Research Article Vol. 27, No. 3 | 4 Feb 2019 | OPTICS EXPRESS 3557

Optics EXPRESS

Light-field compression using a pair of steps and depth estimation

XINPENG HUANG,[1] PING AN,[1,*] FENGYIN CAO,[1] DEYANG LIU,[2] AND QIANG WU[3]

[1] *Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China*
[2] *School of Computer and Information, Anqing Normal University, Anqing 246000, China*
[3] *Faculty of Engineering and Information Technology, University of Technology, Sydney, Ultimo, NSW 2007, Australia*
[*] *anping@shu.edu.cn*

Abstract: Advanced handheld plenoptic cameras are being rapidly developed to capture information about light fields (LFs) from the 3D world. Rich LF data can be used to develop dense sub-aperture images (SAIs) that can provide a more immersive experience for users. Unlike conventional 2D images, 4D SAIs contain both the positional and directional information of light rays; the practical applications of handheld plenoptic cameras are limited by the huge volume of data required to capture this information. Therefore, an efficient LF compression method is vital for further application of the cameras. To this end, the pair of steps and depth estimation (PoS&DE) method is proposed in this paper, and the multiview video and depth (MVD) coding structure is used to relieve the LF coding burden. More specifically, a precise depth-estimation approach is presented for SAIs based on the cost function, and an SAI-guided depth optimization algorithm is designed to refine the initial depth map based on pixel variation tendency. Meanwhile, to reduce running time, intermediate SAI synthesis quality and coding bitrates, including the key SAIs selected and cost-computation steps, are set via extensive statistical experiments. In this way, only a limited number of optimally selected SAIs and their corresponding depth maps must be encoded. The experimental results demonstrate that our proposed LF compression solution using PoS&DE can obtain a satisfied coding performance.

1. Introduction

The rapid development of computational photography has brought great advances and changes to immersive 3D applications. As a computational photography technique, a light field (LF) represents all the light information in a scene, and thus provides a more immersive experience for audiences [1, 2]. Because conventional imaging technologies project a 3D scene onto a 2D plane, there is a loss of stereoscopic sensation. As opposed to conventional cameras, plenoptic cameras have micro-lens arrays in front of their sensor, wherein both the spatial and angular information of light rays can be simultaneously captured in one shot [3]. An image captured by a plenoptic camera is called a light-field image (LFI), which is shown as Fig. 1. As seen, an LFI is composed of massive hexagonal blocks (also called Micro-Images, MIs). Using LFIs, focal plane shifting and dense viewpoint image acquisition can be implemented in postprocessing, and even a viewpoint at a new position can be synthesized [4]. To capture more extra data of target, the light-field-based hardware is improved using the special materials [5]. Due to the additional information within the LFI, it can solve many traditional vision problems, such as image segmentation [6], salience detection [7], depth estimation [8-11], and video stabilization [12] after geometric calibration [13]. Besides, another prevalent application of light field is in microscopy [14-16], where the holographic visualization can explore the potential application of light field in medical and biological researches. The equipment increases field of view, but

#354464 https://doi.org/10.1364/OE.27.003557
Journal © 2019 Received 5 Dec 2018; revised 20 Jan 2019; accepted 20 Jan 2019; published 31 Jan 2019

collects much semi-transparent objects. It is noticed that the existing pixel-based depth estimation algorithms undeniably cannot work for the semi-transparent objects.

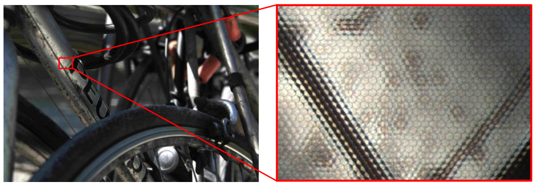


Fig. 1. The raw LFI captured by Lytro-Illum camera (left) and detail of the LFI (right).

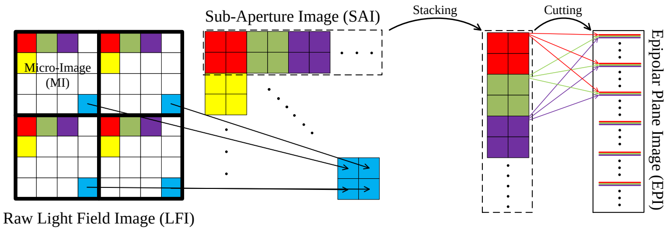


Fig. 2. The various transformations of light field image.

The various transformations of light field image are illustrated in Fig. 2. It descripts that the raw light-field image (LFI) is composed of micro-image (MI); the sub-aperture image (SAI) is derived by rearranging the co-located pixels each MI; the epipolar plane image (EPI) is generated by stacking and cutting SAIs. Additionally, SAIs can provide an immersive experience with the naked eye [17]. A raw LFI with MIs obstructs spatial correlation elimination; it can be observed in Fig. 2 that each pixel from the same MI contributes to different SAIs. Consequently, LFIs with SAIs have been a focus of investigation, as angular correlation is prone to exploitation. However, the huge number of SAIs prevents practical application of LFIs for restoring and transporting. Many technological challenges remain associated with the huge amount of data. To this end, we previously proposed an LF compression method in [18, 19]. This article, based on our earlier work, deeply analyses all combinations of depth computation and sub-aperture images sparsely sampling steps to compensate the computational complexity and the coding efficiency. Although these improvements seem to be unimpressive, they can positively affect depth estimation, bitrates reduction and reconstruction performance.

2. Related works

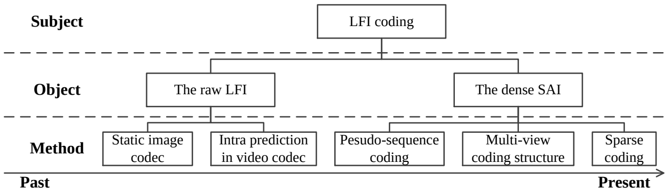


Fig. 3. The related woks about LFI coding.

With regard to LF compression, many solutions have been recently proposed. As shown in Fig. 3, these methods are usually classified into two categories: those that compress raw LFIs and those that encode SAIs extracted from raw LFIs. Moreover, with the improvement of codec, the developing redundancy elimination has been evolving the light field coding approaches. Because there is incoherence among MIs in raw LFIs, the high-efficiency video coding (HEVC) encoder is frequently used for such LFIs. The HEVC codec contains many flexible and novel prediction tools, and a high-order prediction model takes advantage of the HEVC encoder, as proposed in [20], wherein geometric transformations have up to an unprecedented eight degrees of freedom. However, MIs in raw LFIs are difficult to handle because of their enormous quantity and extraordinary low resolution. Some other compression structures have been proposed to improve the compression ratio for raw LFIs using the video codec [21, 22], but the numerous MIs restrict them to exploiting spatial correlation. Because raw LFIs with honeycomb arrangements limit the standard encoder [23], an MI reshaping method is proposed for raw LFIs in [24, 25] for the HEVC encoder. Nevertheless, all these compression solutions require the transmission of camera parameters for further applications, which increases the coding burden on the compressed data stream [26].

LF data formats and coding conditions can strongly impact compression performance, and there is much room for improvement in LFI compression. Another representation of LFIs is the dense-view SAI, which can be applied for autostereoscopic display and has very strong correlation. Consequently, an increasing number of researchers have focused on SAI compression. Two LF compression schemes that obtain a high compression ratio and quality, based on video coding techniques and disparity compensated prediction, respectively, have been proposed in [27]. As reported in [28], each SAI can be considered a random linear combination of angular viewpoints. Consequently, the method in [28] reconstructs all SAIs using only a few SAIs, thus removing the high correlation. The SAI streaming compression scheme proposed in [29] adopts two SAI scan orders (i.e., line scan mapping and rotation scan mapping) to improve compression efficiency. To find an optimal scan order, homography-based low-rank approximation is presented in [30, 31]; this method can adaptively adjust the scan order according to LFI content. Recently, the optimized SAI arrangement based on content correlation introduced in [32] achieved bitrate reduction. All the above approaches convert SAIs into a single-viewpoint pseudo-sequence, which is useful for coding using the video encoder; however, with the increase in the number of SAIs in LFIs, the compression efficiency decreases significantly.

Because there is a high cross-correlation among SAIs, both the horizontal and vertical disparities can be utilized to compress LFIs. Multiview video is transformed using SAIs, and the possibility for efficient SAI compression via multiview video coding (MVC) is verified in [33]. A similar compression algorithm is proposed in [34] to encode LFIs using MVC; this algorithm achieves higher compression performance than that achieved by converting SAIs into a single-view video. Because LFIs encoded by these compression methods are computer-generated, the perfect disparities make efficient inter-view prediction possible. A scheme based on a pseudo-sequence for LFI compression is designed in [35]. Specifically, this method decomposes raw LFIs into multiple SAIs, assigns each SAI a layer, and then encodes all SAIs using a symmetric 2D hierarchical scheme in accordance with the assigned layers. To further improve coding efficiency, the 2D hierarchical reference structure proposed in [36] supplies a motion vector scaling algorithm based on spatial coordinates to enhance the precision of inter-view prediction.

Although compressing SAIs using cross correlation can improve compression efficiency, it still has high computational complexity and requires high bitrates due to the huge number of SAIs in LFIs [37]. To handle this issue, a graph-based representation is presented in [38] to code the LF using geometric information to improve coding performance. However, this method requires extra colour information, which increases the coding burden. In addition, the scalable LF compression scheme described in [39] codes only a subset of SAIs and reconstructs all SAIs using a patch-based restoration method. Moreover, bi-level view compensation with rate-distortion optimization is reported in [40] that first classifies the entire SAIs as either key or non-key SAIs and then performs learning-based angular super-resolution to compensate the non-key SAIs. This approach exploits the intra- and inter-view relationships, but compensation using only super-resolution is unreliable. It should be noted that compression efficiency can be significantly improved using image geometry information, like disparity and depth map. However, because this algorithm cannot reduce the number of SAIs, the problem of high bitrates has not yet been fully solved.

Due to its depth-enhanced format, the advanced multiview video and depth (MVD) map structure can decrease bitrate and save coding time because only a few views and their corresponding depth maps are coded and transmitted while arbitrary views are rendered from the decoded data [41]. According to the literature, methods have been developed recently to compress SAIs with the depth map or disparity. Similar to [40], an LF codec with disparity-guided sparse coding is proposed in [26] that optimally selects several key structural SAIs; all SAIs are reconstructed from the key SAIs and the associated disparities. The depth-image-based rendering (DIBR) technique is employed in [42] to synthesize non-corner SAIs instead of coding them. Coding only the four corner SAIs decreases bitrate, but such corner SAIs have distortions that diminish the results of synthesis. The compression results of these approaches simultaneously verify their validity and confirm that the quality of the depth map or disparity is vital to such an MVD coding scheme. Therefore, an LFI-compression structure using depth estimation and view synthesis was proposed in our previous work. A prototype is proposed in [18], in which the depth map is computed only by horizontal EPI and optimized through crude error pixel repairing, so it generates the unsatisfactory depth map. To obtain a more accurate depth map, the depth estimation of the structure has been improved in [19]. Although the number of SAIs selected and coded is small, an intermediate SAI can still be approximated once the depth map can be precisely generated, which can contribute to low computational complexity and bitrate. Note that this article is based on our earlier work, as described in [18, 19], and the novelties of this article are generalized as follows.

(1) This paper estimates depth maps for SAIs according to the angular domain information of the LF. Based on an epipolar plane image (EPI), the pixel displacement in LFIs is fully exploited. A valid cost function is designed that can precisely determine the optimal slope of the epipolar line for each point.

(2) The slope computation procedure generates massive noise points in flat areas, which present a new challenge for depth optimization. Considering the strong relationship between an SAI and its corresponding depth map, an SAI-guided depth map optimization method is presented that can eliminate noise points according to the pixel variation of the associated SAI.

(3) To reduce computational complexity and bitrate while maintaining depth-estimation precision, an appropriate cost-computation step and a key SAI-selection step are determined based on statistical experiments. Because the pair of steps and depth estimation are embedded in the proposed method (PoS&DE), the coding and reconstruction process can be optimized.

(4) Advanced MVD is employed in the proposed compression structure to encode the selected key SAIs. In this way, unselected SAIs can be synthesized at the decoder using DIBR so that no extra information must be sent to the decoder. Additionally, the proposed PoS&DE confirms that the virtual view synthesis solution can significantly improve LFI-compression efficiency.

The rest of this paper is organized as follows. The 4D light-field model is analysed in section 3. Based on the analysis, the proposed PoS&DE is introduced comprehensively in section 4. Section 5 demonstrates and discusses the experimental results. Finally, section 6 concludes this paper.

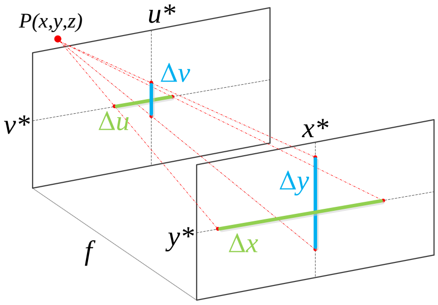


Fig. 4. The related woks about LFI coding.

3. 4D light-field model

From a geometric optics perspective, a light ray in space can be denoted by position (x, y, z), direction (θ, ϕ), wavelength λ and time t such that 7D plenoptic function $P(x, y, z, \theta, \phi, \lambda, t)$ is constructed as the light-field model to describe the set of light rays travelling in every direction through every point in the 3D space. Although the multidimensional function can represent light rays comprehensively, it is difficult to process by computer. Due to concision and intuition, 4D model $P(u, v, x, y)$ is the most common model in computational photography. In the 4D model, (u, v) denotes angular resolution and (x, y) denotes spatial resolution. A diagram of the 4D light-field model is illustrated in Fig. 4. As seen, the distance between the two space-specific planes is represented as f, Δu and Δv denote the geometrical distances between the two virtual cameras in the horizontal and vertical directions, respectively, and Δx and Δy are the distances the corresponding scene points in the image moved in the horizontal and vertical directions, respectively. Therefore, the depth value of each point in every SAI can be derived based on the expression, $Z = \frac{f}{r}(r = \frac{\Delta u}{\Delta x} or \frac{\Delta v}{\Delta y})$ [2].

4. The proposed SAI-compression method

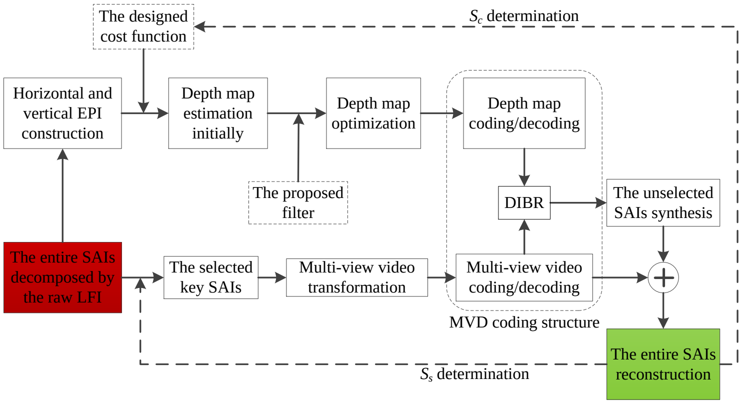

The designed cost function
S_c determination
Horizontal and vertical EPI construction
Depth map estimation initially
Depth map optimization
Depth map coding/decoding
The proposed filter
DIBR
The unselected SAIs synthesis
The entire SAIs decomposed by the raw LFI
The selected key SAIs
Multi-view video transformation
Multi-view video coding/decoding
MVD coding structure
S_s determination
The entire SAIs reconstruction

Fig. 5. The flow chart of the proposed compression method.

Due to a strong relationship, SAIs have been a prevalent topic of study for LF compression. In this section, we propose an improvement to SAI-compression performance based on view synthesis. To reduce computational complexity with satisfactory reconstruction performance, we find a reasonable pair of steps to generate the depth map and encode a small number of selected SAIs. The proposed compression approach encodes only the selected SAIs; the unselected SAIs will be synthesized using DIBR. Meanwhile, the MVD coding scheme is used in our proposed method to fully exploit inter-view (inter-SAI) and inter-component correlations. A flowchart of the proposed compression method is shown in Fig. 5.

4.1. Depth estimation using an epipolar plane image

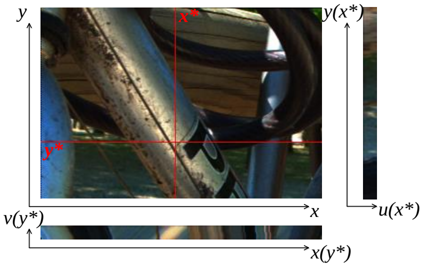


Fig. 6. The flow chart of the proposed compression method.

According to the depth computation expression mentioned in section 3, ratio r plays a crucial role in the depth-estimation process. To obtain ratio r, we use an EPI, which is a famous concept in the field of multiview computer vision [43]. By gathering SAI pixels at a specific spatial coordinate x (or y) with a fixed angular coordinate u (or v), a horizontal (or vertical) EPI can be generated, as illustrated in Fig. 6.

The epipolar lines in each EPI reveal the spatial structure of a 3D scene [11]. More specifically, the slope can be used to deduce the depth value of an associated point in each SAI using a reformulated equation, as follows:

$$Z_{(i)} = \frac{f}{\tan \alpha_{(i)}}. \tag{1}$$

where i is an enumeration value that indicates horizontal or vertical,$\alpha_{(i)}$ specifies the slope of a specific point in the EPI and $\tan \alpha = r$. Therefore, the process to obtain an accurate depth value is to precisely compute all slopes in the horizontal and vertical EPIs.

4.1.1. The optimal slope decision

Using the analysis described above, it is challenging to obtain the depth value that achieves the best slope decision for each epipolar line in an EPI. To this end, the optimal slope decision model, Eq. 2, is constructed to find the best slope from a candidate set for each point. Because it is assumed that points on an epipolar line have pixel values that are nearly the same, during the matching process, the slope of the point with the lowest cost is selected as the optimal slope.

$$\alpha^*_{(i)}(p) = \underset{\alpha \in (-45°, 45°)}{\arg\min} \; C(\alpha_{(i)}, p). \tag{2}$$

Here, $C(\alpha_{(i)}, p)$ denotes the cost of slope α at point $p = (x(y^*), v(y^*))$ (in a horizontal EPI) or point $p = (u(x^*), y(x^*))$ (in a vertical EPI). The tested $\alpha_{(i)}$ is selected from the slope candidate set. $\alpha^*_{(i)}(p)$ indicates the determined best slope for point p in a specific EPI. Since the disparity between two adjacent SAIs is rather tiny in terms of the LFI captured by the plenoptic camera, the range of candidate slopes is restricted from $-45°$ to $45°$. Furthermore, the computational process of $C(\alpha_{(i)}, p)$ is demonstrated as follows.

$$C(\alpha_{(i)}, p) = \frac{\sum_N \omega_n (V_p - V_{\Delta p}^{(\alpha_{(i)})})^2}{\sum_N \omega_n}. \tag{3}$$

For simplicity, V_p denotes the luma value of current point p, and $V_{\Delta p}^{(\alpha_{(i)})}$ represents the luma value of the other point on the epipolar line with candidate slope $\alpha_{(i)}$. In addition, ω_n is a flag that is assigned 0 if the tested point Δp is invalid and is otherwise assigned 1. However, all points on the epipolar line are calculated exhaustively for matching, which is time-consuming. Therefore, a proper computation step should be determined to balance the computational complexity and estimated accuracy, which is analysed in detail in section 5.

By using the proposed model based on the horizontal and vertical EPIs, two preliminarily depth maps are produced for each SAI. Because these two depth maps have the same weight, each depth map is merged by the weighted average of the two depth maps. To be applicable for the MVD coding scheme and maintain the linear structure of the estimated depth map, min-max normalization is adopted in this paper to restrict the depth value to the range from 0 to 255.

4.1.2. SAI-guided depth optimization

Homogenous regions in the EPI that have massive pixels with almost the same value lead to miscalculations in the depth map. There is no dividing line for homogenous areas, and so the value of the computed least cost for one point may be over 2. The proposed cost function fails to select the optimal slope of these points, which are marked as noise points. In this case, the noise points need to be eliminated to optimize the depth map.

Many depth optimization algorithms have been proposed in the literature, including the interpolation method [44], Markov random field (MRF) propagation [8], locally linear embedding (LLE) [2] and sub-pixel-based matching (SPM) [45]. However, these algorithms refine the depth map based on its intrinsic characteristics, which is not suitable for viewpoint images with narrow baselines. Therefore, the correlation between the depth map and its associated texture is considered in the proposed optimization approach, and SAIs are regarded as the reference to refine the estimated depth map.

In general, if points are located on the boundary of the depth map, their corresponding points in the texture image are also located on the boundary. For points astride the boundary in the texture image, their associated points in the depth map may have distinguished depth values. Meanwhile, in the texture image, the homogenous regions contain numerous similar pixels with co-located points in the depth map that vary regularly. Based on these observations, the noise points in the initial depth map can be eliminated according to the variation tendency of their corresponding points in the texture image. To this end, a weighted mean filter is designed in this work that can refine the initial depth map via the generated parameters.

$$Z_e = \frac{1}{|P_e|} \sum_{r \in P_e} \gamma_r Z_r. \tag{4}$$

To repair error depth value Z_e, the proposed weighted mean filter uses a patch. In Eq. 4, P_e is a 3×3 patch centred on a specific noise point, $|P_e|$ represents the cardinality of P_e, Z_r denotes the depth value of the neighbouring points in patch P_e (excluding the noise point), and γ_r implies that the tendency coefficient reflects the pixel variation tendency of the associated SAI. The tendency coefficient can be automatically obtained via Eq. 5:

$$\begin{aligned} &\gamma_r = \frac{V_e}{V_r}, (r \in P_e) \\ &s.t. \begin{cases} 1 - \varepsilon \leqslant |\gamma_r| \leqslant 1 + \varepsilon \\ otherwise, \gamma_r = 0 \end{cases} \end{aligned} \tag{5}$$

where V_e indicates the luma value of the pixel in the SAI corresponding to the noise point in the depth map, and V_r is the luma value of its adjacent point. In this way, tendency coefficient γ_r can be assigned according to the variation tendency of the associated SAI. For validity, the tendency coefficient must be constrained, and ε denotes the offset to control smoothness; the larger this offset, the stronger the smoothness. To balance the smoothness, offset ε is empirically set to 0.012 in this paper. Because SAIs are regarded as the reference images to guide depth optimization, noise points can be removed based primarily on the characteristics of an SAI. Meanwhile, the final depth map can maintain good consistency.

4.2. LF reconstruction and compression with depth maps

The proposed method focuses on compressing LF images captured by a Lytro camera, which is the most frequently used plenoptic camera on the consumer market. With the compact arrangement of the micro-lens, the camera balances the resolution and view number well. However, the increase in resolution and view number for 4D LF requires a much larger storage capacity and transmission bandwidth than a conventional 2D image. Therefore, LF data compression has become a great challenge for its further application. As mentioned in section 2, the estimated depth map is coded with the MVD coding structure from the proposed approach, which can fully exploit the inter-view (inter-SAI) and the inter-component correlations within LFIs.

4.2.1. Pair of steps

Because the MVD coding structure requires only a small number of SAIs and their corresponding depth maps, it can essentially remove the burden associated with coding an LFI. Although reliable depth map estimation fails if there are too few SAIs due to its negative impact on the quality of the synthesized SAIs, having too many SAIs increases data volume, which increases computational complexity and obstructs the usage of the LF. Consequently, this work requires selecting an appropriate key SAI solution to balance the quality of the estimated depth map and the bitrates of the selected SAIs. Additionally, the exhaustive testing for cost computation will achieve a high depth-estimation performance and significantly improve computational complexity. Conversely, sparse testing can save estimating time and decrease depth quality. Furthermore, key SAI selection and computation will jointly impact the reconstruction performance. In the proposed scheme, key SAIs are selected with a proper step, converted into pseudo-multiview sequences and encoded by the 3D extension of the HEVC coding structure. Therefore, a detailed analysis for the pair of steps, including the key SAI-selection and cost-computation steps, are as demonstrated in section 5.

4.2.2. LF coding using MVD

To maximize coding and reconstruction performance, the MVD coding scheme, which can exploit the inter-view (SAI) and inter-component correlations, is used in this work. The advanced MVD structure can encode a small number of texture videos composed by the key SAIs and their corresponding depth maps; then, the coded bitstream packets are multiplexed into a 3D video bitstream [41]. On the decoder side, the view synthesis algorithm proposed in [46] can be employed to render missing SAIs using the decoded depth map and the corresponding key SAIs, and the entire LFI can be reconstructed.

A raw LFI cannot be directly used for display, but after decomposing, the numerous SAIs generated with slightly different disparities can provide an immersive 3D experience. Augmenting SAIs will dramatically increase the size of the LF data. The most conventional solution to handle this issue is to transform all SAIs into a single-viewpoint pseudo-sequence; this coding approach codes the pseudo-sequence using a video encoder. However, the coding for such huge data is complicated and time consuming. To this end, this work converts the key SAIs into multiview pseudo-sequences and encodes them with their associated depth maps using the MVD structure. Note that the appropriate number and order of views will be discussed in section 5.

5. Experimental results

Extensive experiments, including PoS&DE, are conducted in this section to evaluate the proposed compression method. The LF dataset from the ICME 2016 Grand Challenge [47], which contains 12 LFIs captured by the Lytro camera at various scenes, is used in the simulated experiment of the assessment. Two of the 12 LFIs, Color_Chart_1 and ISO_Chart_12, are mainly used for calibration, and their depth varies slightly. Therefore, they are not applied for PoS&DE. The SAI has been decomposed from the raw LFI (with dimensions of 15 × 15 × 624 × 432) using the MATLAB Light Field Toolbox 0.4 [4]. Table 1 shows the names of the 10 selected LFIs; Fig. 8(a) shows the centre SAIs of those LFIs.

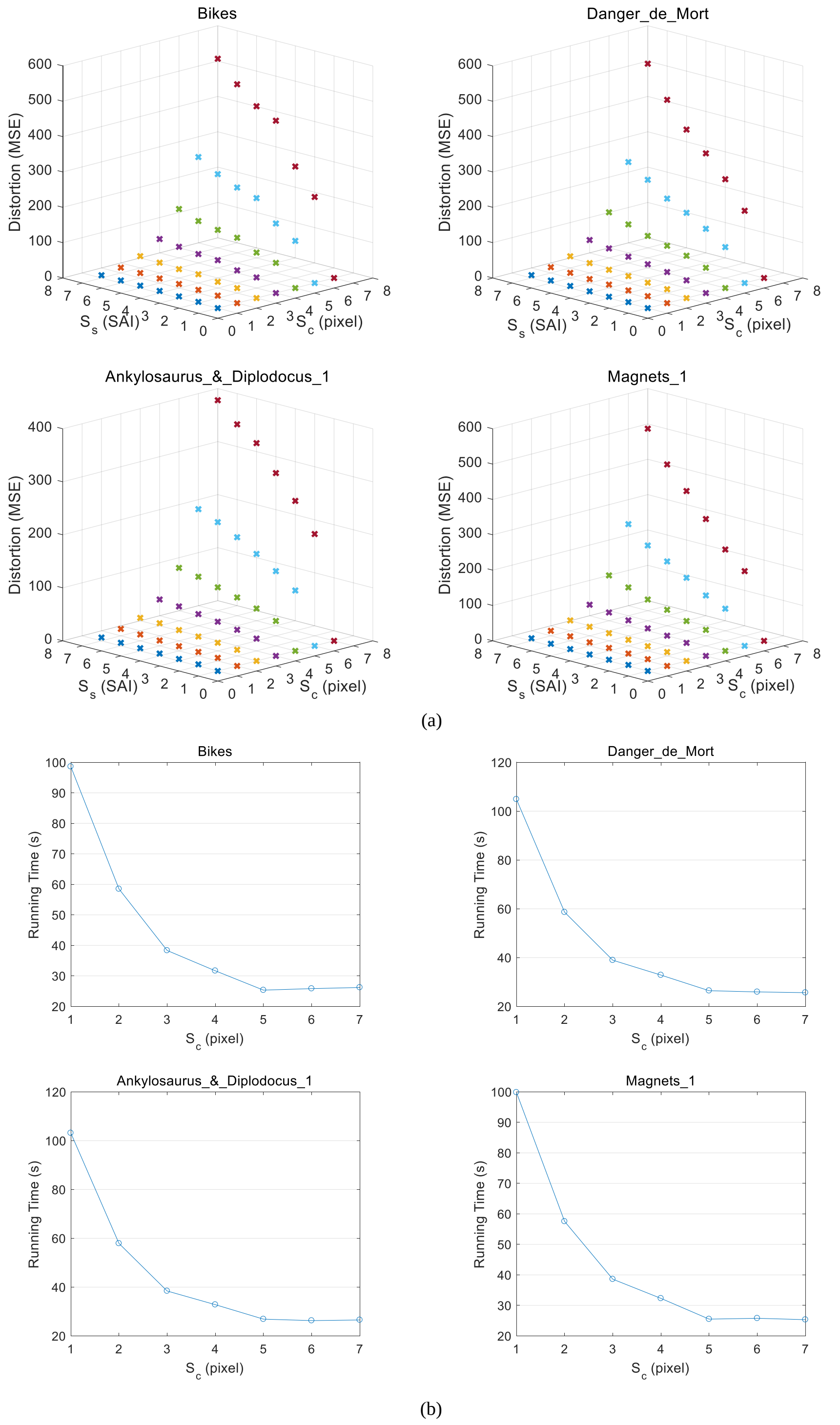
Table 1. Image Test Sequence

ID	Image name	ID	Image name
I01	I01_Bikes	I06	I06_Ankylosaurus_&_Diplodocus_1
I02	I02_Danger_de_Mort	I07	I07_Desktop
I03	I03_Flowers	I08	I08_Magnets_1
I04	I04_Stone_Pillars_Outside	I09	I09_Fountain_&_Vincent_2
I05	I05_Vespa	I10	I10_Friends_1

5.1. Pair of steps determination

The reasonable selection step and computation step play vital roles for time estimation and reconstruction performance. To determine this pair of steps, statistical experiments have been conducted. The experiments investigate the impact of key SAIs selected on the coded bitrates and synthesis quality for four representative LFIs: Bikes, Danger_de_Mort, Ankylosaurus_&_Diplodocus_1 and Magnets_1. Bikes and Danger_de_Mort contain complex textural features captured in natural scenes, while Ankylosaurus_&_Diplodocus_1 and Magnets_1 contain large homogenous regions.

The statistical results are demonstrated in Fig. 7. Note that, because the depth maps are used only to synthesize missing SAIs as non-visual data, the distortion of the reconstructed LFs rendered using the SAIs and depth maps is measured by the mean square error (MSE). The computation step (S_c) indicates the step length to compute the costs for matching epipolar line slopes. In addition, the selection step (S_s) selects the disparity between any two adjacent viewpoint sequences. The pair of steps (S_c, S_s) influences reconstruction distortion. As seen from Fig. 7(a), with the increase of both the computation and selection steps, the reconstructed SAI distortion severely increases. When S_c is fixed, the performance of the estimated depth map is changeless for a specific light field image. The disparity is the unique factor that impacts on reconstruction distortion. The reference SAIs used for synthesis are sampled along vertical direction, so only the horizontal disparity needs to be taken into consideration. Once the depth



(a)

(b)

Fig. 7. The statistical results. (a) The distribution of distortion under different (S_c, S_s); (b) The average estimating time for one SAI under different computation steps.

map is produced, the intermediate SAI synthesis is regarded as the horizontal pixel displacement of the reference SAI according to the depth value. Consequently, with the linear increase of S_s, the error propagation from pixel displacement causes the linear reconstruction distortion increase. Moreover, the every depth point value implies the displacement of the reference SAI during view synthesis. The depth map precision is subject to the size of S_c. The reconstruction via an outstanding depth map is robust whether the disparity is large or not, so the distortion variation under small S_c changes slowly. The distortion variation is very sensitive to the disparity if the estimated depth map contains much error points. Therefore, the distortion slope variation is different for each S_c as shown in Fig. 7(a). When S_s is set to 1, all SAIs are transformed into pseudo-sequences. Because it is unnecessary to synthesize missing SAIs using the depth map, there is no reconstructed SAI evaluation. An increase in S_s leads to a large disparity between the pseudo-sequences composed by the key SAIs, and a larger S_c decreases the quality of the estimated depth map. Therefore, reconstructed SAI distortion is severe and cannot be applied for practical usage in this case. Considering feasibility, reconstructed SAI distortion should be under 30. Within this restriction, the pair of steps (S_c, S_s) should be fewer than $(4, 4)$. Due to the special structure of LFIs, the centre SAI has the most representative features, so it must be contained in the base view sequence in order to maintain the high performance for coding and synthesis. With regard to the 15×15 angular resolution, the number of pseudo-sequences converted by SAIs is 15 at most. Within this restriction, as S_s increases from 1 to 7, the number of pseudo-sequence reduces gradually; it will be fixed to 3 if Ss is more than 3. Additionally, Fig. 7(b) shows the average running time for one depth map estimation under varied S_c. The S_c decides the number of the points to be matched in EPI slope decision. A small S_c signifies more points to be computed, resulting in more running time for depth estimation. On the contrary, a larger S_c reduces the sampling rate of computed points, which saves estimating time. A close relationship is found when running time is compared with S_c, while the image content, however, is irrelevant. As a result, four images with different contents have almost the same time consumption under the same S_c. An over-sparse calculation produces a low-quality depth map, which significantly damages the reconstructed SAI. Based on the aforementioned analysis and the computational complexity statistics, the pair of steps (S_c, S_s) is set to $(3, 4)$.

5.2. Depth-estimation comparison

In this section, typical methods, line-assisted graph cut (LAGC) [9] and accurate depth map estimation (ADME) [10], are considered anchors to compare the estimated depth map with the proposed depth-estimation algorithm. Figure 8 shows the experimental results for 10 LFIs from the LF dataset. Because LAGC estimates the depth map based on stereo-pair matching, it fails to produce an available depth for an LFI with a narrow baseline. It is difficult to match stereo pairs in LFIs with low contrast, and so LAGC cannot distinguish the objects in such LFIs. In comparison with LAGC, depth-estimation method ADME performs better. More specifically, the depth maps generated by ADME are more continuous than those generated by LAGC. Moreover, much of the depth information in various LFIs can be exploited by ADME. The proposed depth-estimation method precisely computes the slope of every epipolar line in the horizontal and vertical EPIs and employs an efficient depth optimization with the proposed SAI-guided algorithm. Therefore, the proposed algorithm can comprehensively generate a continuous depth map and preserve much of the detailed information in the complex and background regions.

5.3. Reconstruction evaluation

To further assess the estimated depth map, the objective and subjective quality of the intermediate SAI is shown in Fig. 9 and Fig. 10, respectively, in which the SAI is synthesized by the algorithm in [46]. The average structural similarity index (SSIM) and multi-scale (MS)-SSIM of the synthesized SAIs are computed as depicted in Fig. 9. Note that LAGC fails to produce a feasible

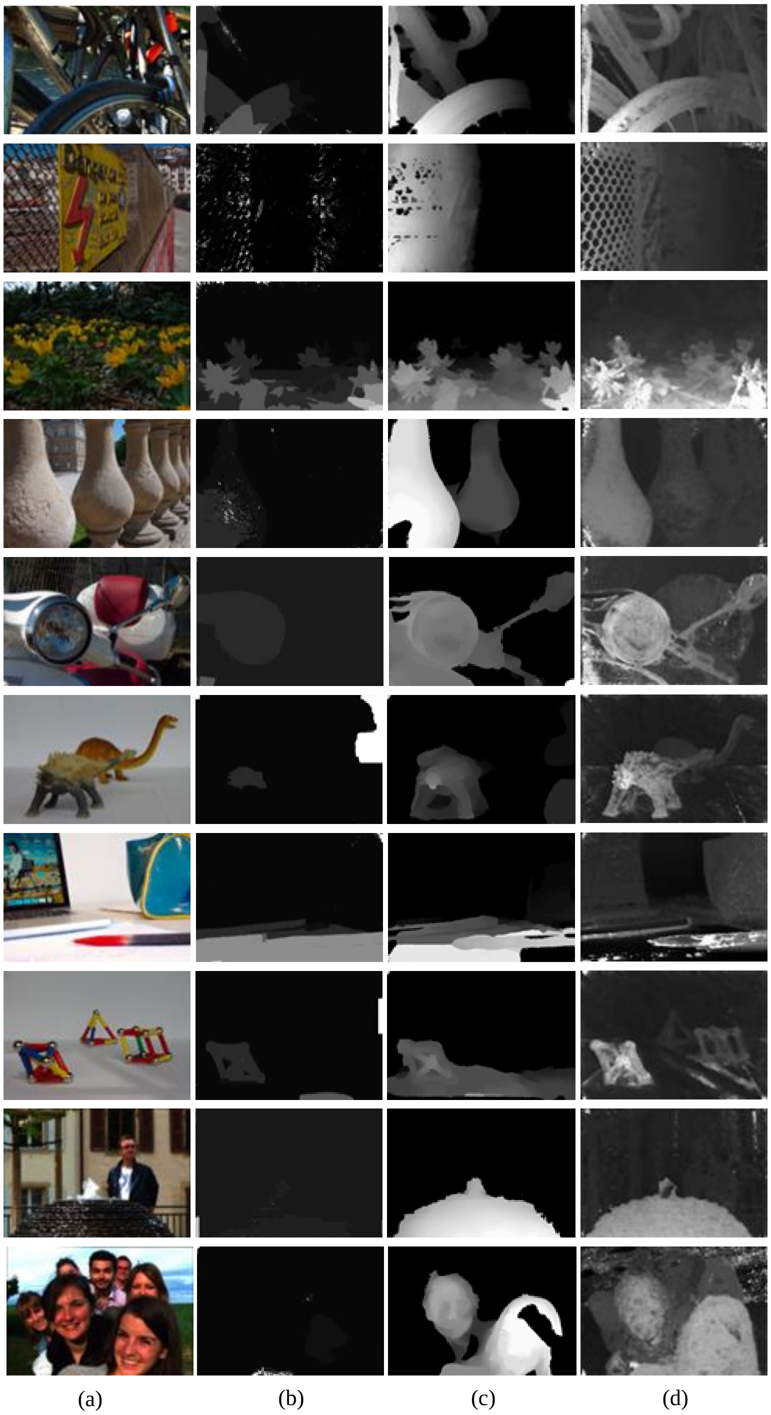


Fig. 8. Experiments on the partial LFIs from Table 1. Top to bottom: I01 to I10. (a) center SAI; (b) depth map by LAGC; (c) depth map by ADME; (d) depth map by the proposed method.

depth map, and so it does not appear in Fig. 9. In general, our proposed method outperforms ADME because it preserves more information than ADME. More specifically, most of the SAIs produced with our depth map can achieve 0.83 in SSIM and 0.9 in MS-SSIM. With regard to I03, too much coding distortion and complex textural information in the associated SAI result in poor synthesis. However, as illustrated in Fig. 10, only the edge of the flowers in the front of Flowers is blurred, which is negligible. Similarly, the wheel in Bikes has some distortion. I06 and I08 contain very sparse foreground objects in the studio with large homogeneous areas, and so they are insensitive to the depth map and their quality can reach 0.95 in SSIM and MS-SSIM. Although there is some noise in their estimated depth maps, the synthesized SAI is almost same as the original SAI. In Fig. 10, both the decoded and the synthesized intermediate SAIs of Ankylosaurus_&_Diplodocus_1 and Magnets_1 look approximately like the original SAI. Consequently, the proposed depth-estimation algorithm can generate a refined depth map for an LFI with a narrow disparity compared with other state-of-the-art approaches.

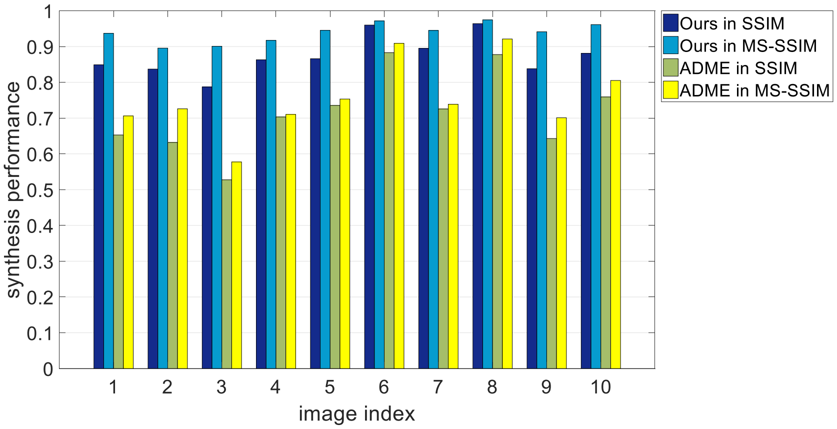


Fig. 9. The synthesis performance of the LFIs from Table 1. The horizontal axis represents the image index and the vertical axis implies the synthesized SAI quality in SSIM.

5.4. Coding performance

As mentioned in section 5.1, the selected step is set to 4, and so a 3-viewpoint pseudo-sequence is developed from partial SAIs. In Fig. 11, the red arrow denotes the base viewpoint and its scan order, and the two neighbouring blue arrows represent the non-base viewpoints and their scan order. The border SAIs (in white) suffer severe geometric distortion and blurring, resulting in less value for usage. Therefore, they are not coded in the proposed PoS&DE. The SAIs in red represent the missing SAIs that can be synthesized using the decoded multiview sequence and the associated depth map using the view synthesis method in [46]. Consequently, coding performance is jointly measured using average peak signal to noise ratio (PSNR) for the $13 \times 13 = 169$ viewpoint images (including the $13 \times 3 = 39$ coded viewpoint images and the 130 synthesized viewpoint images, as shown in Eq. 6) and the total bitrate (viewpoints and depth maps). Coding performance is jointly measured by the PSNR and the bitrate change, which are calculated by the Bjontegaard delta PSNR (BD-PSNR) and Bjontegaard delta bitrate (BD-BR), respectively [48].

$$PSNR_{avg} = \frac{1}{169}(\sum_{39} PSNR_{coded}^{(u,v)} + \sum_{130} PSNR_{syn}^{(u,v)}). \tag{6}$$

Table 2 compares the results for SC-SKV and the proposed LF compression method, PoS&DE,

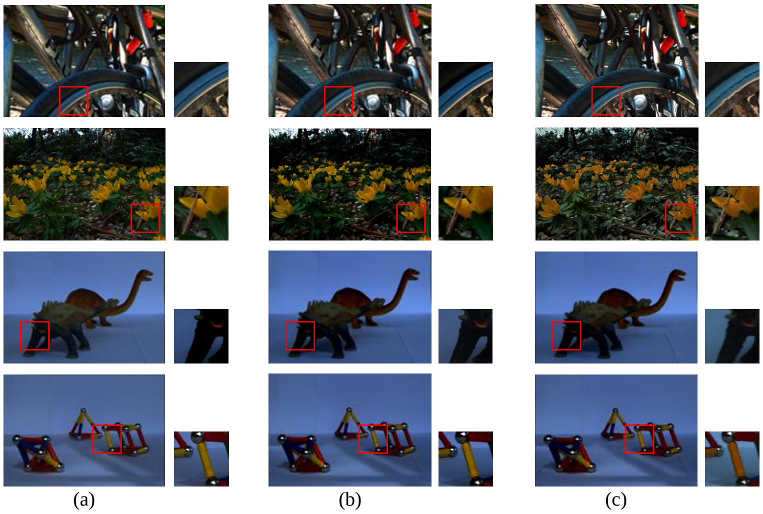


Fig. 10. The perceptual quality comparison for four SAIs. Top to bottom: Bikes, Flowers, Ankylosaurus_&_Diplodocus_1 and Magnets_1. (a) the original intermediate SAI; (b) the decoded intermediate SAI with QP=22; (c) the synthesized intermediate SAI using our produced depth map.

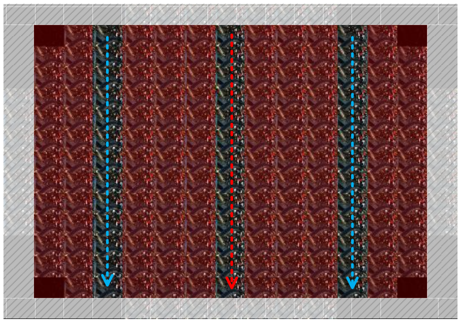


Fig. 11. The diagram of multi-view videos selection and their transformation from SAIs.

with 7- and 3-viewpoint pseudo-sequences. The HEVC codec under the “Low-Delay P-main" configuration is regarded as the benchmark, and the SAIs are arranged as pseudo-sequences to be encoded. Moreover, a lower BD-BR implies that the bitrates are saved more significantly, and a higher BD-PSNR denotes a higher reconstruction quality. SC-SKV uses disparity to guide non-key SAI coding, which reduces the residual of the non-key SAIs. Thus, SC-SKV can achieve low bitrates compared with the HEVC codec, as shown in Table 2 (the bitrate reduction is 51% and the PSNR increase is 1.76 dB, on average). However, due to the advanced 3D-HEVC structure, PoS&DE is more robust than SC-SKV. Because PoS&DE requires only a small number of SAIs and their corresponding depth maps, much higher bitrates can be saved. Furthermore, an average 4.58-dB PSNR gain and an 86.1% bitrate reduction can be obtained by PoS&DE with 3 viewpoints over PoS&DE with 7 viewpoints (2.35 dB and 64.3% for PSNR gain and bitrate reduction, respectively). Because 4-viewpoint sequences are omitted, PoS&DE with 3 viewpoints outperforms that with 7 viewpoints in terms of bitrate reduction. As analysed previously, there is no clear difference between the 2- and 4-key SAI-selection steps. Thus, under the same bitrate reduction, PoS&DE with 3 viewpoints achieves better reconstruction performance.

Table 2. Coding Results Comparison.

LFI name	SC-SKV		Ours(7-viewpoint)		Ours(3-viewpoint)	
	BD-PSNR	BD-BR	BD-PSNR	BD-BR	BD-PSNR	BD-BR
I01	2.27 dB	-61.6%	2.85 dB	-63.5%	5.96 dB	-87.1%
I02	2.32 dB	-64.8%	2.76 dB	-64.9%	6.07 dB	-88.0%
I03	2.54 dB	-70.1%	2.79 dB	-67.7%	5.53 dB	-87.8%
I04	2.31 dB	-76.0%	2.36 dB	-68.5%	3.61 dB	-86.8%
I05	1.59 dB	-58.3%	1.76 dB	-56.7%	3.27 dB	-82.6%
I06	0.73 dB	-13.9%	1.84 dB	-66.3%	3.35 dB	-85.8%
I07	1.32 dB	-46.5%	2.14 dB	-63.2%	4.36 dB	-85.6%
I08	-0.13 dB	2.9%	2.06 dB	-65.4%	3.67 dB	-84.8%
I09	3.41 dB	-89.5%	2.31 dB	-59.4%	5.28 dB	-86.4%
I10	1.28 dB	-32.5%	2.63 dB	-66.9%	4.68 dB	-86.1%
Average	1.76 dB	-51.0%	2.35 dB	-64.3%	4.58 dB	-86.1%

To evaluate coding efficiency intuitively, the typical LF compression solution, the multiple view structure (MVS)[33], is added for comparison, as illustrated in Fig. 12. As seen, in general, PoS&DE consumes fewer bits than the other two methods. Only for some manually constructed scenes, such as I06, I07 and I08, PoS&DE can achieve nearly the same reconstruction performance as the other two methods at low bitrates. We believe a better reconstruction solution will improve the overall performance, which will be investigated in our future work. Additionally, for the high-reconstruction case, PoS&DE can greatly decrease bits, especially for 3 viewpoints. Therefore, these results confirm that PoS&DE can maintain acceptable quality while drastically improving compression performance.

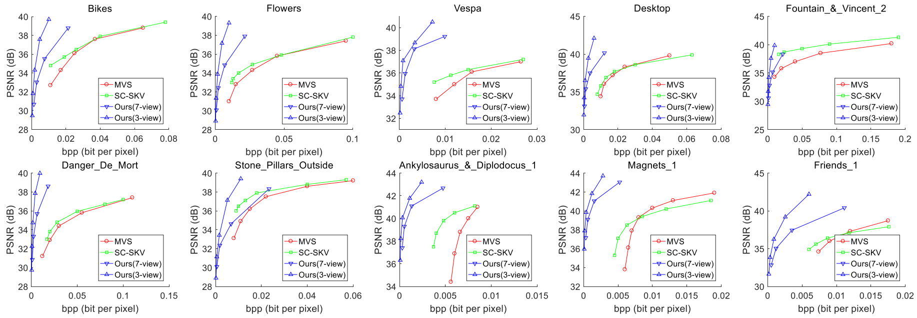


Fig. 12. The rate-distortion curves are demonstrated for MVS, SC-SKV and PoS&DE with 7 and 3 viewpoints sequences. The horizontal axis indicates the bit per pixel (bpp), and the vertical axis indicates the reconstruction PSNR (dB).

6. Conclusion

In this paper, we propose a LF compression solution that significantly reduces coding bitrate by coding a subset of selected SAIs with their corresponding depth map, synthesizing the remaining unselected SAIs and ultimately reconstructing all SAIs. By computing the cost of each point on the epipolar line in the horizontal and vertical EPI, an initial depth map can be obtained. SAI-guided depth optimization is designed to refine the noise in the initial depth map in accordance with the pixel variation in the associated SAI. The MVD coding structure is adopted to fully exploit the relationships among SAIs. Based on statistical experimentation, depth estimation is accelerated using an optimal cost-computation step. Meanwhile, the key SAI-selection step is determined to decrease the number of coded SAIs. Experiments have been conducted on the benchmark LF dataset and the results demonstrate that the proposed PoS&DE LF compression scheme can generate an accurate depth map to improve coding performance compared to other LF compression methods. This method results in a 4.58-dB PSNR gain and an 86.1% bitrate reduction, on average, over the HEVC codec.

Funding

National Natural Science Foundation of China (NSFC) (61571285, 61828105); Project of Science and Technology Commission of Shanghai Municipality (18XD1423900); National Natural Science Foundation of China (NSFC) (61801006).

Acknowledgments

The authors thank the reviewers for their helpful comments.

References

1. I. Ihrke, J. Restrepo, and L. Mignard-Debise, “Principles of Light Field Imaging: Briefly revisiting 25 years of research," IEEE Signal Processing Magazine **33**, 59-69 (2016).
2. G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, “Light Field Image Processing: An Overview," IEEE Journal of Selected Topics in Signal Processing **11**, 926-954 (2017).
3. N. Ren, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, “Light Field Photography with a Hand-Held Plenopic Camera," Computer Science Technical Report CSTR (2005).
4. D. G. Dansereau, O. Pizarro, and S. B. Williams, “Decoding, Calibration and Rectification for Lenselet-Based Plenoptic Cameras," in 2013 IEEE Conference on Computer Vision and Pattern Recognition, 1027-1034 (2013).
5. X. Kaikai, “Monolithically integrated Si gate-controlled light-emitting device: science and properties," Journal of Optics **20**, 024014 (2018).
6. M. Hog, N. Sabater, and C. Guillemot, “Superrays for Efficient Light Field Processing," IEEE Journal of Selected Topics in Signal Processing **11**, 1187-1199 (2017).
7. N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, “Saliency Detection on Light Field," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2806-2813(2014).
8. M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, “Depth from Combining Defocus and Correspondence Using Light-Field Cameras," in 2013 IEEE International Conference on Computer Vision, 673-680 (2013).
9. S. Wanner and B. Goldluecke, “Globally consistent depth labeling of 4D light fields," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 41-48 (2012).
10. Z. Yu, X. Guo, H. Ling, A. Lumsdaine, and J. Yu, “Line Assisted Light Field Triangulation and Stereo Matching," in 2013 IEEE International Conference on Computer Vision, 2792-2799 (2013).
11. H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, and I. S. Kweon, “Accurate depth map estimation from a lenslet light field camera," in 2015 IEEE Conference on Computer Vision and Pattern Recognition, 1547-1555 (2015).
12. B. M. Smith, L. Zhang, H. Jin, and A. Agarwala, “Light field video stabilization," in 2009 IEEE 12th International Conference on Computer Vision, 341-348 (2009).
13. Y. Bok, H. G. Jeon, and I. S. Kweon, “Geometric Calibration of Micro-Lens-Based Light Field Cameras Using Line Features," IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 287-300 (2017).
14. M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, “Light field microscopy," ACM Trans. Graph. **25**, 924-934 (2006).
15. X. Lin, J. Wu, G. Zheng, and Q. Dai, “Camera array based light field microscopy," Biomed. Opt. Express **6**, 3179-3189 (2015).
16. R. Prevedel, Y.-G. Yoon, M. Hoffmann, N. Pak, G. Wetzstein, S. Kato, T. SchrÃűdel, R. Raskar, M. Zimmer, E. S. Boyden, and A. Vaziri, “Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy," Nature Methods **11**, 727 (2014).
17. F. Pereira and E. A. B. d. Silva, “Efficient plenoptic imaging representation: Why do we need it?" in 2016 IEEE International Conference on Multimedia and Expo, 1-6 (2016).
18. X. Huang, P. An, L. Shen, and K. Li, “Light Field Image Compression Scheme Based on MVD Coding Standard," in The Pacific-Rim Conference on Multimedia, 1-10 (2017).

19. X. Huang, P. An, L. Shan, R. Ma, and L. Shen, “View Synthesis for Light Field Coding Using Depth Estimation," in 2018 IEEE International Conference on Multimedia and Expo, 1-6 (2018).
20. R. J. Monteiro, P. Nunes, N. Rodrigues, and S. M. M. d. Faria, “Light Field Image Coding using High Order Intra Block Prediction," IEEE Journal of Selected Topics in Signal Processing **11**, 1120-1131 (2017).
21. C. Conti, L. D. Soares, and P. Nunes, “HEVC-based 3D holoscopic video coding using self-similarity compensated prediction," Signal Processing: Image Communication **42**, 59-78 (2016).
22. Y. Li, M. SjÃűstrÃűm, R. Olsson, and U. Jennehag, “Scalable Coding of Plenoptic Images by Using a Sparse Set and Disparities," IEEE Transactions on Image Processing **25**, 80-91 (2016).
23. C. Perra and P. Assuncao, “High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in 2016 IEEE International Conference on Multimedia & Expo Workshops, 1-4 (2016).
24. H. Han, X. Jin, and Q. Dai, “Lenslet image compression based on image reshaping and macro-pixel Intra prediction," in 2017 IEEE International Conference on Multimedia and Expo, 1177-1182 (2017).
25. X. Jin, H. Han, and Q. Dai, “Image Reshaping for Efficient Compression of Plenoptic Content," IEEE Journal of Selected Topics in Signal Processing **11**, 1173-1186 (2017).
26. J. Chen, J. Hou, and L. P. Chau, “Light Field Compression With Disparity-Guided Sparse Coding Based on Structural Key Views," IEEE Transactions on Image Processing **27**, 314-324 (2018).
27. M. Magnor and B. Girod, “Data compression for light-field rendering," IEEE Transactions on Circuits and Systems for Video Technology **10**, 338-343 (2000).
28. S. D. Babacan, R. Ansorge, M. Luessi, P. R. Mataran, R. Molina, and A. K. Katsaggelos, “Compressive Light Field Sensing," IEEE Transactions on Image Processing **21**, 4746-4757 (2012).
29. F. Dai, J. Zhang, Y. Ma, and Y. Zhang, “Lenselet image compression scheme based on subaperture images streaming," in 2015 IEEE International Conference on Image Processing, 4733-4737 (2015).
30. X. Jiang, M. L. Pendu, R. A. Farrugia, S. S. Hemami, and C. Guillemot, “Homography-based low rank approximation of light fields for compression," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, 1313-1317 (2017).
31. X. Jiang, M. L. Pendu, R. A. Farrugia, and C. Guillemot, “Light Field Compression With Homography-Based Low-Rank Approximation," IEEE Journal of Selected Topics in Signal Processing **11**, 1132-1145 (2017).
32. C. Jia, Y. Yang, X. Zhang, X. Zhang, S. Wang, S. Wang, and S. Ma, “Optimized inter-view prediction based light field image compression with adaptive reconstruction," in 2017 IEEE International Conference on Image Processing, 4572-4576 (2017).
33. R. Olsson, M. Sjostrom, and Y. Xu, “A Combined Pre-Processing and H.264-Compression Scheme for 3D Integral Images," in 2006 International Conference on Image Processing, 513-516 (2006).
34. S. Shi, P. Gioia, and G. Madec, “Efficient compression method for integral images using multi-view video coding," in 2011 18th IEEE International Conference on Image Processing, 137-140 (2011).
35. D. Liu, L. Wang, L. Li, X. Zhiwei, W. Feng, and Z. Wenjun, “Pseudo-sequence-based light field image compression," in 2016 IEEE International Conference on Multimedia & Expo Workshops, 1-4 (2016).
36. L. Li, Z. Li, B. Li, D. Liu, and H. Li, “Pseudo Sequence Based 2-D Hierarchical Coding Structure for Light-Field Image Compression," in 2017 Data Compression Conference, 131-140 (2017).
37. K. Muller, P. Merkle, and T. Wiegand, “3-D Video Representation Using Depth Maps," Proceedings of the IEEE **99**, 643-656 (2011).
38. X. Su, M. Rizkallah, T. Maugey, and C. Guillemot, “Graph-based light fields representation and coding using geometry information," in 2017 IEEE International Conference on Image Processing, 4023-4027 (2017).
39. F. Hawary, C. Guillemot, D. Thoreau, and G. Boisson, “Scalable light field compression scheme using sparse reconstruction and restoration," in 2017 IEEE International Conference on Image Processing, 3250-3254 (2017).
40. J. Hou, J. Chen, and L. P. Chau, “Light Field Image Compression Based on Bi-Level View Compensation with Rate-Distortion Optimization," IEEE Transactions on Circuits and Systems for Video Technology **PP**, 1 (2018).
41. K. MÃijller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F. H. Rhee, G. Tech, M. Winken, and T. Wiegand, “3D High-Efficiency Video Coding for Multi-View Video and Depth Data," IEEE Transactions on Image Processing **22**, 3366-3378 (2013).
42. X. Jiang, M. L. Pendu, and C. Guillemot, “Light field compression using depth image based view synthesis," in IEEE International Conference on Multimedia & Expo Workshops, 19-24 (2017).
43. R. C. Bolles, H. H. Baker, and D. H. Marimont, “Epipolar-plane image analysis: An approach to determining structure from motion," International Journal of Computer Vision **1**, 7-55 (1987).
44. J. Navarro and A. Buades, “Robust and Dense Depth Estimation for Light Field Images," IEEE Transactions on Image Processing **26**, 1873-1886 (2017).
45. Y. Pan, R. Liu, B. Guan, Q. Du, and Z. Xiong, “Accurate Depth Extraction Method for Multiple Light-Coding-Based Depth Cameras," IEEE Transactions on Multimedia **19**, 685-701 (2017).
46. J. Dai and T. Nguyen, “View synthesis with hierarchical clustering based occlusion filling," in 2017 IEEE International Conference on Image Processing, 1387-1391 (2017).
47. M. Rerabek, L. Yuan, L. A. Authier, and T. Ebrahimi, “[ISO/IEC JTC 1/SC 29/WG1 contribution] EPFL Light-Field Image Dataset," (ISO/IEC JTC 1/SC 29/WG1, 2015).
48. G. Bjontegaard, “Calculation of Average PSNR Differences between RD-curves," (2001).