Light-Field Depth Estimation via Epipolar Plane Image Analysis and Locally Linear Embedding

Yongbing Zhang, Huijin Lv, Yebin Liu, Haoqian Wang, *Member, IEEE*, Xingzheng Wang, Qian Huang, Xinguang Xiang, and Qionghai Dai

***Abstract*— In this paper, we propose a novel method for 4D light-field (LF) depth estimation exploiting the special linear structure of an epipolar plane image (EPI) and locally linear embedding (LLE). Without high computational complexity, depth maps are locally estimated by locating the optimal slope of each line segmentation on the EPIs, which are projected by the corresponding scene points. For each pixel to be processed, we build and then minimize the matching cost that aggregates the intensity pixel value, gradient pixel value, spatial consistency, as well as reliability measure to select the optimal slope from a predefined set of directions. Next, a subangle estimation method is proposed to further refine the obtained optimal slope of each pixel. Furthermore, based on a local reliability measure, all the pixels are classified into reliable and unreliable pixels. For the unreliable pixels, LLE is employed to propagate the missing pixels by the reliable pixels based on the assumption of manifold preserving property maintained by natural images. We demonstrate the effectiveness of our approach on a number of synthetic LF examples and real-world LF data sets, and show that our experimental results can achieve higher performance than the typical and recent state-of-the-art LF stereo matching methods.**

***Index Terms*— Depth estimation, epipolar plane image (EPI), light field (LF), locally linear embedding (LLE).**

I. Introduction

LIGHT field (LF) is a function that describes the amount of light flowing in every direction through every point in space. Unlike traditional 2D images, an LF contains information about not only the accumulated intensity at each image point but also the separated intensity value of light rays in all directions, which allows a wide range of applications, especially in computer graphics, e.g., LF rendering, scene reconstruction, synthetic aperture photography, or 3D display.

LFs are typically produced either by rendering a 3D model or by photographing a real scene. In either case, a large collection of viewpoints must be obtained to produce the LF views. Nowadays, there are many devices for capturing LFs photographically such as camera arrays or a gantry consisting of a single moving camera [1]. However, the camera arrays are hardware-intensive and need a complex calibration procedure, and the less expensive gantry consisting of a single moving camera is limited to static scenes. Recently, plenoptic cameras, such as Lytro [2] and Raytrix [3], are becoming commercially available, making it available to acquire a large number of LFs for various scenes and can be applied in many specific applications, in particular depth estimation.

The quality of depth maps has a significant influence in the LF-related applications; however, it is a great challenge to obtain a dense and accurate depth map due to its large number of views in LF. To derive accurate and reliable depth maps, many pioneering works for the LF depth estimation have been done in the literature. According to whether to use the epipolar plane image (EPI, 2D slices of constant angle and spatial direction) or not, the LF depth estimation can be simply divided into two categories.

A. Depth Estimation Approaches Employing EPI

To the best of our knowledge, the first attempt to utilize the EPI for depth estimation was presented by Bolles *et al.* [4], who detect edges in an EPI and fit straight-line segments to the edges afterward to estimate the 3D structure. However, the basic line fitting is not robust enough, and consequently, the quality of reconstruction is sparse and noisy. Another approach was proposed by Criminisi *et al.* [5], who decomposed the scene into a set of spatiotemporal layers and obtained the disparities by exploiting the high degree of regularity in the EPI volume. To achieve higher quality, Wanner and Goldluecke [6], [7] applied a structure tensor to yield high quality depth maps from 4D LFs. It enables the generation of depth maps with higher accuracy; however, the global optimization process is always computational expensive, which hampers its practical usage.

B. Depth Estimation Approaches Without Employing EPI

Yu *et al.* [8] encoded 3D line constraints and applied the constrained Delaunay triangulation to implement the LF stereo

Manuscript received September 25, 2015; revised February 18, 2016; accepted March 24, 2016. Date of publication April 21, 2016; date of current version April 3, 2017. This work was supported in part by the National High-Tech Research and Development Program of China (863 Program) under Grant 2015AA015901 and in part by the National Natural Science Foundation of China under Grant 61571254, Grant 61571259, Grant U1301257, and Grant 61522111. The work of Q. Huang was supported in part by the National Natural Science Foundation of China under Grant 61300122 and in part by the Fundamental Research Funds for the Central Universities under Grant 2013B01814. This paper was recommended by Associate Editor P. Eisert. *(Corresponding author: Haoqian Wang.)*

Y. Zhang, H. Lv, H. Wang, and X. Wang are with the Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: zhang.yongbing@sz.tsinghua.edu.cn; lvhj13@mails.tsinghua.edu.cn; wanghaoqian@tsinghua.edu.cn; xingzheng.wang@sz.tsinghua.edu.cn).

Y. Liu and Q. Dai are with the Tsinghua National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: liuyebin@tsinghua.edu.cn; qhdai@tsinghua.edu.cn).

Q. Huang is with the College of Computer and Information, Hohai University, Nanjing 210098, China (e-mail: huangqian@hhu.edu.cn).

X. Xiang is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210044, China (e-mail: xgxiang@njust.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2555778

matching; however, this comes at a very high memory cost and is vulnerable to severe occlusions. Chen *et al.* [9] introduced a cost aggregation method based on the a bilateral consistency metric (BCM) on the surface camera (SCam) [10]. However, since Chen *et al.* [9] utilized the color of the reference pixel as the mean of the bilateral filter, it is biased toward to the reference view, and consequently has poorer performance when the input images are noisy. Kim *et al.* [11] leveraged coherence in massive LFs to reconstruct the depth. They first computed the depth around object boundaries and then dealt with the homogeneous interior regions with a fine-to-coarse procedure instead of the standard coarse-to-fine (CTF) approaches, which can yield precise object contours and ensure smoothness in less detailed areas. However, the performance of Kim *et al.* [11] would get worse for the surfaces with spatially varying reflectance, since they violate the assumptions behind the radiance density estimation. In addition, Tao *et al.* [12] proposed a dense depth estimation algorithm by combining both the defocus and correspondence depth cues simultaneously. However, it will result in incorrect depth estimations for the regions that are too far from the main lens's focus plane.

To improve the accuracy of depth maps while maintaining relatively lower computational complexity, we present a new depth estimation algorithm by analyzing the structure of the EPI in this paper. In the EPI, every corresponding pixel is projected onto a line, whose slope reflects the depth (or inverse disparity) of the corresponding scene point. Inspired by this, we present an efficient way to select the optimal slope of the corresponding line, from a given candidate angle set, as the one minimizing the devised matching cost along the linear trace, which is more robust and much simpler than the point correspondences in 2D images. The proposed cost aggregation approach incorporates the intensity pixel value, gradient pixel value, spatial smoothness consistency, as well as reliability measure to improve the accuracy of the slope while reducing noises in homogeneous regions. Next, we devise a subangle estimation method to further refine the obtained optimal slope value of each pixel, utilizing the selected best angle as well as the two closest candidates in reverse direction. Furthermore, for each pixel, we observe its matching cost curve to identify its disparity reliable or not. Finally, we employ the locally linear embedding (LLE) method to estimate disparity of unreliable pixels. Since the disparity is calculated by locating the orientation of each line segmentation, the slope can range from 0 to ∞, which has a larger scale of disparity range. Besides, the computational complexity of our method is greatly reduced compared with other methods, since no global optimization is imposed.

To demonstrate the effectiveness of our proposed method, we carry out the LF depth estimations over a large range of data sets, from synthetic LFs up to real-world examples from several sources. All the experimental results verify the superiority of our method over other existing LF depth estimation methods.

The reminder of this paper is organized as follows. Section II provides a brief introduction of the LF structure analysis. In Section III, the proposed local depth estimation method is investigated in detail. Section IV provides the detailed description of disparity labeling and enhancement followed by the experimental results given in Section V. Finally, the conclusions are given in Section VI.

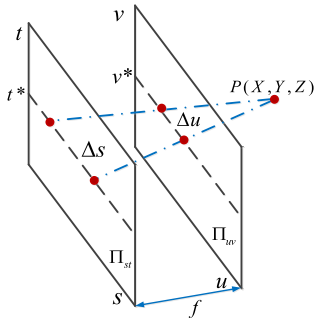


Fig. 1. 2PP of a 4D LF by coordinates (u, v) in the image plane Π_{uv} and coordinates (s, t) in the camera plane Π_{st}, which describes the projection of every 3D point P into every camera.

II. Light-Field Structure Analysis

A number of ways have been proposed to represent LF, e.g., two-plane parameterization (2PP) and the sphere–sphere and sphere–plane parameterizations. In this paper, we adopt the 2PP, as shown in Fig. 1, for depth estimation due to its simple structure and high efficiency. In 2PP, the direction is parameterized using two paralleled planes, i.e., the camera plane Π_{st} and the image plane Π_{uv}. The camera plane Π_{st} is at $z = 0$ and the image plane Π_{uv} is at $z = 1$, and thus, any point in the 4D LF can be identified by its four coordinates $[s, t, u, v]$, and the coordinates of a ray pierces the first plane at $(s, t, 0)$ and intersects the second plane at $(u, v, 1)$.

To better exploit the efficiency of 2PP, we consider the structure of the 3D LF, i.e., a set of photographs captured along a linear path (by a linearly translating camera). To visualize the positional changes in Π_{uv} caused by the changing of camera position, we draw out the horizontal line of constant v^* in the image plane and a constant camera coordinate t^*, resulting the map called an EPI, which is shown in Fig. 2. It should be noted that here we take the EPI by drawing a horizontal line of each corresponding image as an example, and EPI can also be formed by drawing out a vertical line of constant u^* in the image plane and a constant camera coordinate s^*. It can be easily observed that the EPI consisted of simple linear structures, which are projected by corresponding scene points, even though the photographs contain quite complex shapes and intensity changes.

Given the geometry of Fig. 1, if we vary s, the coordinate u changes as follows:

$$\Delta u = -\frac{f}{Z} \cdot \Delta s \tag{1}$$

where Δs is the geometrical distance between the two cameras along the line and Δu is the distance between the scene points moved in the image plane. Equation (1) can also be reformulated as

$$Z = -f\frac{\Delta s}{\Delta u} \tag{2}$$

Fig. 2. Definition of an EPI. The depicted EPI is related to the horizontal lines in the corresponding images.

where $\Delta s/\Delta u$ is the slope of lines in the EPI. This expression indicates that the real metric depth value Z is inversely proportional to the slope of its line in the EPI. In other words, the orientation corresponds to the depth (or inverse disparity) of the corresponding scene point, which is defined as the depth-slope relationship in this paper. Based on this principle, we proposed our depth estimation algorithm, which is introduced in detail in Section III.

III. Local Depth Estimation

Given the depth-slope relationship, the strategy of the proposed LF depth estimation is outlined as follows.

1) Form an EPI for each row and column.
2) Define a cost aggregation method for each pixel in EPI generated along the horizontal and vertical directions.
3) Refine the cost volume.
4) Select the optimal direction with the minimal cost from the EPIs generated along the horizontal and vertical directions.
5) Detect the unreliable pixels and fill them using the LLE-based depth propagation method.

It should be noted that for the LFs only containing images in the horizontal row or the vertical column, we only employ the EPI generated along the horizontal or vertical direction, and then select the optimal direction with the minimal cost from the corresponding EPI. The pipelines of each step are given in Sections III-A–III-C.

A. Optimal Orientation Selection

We locate the orientation of lines in the EPIs using the matching cost aggregation method and the winner-take-all method. Please note that the EPI consisted of homogeneous regions bounded by straight lines, and the disparity can be estimated by referring the orientation of the corresponding straight lines. Inspired by this, the optimal orientation can be selected from a given candidate angle set for each pixel to be processed in the EPI image, as shown in Fig. 3. In particular, given the EPI image E_i along the direction of i, with i being horizontal or vertical, for each pixel located at $o = (u, v)$, the best angle can be selected as

$$\theta^*(o, E_i) = \arg\min_{\theta \in \Theta} C_\theta(o, E_i) \tag{3}$$

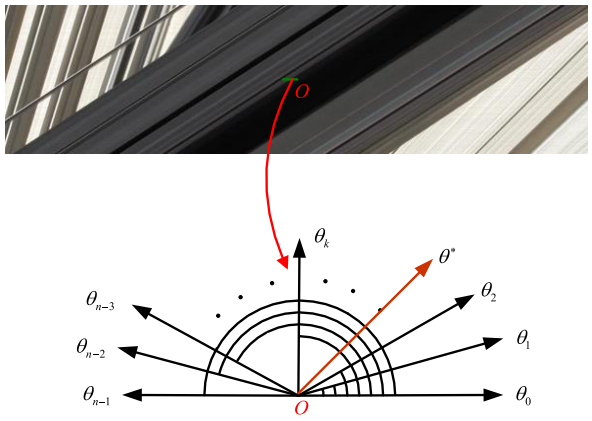


Fig. 3. Illustration of each candidate angle in terms of the pixel to be processed.

where Θ represents the candidate angle set and $C_\theta(o, E_i)$ represents the cost induced by selecting the candidate angle θ in E_i. Throughout this paper, $\Theta = \{0°, 1°, \ldots, 180°\}$ and i is an enumeration value, meaning horizontal when i is 0 and vertical when i is 1 or vice versa.

The optimal angle can then be selected as

$$\theta^*(o) = \arg\min_{i=0,1} C_{\theta^*(o,E_i)}(o, E_i). \tag{4}$$

In case there is only horizontal EPI or vertical EPI, the optimal angle is computed immediately via (3). The corresponding depth value can then be computed as $d(o) = -f \cdot \tan(\theta^*(o))$ according to (2).

Obviously, the core of (3) is the definition of cost function $C_\theta(o, E_i)$. Note that pixel intensity values exhibit quite similar along the optimal direction within both the grayscale image and the gradient image along the x- and y-axes, respectively, as shown in Fig. 4. Here, we depict the EPI image generated along the horizontal direction, and the corresponding gradient images, obtained by Sobel operator, along both the x- and y-axes of the EPI image, respectively. Obviously, there is a high consistency between the grayscale and gradient images, especially along the significant edge of the linear structure within the EPI. Based on such an observation,

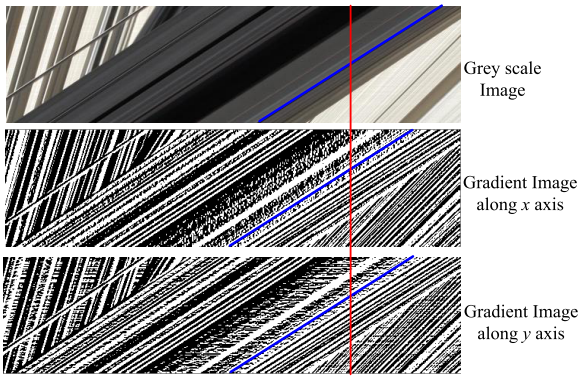


Fig. 4. Illustration of intensity difference along the optimal direction within the grayscale image and the gradient image along the x- and y-axes, respectively.

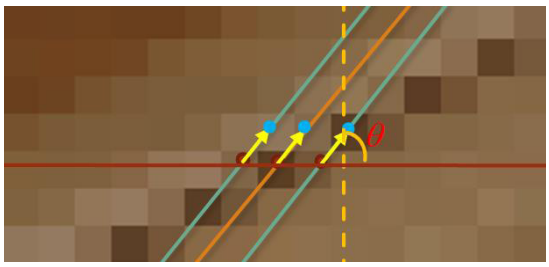


Fig. 5. Process of extending the pixel set $S(\theta)$.

the cost function $C_\theta(o, E_i)$ is defined as

$$C_\theta(o, E_i) = \frac{1}{|S(\theta)|}\sum_{q\in S(\theta)}\begin{pmatrix}(1-\alpha)\min(\|I(q)-I(o)\|^l, \tau_1)\\ +\alpha\min\begin{pmatrix}\|I_x(q)-I_x(o)\|^l\\ +\|I_y(q)-I_y(o)\|^l, \tau_2\end{pmatrix}\end{pmatrix} \quad (5)$$

where $S(\theta)$ represents the pixel set along the candidate angle θ, $|S(\theta)|$ represents the cardinality of $S(\theta)$, I is the intensity of the grayscale image, $I_x(o)$ and $I_y(o)$ represent the intensity value located at position o within the horizontal and vertical gradient images, α is the weighting factor to balance the energy between the two terms, and it is set to 0.7 throughout this paper, and τ_1 and τ_2 are the truncated sum of absolute differences and set to 1 and 3, respectively. Taking the gradient images into consideration, the cost function in (5) is able to reflect the linear structure of the EPI more faithfully, and consequently able to improve the robustness of the proposed method.

More generally, the more samples are considered, the more robust the performance would be. However, in practice, there are usually limited number of views within LF, which inhibits the performance. To improve the accuracy of the estimated optimal direction, we extend the pixel set $S(\theta)$ by adding some lines on both the left-hand and right-hand sides of the processed pixel having the same angle with the candidate one, which is shown in Fig. 5. The matching cost for each pixel within the extended set can be redefined as

$$C_{\theta,e}(o, E_i) = \sum_{\Delta\in X}\exp\left(-\frac{\|\Delta\|^l}{\sigma_s}-\frac{\|I(o)-I(o+\Delta)\|^l}{\sigma_c}\right) \times C_\theta(o+\Delta, E_i) \quad (6)$$

where vector $X = [(-L, 0), (-L+1, 0), \ldots, (L, 0)]$, where L is set to be 2 or 3 in this paper, and σ_s and σ_c are the two constants used to adjust the spatial similarity and color similarity, respectively. Here, σ_s and σ_c are set to 5 and 1, respectively. The weight of each extended sample depends on not only the spatial distance but also the intensity difference between the sample and the target pixel, which is able to reflect the similarity between each sample and the target pixel to the greatest extent. It should be noted that (6) employed bilateral filtering [13] to smooth the plain regions while preserving the edge regions within the cost function; consequently, it will select the more reliable slope.

In (6), different values of error norm l can be tried, and the comparison of choosing different error norms as cost function is shown in Fig. 6. Here, we select four pixels (two unoccluded points and two occluded points) within the EPI image and then depict the corresponding $c_d - d$ curves for different scene points. It can be observed that when l is equal to 1, it always coincides with the ground truth for the unoccluded case. Consequently, l is set to be 1 throughout this paper.

Utilizing (6), the optimal angular can be easily selected. However, by selecting $\theta^*(o, E_i)$ with the minimal cost, we found that the disparity map is very noisy in homogeneous regions. To address this issue and make the results more robust simultaneously, we add the spatial smoothness constraint in the cost volume construction process as

$$C^A_{\theta,e}(o, E_i) = (1-\lambda)C_{\theta,e}(o, E_i) + \lambda\frac{1}{|\Lambda|}\sum_{\Delta\in\Lambda}\omega_{\theta^*}(o, \Delta) \times|\tan(\theta^*(o+\Delta, E_i)) - \tan(\theta(o, E_i))| \quad (7)$$

where $\Lambda = [(0, -1), (-1, -1), (-1, 0)]$, $\theta^*(\Delta, E_i)$ represents the optimal angle at position Δ, $\theta(o, E_i)$ represents the candidate angle at position o being processed, and λ is the relative weight to balance the energy between the two terms. It should be noted that since the raster scanning mode is applied in this paper, the optimal angle at the positions in Λ is available. Here, $\omega_{\theta^*}(o, \Delta)$ can be formulated as

$$\omega_{\theta^*}(o, \Delta) = \exp\left(-\frac{|I(o)-I(\Delta)|}{\sigma_c}\right)\cdot r_{\theta^*}(\Delta) \quad (8)$$

where $r_{\theta^*}(\Delta)$ is the local reliability measure, which is described in Section III-B.

B. Local Reliability Measure

The optimal orientation generated by the simple winner-take-all principle may be not accurate enough, especially around the occlusion boundaries. To solve this problem, we propose a reliability measure based on the SCam [10]. In short, the SCam is an image that consists of all the projections $I_{s,t}(u, v)$ of a 3D scene point o in each camera.

Here, we take the EPI generated by arranging images along the horizontal line as an example, and it can be easily extended

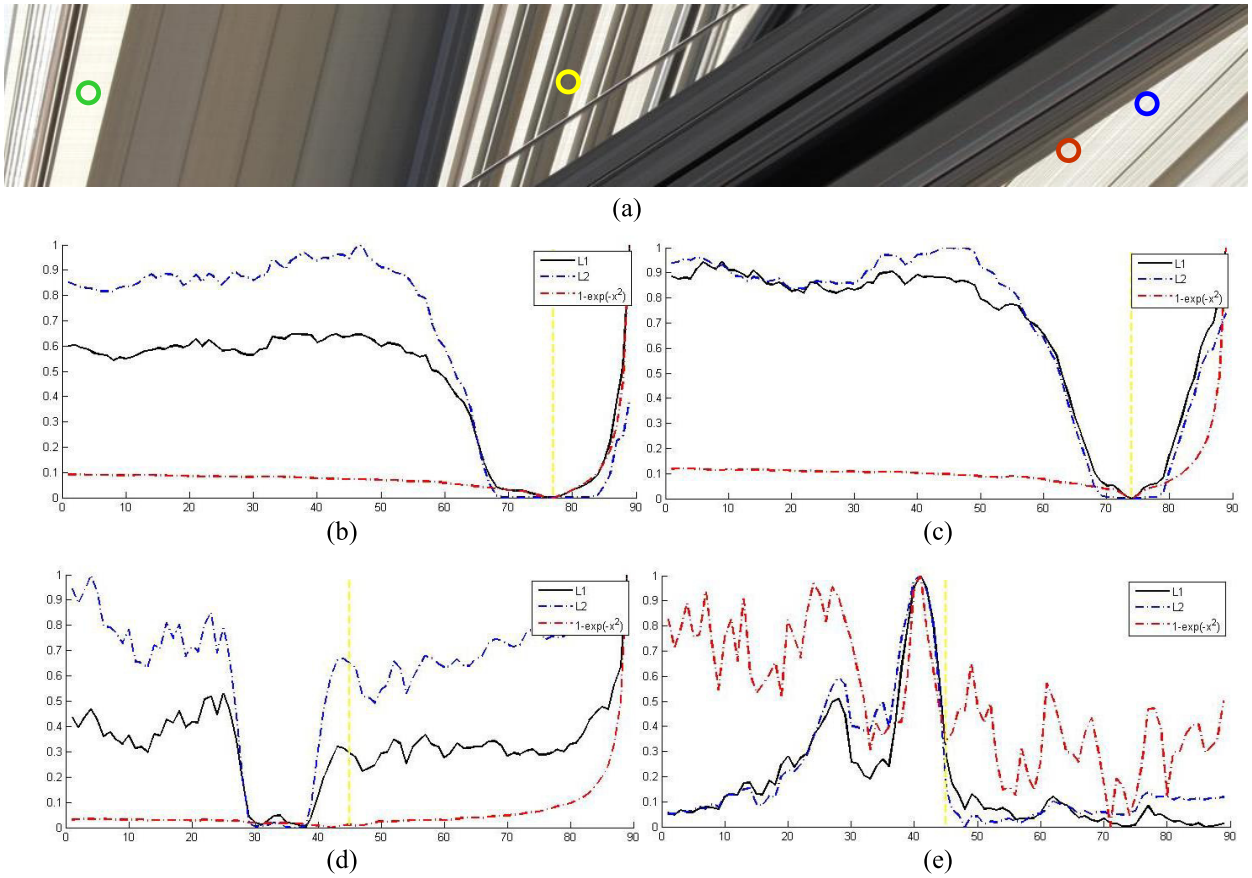


Fig. 6. $c_d - d$ curves for different scene points. Yellow lines: true orientation. The ground truth orientation always corresponds to the global minimum when using error norm 1 while other functions do not. (a) Four different scene points within the EPI. (b) $c_d - d$ curve (green pixel). (c) $c_d - d$ curve (yellow pixel). (d) $c_d - d$ curve (red pixel). (e) $c_d - d$ curve (blue pixel).

to the EPI by arranging images along the vertical line. Inspired by [10], we generate SCam for each pixel o in single EPI E_0 based on its depth $d(o) = -f \cdot \tan(\theta^*(o, E_0))$. Here, we use o_d to represent the 3D point (u, v, d), where d is the depth along ray $[s_r, t^*, u, v]$, with s_r being the view that o_d locates at. For every 3D point o_d, we back project it to every camera (s, t^*) in the LF and denote the SCam as $A_o(s_r, t^*)$, representing, at each (s, t^*) camera a ray (pixel) that passes through o_d.

Then, we define the reliability measure of o as

$$r_{\theta^*}(o) = \frac{1}{|\Omega|} \sum_{s \in \Omega} 1 - \exp\left(-\frac{(A_{o_d}(s, t^*) - A_{o_d}(s_r, t^*))^2}{2\sigma^2}\right) \tag{9}$$

where s_r is the reference view at which o_d located at and Ω is the set of pixels in the SCam that are not occluded.

However, since Ω in (9) is unknown, we employ the BCM [9] to roughly estimate the probability of each pixel in $A_{o_d}(s, t^*)$ belonging to Ω as

$$P(A_{o_d}, s, t^*) = \exp\left(-\frac{(A_{o_d}(s, t^*) - A_{o_d}(s_r, t^*))^2}{2\sigma_c^2} - \frac{(s - s_r)^2}{2\sigma_s^2}\right). \tag{10}$$

If we assume that the size of Ω is at least N, then we have

$$\Omega = \{s | P(A_{o_d}, s, t^*) \geq \min(P_{\text{threshold}}, P^N)\} \tag{11}$$

where $P_{\text{threshold}}$ is a predefined threshold and P^N is the N highest BCM. In this paper, $P_{\text{threshold}}$ is set to 0.5 and N is set as half of the total number of views.

C. Subangle Estimation

Since the cardinality of the candidate angle set is limited, the generated disparity maps may be discontinuous for some regions. To address the problem caused by quantization in the orientation hypothesis selection process, we utilize a subangle estimation algorithm based on quadratic polynomial interpolation [14] to further improve the estimation accuracy. In particular, we have

$$f(\theta) = a\theta^2 + b\theta + c \tag{12}$$

and

$$\theta_{\min} = d - \frac{-b}{2a} \tag{13}$$

where $f(\theta_{\min})$ is the minimum of function $f(\theta)$. Given the optimal angle value θ^* obtained in (3), $f(\theta^*)$, $f(\theta_-^*)$, and $f(\theta_+^*)$, the parameters a and b of the continuous cost function can be calculated. Consequently, we have

$$\theta_{\min} = \theta^* - \frac{f(\theta_+^*) - f(\theta_-^*)}{2(f(\theta_+^*) + f(\theta_-^*) - 2f(\theta^*))} \tag{14}$$

where θ_-^* and θ_+^* represent the two closest candidates in the reverse direction within the candidate angle set. Using the

refined disparity, the overall procedure is applied again for better results. In this paper, four iterations are sufficient for appropriate results.

IV. Disparity Labeling and Enhancement

To further improve the accuracy of the obtained depth image, we also propose an enhancement method through disparity labeling and disparity propagation, which is detailed in this section.

A. Disparity Labeling

The calculation of local depth estimation only takes into account the local structure of the LF, which may lead to some inaccurate pixels. In this section, we will show how we distinguish whether it is reliable or not. For each pixel, if the variance of the cost over this pixel is smaller than a threshold τ_{reject}, this pixel is regarded as unreliable, because it does not have distinctive minimum values, namely, it is within the textureless region. Moreover, if the reliability measure is smaller than a threshold r_{reject}, this pixel is also regarded as unreliable. In this paper, τ_{reject} and r_{reject} are set to 0.01 and 0.5, respectively, when detecting the unreliable pixels.

B. Disparity Propagation

To obtain the accurate values of unreliable pixels, in this section, we propose a disparity propagation method. The proposed propagation is performed with two assumptions. First, the reliability of the estimated disparity should remain unchanged or similar before and after disparity propagation. Second, we should maintain the manifold structure formed by the pixels in some feature space, for example, suppose the grayscale value at pixel A can be obtained by linearly combining those at pixels B and C, and then we deem the disparity at pixel A also have this linear combination with disparities at pixels B and C.

Our method is inspired by the LLE [15] and the work of Chen *et al.* [16]. LLE can project data from a high dimensional space to a low-dimensional manifold based on the simple intuition that each sample can be represented by a linear combination of its neighbors. In this paper, we define the feature vector X_i as the grayscale value I and spatial coordinate (x, y) to represent a pixel i. Given that the feature vectors of all the pixels $X_i, \ldots, X_N$, for each pixel, we find its K nearest neighbors, namely, $X_{i1}, \ldots, X_{iK}$, and then we compute linear coefficients ω_{ij} that reconstruct each data point from its neighbors by minimizing

$$\min_{\omega_{ij}} \sum_{i=1}^{N} \left\| X_i - \sum_{j=1}^{K} \omega_{ij} X_{ij} \right\|^2, \quad \text{s.t.} \sum_{j=1}^{K} \omega_{ij} = 1. \tag{15}$$

These coefficients ω_{ij} can be derived according to [15]. In our disparity propagation process, we should seek to maintain the manifold structure by requiring $d_i = \sum_{j=1}^{K} w_{ij} d_{ij}$ in the depth map. Here, d_i is the depth value at pixel i.

We define the pixels regarded as reliable in Section III as a set $\Re$, and then propagate them to the whole depth map by minimizing the following energy function:

$$E = \sum_{i \in \Re} (d_i - g_i)^2 + \beta \sum_{i=1}^{N} \left(d_i - \sum_{z_j \in N_i} \omega_{ij} d_j \right)^2 \tag{16}$$

where g_i is the depth value of pixel in $\Re$ and N_i is the set of K nearest neighbors of pixel i. The first term ensures the final result to be close to the estimated depth in $\Re$, while the second term maintains the manifold structure in the feature space.

The energy can be further written in a matrix form as

$$E = (D - G)^T \Lambda (D - G) + D^T (I - W)^T (I - W) D \tag{17}$$

where D is a vector formed by concatenating all the reliable d_i's, I is the identity matrix, Λ is a diagonal matrix, and G is a vector with

$$\Lambda_{ii} = \begin{cases} \lambda, & i \in \Re \\ 0, & \text{otherwise} \end{cases} \quad G_{ii} = \begin{cases} g_i, & i \in \Re \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Equation (17) is a quadratic function about D, which can be minimized by solving the linear equation

$$[(I - W)^T (I - W) + \Lambda] D = \Lambda G. \tag{19}$$

Equation (19) is a sparse linear system and can be solved efficiently.

V. Experimental Results

In this section, we verify the effectiveness of our proposed depth estimation algorithm and compare our results with some recent typical LF depth estimation approaches. We will first give the visual comparison of our method and the existing methods, and then provide the computational complexity.

A. Visual Comparison

Since the pipeline of the proposed method is composed of several steps, we will first provide the step-by-step comparison in Fig. 7. We provide the intermediate results by using the initial cost volume, the reliability map, the detected unreliable pixels, and the estimated depth map after enhancement. In Fig. 7(b), many small gray regions with irregular shapes can be observed, which indicate that the depth estimation via the cost volume is not reliable and contains many noise. Fig. 7(c) shows the reliability map based on (8), where the white pixels denote the regions with higher reliability, while the black pixels represent the regions with lower reliability. It can be observed that the pixels in plain regions with a similar texture have a higher reliability, while the pixels with abrupt texture changes or around the edge exhibit a lower reliability. Fig. 7(d) shows the detected reliable and unreliable pixels. The unreliable pixels, indicated by white pixels are further processed by the proposed depth propagation by LLE. Fig. 7(e) shows the final estimated depth after disparity labeling and enhancement. It can be observed that the visual quality of the estimated depth gets improved gradually, especially around

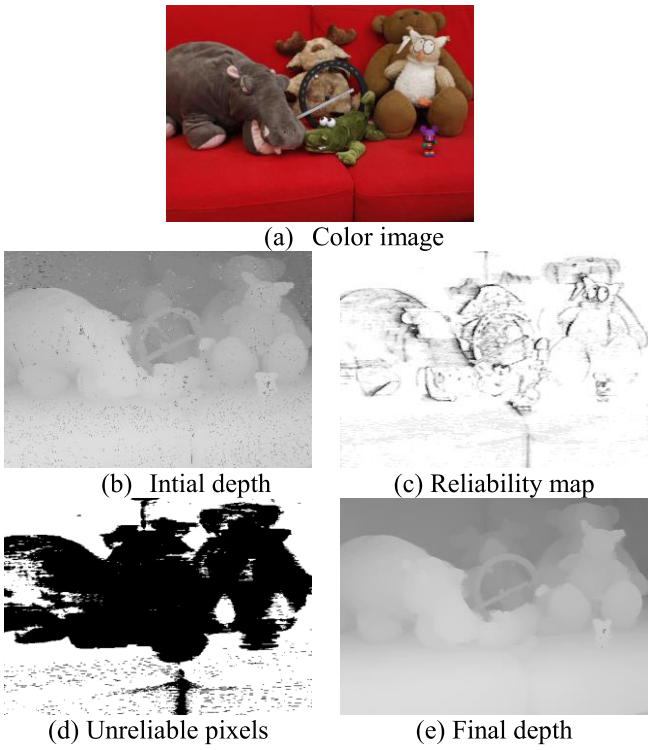


Fig. 7. Results at different steps of our proposed algorithm. (a) Color image. (b) Estimated map based on the initial cost volume (winner-takes-all strategy). (c) Reliability map based on (8). (d) Unreliable pixels (the white pixels indicate the detected unreliable pixels). (e) Estimated depth map after disparity labeling and enhancement.

the edge region or the contour of the different objects, which demonstrates the effectiveness of the proposed method.

We also compare our algorithm with other existing methods. Here, we evaluate our algorithm not only on the synthetic LF data sets used in [17] and [18] but also on the natural LF examples used in [11]. The existing LF estimation approaches we used in this paper are the globally consistent depth labeling (GCDL) [6] and line-assisted graph cut (LAGC) [8] as well as the CTF method [11]. We implement the code of CTF and use the source codes of GCDL and LAGC from the authors to carry out the depth estimation process. Moreover, we compare our results with the classical multi-view graph cut (MVGC) [19] and the efficient large scale stereo (ELAS) [20], for which the source code of binocular stereo matching is available in the authors' homepage. It is worth noting that LAGC only uses the images of $\sim$30 view-points due to the limitation of memory capacity and the computation complexity.

Figs. 8 and 9 show the visual results of our algorithm and the existing methods on real LF data sets used in [11]. For Couch, over the continuous regions, such as the hippo, GCDL and LAGC, produces smooth disparities, whereas some homogeneous regions are missing, which can be observed around the ears. However, the results of CTF and our proposed method are more accurate and particularly good at preserving edges. For the Statue in Fig. 9, GCDL and LAGC can yield sharp edges, while both of them miss some information. However, CTF and our proposed method achieve almost the same performance.

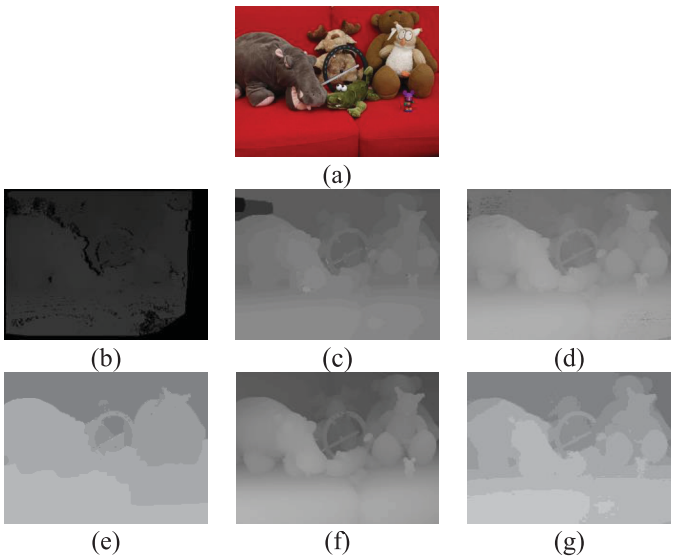


Fig. 8. Qualitative comparison of the Couch image used in [11]. (a) Center view. (b) ELAS. (c) MVGC. (d) GCDL. (e) LAGC. (f) CTF. (g) Ours.

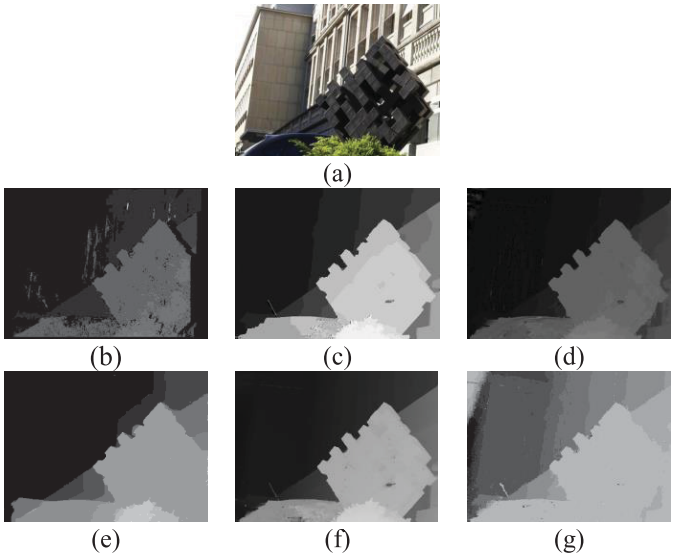


Fig. 9. Qualitative comparison of the Statue image used in [11]. (a) Center view. (b) ELAS. (c) MVGC. (d) GCDL. (e) LAGC. (f) CTF. (g) Ours.

We also test our method on several synthetic LF examples, as shown in Figs. 10 and 11. In Fig. 10, there are wrong repeated patterns around the contour of the Vase in the result generated by ELAS, due to the large amount of disocclusion. Besides, it is hard to observe the base. In the result generated by MVGC, the base cannot be easily found and the background is wrong. In the result generated by GCDL, there is serious blurring artifact around the Vase and base regions. In the result of CTF, chaos background can be observed, and the contour of the Vase is contaminated by the noises, while in the result generated by our proposed method, sharp edges can be perceived in both the Vase and base regions. For the Bonsai LF scene in Fig. 11, it is hard to distinguish the foreground and background regions and has the poorest performance in the result of ELAS. In MVGC result, there are significant artifacts around the right-hand side wall and the left-hand

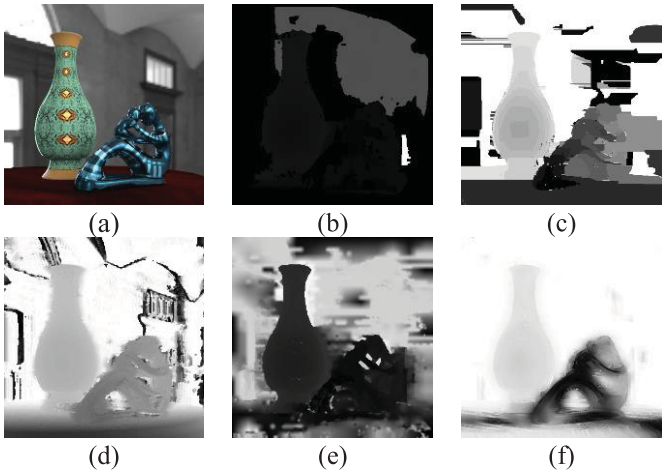


Fig. 10. Qualitative comparison on the synthesized scenes (Lobby). (a) Center view. (b) ELAS. (c) MVGC. (d) GCDL. (e) CTF. (f) Ours.

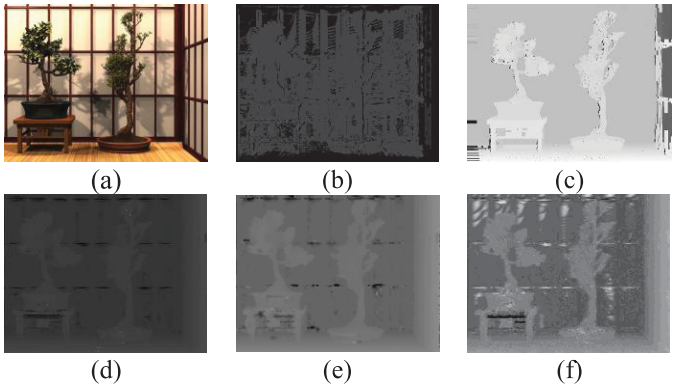


Fig. 11. Qualitative comparison on the synthesized scenes (Bonsai). (a) Center view. (b) ELAS. (c) MVGC. (d) GCDL. (e) CTF. (f) Ours.

TABLE I

AVERAGE PROCESSING TIMES (S) ON A TYPICAL COMPUTER (3.4 GHz INTEL DUAL CORE, 10G MEMORY)

Test LF	ELAS	MVGC	GCDL	LAGC	CTF	Ours
Number of views	2	2	101	21	101	101
Running time	0.29	87.5	300	59.3	216.7	39.5

side of the desk. In GCDL result, although the contour of the foreground can be perceived, it is difficult to differentiate the foreground and background regions, since the gray values are very similar. In the results generated by CTF and our method, both the foreground and the background can be clearly observed and the best performance can be achieved.

B. Computational Complexity

Table I lists the average processing times (seconds/frame) of each existing method and the proposed method when conducting the depth estimation over the LF image Couch (with a resolution of 360×512 and a view number of 101). It should be noted that ELAS, MVGC, and CTF are executed on CPU in C language, GCDL is executed on Graphic Processing Unit (GPU) in C language, and ours is executed on CPU in MATLAB. Here, number of views represent the number of views employed during the depth estimation for each view. Both the ELAS and MVGC utilize two views to perform depth estimation, among which ELAS has the fastest speed, requiring only 0.29 s to get the depth image of each view. Since MVGC uses graph cuts to minimize the energy function, the complexity is much higher than ELAS, and needs ∼87.5 s to finish the depth estimation for each view. Similarly, since GCDL employs the global optimization to enforce the smooth transition among neighboring regions, it has the largest computational complexity, requiring 300 s to obtain the depth image for each view. In LAGC, we only use 21 views to estimate the depth image for each view, since the exe file downloaded from the authors' homepage will collapse if the number of utilized view exceeds 21 when the system memory is 10G. The average processing time of LAGC is 59.3 s. In CTF, it takes ∼216.7 s to finish the LF depth estimation. In our proposed method, we fully use all the views of the LF, and the processing time is ∼39.5 s, which is the fastest except ELAS. However, the output of ELAS is much worse than the proposed method. Since we employ the EPI structure of LF, it is very easy to run the algorithm in parallel, and the expected processing time can be further reduced. Overall, it can observed that the proposed method is able to achieve the best visual quality while spending much less time compared with the majority of the existing methods.

VI. CONCLUSION

In this paper, we have presented an LF depth estimation framework by taking into account the geometry structure of EPI for LF. The relationship between the depth and slope of the linear structure in the EPI is studied, inspired by which a novel depth estimation approach is proposed by locating optimal orientation. During the orientation selection procedure, a spatial smoothness constraint is added to help preserve the consistency in homogeneous regions. And then, we introduce a scheme to detect and handle the unreliable pixels to enhance the quality of the final depth map.

For depth estimation, we have experimented on both the synthetic and real-world LF data sets and demonstrate that our method has better performance than the existing methods in the homogeneous areas as well as the edges. In addition, the computation complexity of our approach is very low without the need for global optimization.

The key component of our approach is the optimal orientation selection procedure; however, this process is applicable only for LFs with large numbers of densely sampled views. When the number of views is fewer than 20, there is a noticeable degradation in quality. In the future, we will try to improve our framework using fewer views.

REFERENCES

[1] *The (New) Stanford Light Field Archive*. [Online]. Available: http://lightfield.stanford.edu
[2] *Lytro*. [Online]. Available: https://illum.lytro.com/
[3] C. Perwass and L. Wietzke. (2010). *The Next Generation of Photography*. [Online]. Available: http://www.raytrix.de
[4] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Int. J. Comput. Vis.*, vol. 1, no. 1, pp. 7–55, 1987.
[5] A. Criminisi, S. B. Kang, R. Swaminathan, R. Szeliski, and P. Anandan, "Extracting layers and analyzing their specular properties using epipolar-plane-image analysis," *Comput. Vis. Image Understand.*, vol. 97, no. 1, pp. 51–85, 2005.

[6] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 41–48.
[7] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.
[8] Z. Yu, X. Guo, H. Ling, A. Lumsdaine, and J. Yu, "Line assisted light field triangulation and stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2792–2799.
[9] C. Chen, H. Lin, Z. Yu, S. B. Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1518–1525.
[10] J. Yu, L. McMillan, and S. Gortler, "Scam light field rendering," in *Proc. 10th Pacific Conf. Comput. Graph. Appl.*, 2002, pp. 137–144.
[11] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Art. no. 73.
[12] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2013, pp. 673–680.
[13] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE 6th Int. Conf. Comput. Vis.*, Jan. 1998, pp. 839–846.
[14] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[15] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[16] X. Chen, D. Zou, Q. Zhao, and P. Tan, "Manifold preserving edit propagation," *ACM Trans. Graph.*, vol. 31, no. 6, 2012, Art. no. 132.
[17] A. Jarabo, B. Masia, A. Bousseau, F. Pellacini, and D. Gutierrez, "How do people edit light fields?" *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 146.
[18] *POV-Ray*. [Online]. Available: http://www.povray.org/
[19] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proc. 7th Eur. Conf. Comput. Vis.*, 2002, pp. 82–96.
[20] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Proc. IEEE Asian Conf. Comput. Vis.*, Nov. 2011, pp. 25–38.

Yongbing Zhang received the B.A. degree in English and the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively.

He joined the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, in 2010, where he is currently an Associate Professor. His current research interests include video processing, image and video coding, video streaming, and transmission.

Dr. Zhang was a recipient of the Best Student Paper Award at the IEEE International Conference on Visual Communication and Image Processing in 2015.

Huijin Lv received the B.E. degree from the School of Information Science and Technology, University of Science and Technology of China, Hefei, China, in 2013. She is currently pursuing the M.E. degree with the Department of Automation, Tsinghua University, Beijing, China.

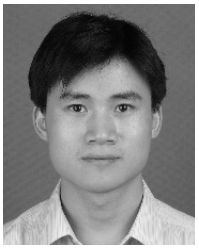
Her current research interests include superresolution and light field stereo matching.

Yebin Liu received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2002, and the Ph.D. degree from the Automation Department, Tsinghua University, Beijing, in 2009.

He was a Research Fellow with the Computer Graphics Group, Max Planck Institute for Informatik, Saarbrücken, Germany, in 2010. He is currently an Associate Professor with Tsinghua University. His current research interests include computer vision and computer graphics.

Haoqian Wang (M'13) received the B.S. and M.E. degrees from Heilongjiang University, Harbin, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Harbin Institute of Technology, Harbin, in 2005.

He was a Post-Doctoral Fellow with Tsinghua University, Beijing, China, from 2005 to 2007. He has been a Faculty Member with the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, since 2008, where he has also been an Associate Professor since 2011. His current research interests include video communication and signal processing.



Xingzheng Wang received the B.Sc. degree in mechanical engineering and automation from the North China University of Technology, Beijing, China, in 2004, the M.Sc. degree in mechanical engineering from Tsinghua University, Beijing, in 2007, and the Ph.D. degree in computer science from Hong Kong Polytechnic University, Hong Kong, in 2013.

He has been a Post-Doctoral Fellow with the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, since 2013. His current research interests include computational imaging, computer vision, and pattern recognition.

Qian Huang received the B.Sc. degree in computer science from Nanjing University, Nanjing, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2010.

He was a Deputy Technical Manager of Mediatek Inc., Beijing, from 2010 to 2012. Since 2012, he has been with Hohai University, Nanjing, where he serves as an Associate Professor of Computer Science. He has authored a dozen of technical articles in refereed journals and conference proceedings. Some of his techniques have been adopted by the international video coding standard HEVC. His current research interests include multimedia big data processing, cloud computing, and machine learning.

Xinguang Xiang received the B.A., M.S., and Ph.D. degrees from the Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2005, 2007, and 2011, respectively, all in computer science.

He is currently with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His current research interests include video coding, media analysis and understanding, and face recognition.

Qionghai Dai received the M.S. and Ph.D. degrees in computer science and automation from Northeastern University, Shenyang, China, in 1994 and 1996, respectively.

He is currently a Professor with the Department of Automation and the Director of the Broadband Networks and Digital Media Laboratory with Tsinghua University, Beijing, China. He has authored or co-authored over 200 conference and journal papers and two books. His current research interests include computational photography and microscopy, computer vision and graphics, and intelligent signal processing.

Dr. Dai is an Associate Editor of the *Journal of Visual Communication and Image*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and the IEEE TRANSACTIONS ON IMAGE PROCESSING.