Computer Vision and Image Understanding 000 (2017) 1–11

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Optimized sampling for view interpolation in light fields using local dictionaries

David C. Schedl*, Clemens Birklbauer, Oliver Bimber

Johannes Kepler University Linz, Institute of Computer Graphics, Altenberger Strasse 69, Linz 4040, Austria

ARTICLE INFO

Article history:
Received 10 February 2017
Revised 12 May 2017
Accepted 21 June 2017
Available online xxx

Keywords:
Light fields
Sampling
View interpolation
Superresolution
Compressed sensing

ABSTRACT

We present an angular superresolution method for light fields captured with a sparse camera array. Our method uses local dictionaries extracted from a sampling mask for upsampling a sparse light field to a dense light field by applying compressed sensing reconstruction. We derive optimal sampling masks by minimizing the coherence for representative global dictionaries. The desired output perspectives and the number of available cameras can be arbitrarily specified. We show that our method yields qualitative improvements compared to previous techniques.

1. Introduction and contributions

Compared to standard digital photography, light fields offer various new options, such as refocussing, perspective changes, and 3D filtering as a postprocess. However, capturing them at an adequate resolution remains challenging. Popular approaches typically multiplex the 4D information onto a single 2D sensor, which results either in low spatial resolution, low angular resolution, or both. Multiple sensors (e.g., a camera array) can be used to overcome the aforementioned issue. However, achieving an adequate resolution in the angular domain requires a vast number of cameras, resulting in high construction costs and complexity.

In this paper, we present an angular superresolution approach for light fields captured with sparse camera arrays. We apply compressed sensing theory for reconstruction and find optimal sampling masks for a desired number of cameras and sampling grid resolution. In contrast to related work, we avoid the need for depth reconstruction, which often fails for non-Lambertian scenes. Compressed sensing has previously been applied to light fields (e.g., in Cao et al., 2014; Marwah et al., 2013). One of our contributions is the use of online learned local dictionaries extracted directly from the scene sampled with an optimized mask instead of using global dictionaries that are learned offline from a set of representative pre-recorded light fields. Therefore, our method yields superior reconstruction results compared to related techniques. A second contribution is that, in contrast to previous work (including our own Schedl et al., 2015), the number of samples is not constrained to the sampling pattern. Thus, our new approach allows to determine sampling masks for an arbitrary number of cameras. We compute coherence values for representative global dictionaries that provide a formal basis for estimating the reconstruction quality of a given sampling pattern. We find sampling masks by minimizing the coherence. Corresponding sampling masks are optimal with respect to the representative light fields used for training the global dictionary. Our method can be applied in situations where a high angular light-field resolution is desired, but camera arrays can only be constructed with a limited number of cameras (e.g., due to bandwidth limitations or high hardware costs). Our method is not suitable for light-field camera designs that do not support angular subsampling (e.g., single sensor microlens-array-based cameras).

The remainder of the paper is organized as follows: After discussing related and previous work in Section 2, we introduce mathematical notations and revisit compressive light-field reconstruction in Section 3. Section 4 describes the proposed coherence-based quality metric, the sampling pattern optimization with offline learned global dictionaries, and the reconstruction with online learned local dictionaries. While Section 5 focusses on parameter choices and implementation details, Section 6 is devoted to experimental results and evaluation. We conclude this article in Section 7 with a summary of limitations and future work.

* Corresponding author.
E-mail address: david.schedl@jku.at (D.C. Schedl).

http://dx.doi.org/10.1016/j.cviu.2017.06.009
1077-3142/

2. Related work

Compact light-field cameras often multiplex spatial and angular information on a single 2D sensor and thus suffer from either low spatial resolution, low angular resolution, or both. As a consequence spatial super-resolution methods for light fields have been proposed (Bishop et al., 2009; Boominathan et al., 2014; Georgiev and Lumsdaine, 2009). For camera arrays with multiple image sensors, spatial resolution is usually not an issue. However, high angular resolution requires a vast number of cameras, incurring high costs and complexity.

Angular super-resolution methods reduce the number of required cameras by reconstructing missing camera perspectives. Upsampling is applied to avoid undersampling artefacts and to enable smooth view transitions. For Lambertian scenes, depth reconstruction and subsequent view interpolation can be applied (Di et al., 2012; Kalantari et al., 2016; Kim et al., 2013; Mitra and Veeraraghavan, 2012; Pujades et al., 2014; Wanner and Goldluecke, 2014; Zhang et al., 2015).

Depth reconstruction works well for adequately textured isotropic content, but can fail for more realistic scenes with non-Lambertian, anisotropic, or completely uniform objects. Non-Lambertian content cannot be described sufficiently in 3D but requires additional information, as provided in 4D light-field recordings. Thus, we compare our approach to upsampling methods that do not rely on explicit depth reconstruction.

In Levin and Durand (2010), an approach called linear view synthesis was presented that can calculate novel views from a focal stack without depth information. However, it is limited mainly to Lambertian scenes, since a focal stack covers only a 3D subset of a full 4D light field. The same restriction applies to the method presented in Kubota et al. (2004), where a focal stack is computed for each new perspective, and an all-in-focus image is then extracted from the focal stack.

The approach described in Vagharshakyan et al. (2017) uses a shearlet transform to reconstruct sub-sampled epipolar-plane images of a light field, which does not require explicit depth reconstruction. However, reconstruction is still based on a Lambertian scene model; the authors discussed possible extensions to non-Lambertian scenes only as part of future work. Furthermore, their sampling mask is regular, while we optimize our mask and allow arbitrary irregular patterns.

In Shi et al. (2014), a method specifically targeted at non-Lambertian scenes was introduced which uses sparsity in the continuous Fourier domain to reconstruct light fields from a small number of 1D viewpoint trajectories in a camera array. Although the sampling mask is sparse, the method requires very specific sampling patterns with a fixed number of cameras for capturing. In contrast, we describe how to find an optimal sampling pattern for an arbitrary number of cameras and also show that we achieve higher reconstruction quality with the same number of cameras.

Recently, learning-based methods for light-field superresolution have been presented (Flynn et al., 2016; Kalantari et al., 2016; Yoon et al., 2015). The approach introduced in Yoon et al. (2015), for example, trains convolutional neural networks to upsample a light field in the spatial and angular domains. However, it requires a relatively dense and regularly sampled input, while our method supports sparse and irregular samples.

Methods in Flynn et al. (2016) and Kalantari et al. (2016) use sparse input samples but rely on depth layers or depth reconstruction. In Kalantari et al. (2016) two convolutional neural networks are applied—one for disparity estimation, and one for view interpolation. Therefore, these methods are limited to Lambertian scenes. Furthermore, in comparison to our approach, these learning-based techniques do not optimize sampling masks, but rely on manually defined sampling patterns.

The aforementioned methods can upsample sparse light fields but require regular sampling masks. Compressed sensing approaches use irregular sampling masks to encode additional information in a low-resolution recording. The methods presented in Marwah et al. (2013), Babacan et al. (2012), Ashok and Neifeld (2010), Mitra et al. (2014), Chen and Chau (2016), Miandji et al. (2015), Kamal et al. (2016) and YAO et al. (2014) place sampling masks in the optical path of standard cameras or compact microlens-based plenoptic cameras. Reconstructions of full light fields from the recordings are computed with sparse bases (e.g., DCT, trained global dictionaries, or Gaussian mixture models) and sparsity-aware optimization methods. We also use compressed sensing theory for reconstruction, but optimize the binary angular sampling pattern of a camera array instead of using (often continuous) optical sampling masks (which affect the spatial and angular domains). Compressed sensing in the spatial domain for camera arrays was presented in Kamal et al. (2012). Lambertian Gaussian mixture models, as used in Mitra and Veeraraghavan (2012) and Mitra et al. (2014), ignore anisotropic effects and transparencies. Corresponding methods require disparity estimations as an aditional preprocessing step. While it might be possible to reformulate the approach proposed in Mitra et al. (2014) to address the problem of choosing optimal camera sample locations, it is still limited to Lambertian scenes.

The methods in Cao et al. (2014) and Schedl et al. (2015) are the closest to our approach. Similarly, these techniques upsample light fields captured with a sparse camera array while avoiding depth information. Like the approach in Cao et al. (2014), our method uses compressed sensing techniques for reconstruction. However, we extended this idea by using local dictionaries extracted from a sub-sampled light field for reconstruction. Furthermore, we present methods for computing optimal sampling masks for an arbitrary number of cameras and sampling grid sizes.

Our previous method Schedl et al. (2015) already presented the idea of using higher-resolution guidance areas to support upsampling. In this article, we improved the reconstruction quality by using compressed sensing. Additionally, we present a method for computing optimal sampling configurations based on coherence values in a global dictionary and for an arbitrary number of cameras. In Schedl et al. (2015) we applied (empirically found) rules for estimating sampling masks that supported only specific numbers of cameras.

3. Mathematical notation and sparse light-field reconstruction

In this section, we introduce the mathematical notations that we will use throughout this article and revisit sparse light-field reconstruction with global dictionaries (e.g., Marwah et al., 2013).

We consider light fields captured with camera arrays and described by a regular two-plane parametrization, as discussed in Levoy and Hanrahan (1996). Thus, rays are parametrized by their intersections with two parallel planes: the camera plane UV (representing the angular domain), where the cameras are located, and the common image plane ST (representing the spatial domain), placed at a fixed distance from UV towards the objects to be captured. The indices u, v describe different camera positions on UV, and s, t address pixels in the captured perspective images $I_{u,v}$. We assume the light field to be regularly discretized and describe the ray intensities with the 4D matrix $\mathbf{L}$ (of size $S \times T \times U \times V$) or its vectorized 1D counterpart $\mathbf{l} = vec(\mathbf{L}) = [i_{0,0}, i_{0,1}, \ldots, i_{U,V}]^\top$), which contains a sequence of vectorized 1D versions of the captured perspectives ($i_{u,v} = vec(I_{u,v})$).

The goal of upsampling is to reconstruct a full light field $\mathbf{l}$ from its sub-sampled counterpart $\mathbf{l}' = \mathbf{\Phi l}$, which only contains a subsection of all captured perspective images $I_{u,v}$ described by the sampling matrix $\mathbf{\Phi}$. Since the size of $\mathbf{l}'$ is much lower than that of $\mathbf{l}$,

Please cite this article as: D.C. Schedl et al., Optimized sampling for view interpolation in light fields using local dictionaries, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.06.009