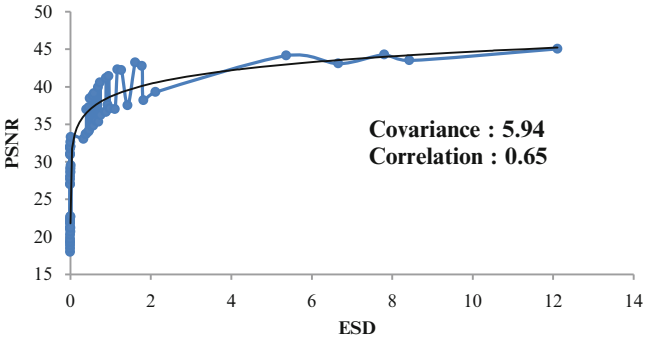

Fig. 2.20 Theoretical calculated ESD from Fig. 2.9 vs. experimental PSNR from Fig. 2.10, both obtained by changing the object depth ($\overline{d}$ from 10 to 50 m)

an analysis was conducted of ESD vs. its counterpart PSNR, i.e. pairs of Figs. (2.9, 2.10), (2.16, 2.17) and (2.18, 2.19).

Figure 2.20 shows the average experimental PSNR from Fig. 2.10 vs. theoretical ESD from Fig. 2.9, both obtained by changing the object depth $\overline{d}$. The trendline, covariance, and correlation of PSNR vs. ESD are also shown in Fig. 2.20.

Similarly, Fig. 2.21 demonstrates the observed PSNR from Fig. 2.17 vs. calculated ESD from Fig. 2.16, both obtained by changing the camera density. Again, the trendline, covariance, and correlation of PSNR vs. ESD are shown.

Figure 2.22 shows the observed PSNR from Fig. 2.19 vs. calculated ESD from Fig. 2.18, both obtained by changing the camera resolution.

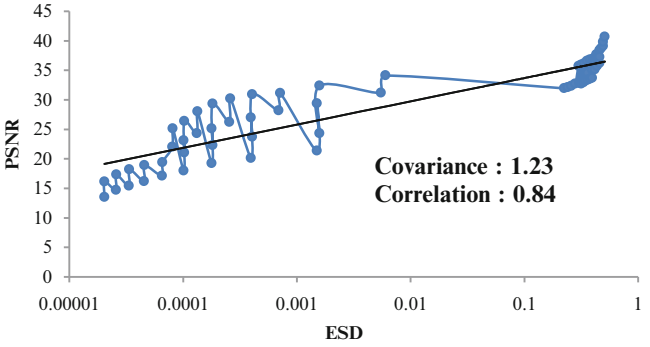

Fig. 2.21 Theoretical calculated ESD from Fig. 2.16 vs. experimental PSNR from Fig. 2.17, both obtained by changing the camera density (k from 1 to 9 m)

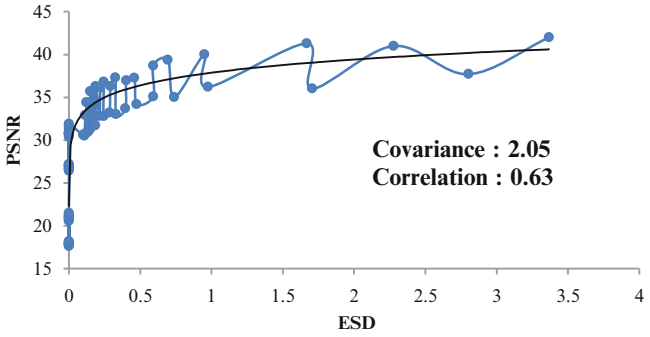

Fig. 2.22 Theoretically calculated ESD from Fig. 2.18 vs. experimental PSNR from Fig. 2.19, both obtained by changing the resolution (l from 0.02 to 0.1 cm)

Figures 2.20, 2.21, and 2.22 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.

2.6 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 MSE is a function of $\overline{\text{ESD}}$ for each given LF rendering method, denoted by $\text{ESD}_{\text{method}}$, and for a given fixed scene, i.e. $\text{MSE} = f(\text{ESD}_{\text{method}})$. In general, empirical f can be formulated as

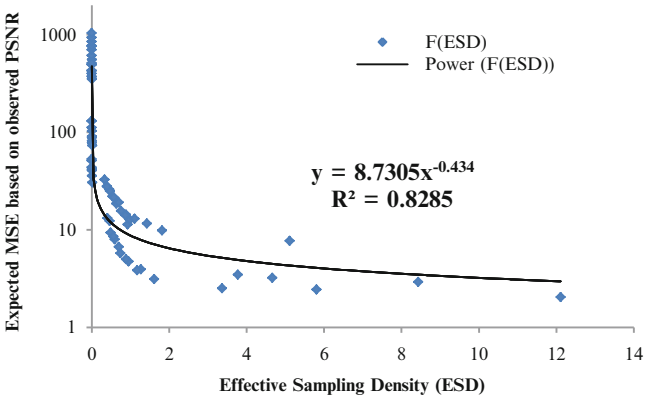

Fig. 2.23 A general curve fitting for f(ESD) estimation based on calculated $\overline{\text{ESD}}$ vs. expected MSE

$$f\left(\text{ESD}_{\text{method}}\right) = Q \times \text{ESD}_{\text{method}}{}^{P} \tag{2.12}$$

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:

$$f\left(\text{ESD}_{\text{method}}\right) = \text{Expected MSE} = \frac{255^2}{10^{\left(\frac{\text{Observed PSNR}}{10}\right)}} \tag{2.13}$$

The data presented in Figs. 2.9 and 2.10 (theoretical and experimental results based on changing the object depth) is used as the training set and data demonstrated in Figs. 2.16, 2.17, and Figs. 2.18, 2.19 for validation. Figure 2.23 demonstrates the overall curve fitting. This curve fitting is done on all the data and without clustering the data based on the rendering methods. Figure 2.24 shows the curve fitting for each LF rendering method separately (method dependent). The optimum value for $f(\text{ESD}_{\text{method}})$ for best estimation is when it is equal to expected MSE.

Figure 2.25 shows a summary of curve fitting and validation errors of PSNR estimation for all LF rendering methods. As it can be seen from Fig. 2.25, 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 $f(\text{ESD}_{\text{method}})$ are accurate to indicate the rendering distortion in terms of PSNR. These equations offer a way to directly estimate the overall rendering distortion of an LF-based FVV system from the calculated ESD without implementation and experiments.

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 optimise the LF acquisition

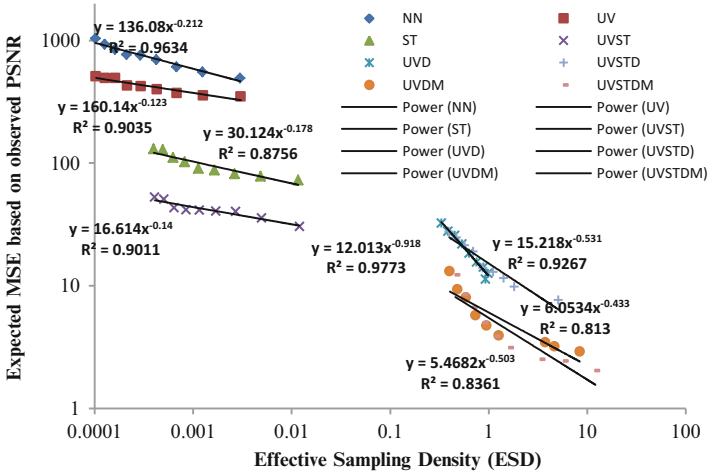

Fig. 2.24 Method-dependent curve fittings for $f(\mathrm{ESD}_{\mathrm{method}})$

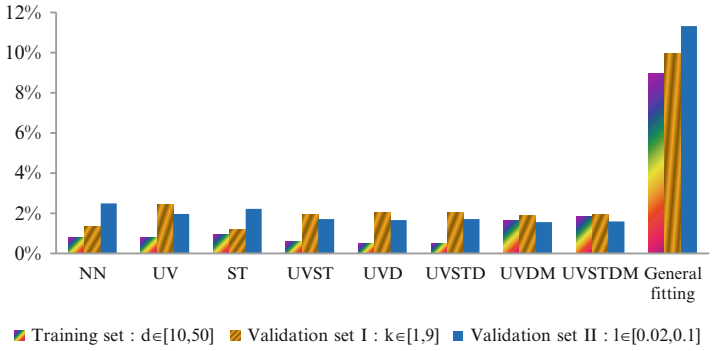

Fig. 2.25 Summary of curve fitting training and validation errors of PSNR estimation

and LF rendering components without exhaustive experimental implementation of each configuration. For instance, for a general UVDM$(d, \Delta d, k, l, |\omega|)$ method, by applying the ESD from Eq. (2.11), the rendering distortion can be directly calculated as

$$\mathrm{PSNR}_{\mathrm{UVDM}(d,\Delta d,k,l,|\omega|)} \cong 20\log_{10} \frac{255}{\sqrt{3.4545\left(\frac{|\omega|}{\left[l(d+\Delta d)+\frac{\Delta d \times k}{d}\left(\sqrt{|\omega|}-1\right)\right]^2}\right)^{-0.256}}} \tag{2.14}$$

Table 2.5 summarises the empirical boundaries of Q and P for different LF rendering methods, estimated for different scenes and acquisitions.

Table 2.5 Empirical boundaries of $\boldsymbol{P}$ and $\boldsymbol{Q}$

LF rendering method type	LF rendering method	Q	P
LF rendering methods with no depth information $10 < Q < 300$ $-0.3 < P < -0.1$	NN	$50 < Q_{NN} < 300$	$-0.3 < P_{NN} < -0.2$
	ST	$20 < Q_{ST} < 200$	$-0.2 < P_{ST} < -0.1$
	UV	$20 < Q_{UV} < 250$	$-0.25 < P_{UV} < -0.1$
	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 $f(\mathrm{ESD}_{\mathrm{method}})$ 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 $\overline{\mathrm{ESD}}$.

2.7 Subjective Assessment

While previous section discussed the correlation between ESD and output video distortion in terms 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 standardisation and guidelines on "subjective video quality assessment methods for multimedia applications" [25] and using degradation category rating (DCR) method was carried out. The test procedure is based on recommendations proposed in VQEG reports [59, 60]. 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 [61] was used as reference images. The original Stanford camera grid to capture real scenes is 17×17, 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 8×8 camera grid was selected for acquisition component and a subset of other cameras were used as ground truth. Eighteen subjects participated in the test. For each of the 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 [25, 62]. 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 Tables 2.1 and 2.3. There are totally 12 values for ESD (4 scenes and 3 rendering methods). Each value of ESD is corresponded to eight different views.

Figure 2.26 shows samples of the test sequences, presented to the subject panel. Note that Fig. 2.26 shows 12 different pairs out of 96 test sequences which were presented to each subject.

Figure 2.27 illustrates the results of the subjective test for each rendering method. The average and variance of the impairment for each rendering method were 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

Fig. 2.26 Samples of test sequences used in the subjective assessment

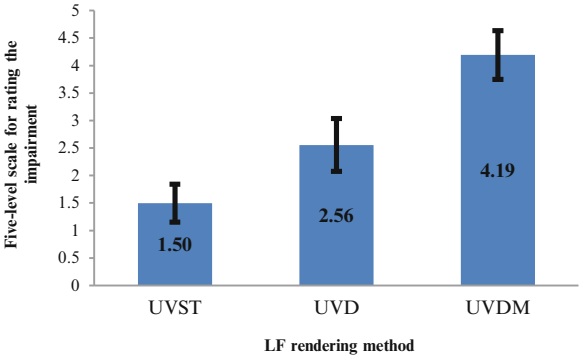

Fig. 2.27 Subjective assessment of three LF rendering methods by using degradation category rating (DCR), showing the mean and variance of rating from 576 collected scores for each method (32 test sequences among 18 subjects) with a five-level scale for rating the impairment

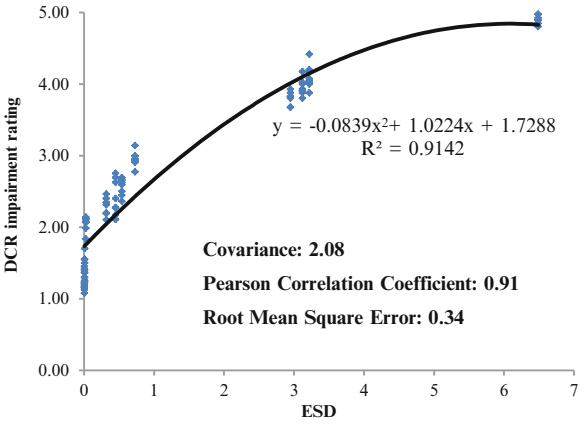

Fig. 2.28 DCR impairment rating for subjective assessment vs. theoretical ESD and the empirical relationship between these two parameters

proposed in VQEG reports [59, 60] was employed. Figure 2.28 shows the scatter plot for the ESD-DCR couples for all 96 test sequences. Please note that for each eight 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 calculated ESD which is technically satisfactory.

Figure 2.29 shows an outdoor scene rendered with the proposed FVV system for subjective comparison of ground truth with the rendered output.

2.8 Application of ESD

2.8.1 Calculating the Minimum Number of Cameras

Regular camera grids are widely used for FVV acquisition. Several studies are reported to calculate the minimum number of cameras for regular grids which can be categorised into three main approaches: (a) plenoptic signal spectral analysis [3, 24] and the light-field spectral and frequency analysis [4, 5], (b) view interpolation geometric analysis such as [6], and (c) optical analysis of light field [14, 35, 36]. However, these methods are essentially based on several simplifying assumptions (e.g. Lambertian scene, no occlusion, linear interpolation over 4 or 16 rays, and calculating the Nyquist sampling rate without considering under-sampling), and also suggest an impractically high number of cameras [28, 29].

In contrast, using ESD to address this problem has several advantages such as studying under-sampled light field under realistic conditions (non-Lambertian

Fig. 2.29 An outdoor scene, ground truth, and the rendered output for subjective comparison

reflections and occlusions) and rendering with complex interpolations. The optimisation method based on ESD proposed in [28, 29] is summarised here.

In $\mathrm{ESD}_{\mathrm{UVDM}(d,\Delta d,k,l,|\omega|)}$ expression given as Eq. (2.11), d is given by scene geometry and Δd is determined by the depth estimation method and cannot be altered by us. Changing the other three parameters could potentially improve the rendering quality. By assuming a given camera resolution, i.e. a fixed value of l, two other parameters can be tuned to compensate for the depth estimation error while maintaining the rendering quality. These parameters include k as a measure of density of cameras during acquisition and $|\omega|$ as an indicator of complexity of rendering method. ESD is proportional to $|\omega|$ and inversely proportional to k, i.e. higher camera density (smaller k) and employing more rays for interpolation results in higher ESD. The optimisation of k is summarised here and optimisation of $|\omega|$ will be discussed in next subsection.

The problem of calculating the minimum number of cameras can be expressed in terms of minimum camera density, i.e. maximum k to provide required ESD in each point of the scene to compensate for the adverse effect of depth map estimation errors. This minimum required ESD can be calculated for the ideal case when there is no error in depth estimation and there are n rays employed for interpolation. Hence the optimisation method can be written as follows:

Find the maximum k to satisfy

$$\mathrm{ESD}_{\mathrm{UVDM}(d,\Delta d,k,l,|\omega|)} = \mathrm{ESD}_{\mathrm{Ideal}} \rightarrow \mathrm{ESD}_{\mathrm{UVDM}(d,\Delta d,k,l,|\omega|)} = \mathrm{ESD}_{\mathrm{UVDM}(d,0,k,l,n)} \rightarrow$$

$$k = \frac{ld\left(d\sqrt{\frac{|\omega|}{n}} - d - \Delta d\right)}{\Delta d\left(\sqrt{|\omega|} - 1\right)} = \frac{l\left(\left(\sqrt{\frac{|\omega|}{n}} - 1\right)d^2 - d\Delta d\right)}{\Delta d\left(\sqrt{|\omega|} - 1\right)} \tag{2.15}$$

where

$$\Delta d > 0 \text{ and } |\omega| > n\left(\frac{d + \Delta d}{d}\right)^2$$

Figure 2.30 shows the summary of theoretical expectations and experimental results for the optimisation process. Figure 2.30a, b illustrates the theoretical expectations. It is assumed that $l = 0.01$, average depth of scene $\overline{d} = 100$, relative depth map error $\frac{\Delta d}{d}$ between 1 % and 20 %, and $|\omega|$ is calculated as follows to satisfy the condition of Eq. (2.15): $|\omega| > 4\left(\frac{100+20}{100}\right)^2 > 5.76 \rightarrow |\omega| = 6$. For any given depth estimation error $\Delta d \leq 20\%$, k is calculated directly from Eq. (2.15) to maintain $\overline{\mathrm{ESD}}$ at 4.00, the ideal ESD calculated for $n = 4$ and $\Delta d = 0$. Figure 2.30a demonstrates the ESD for fixed $k = 14.4$ and optimum k calculated from Eq. (2.15). Figure 2.30b shows the calculated k in such a scenario. The corresponding point for 10 % error in depth estimation is highlighted in Fig. 2.30a, b, respectively, to show the relation of these two figures. Figure 2.30c shows that the rendering PSNR is maintained at a prescribed value (for instance 50 dB) with calculated k in contrast with the average

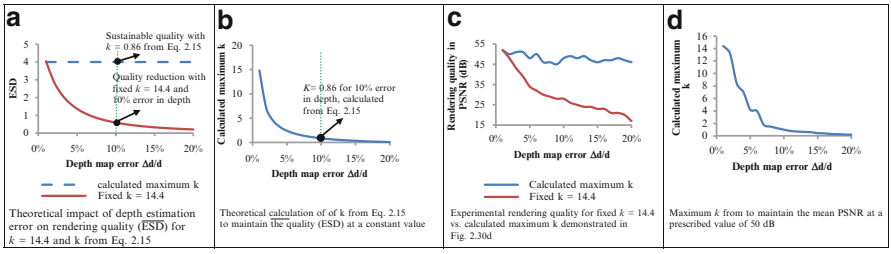

Fig. 2.30 Summary of theoretical and experimental optimisation of k (camera density) based on ESD

PSNR for fixed $k = 14.4$; the required k to maintain the quality is demonstrated in Fig. 2.30d. Figure 2.30 shows that for high error rates, changing k using Eq. (2.15) results in significant improvements over the fixed camera density and can maintain the quality around the prescribed 50 dB.

2.8.2 Calculating the Minimum Interpolation Complexity

The number of rays selected by *ray selection* process of a given rendering method is an important parameter of the rendering complexity. On the one hand, increasing the number of rays results in increasing ESD in each point of the scene resulting in higher output quality. On the other hand, this also increases the interpolation complexity resulting in slower rendering which might not be acceptable in real-time applications. To calculate the optimum number of rays for interpolation to satisfy both required rendering quality and rendering efficiency, an optimisation method is proposed in [28, 31].

With the same approach as in previous subsection the minimum $|\omega|$ to avoid quality deterioration due to errors in depth maps can be calculated as

$$\begin{aligned} &\text{Find the minimum } |\omega| \text{ to satisfy} \\ &\text{ESD}_{\text{UVDM}(d,\Delta d,k,l,|\omega|)} = \text{ESD}_{\text{Ideal}} \rightarrow \text{ESD}_{\text{UVDM}(d,\Delta d,k,l,|\omega|)} \\ &= \text{ESD}_{\text{UVDM}(d,0,k,l,n)} \rightarrow |\omega| = \left(\frac{l\,(d+\Delta d) - \frac{\Delta d \times k}{d}}{\frac{ld}{\sqrt{n}} - \frac{\Delta d \times k}{d}} \right)^2 \end{aligned} \tag{2.16}$$

where

$$k < \frac{ld^2}{\Delta d \sqrt{n}}$$

Figure 2.31 shows the summary of theoretical expectations and experimental results for the optimisation process.

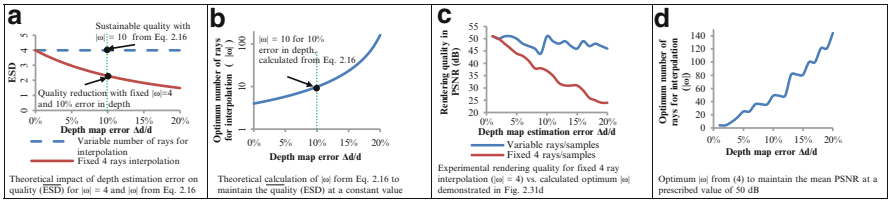

Fig. 2.31 Summary of theoretical and experimental optimisation of $|\omega|$ (number of rays employed in interpolation)

Figure 2.31a, b shows the theoretical expectations for this optimisation model. l, $\overline{d}$, and $\overline{\Delta d}$ are the same as in Fig. 2.30. k is calculated as follows to satisfy the condition of Eq. (2.16): $k < \frac{0.01 \times 100^2}{20\sqrt{4}} < 2.5 \rightarrow k = 2.2$. For any $\Delta d < 20\%$, $|\omega|$ is calculated directly from Eq. (2.16) to maintain $\overline{\text{ESD}}$ at 4.00, the ideal ESD calculated for $n = 4$. Figure 2.31a demonstrates the ESD for fixed four-ray interpolation and for optimum number of rays calculated from Eq. (2.16). Figure 2.31b shows the actual number of rays $|\omega|$, employed in interpolation in such a scenario. The corresponding point for 10 % error in depth estimation is highlighted in Fig. 2.31a, b, respectively, to show the relation of these two figures. Figure 2.31c shows that the rendering PSNR is maintained at a prescribed value (for instance 50 dB) with calculated optimum number of rays $|\omega|$ in contrast with the average PSNR for conventional fixed four-ray interpolation, and calculated number of rays $|\omega|$ is demonstrated in Fig. 2.31d. Figure 2.31 shows that for high level of error in depth, the use of optimum $|\omega|$ using Eq. (2.16) results in significant improvements over the conventional fixed four-ray interpolation and can maintain the rendering quality around the prescribed 50 dB.

2.8.3 *Irregular Acquisition Based on the Scene Complexity*

As noted before, FVV acquisition is typically performed by using a regular camera grid. While a regular acquisition itself results in non-uniform sampling density, this non-uniformity does not match the scene complexity and frequency variations. The simplest non-uniform acquisition can be done by using an irregular camera grid. The problem is then to find the positions and orientations of the camera in the gird to provide higher ESD in the parts of the scene with higher complexity and vice versa. The theory of irregular/non-uniform signal sampling has been widely investigated and it is shown that irregular sampling can reduce the number of required samples for perfect reconstruction of the signal. However to the best of our knowledge, this property has not been explored for FVV acquisition and rendering. An optimisation method based on ESD for this problem is proposed in [28, 30]. It is shown that ESD can be regarded as a set of utility functions U_h(ESD) based on the given scene complexity factor h. The higher the scene complexity, more ESD would be required for a given reconstruction fidelity. Each acquisition configuration and rendering method result in an ESD pattern, which varies in the scene space. Assume that the scene could be partitioned into a number of smaller 3D regions or blocks, each having a fixed average complexity h, determined from the highest frequency components of the block computed by applying DCT transform. Then, the aim of the optimisation problem could be to find the optimum acquisition configuration which provides the minimum required ESD for all blocks. This optimisation problem is discussed in [28, 30] and is shown that an analytical dynamic programming solution is available to compute the optimum irregular camera grid.

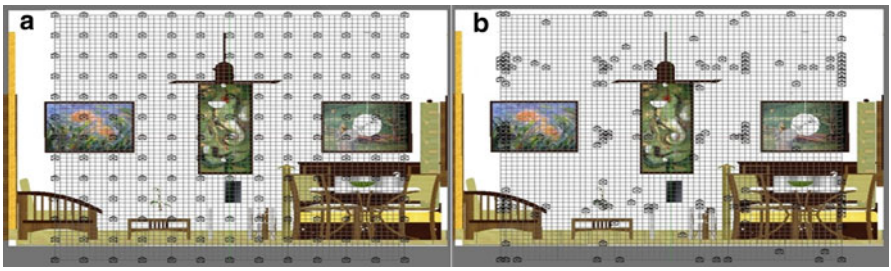

Fig. 2.32 (**a**) Regular camera grid with 169 (13 × 13) cameras; (**b**) optimum irregular camera grid for 169 cameras

Theoretical analysis and experimental validation showed that the output video quality can be significantly improved (around 20 % in mean PSNR) by employing the proposed irregular acquisition compared with the regular camera grid. Figure 2.32 shows the initial regular camera grid and the optimum irregular camera grid for 169 cameras. The average of rendering PSNR from 1000 virtual cameras was improved from 39.10 dB for regular grid to 46.60 dB for optimum irregular grid.

2.9 Conclusion

This chapter has discussed the concept of ESD and its application in FVV quality assessment, and comparison, evaluation, and optimisation of FVV acquisition and rendering subsystems. 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 analysed through ESD and simulation. The results have shown that ESD is an effective indicator of distortion that can be obtained directly from 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. Finally several problems on FVV evaluation and optimisation have been approached by using ESD. This has been done by analysing the impact of depth estimation errors on ESD and optimisation of ESD with respect to the *camera density* and *ray selection complexity* for a given output quality. Although this chapter focuses on the overall distortion of an LF-based FVV system, the concept is readily extended to measure the rendering quality at a specific location or part of the scene.

2.10 Biography

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 Ph.D. 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 optimisation.

Farzad Safaei graduated from the University of Western Australia with the degree of Bachelor of Engineering (Electronics) and obtained his Ph.D. 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, Melbourne, Australia. His research interests include immersive multimedia communications and free viewpoint TV.

Wanqing Li received his Ph.D. 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, Redmond, USA, 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.

References

1. Tanimoto M, Tehrani MP, Fujii T, Yendo T (2011) Free-viewpoint TV. IEEE Signal Process Mag 28:67–76
2. Tanimoto M (2012) FTV: free-viewpoint television. Signal Process Image Comm 27:555–570
3. Chai JX, Tong X, Chan SC, Shum HY (2000) Plenoptic sampling. Proc SIGGRAPH (ACM Trans Graphics) 307–318
4. Zhang C, Chen T (2003) Spectral analysis for sampling image-based rendering data. IEEE Trans Circ Syst Video Technol 13:1038–1050
5. Zhang C, Chen T (2006) Light field sampling. Synth Lect Image Video Multimed Process 2:1–102
6. Zhouchen L, Heung-Yeung S (2004) A geometric analysis of light field rendering. Int J Comput Vision 58:121–138
7. King-To N, Zhen-Yu Z, Chong W, Shing-Chow C, Heung-Yeung S (2012) A multi-camera approach to image-based rendering and 3-D/multiview display of ancient Chinese artifacts. IEEE Trans Multimed 14:1631–1641

8. Safaei F, Mokhtarian P, Shidanshidi H, Li W, Namazi-Rad M, Mousavinia A (2013) Scene-adaptive configuration of two cameras using the correspondence field function. In: IEEE international conference on multimedia and expo (ICME). pp 1–6
9. Takahashi K, Naemura T (2006) Layered light-field rendering with focus measurement. Signal Process Image Comm 21:519–530
10. Daniel NW, Daniel IA, Ken A, Brian C, Tom D, David HS et al (2000) Surface light fields for 3D photography. In: 27th annual conference on computer graphics and interactive techniques
11. Jingyi Y, McMillan L, Gortler S (2002) Scam light field rendering. In: 10th pacific conference on computer graphics and applications. pp 137–144
12. Shum HY, Sun J, Yamazaki S, Lin Y, Tang CK (2004) Pop-up light field: an interactive image-based modeling and rendering system. ACM Trans Graphics 23:143–162
13. Wen W, Jiang Zhang Z, Yao Si C, Zeng D (2010) An efficient method for all-in-focused light field rendering. In: 3rd IEEE international conference on computer science and information technology (ICCSIT). pp 399–404
14. Aaron I, Leonard M, Steven JG (2000) Dynamically reparameterized light fields. In: 27th annual conference on computer graphics and interactive techniques
15. Hansung K, Guillemaut JY, Takai T, Sarim M, Hilton A (2012) Outdoor dynamic 3-D scene reconstruction. IEEE Trans Circ Syst Video Technol 22:1611–1622
16. Liu SX, An P, Zhang ZY, Zhang Q, Shen LQ, Jiang GY (2009) High quality virtual view synthesis based on corrected surface mapping and image fusion. Electron Lett 45:30–32
17. Ekmekcioglu E, Velisavljevic XV, Worrall ST (2011) Content adaptive enhancement of multi-view depth maps for free viewpoint video. IEEE J Selected Topics Signal Process 5:352–361
18. Scandarolli T, de Queiroz RL, Florencio DA (2013) Attention-weighted rate allocation in free-viewpoint television. IEEE Signal Process Lett 20:359–362
19. Qifei W, Xiangyang J, Qionghai D, Naiyao Z (2012) Free viewpoint video coding with rate-distortion analysis. IEEE Trans Circ Syst Video Technol 22:875–889
20. Zhun H, Qionghai D (2007) A new scalable free viewpoint video streaming system over IP network. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). pp II-773-II-776
21. Adelson EH, Bergen JR (1991) The plenoptic function and the elements of early vision. In: Computational models of visual processing. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, pp 3–20
22. Levoy M, Hanrahan P (1996) Light field rendering. Proc SIGGRAPH (ACM Trans Graphics) 31–42
23. Gortler SJ, Grzeszczuk R, Szeliski R, Cohen MF (1996) The lumigraph. Proc SIGGRAPH (ACM Trans Graphics) 43–54
24. Do MN, Marchand-Maillet D, Vetterli M (2012) On the bandwidth of the plenoptic function. IEEE Trans Image Process 21:708–717
25. ITU-T Recommendation P (1999) Subjective video quality assessment methods for multimedia applications
26. Shidanshidi H, Safaei F, Li W (2011) Objective evaluation of light field rendering methods using effective sampling density. In: IEEE international workshop on multimedia signal processing (MMSP). pp 1–6
27. Shidanshidi H, Safaei F, Li W (2015) Estimation of signal distortion using effective sampling density for light field based free viewpoint video. IEEE Trans Multimed 17(10):1677–1693
28. Shidanshidi H (2014) Effective sampling density for quality assessment and optimization of light field rendering and acquisition. Doctor of Philosophy Thesis, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong
29. Shidanshidi H, Safaei F, Li W (2013) A method for calculating the minimum number of cameras in a light field based free viewpoint video system. In: IEEE international conference on multimedia and expo (ICME). pp 1–6
30. Shidanshidi H, Safaei F, Zamani-Farahani A, Li W (2013) Non-uniform sampling of plenoptic signal based on the scene complexity variations for a free viewpoint video system. In: IEEE international conference on image processing (ICIP). pp 3147–3151

31. Shidanshidi H, Safaei F, Li W (2015) Optimization of the number of rays in interpolation for light field based free viewpoint systems. In: IEEE international conference on multimedia and expo (ICME). pp 1–6
32. Shidanshidi H, Safaei F, Li W (2015) Effective sampling density and its applications to the evaluation and optimization of free viewpoint video systems. IEEE COMSOC MMTC E-Lett 10(2):21–25
33. Shidanshidi H, Safaei F, Li W (2016) Optimization of free viewpoint video acquisition and rendering subsystems by using effective sampling density. IEEE Trans Multimedia TBA
34. Camahort E, Lerios A, Fussell D (1998) Uniformly sampled light fields. Rendering Tech 98:117–130
35. Feng T, Shum HY (2000) An optical analysis of light field rendering. In: Fifth Asian conference on computer vision. pp 394–399
36. Lumsdaine A, Georgiev T (2008) Full resolution lightfield rendering. Indiana Univ Adobe Syst Tech Rep
37. Stewart J, Yu J, Gortler SJ, McMillan L (2003) A new reconstruction filter for undersampled light fields. In: 14th Eurographics workshop on rendering, Leuven, Belgium
38. Wenfeng L, Jin Z, Baoxin L, Sezan MI (2009) Virtual view specification and synthesis for free viewpoint television. IEEE Trans Circ Syst Video Technol 19:533–546
39. Zitnick CL, Kang SB, Uyttendaele M, Winder S, Szeliski R (2004) High-quality video view interpolation using a layered representation. Proc Siggraph (ACM Trans Graphics) 600–609
40. Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR. pp 519–528
41. Kilner J, Starck J, Guillemaut JY, Hilton A (2009) Objective quality assessment in free-viewpoint video production. Image Commun 24:3–16
42. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15:430–444
43. Pons A, Malo J, Artigas J, Capilla P (1999) Image quality metric based on multidimensional contrast perception models. Displays 20:93–110
44. Winkler S (1998) A perceptual distortion metric for digital color images. In: ICIP, vol 3. pp 399–403
45. Brandão T, Queluz P (2006) Towards objective metrics for blind assessment of images quality. In: IEEE international conference on image processing (ICIP). pp 2933–2936
46. Seshadrinathan K, Bovik AC (2007) A structural similarity metric for video based on motion models. In: IEEE international conference on acoustics, speech, and signal processing, vol 1, pp I-869–I-872
47. Winkler S (2007) Video quality and beyond. In: European signal processing conference. pp 3–7
48. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13:600–612
49. Eskicioglu AM, Fisher PS (1995) Image quality measures and their performance. IEEE Trans Commun 43:2959–2965
50. Avcıbaş İ, Sankur B, Sayood K (2002) Statistical evaluation of image quality measures. J Electron Imag 11:206
51. Bosc E, Pepion R, Le Callet P, Koppel M, Ndjiki-Nya P, Pressigout M et al (2011) Towards a new quality metric for 3-D synthesized view assessment. IEEE J Selected Topics Signal Process 5:1332–1343
52. Bosc E, Koppel M, Pepion R, Pressigout M, Morin L, Ndjiki-Nya P et al (2011) Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols? In: 18th IEEE international conference on image processing (ICIP). pp 2597–2600
53. Raskar R, Agrawal AK (2010) 4D light field cameras. Google Patents (ed)
54. Takahashi K (2012) Theoretical analysis of view interpolation with inaccurate depth information. IEEE Trans Image Process 21:718–732

55. Shidanshidi H, Safaei F, Li W (2011) A quantitative approach for comparison and evaluation of light field rendering techniques. In: IEEE international conference on multimedia and expo (ICME). pp 1–4
56. Schwarz S, Olsson R, Sjostrom M (2013) Depth sensing for 3DTV: a survey. IEEE Multimed 20:10–17
57. Khoshelham K, Elberink SO (2012) Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors 12:1437–1454
58. Pattinson T (2010) Quantification and description of distance measurement errors of a time-of-flight camera. M.Sc. Thesis, University of Stuttgart, Stuttgart, Germany
59. (2001) Methodological framework for specifying accuracy and cross-calibration of video quality metrics. Tech. Rep. T1.TR.72-2001
60. Brill MH, Lubin J, Costa P, Wolf S, Pearson J (2004) Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1A1. Signal Process Image Comm 19:101–107
61. The (new) Stanford light field archive. Stanford University Computer Graphics Laboratory, [Online]. http://lightfield.stanford.edu/lfs.html
62. Mantiuk RK, Tomaszewska A, Mantiuk R (2012) Comparison of four subjective methods for image quality assessment. In: Computer graphics forum, vol 31 (no. 8). Blackwell Publishing Ltd., pp 2478–2491

http://www.springer.com/978-1-4939-4024-0

Connected Media in the Future Internet Era

Kondoz, A.; Dagiuklas, T. (Eds.)

2017, V, 224 p. 108 illus., 87 illus. in color., Hardcover
ISBN: 978-1-4939-4024-0