Light Field from Micro-baseline Image Pair

Zhoutong Zhang, Yebin Liu, Qionghai Dai
Beijing Key Laboratory of Multi-dimension & Multi-scale Computational Photography (MMCP),
Tsinghua University, Beijing 100084 China

Abstract

We present a novel phase-based approach for reconstructing 4D light field from a micro-baseline stereo pair. Our approach takes advantage of the unique property of complex steerable pyramid filters in micro-baseline stereo. We first introduce a Disparity Assisted Phase based Synthesis (DAPS) strategy that can integrate disparity information into the phase term of a reference image to warp it to its close neighbor views. Based on the DAPS, an "analysis by synthesis" approach is proposed to warp from one of the input binocular images to the other, and iteratively optimize the disparity map to minimize the phase differences between the warped one and the ground truth input. Finally, the densely and regularly spaced, high quality light field images can be reconstructed using the proposed DAPS according to the refined disparity map. Our approach also solves the problems of disparity inconsistency and ringing artifact in available phase-based view synthesis methods. Experimental results demonstrate that our approach substantially improves both the quality of disparity map and light field, compared with the state-of-the-art stereo matching and image based rendering approaches.

1. Introduction

As an alternative to traditional 2D images, 4D light fields [14, 8] allow a wide range of applications by coding the spatial and angular information of a scene, which implicitly captures 3D scene geometry and reflectance properties. Despite its rapidly gaining popularity, high quality light field acquisition from real scenes is still an open problem. Expensive and sophisticated camera arrays have been proposed to capture dynamic light fields [33, 16], while a public repository for static light fields was captured with a custom-made gantry system using a single camera [1]. Integrating a microlens array [22] in a camera system has recently enabled consumer-grade light field capture (e.g. the Lytro camera), and a recent dictionary-based compressive sensing approach or a coded approach has been introduced [18, 34]; however, the quality of a captured light field is still a tradeoff between spatial and angular resolution.

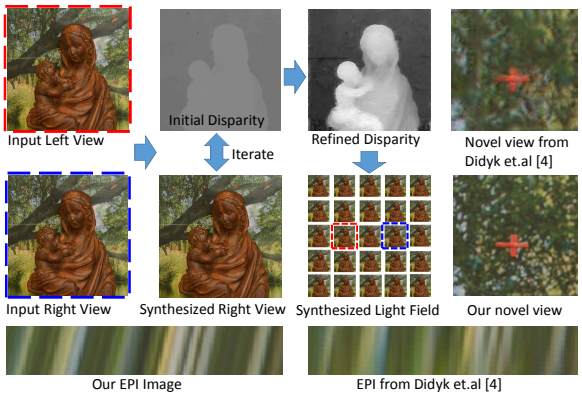

Figure 1. We reconstruct densely spaced 4D light field from a micro-baseline stereo pair. Our light field synthesis method iterates between disparity estimation and view synthesis in the phase domain using a complex steerable pyramid filter. The final disparity map is used in disparity assisted phase based synthesis of the target 4D light field with expanded view points. The EPI image of our reconstructed light field is free of ringing artifacts and shows more clear structures, compared with [4].

We propose a novel phase-based light field synthesis architecture that allows high quality densely sampled light fields to be reconstructed from a micro-baseline stereo pair. Micro-baseline in this work refers to a stereo image pair with a disparity less than 5 pixels [11]. Such an image pair can be efficiently captured by vibrating a static camera or applying small motion to a hand-held camera [35]. Inspired by the recent researches on phase based video magnification [29] and view expansion for generating automultiscopic content [4], we bring the complex steerable pyramid filter to the problem of light field reconstruction from a micro-baseline stereo pair. The complex steerable pyramid filter possess a particular property that most other filters do not have, i.e. it is Gabor-like (each decomposed band has both limited spatial and frequency support), translation invari-

ant and alias-free, which allows for perfect reconstruction. Because of this property, it is especially suitable for phase domain reconstruction of images that are very close to the input. The variance between views in a densely sampled light field is very small given a Lambertian scene assumption, which makes the complex steerable pyramid filter a good choice to reconstruct light field from micro-baseline stereo pair.

One of the main issues of the state-of-the-art phase based view expansion approach [4] is that its output suffers from ringing artifacts, as shown in the upper right of Fig.1. Moreover, it cannot satisfy the basic light field structure, i.e. the resulting LF is not structured with equal steps between views. As shown in the bottom right of Fig.1, the resulting epipolar-plane image (EPI) from the view expansion approach [4] contains obvious serrated structures, which indicates inconsistent disparity across views. Against this backdrop, we explicitly leverage disparity information under a phase based processing framework and propose a disparity assisted, phased based synthesis (DAPS) strategy to calculate the phase differences caused by the disparities, and compensate those into each decomposed bands. With a plausible disparity information, our DAPS approach can synthesize disparity consistent light fields, as shown in the bottom left EPI of Fig.1.

Furthermore, to have high quality disparity information, we take advantage of the property that the complex steerable pyramid allows a faithful reconstruction of close views in the phase domain, and propose an *analysis by synthesis* scheme to iteratively optimize the disparity information between the two input views in the micro-baseline pair, so that the phase terms of the right view synthesized from the left are well matched. Fig.1 also demonstrates the significant quality improvement of the disparity map.

We evaluate the proposed method quantitatively and qualitatively in terms of both disparity quality and light field image synthesis quality, which outperforms the state-of-the-art stereo matching method [24] and image based rendering method [23], respectively. Our recovered light fields are successfully used in novel view rendering and scene refocus, and may also be used as input for light field displays (see [19] for a recent overview of automultiscopic display technology). We hope that this work will open the door for further analysis and processing of light fields using a phase based framework. The source code of our work will be made public.

2. Related Work

Densely sampled light fields are important for many vision algorithms [15] and applications [10, 32]. Here we review light field acquisition methods and other techniques related with this work.

Dense Light Field Acquisition. Light field acquisition requires special systems and densely sampled ones are much more expensive to acquire. For consumer-grade acquisition systems like the Lytro (based on the seminal work by Ng [22]) and sparse sensing acquisition systems [18], the trade off between spatial and angular resolution exists. A dimensionality gap has been identified in light fields of Lambertian scenes with modest depth continuities [21, 13], which indicates redundancy and sparsity in the dense light field structure. Levin and Durand [12] leverage this dimensionality gap in light fields with specific priors, and reconstruct light fields using focal stack images.

Image-based Rendering. Most light field synthesis methods fall in the the category of image-based rendering (IBR), the core task of which is to synthesize new views given set of images. According to [27], image-based rendering algorithms can be categorized according to the dependency on geometry. Methods that use explicit geometry, such as view-dependent texture mapping (VDTM)[3], 3D warping [17] , layered depth images [26], are sensitive to inaccurate depth or disparity estimation, yet high quality depth maps are hard to acquire as well. Other methods, such as light field rendering [14, 2] require little or even no geometry information, but need a dense set of images as input. Recently, Pujades *et al*. [23] propose a view synthesizing method by optimizing a novel cost function with a Bayesian formulation, which improves both the spatial and angular disparity of light fields. Such methods, though greatly reduce the sampling rate for a dense light field, still require multiple input views to overcome the geometry uncertainty. Some other works[20, 36] devised methods to convert stereo pairs to light fields, but they emphasize more on scene refocusing using generated lightfield therefore do not provide a detailed examination of their reconstruction quality.

Phase-based Algorithms. Previous works [25, 7] have shown that the phase information of complex gabor filtering can be used to compute disparities and optical flow. Since the phase term is assumed to be linear in terms of displacement, the result suffers from the inaccuracy of this assumption. Sanger [25] gives a error analysis of disparity estimations using Gabor filters. Recently, Didyk *et al*. [4] use the phase information from a complex steerable pyramid decomposition [28] to expand view positions for 3D displays. Their method does not require disparity estimation, but results suffer from synthesized ringing artifacts and fail to follow the basic light field structure because of phase warping.

In contrast, our method is robust to low quality disparity estimations, follows the disparity consistency required by light field structures. In addition, our method is able to render good quality disparity maps and light field images

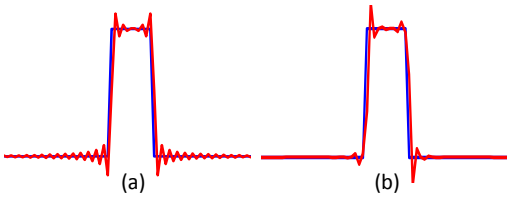

Figure 3. An illustration of the advantages of our method.(a) A non-bandlimited signal (blue) shifted by 0.5 pixel (red) by applying Fourier Shift theorem. (b) The same signal (blue) shifted by 0.5 pixel using our DAPS method (red). Note here how our result approximates the original high frequency signal without obvious ringing artifacts compared with (a).

from only two very small baseline images.

3. Overview

The input to our algorithm is a micro-baseline stereo pair after epipolar rectification [6]. Such data can be conveniently captured using a closely packed binocular camera or from images captured through tiny horizontal translation of a single camera. We first compute a raw disparity map between the two input images using an available dense stereo matching algorithm [24]. We then iteratively optimize the disparity map so that the difference between the input right view and the synthesized right view image is minimized. The key components of this optimization are a disparity-assisted, phase-based synthesis (DAPS) module and a phase-based disparity refinement module. The light field images are finally synthesized using our proposed DAPS algorithm.

As shown in Fig.2, we decompose the left view and right view using a complex steerable pyramid (Fig.2(a) and Fig.2(b)). By integrating phase and disparity information (Fig.2(c)) in different bands of the pyramid, our DAPS algorithm is able to synthesize novel views with sub-pixel accuracy by calculating the corresponding bands of the synthesized view (Fig.2(d)). This accuracy enables disparity refinement (Fig.2(f)) since the systematic error introduced in the view synthesis is relatively small compared to disparity errors. Then the phase difference between the synthesized bands and the corresponding ground truth bands is calculated (Fig.2(g)). Using cosine fitting, the phase difference is converted to estimated disparity error and is added to the initial estimation. Since the phase difference is noisy in nature, a filtering process [9] is required for the disparity estimation before it can be used in the new iteration. The iteration stops once the disparity improvement is lower than a threshold, and light field images are synthesized based on the optimized disparity map using DAPS (Fig.2(h)).

The advantages of our synthesis method can be interpreted in both the spatial and frequency domains. According to the Fourier shift theorem for discrete signals, adding an extra phase would accurately shift the signal by subpixel displacement if the input signal is band limited. However, most images are not band limited signals and adding the phase term in the Fourier domain will cause artifacts because of aliasing. Instead, we decompose the image into different band limited sub-bands using the complex steerable pyramid, and then shift all the bands. The only exception is the high residue band that contains a high frequency component and is not band limited. However, since most of the bands are accurately shifted, the reconstructed image would have less ringing artifacts overall. Fig.3 provides a simple illustration. The shifted signal can be reconstructed by collapsing all bands. In the spatial domain, this process can be intuitively interpreted as moving the image by patches with different sizes. A pixel in the target image is the weighted average of different image patches covering that pixel. This averaging process assures that our method is insensitive to bad disparity estimations.

4. Background

The first step of the proposed pipeline is to decompose the two input images using the complex steerable pyramid [28]. For the sake of clarity, we first review its mathematical background. The complex steerable pyramid is first and foremost a set of Gabor-like filters, which have limited support in both the frequency and spatial domains [5].

Gabor and Gabor-like filters. Previous phase-based methods [25, 7] use Gabor or Gabor-like filters to extract features, or calculate disparities and flow vectors. A Gabor filter is defined as a Gaussian-enveloped sinusoidal plane wave in the spatial domain, i.e. $e^{-\frac{\mathbf{x}^2}{2\sigma^2}} cos(\boldsymbol{\omega} \cdot \mathbf{x})$, where $\mathbf{x}$ stands for the 2-D position vector and $\boldsymbol{\omega}$ is the 2-D spatial frequency vector. For phase-based methods, complex Gabor filters are frequently used, which only include the positive frequency components of regular Gabor filters. In the frequency domain, the complex Gabor filter can be viewed as a 2-D Gaussian centered at frequency $\boldsymbol{\omega}$. Since complex Gabor filters have limited frequency support, a given image can be decomposed by convolving complex Gabor filters tuned at different frequencies to cover all frequency components of the image. However, such decomposition does not support perfect reconstruction since aliasing exists between filters tuned at different frequencies.

Complex Steerable Pyramid. The complex steerable pyramid decomposes an image into sub-bands with different scale and orientation by a set of Gabor-like filters. We denote the frequency response of a given image I at band i as $B_I^i(\boldsymbol{\omega}) = I_\omega(\boldsymbol{\omega})G_i(\boldsymbol{\omega})$, where $G_i(\boldsymbol{\omega})$ is the filter for band i in the frequency domain and I_ω denotes

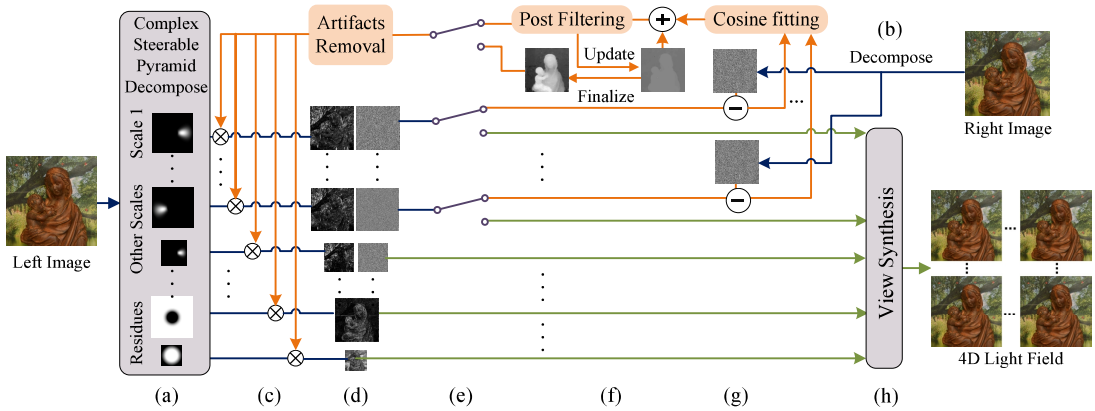

Figure 2. The pipeline of the whole proposed framework. (a) and (b) The two input images are decomposed using the complex steerable pyramid. (c) The disparity information is integrated into each band. (d) Synthesized bands for the target views (choose between the right view and the novel views). (e) Switches represent the choice between disparity refinement (orange lines) and novel view synthesis (green lines). (f) The disparity is iteratively refined and the final result is used to generate a mapping function for the novel views. (g) The band of the right view is used as ground truth, and the phase difference is computed. (h) Synthesis of novel views in the light field using DAPS.

the Fourier transform of I. All the bands, except for the high residue band, are band-limited since the corresponding $G_i(\boldsymbol{\omega})$ has limited frequency support. In addition, the complex steerable pyramid is translation-invariant, which means that a translation in the input image causes the same translation in all of its bands. That is, if the input image becomes $I(\mathbf{x}+\boldsymbol{\Delta x})$ instead of $I(\mathbf{x})$, all the $b_I^i(\mathbf{x})$ will become $b_I^i(\mathbf{x}+\boldsymbol{\Delta x})$, where b_I^i is the inverse transform of B_I^i. Besides, unlike Gabor decomposition, the complex steerable pyramid is designed to be self-inverting, and has non-aliased sub-bands since the aliasing terms between different bands cancel when collapsing the pyramid [5]; therefore the reconstruction can be noted as:

$$I_\omega(\boldsymbol{\omega}) = \sum_{i=1}^{N} B_I^i(\boldsymbol{\omega}) G_i(\boldsymbol{\omega}) = \sum_{i=1}^{N} I(\boldsymbol{\omega}) G_i^2(\boldsymbol{\omega}), \quad (1)$$

where N is the number of bands.

5. Disparity Assisted Phased based Synthesis

Given a rectified stereo pair $l(\mathbf{x})$ and $r(\mathbf{x})$, a disparity map $\mathbf{d}_r(\mathbf{x})$ from $l(\mathbf{x})$ to $r(\mathbf{x})$ can be initialized by available stereo matching algorithms (we use [24]). Note that the direction of $\mathbf{d}_r(\mathbf{x})$ is horizontal since the given pair is rectified. To synthesize a target view $n(\mathbf{x})$, which lies in the same image plane (light field plane) as the stereo pair, a corresponding 2D disparity estimation $\mathbf{d}(\mathbf{x})$ is first calculated by $\alpha d_r(\mathbf{x})\hat{\mathbf{x}} + \beta d_r(\mathbf{x})\hat{\mathbf{y}}$, where $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ are unit vectors for x and y axis; α, β are two arbitrary constants, controlling the novel view's position in the 2D plane. Here, it should be noted that in the disparity refinement step, the target view we synthesize is the right view $r(\mathbf{x})$, while in the final light field synthesis step, it is the novel views in the light field. The disparity $\mathbf{d}$ can be viewed as a pixel-wise mapping function, i.e. $l(\mathbf{x}) = n(\mathbf{x}+\mathbf{d}(\mathbf{x}))$. For simplicity, we use $\mathbf{f}(\mathbf{x}) = \mathbf{x} + \mathbf{d}(\mathbf{x})$ to denote the mapping function. It is important to note that $\mathbf{f}$ may not be invertible because of the existence of occluded points in $n(\mathbf{x})$ or in $l(\mathbf{x})$. A simple illustration can be found in Fig.4(c). In the following, for the sake of clarity, we assume that a bijection mapping function of $\mathbf{f}$ and $\mathbf{f}^{-1}(\mathbf{x})$ exists, and formulate our disparity assisted phase based view synthesis strategy. Fig.4(b) shows an example of a monotonically increasing function that satisfies this assumption, and Fig.4(a) shows its inverse function. Later, we will discuss the artifacts caused by breaking this assumption and how to remove them.

5.1. Formulation

According to the above assumption, a specific band b_n^i in the complex steerable pyramid of the target view n can be derived by:

$$b_n^i(\mathbf{x}) = n(\mathbf{x}) * g_i(\mathbf{x}) = \int n(\mathbf{k}) g_i(\mathbf{x}-\mathbf{k})\, d\mathbf{k}, \quad (2)$$

where g_i is the decomposition filter of band i in the complex steerable pyramid. Since $l(\mathbf{x}) = n(\mathbf{f}(\mathbf{x}))$, we have $l(\mathbf{f}^{-1}(\mathbf{x})) = n(\mathbf{x})$ and Eqn. 2 becomes:

$$b_n^i(\mathbf{x}) = \int l(\mathbf{f}^{-1}(\mathbf{k})) g_i(\mathbf{x}-\mathbf{k})\, d\mathbf{k}. \quad (3)$$

By representing l by its Fourier transform $L(\omega)$, the above equation becomes:

$$b_n^i(\mathbf{x}) = \frac{1}{2\pi} \int L(\boldsymbol{\omega}) \int g_i(\mathbf{x}-\mathbf{k}) e^{j\boldsymbol{\omega}\cdot\mathbf{f}^{-1}(\mathbf{k})}\, d\mathbf{k}\, d\boldsymbol{\omega}. \quad (4)$$

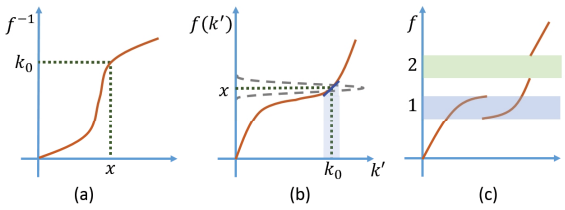

Figure 4. An illustration of $\mathbf{f}$, $\mathbf{f}^{-1}$, $\mathbf{x}$ and $\mathbf{k}_0$. $\mathbf{f}$ denotes the forward mapping function, which maps pixels from the reference image to the target image. (a) $\mathbf{f}^{-1}$ is the inverse mapping function. (b) $\mathbf{x}$ denotes the pixel positions in the target image and $\mathbf{k}_0$ is its corresponding pixel in the reference image. The segment is the approximation used to derive Eqn.6, which is centered at $\mathbf{k}_0$ with a slope of 1. (c) A illustration of $\mathbf{f}$ when occlusion regions exists. Region 1 denotes the newly occluded region in the target view, and region 2 denotes the disoccluded region.

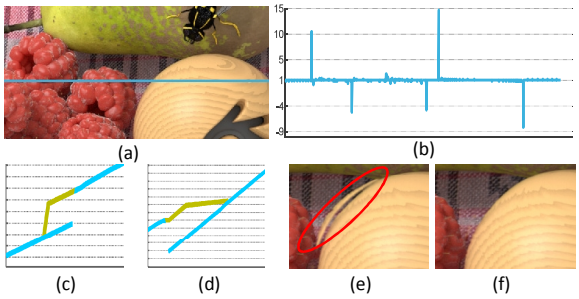

Figure 5. Illustration for artifacts removal when $\mathbf{f}(\mathbf{x})$ is along the baseline.(a) The line where the example $f(x)$ is defined. (b) The gradient of $f(x)$. Spikes above 1 indicates a disoccluded region, and spikes below 1 indicates a newly occluded region. The deviations from one indicates the size of the region. (c) Modifications(green) made to $f(x)$ in disoccluded regions. (d) Modifications(green) made to $f(x)$ in newly occluded regions. (e) New view synthesized without artifacts removal, with artifacts marked in the red ellipse. (f) New view with artifacts removal.

where $j = \sqrt{-1}$. By denoting $\mathbf{k}' = \mathbf{f}^{-1}(\mathbf{k})$, we have

$$\frac{1}{2\pi}\int L(\boldsymbol{\omega})\int g_i(\mathbf{x}-\mathbf{f}(\mathbf{k}'))e^{j\boldsymbol{\omega}\cdot\mathbf{k}'}\frac{d(\mathbf{f}(\mathbf{k}'))}{d\mathbf{k}'}\,d\mathbf{k}'\,d\boldsymbol{\omega}. \quad (5)$$

Since g_i is limited in the spatial domain, with a given $\mathbf{x}$, g_i returns non-zero values only when $\mathbf{f}(\mathbf{k}')$ approximates x. As shown in Fig. 4(b), we define $k_0 = \mathbf{f}^{-1}(\mathbf{x})$ that satisfies $\mathbf{f}(\mathbf{k}_0) = \mathbf{x}$, so $\mathbf{f}(\mathbf{k}')$ can be approximated as $\mathbf{d}(\mathbf{k}_0) + \mathbf{k}'$, which is a linear approximation to the function $\mathbf{f}(\mathbf{x})$ centered at k_0. Note that the accuracy of this approximation is determined by the spatial bandwidth of g_i. Under this approximation, $\frac{d(\mathbf{f}(\mathbf{k}'))}{d\mathbf{k}'}$ becomes 1, and Eqn.5 turns to

$$b_n^i(\mathbf{x}) = \frac{1}{2\pi}\int L(\boldsymbol{\omega})\int g_i(\mathbf{x}-\mathbf{d}(\mathbf{k}_0)-\mathbf{k}')e^{j\boldsymbol{\omega}\cdot\mathbf{k}'}\,d\mathbf{k}'\,d\boldsymbol{\omega}. \quad (6)$$

Substituting $\mathbf{x}-\mathbf{d}(\mathbf{k}_0)-\mathbf{k}'$, we get

$$b_n^i(\mathbf{x}) = \frac{1}{2\pi}\int L(\boldsymbol{\omega})G_i(\boldsymbol{\omega})e^{-j\boldsymbol{\omega}\cdot\mathbf{d}(\mathbf{k}_0)}e^{j\boldsymbol{\omega}\cdot\mathbf{x}}\,d\boldsymbol{\omega}, \quad (7)$$

where $\mathbf{d}(\mathbf{k}_0)$ is already known, and the novel view is synthesized by collapsing all the bands b_n^i.

Eqn.7 interprets the advantages we stated previously. First, since $G_i(\boldsymbol{\omega})$ has limited frequency support, $L(\boldsymbol{\omega})G_i(\boldsymbol{\omega})$ is therefore band-limited. By adding an extra phase term $e^{i\boldsymbol{\omega}\cdot\mathbf{d}(\mathbf{k}_0)}$ to $L(\boldsymbol{\omega})G_i(\boldsymbol{\omega})$, we can effectively shift the inverse Fourier transform of $L(\boldsymbol{\omega})G_i(\boldsymbol{\omega})$, which is b_l^i (the corresponding band of the left image), accurately by a subpixel displacement. Therefore, we avoid shifting the given image directly but shift its bands instead, since this band shift is accurate and the bands can be reconstructed perfectly. Second, since we are shifting b_l^i, which has different scales, we can interpret our method intuitively as moving small patches of the given image to the desired position. A pixel in the target image is a weighted average of all the patches that contain it.

5.2. Artifacts Removal

Regions where $\mathbf{f}^{-1}(\mathbf{x})$ is not a bijection mapping would cause artifacts, since $\mathbf{d}(\mathbf{k}_0)$ itself is not well defined. Another more intuitive explanation is that since we are moving pixels as patches, at significant disparity discontinuities, the image is locally stretched when the region is disoccluded and contracted when the region is newly occluded.

To remove those artifacts, we process the disparity estimation by the following procedure before synthesizing. We first detect those regions that are larger than a user specified size (1.5 pixels by default). To detect those regions, the gradient of $\mathbf{f}(\mathbf{x})$ and its direction are computed. Note that if there is no occlusions between the reference and target images, the disparity should be continuous and smoothly varying, therefore the gradient of $\mathbf{f}(\mathbf{x})$ should be around 1 without significant spikes. If the gradient has spikes that deviate 1 pixel from the user-specified size, a region of occlusion is located at the spike position, as shown in Fig.5(b). Pixels within the region are assigned the same disparity values as the nearest foreground content (a simple illustration showing that $\mathbf{f}(\mathbf{x})$ is horizontal can be found in Fig.5). By doing so, the background content is stretched to fill in the disoccluded regions and contracted in the occluded regions, therefore the artifacts are effectively removed.

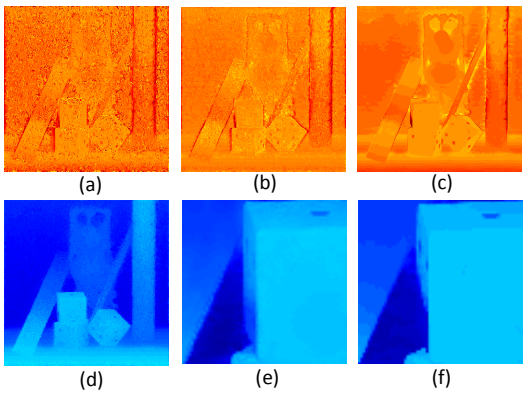

Figure 6. Disparity refinement. (a) The phase difference between the synthesized band and the corresponding band of the target image, note that the phase difference is very noisy. (b) The estimated disparity error by cosine fitting, which greatly reduce the noise in (a). (c) The ground truth disparity error, note its similarity to (b). (d) The refined disparity before filtering. (e) A zoom-in on the refined disparity after filtering. (f) A zoom-in on the initial disparity.

6. Disparity Refinement

Our synthesis method shares a common weakness with others depth-based rendering methods: The inaccuracy of the disparity estimation would cause the synthesis result to be inaccurate. Therefore, it is crucial to have a good disparity estimation from the given stereo pair. However, when the the baseline of the given pair is very small, an accurate disparity estimation becomes very challenging. Our disparity refinement method adopts an iterative optimization based on our proposed DAPS strategy, which substantially improves the initial disparity quality and enables high quality novel views synthesize.

We assume that in Eqn.7, $G(\boldsymbol{\omega})$ can be approximated by a complex Gabor filter tuned at frequency $\boldsymbol{\omega}_0$. Therefore, similar to traditional phase-based methods [25, 7], we first set the target image to be the right view, i.e. set $\mathbf{d} = \mathbf{d}_r$, and then approximate Eqn.7 by:

$$b_n^i(\mathbf{x}) = \frac{1}{2\pi} e^{-j\boldsymbol{\omega}_0 \cdot \mathbf{d}_r(\mathbf{k}_0)} \int L(\boldsymbol{\omega}) G(\boldsymbol{\omega}) e^{j\boldsymbol{\omega}\cdot\mathbf{x}}\, d\boldsymbol{\omega} \quad (8)$$

We also assume that there exists an accurate disparity $\mathbf{d}_{true}$ that maps the left image to the right image. Similar to the above equation, we can note:

$$b_r^i(\mathbf{x}) = \frac{1}{2\pi} e^{-j\boldsymbol{\omega}_0 \cdot \mathbf{d}_{true}(\mathbf{k}_0)} \int L(\boldsymbol{\omega}) G(\boldsymbol{\omega}) e^{j\boldsymbol{\omega}\cdot\mathbf{x}}\, d\boldsymbol{\omega}. \quad (9)$$

Since the synthesized band b_n^i and the ideal corresponding band b_r^i of the right image are known, the disparity estimation error $\mathbf{d}_{true} - \mathbf{d}_r$ can be calculated from the phase difference between b_n^i and $b_r^i(\mathbf{x})$. Traditional phase-based methods estimate the true disparity $\mathbf{d}_{true}$ directly from the phase difference in the corresponding bands of the left and right images. However, this estimation has three major limitations: First, the disparity has to be small enough, so that the corresponding points fall in the spatial bandwidth of the filter and the phase term does not wrap up; Second, the phase response is sensitive and unreliable when the magnitude of the band approaches zero, since small variations of the norm around zero would cause phase variations of $\pm\pi$. And third, the approximation is not accurate and therefore introduces significant noise to the estimated disparity. An illustration of this inaccuracy can be find in Fig.6(a).

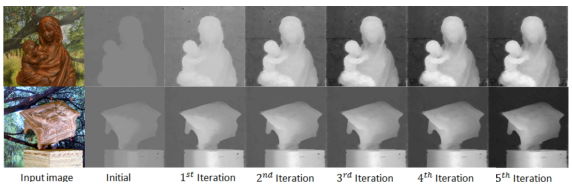

Figure 7. Disparity refinement result. It can be seen that after five iterations, the disparity estimation is greatly improved.

Our disparity refinement method circumvents those limitations. We first synthesize the bands of the right image using the initial disparity estimation to make sure that the disparity between them is small enough. We then compute the phase difference in 16 different orientations at the highest level, since the lower levels are decimated and therefore the estimated disparity has to be up-sampled, which introduces significant inaccuracies. For each pixel at a specific band, we measure the confidence of its phase term by the value of its magnitude term, where larger magnitude would infer a more reliable phase. Therefore, for each pixel, we pick 8 phase from 16 orientations with the largest corresponding magnitudes and compute the phase differences. Since the given images are rectified, which means that $\mathbf{d}$ and $\mathbf{d}_{ture}$ have only horizontal component, the phase difference can be written as $\boldsymbol{\omega}_0 cos(\theta_i) d_{err}$, where d_{err} stands for $\mathbf{d}_{true} - \mathbf{d}_r$, and θ_i is the orientation angle for band i. We then compute d_{err} by cosine fitting with minimum square error. A refined estimation is calculated by adding d_{err} back to the initial estimation. To reduce the introduced noise, we simply smooth the refined estimation by guided filtering [9] using $l(\mathbf{x})$ as reference. An illustration can be found in Fig.6.

When complex steerable pyramid filters are approximated by Gabor filters, we assume $\boldsymbol{\omega}_0$ to be the frequency where the pyramid filters have the highest frequency re-

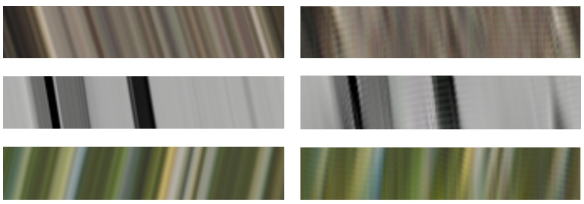

Figure 9. A comparison of the EPI image from the generated light field. (a) EPI image using our method (b)EPI image by Didyk *et al.*[4] Note that our result is sharper and shows a much clearer structure.

	Synthetic Scenes		Real-World Scenes			
α	still life	buddha	maria	couple	truck	gum
1	**34.81**	**44.71**	**40.81**	**32.07**	**37.79**	**40.22**
2	32.31	41.34	39.15	29.61	36.01	39.15
3	30.39	38.54	37.63	28.21	34.37	35.34
4	28.96	36.81	36.69	27.14	32.27	33.06
[23]	30.45	42.37	40.06	28.50	33.78	31.93
[4]	28.16	42.66	39.05	29.85	36.97	37.45

Table 1. Accuracy evaluation of our method for both synthetic and real scenes. The reference result is given by Pujades et al.[23] with disparity estimated from the entire light field, and available phase based view expansion approach [4]. The quality of the synthesized views is measured by the PSNR against the ground truth image. The best value for each scene is highlighted in bold. Only the best results of [4] are listed in the table. Our method performs consistently better when α is small.

sponse. In practice, this approximation would result in a d_{err} smaller than expected, and therefore we iteratively refine our disparity estimation until convergence. A convergence analysis can be found in the supplementary materials. This refinement works because our synthesize method DAPS has sub-pixel accuracy so that the systematic error introduced in warping can be neglected compared with the disparity estimation error. the iterative disparity refinement results are shown in Fig.7.

7. Results

We evaluate our light field synthesis method with current state-of-the-art methods in terms of disparity consistency, accuracy of the novel synthesized views, and ability to recover disparity maps from very close viewpoints.

Experiment setup In all the experiments, we set the left image as reference to generate novel views. Our method is capable of synthesizing a 4D light field (see the supplemental material for the results) , however, for qualitative and quantitative comparison with available methods, we constrain the novel view points to be collinear with view points of the given stereo pairs.

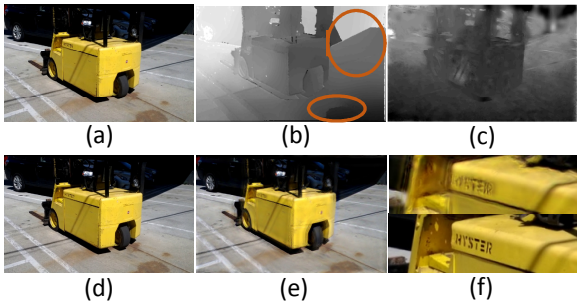

Figure 10. Experiment using micro-baseline images from a data set of [35]. The input images are captured by a handheld camera with tiny hand motion. (a) Reference frame. (b) Disparity estimation result from [35]. Note that the input to their algorithm is a 100-frame video, while we only use a stereo pair. Circled regions are obvious wrong estimations affected by the color of the reference image. (c) Our disparity estimation. (d) A synthesized view using our disparity estimation in (c) with $\alpha = 4$. (e) The same view synthesized using method from [4]. (f) Zooming in showing the detailed differences, with upper from (e) and bottom from ours (d).

That is, according to our previous formulation, we set the corresponding disparity $\mathbf{d}$ for synthesis to be $-\alpha\mathbf{d}_r$, where $\mathbf{d}_r$ is the disparity estimation of the given image pair. α is a constant that controls the distance between the synthesized view and the reference image. The minus sign denotes the novel view is set in the opposite direction to the reference image than the right image. To validate our algorithm, we use both synthetic and real scenes from the HCI datasets [31], Stanford light field datasets[1] and small motion datasets [35]. A quality evaluation of generated 4D light field can be found in supplementary materials.

In the first set of experiments, we generate 26 views with identical spacing by setting α from 0.3 to 7.8 linearly. We compare our resulting EPI image with the one by Didyk *et al.*[4]. As shown in Fig.9, our method generates highly linear EPI images and therefore follows the disparity constrain of light field.

In the second set of experiments, we first compare our results qualitatively by generating views with different α and compare with the ground truth views and results from state-of-art image-based rendering algorithm[23]. A set of zoomed in images are shown in Fig.8. Then we evaluate our method quantitatively by using 2 nearby views in the light field as input and synthesize the same view (center of the light field) with different α. We evaluate our accuracy by calculating PSNR between ground truth view and the synthesized views and use the results from Pujades *et al.* [23] for comparison. The results are shown in Table 1. It is important to note that Pujades *et al.* [23] use multiple input views and a depth estimation from a state-of-art light field

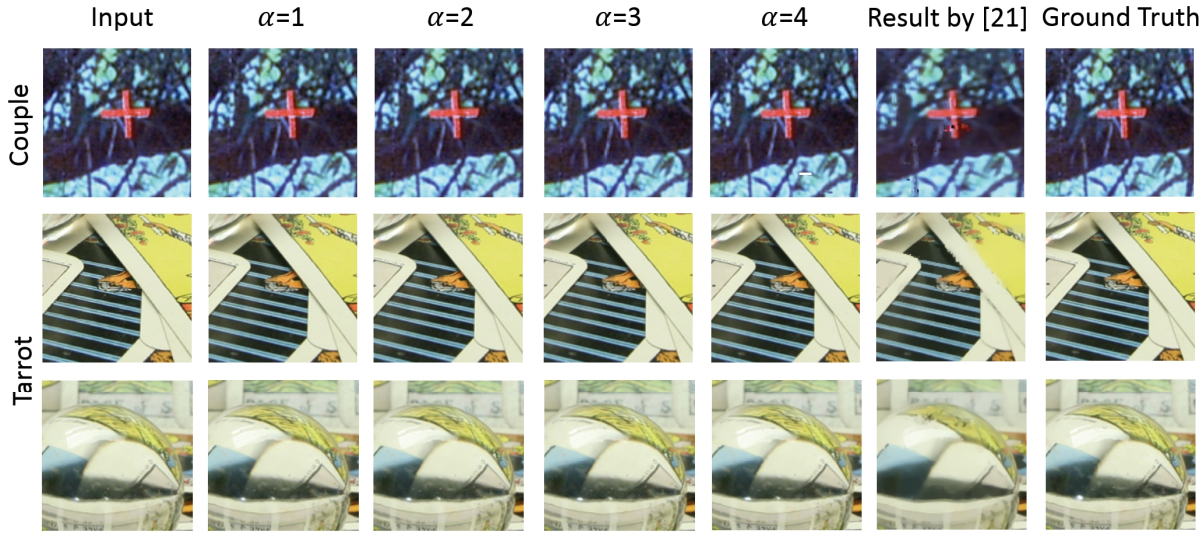

Figure 8. A zoomed in comparison of our result and Pujades *et al*. [23]. The results from [23] and the ground truth images have the same view points as our $\alpha = 4$ results. Out result is more accurate and contains less artifacts. Full resolution images can be found in supplementary martials.

depth estimation algorithm [30]. Even with just a stereo pair as input, our method is more accurate when synthesized view point is close to the input(small α). The state-of-art phase-based method given by Didyk *et al*.[4] is also used as reference.

We also compare the quality of our disparity map with the result of state-of-art algorithm for accidental motion videos [35]. We use two frames from an input video and rectify the frames with an uncalibrated rectification algorithm [6]. Our estimated disparity, shown in Fig.10(c), seems initially less pleasant because of large textureless regions and inaccuracies from rectification. However, the result by Yu and Gallop [35]contains large wrongly estimated regions due to the dense reconstruction process, where the color information of the reference view affects the estimation and our method is more accurate in contrast. Despite this, our final synthesized view is more accurate than the result of state-of-art phase-based view expansion algorithm [4], which demonstrates that our method is insensitive to low quality disparity estimation.

The computational cost of our method is mainly determined by the integration in Eqn.7. In the supplemental material we prove that for an image of $n \times n$ pixels, the temporal complexity is of $O(n^4)$. If the disparity $\mathbf{d}$ is in the direction of the x or y axes, the time complexity can be reduced to $O(n^3)$.

8. Discussions and Conclusions

The limitation of our work is that the novel view points are restricted to the neighbor of input view points. When the novel view point is relatively far from the input, the assumption of Lambertian scenes becomes less accurate and the occlusion regions become larger. Therefore, the quality of novel view synthesised would degrade. The main contribution of our work is a novel phase-based framework to render high quality 4D lightfield from very small baseline stereo pairs. Our framework is also capable to give a quality disparity estimation under such small baselines, which is very challenging for traditional stereo matching algorithms. Besides, our work enables small baseline stereo pairs for light field algorithms and serves as a potential bridge to connect between the two. It also has the potential of transforming traditional cameras into light field cameras without installing extra hardware.

Future work should extend our framework for unstructured input, which would further reduce the complexity of acquiring dense light field. In addition, it is also possible to consider using multiple input images to further improve the quality of synthesized light field in terms of accuracy and occlusion handling, which is the bottle neck for generating distant views. Also, extending the assumption of Lambertian scenes is also crucial for rendering distant views but may be very challenging; one should consider using multiple input images in order to sample the BRDF of scene surfaces effectively.

Acknowledgement This work was supported by the Na-

tional key foundation for exploring scientific instrument No. 2013YQ140517, the 863 Program (No.2013AA01A604) and the open funding project of state key laboratory of virtual reality technology and systems, Beihang University (Grant No. BUAA-VR-14KF-08).

References

[1] The (new) stanford light field archive, 2008. http://lightfield.stanford.edu. 1, 7

[2] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *ACM SIGGRAPH*. 2

[3] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *SIGGRAPH*, pages 11–20, 1996. 2

[4] P. Didyk, P. Sitthi-amorn, W. T. Freeman, F. Durand, and W. Matusik. Joint view expansion and filtering for automultiscopic 3d displays. *ACM Trans. Graph.*, 32(6):221, 2013. 1, 2, 7, 8

[5] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE T.PAMI*, 13(9):891–906, 1991. 3, 4

[6] A. Fusiello and L. Irsara. Quasi-euclidean epipolar rectification of uncalibrated images. *Machine Vision and Applications*, 22(4):663 – 670, 2011. 3, 8

[7] T. Gautama and M. M. V. Hulle. A phase-based approach to the estimation of the optical flow field using spatial filtering. *IEEE Trans. Neural Networks*, 13(5):2002, 2002. 2, 3, 6

[8] S. J. Gortler and M. F. Cohen. Hierarchical and variational geometric modeling with wavelets. In *SI3D*, pages 35–42, 205, 1995. 1

[9] K. He, J. Sun, and X. Tang. Guided image filtering. In *ECCV*, 2010. 3, 6

[10] A. Jarabo, B. Masia, A. Bousseau, F. Pellacini, and D. Gutierrez. How do people edit light fields? *ACM Trans. Graph.*, 33(4), 2014. 2

[11] N. Joshi and C. L. Zitnick. Micro-baseline stereo. Technical Report MSR-TR-2014-73, May 2014. 1

[12] A. Levin and F. Durand. Linear view synthesis using a dimensionality gap light field prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2010. 2

[13] A. Levin, S. W. Hasinoff, P. Green, F. Durand, and W. T. Freeman. 4d frequency analysis of computational cameras for depth of field extension. *ACM Trans. Graph.*, 28(3), 2009. 2

[14] M. Levoy and P. Hanrahan. Light field rendering. In *ACM. SIGGRAPH*, pages 31–42, 1996. 1, 2

[15] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu. Saliency detection on light field. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2806–2813, 2014. 2

[16] Y. Liu, Q. Dai, and W. Xu. A real time interactive dynamic light field transmission system. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 2173–2176. IEEE, 2006. 1

[17] W. R. Mark, L. McMillan, and G. Bishop. Post-rendering 3d warping. In *SI3D*, pages 7–16, 180, 1997. 2

[18] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Trans. Graph.*, 32(4):46, 2013. 1, 2

[19] B. Masia, G. Wetzstein, P. Didyk, and D. Gutierrez. A survey of computational displays. pushing the boundaries of optics, computation and perception. 37(8). 2

[20] K. Mitra and A. Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. pages 22–28, 2012. 2

[21] R. Ng. Fourier slice photography. *ACM Trans. Graph.*, 24(3):735–744, 2005. 2

[22] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11), 2005. 1, 2

[23] S. Pujades, F. Devernay, and B. Goldluecke. Bayesian view synthesis and image-based rendering principles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3906–3913, 2014. 2, 7, 8

[24] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2, 3, 4

[25] T. D. Sanger. Stereo disparity computation using gabor filters. *Biological Cybernetics*, 59(6):405–418, 1988. 2, 3, 6

[26] J. Shade, S. Gortler, L.-w. He, and R. Szeliski. Layered depth images. In *ACM SIGGRAPH*. 2

[27] H. Shum and S. B. Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing 2000*, pages 2–13, 2000. 2

[28] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *ICIP*, pages 444–447, 1995. 2, 3

[29] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman. Phase-based video motion processing. *ACM Trans. Graph.*, 32(4):80, 2013. 1

[30] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4D lightfields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 8

[31] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modelling and Visualization (VMV)*, 2013. 7

[32] G. Wetzstein, D. Lanman, M. Hirsch, W. Heidrich, and R. Raskar. Compressive light field displays. *IEEE Computer Graphics and Applications*, 32(5):6–11, 2012. 2

[33] B. Wilburn, N. Joshi, V. Vaish, E. Talvala, E. R. Antúnez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776, 2005. 1

[34] J. Wu, X. Lin, Y. Liu, J. Suo, and Q. Dai. Coded aperture pair for quantitative phase imaging. *Optics letters*, 39(19):5776–5779, 2014. 1

[35] F. Yu and D. Gallup. 3d reconstruction from accidental motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3986–3993, 2014. 1, 7, 8

[36] Z. Yu, C. Thorpe, X. Yu, S. Grauer-Gray, F. Li, and J. Yu. Dynamic depth of field on live video streams: a stereo solution. *Proc. of the 2011 Computer Graphics Int. Conf., CGI 2011*, 2011. 2