Unsupervised extraction of coherent regions for image based rendering

Jesse Berent and Pier Luigi Dragotti
Electrical and Electronic Engineering Department
Imperial College London
Exhibition Road, SW7 2AZ, UK
{jesse.berent, p.dragotti}@imperial.ac.uk

Abstract

Image based rendering using undersampled light fields suffers from aliasing effects. These effects can be drastically reduced by using some geometric information. In pop-up light field rendering [18], the scene is segmented into coherent layers, usually corresponding to approximately planar regions, that can be rendered free of aliasing. As opposed to the supervised method in the pop-up light field, we propose an unsupervised extraction of coherent regions. The problem is posed in a multidimensional variational framework using the level set method [16]. Since the segmentation is done jointly over all the images, coherence can be imposed throughout the data. However, instead of using active hypersurfaces, we derive a semi-parametric methodology that takes into account the constraints imposed by the camera setup and the occlusion ordering. The resulting framework is a global multidimensional region competition that is consistent in all the images and efficiently handles occlusions. We show the validity of the method with some captured multi-view datasets. Other special effects by coherent region manipulation are also demonstrated.

1 Introduction

The notion of plenoptic function [1] is central to image based rendering (IBR). It was introduced by Adelson and Bergen as a way of parameterizing the visual information available from any viewpoint, in any direction at any time and wavelength. IBR is essentially the problem of sampling and interpolating the plenoptic function. It is far from trivial to deal with the seven dimensions therefore several parameterizations have been proposed that reduce the dimensionality. The light field [13] or the lumigraph [9] are popular four dimensional parameterizations that limit the viewing space to a bounding box. Each light ray is usually parameterized by its intersection with two planes namely the image plane (u,v) and the camera location plane (s,t). Further reducing the dimensionality to three leads to a simplified light field [7] where the cameras are constrained to line. This setup is also known as the epipolar plane image (EPI) volume [4].

In conventional light field rendering, the data is highly sampled and a single depth plane is enough to render views with little or no aliasing. However, more geometric information is needed when the data is more sparsely sampled and occlusions occur. The tradeoff between geometry and sampling density was studied in [6]. Some IBR methods use complex geometry for rendering single objects such as in [9, 5]. However, these

BMVC 2007 doi:10.5244/C.21.28

models are sometimes difficult to obtain in scenes containing numerous objects and occlusions.

In pop-up light field rendering [18], the multi-view data is segmented into coherent layers. Each coherent region is represented with a collection of corresponding layers in each of the input images. These coherent layers are defined as regions in the light field that can be individually rendered free of aliasing. In [18], the segmentation is performed in a supervised fashion where the contours of layers are semi-manually defined on a key frame. These contours are then propagated to all the views using a user defined planar depth map. By performing the extraction in this way, the layers across all the image are coherent which enables a consistent rendering.

In this paper, we propose an unsupervised multidimensional analysis to extract coherent regions in multi-view data. Our method is based on the fact that objects in the scene carve out EPI-tubes [8] in EPI volumes and hypervolumes in light fields. Our approach therefore imposes coherence in all dimensions in order to extract these volumes or hypervolumes and generates a regular and approximate decomposition of the scene. The method presented is a novel multidimensional variational framework that explicitly takes into account the structure of the data. That is, the contours of the hypervolumes in the (u,v) dimensions (i.e. the images) are extracted using the level set method [16] which is a non parametric approach. However the dependencies between views (i.e. the two other dimensions (s,t)) and occlusions are constrained by the camera setup. They can therefore easily be parameterized. The resulting framework is a semi-parametric region competition that is global and hence uses all the available data jointly. This in turn insures that the segmentation is coherent across all the images and occlusions are efficiently and naturally handled. The main contributions of this work are the extraction of multiple coherent regions using a global variational framework and the formalization the method described in [3] using a more rigorous derivation of the gradient of the energy functional. Furthermore, the method is generalized to four dimensional light fields. Several other unsupervised layer extraction schemes have been proposed in the literature (e.g. [21, 2, 14, 22]). Not all these methods are scalable to higher dimensions and treat all the images equally. Moreover, most of these methods focus on the extraction of layers rather than the coherent volumes or hypervolumes carved out in multi-view data.

The paper is organized as follows: In Section 2 we study the shape constraints of the hypervolumes carved out by coherent regions in light fields. Section 3 derives a multidimensional variational framework based on active contours [12] for the extraction of the coherent regions and Section 4 shows some applications in image based rendering. In particular, view interpolation and some special effects are demonstrated. We conclude in Section 5.

2 Problem statement

Let $\vec{X} = (X,Y,Z)$ be a point in space, (u,v) be the image coordinates and (s,t) the position of the camera on a plane. Then, assuming a pinhole camera model [10], it can be shown that $\vec{X}$ is mapped onto the light field $L(u,v,s,t)$ according to

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \mapsto \begin{pmatrix} u \\ v \\ s \\ t \end{pmatrix} = \begin{pmatrix} X/Z - s/Z \\ Y/Z - t/Z \\ s \\ t \end{pmatrix}, \tag{1}$$

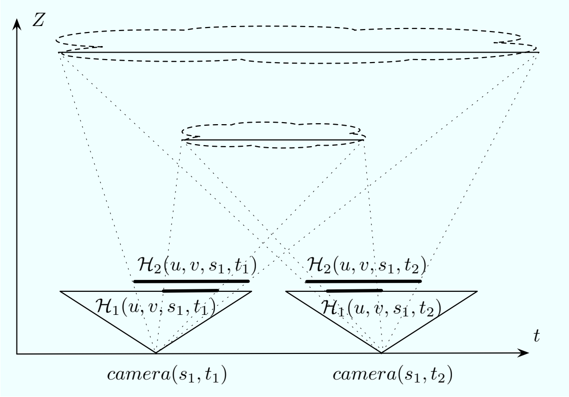

Z
$\mathcal{H}_2(u,v,s_1,t_1)$
$\mathcal{H}_2(u,v,s_1,t_2)$
$\mathcal{H}_1(u,v,s_1,t_1)$
$\mathcal{H}_1(u,v,s_1,t_2)$
t
$camera(s_1,t_1)$
$camera(s_1,t_2)$

Figure 1: Foreground and background regions in the scene are projected on to the image planes (u,v) as a function of the camera location (s,t). Both regions carve out hypervolumes $\mathscr{H}_1$ and $\mathscr{H}_2$ in the light field $L(u,v,s,t)$. The intersection of these hypervolumes with a plane in $(s,t) = constant$ corresponds to the view of the layer in that image.

where we assume that the focal length is unity. That is, a point in space is mapped onto a four dimensional trajectory in the light field hypervolume. Note that this trajectory as a function of (s,t) is a line. Under the assumption of Lambertian surfaces, the intensity along the line will remain fairly constant unless there is an occlusion. The effect of specularities in the EPI setup has been studied in [8, 20] although we will not take these effects into account here. Objects or layers that are made of neighboring points in space will be mapped onto neighboring trajectories. Gathering these trajectories generates a multidimensional hypervolume $\mathscr{H}_n$ in which the information is highly regular. Assume approximately constant depth layers in the scene. Then the boundary of the hypervolume $\vec{\Gamma}_n = \partial \mathscr{H}_n$ carved out by the layer in the light field can be parameterized as

$$\vec{\Gamma}_n(l,s,t) = \begin{pmatrix} u_n(l) - s/Z_n \\ v_n(l) - t/Z_n \\ s \\ t \end{pmatrix}, \tag{2}$$

where $\vec{\gamma}_n(l) = (u_n(l), v_n(l))$ is the contour of the layer on the image in $s = t = 0$ and Z_n is the depth. As illustrated in Figure 1, the intersection of the hypervolume with an image in $(s,t) = constant$ is the layer on that image. Note that, due to occlusions, the full hypervolume $\mathscr{H}_n$ is not always available. Therefore, we denote with $\mathscr{H}_n^\perp$ the available coherent region. From (1), we see that the slopes of the lines constituting the region are inversely proportional to the depth of the points. Hence a line with a higher slope will always occlude a line with a smaller slope. Assuming the hypervolumes are ordered from front $(n = 1)$ to back $(n = N)$, we can write

$$\mathscr{H}_n^\perp = \mathscr{H}_n \cap \sum_{i=1}^{n-1} \overline{\mathscr{H}_i^\perp}, \tag{3}$$

where $\overline{\mathscr{H}_n^\perp}$ denotes the outside of the region and $\mathscr{H}_n^\perp$ is the hypervolume geometrically orthogonalized with the other hypervolumes occluding it (i.e. where i goes from 1 to $n-1$). The problem of extracting coherent regions in the light field consists in estimating all the $\mathscr{H}_n^\perp$ using the available images. Note that the case of the simplified light field is straightforwardly obtained by posing $s = 0$. In the next section, we derive a fully unsupervised method to extract these hypervolumes in a global manner that takes advantage of the particular nature of the data.

3 A variational approach for extracting coherent regions in light fields

Since the seminal work of Kass et al. [12], active contours have been extensively used for numerous computer vision applications. Part of their success is due to the introduction of the level set method [16] which solved some of the issues such as numerical stability and topology dependance. Moreover, contrary to some other methods, the level set framework scales naturally to any number of dimensions [19]. This is a useful trait to analyze the data in light fields since it is parameterized with four dimensions.

Following this variational framework, we set the extraction of the coherent regions as an energy minimization problem. The initialization is performed in an unsupervised fashion by finding local directions in the EPIs and merging regions with similar slopes. The directions are found by locally estimating the angles of the EPI lines using a simple variance based method. The advantage of this approach consists in the fact that no reference image is used. The coherent regions $\mathscr{H}_n^\perp$ are then made dependent on an evolution parameter τ such that the derivative can be computed in order to evolve the hypervolumes in a steepest descent fashion. The total energy to minimize can be written in the form of a global region competition

$$E_{tot}(\tau) = \sum_{n=1}^{N} \int_{\mathscr{H}_n^\perp(\tau)} d_n(\vec{x})d\vec{x},$$

where the scene is segmented into N regions, $\vec{x} = (u,v,s,t)$ and the $d_n(\vec{x})$ are descriptors measuring the consistency with a depth Z_n. One way to minimize the E_{tot} consists in iteratively evolving one hypervolume while keeping the others fixed. In this case, the energy for each iteration is written as

$$E_{tot}(\tau) = \int_{\mathscr{H}_n^\perp(\tau)} d_n(\vec{x})d\vec{x} + \int_{\overline{\mathscr{H}_n^\perp(\tau)}} d_{out}(\vec{x})d\vec{x},$$

where all the other regions are gathered in $\overline{\mathscr{H}_n^\perp(\tau)}$ and $d_{out}(\vec{x}) = d_i(\vec{x})$ when $\vec{x} \in \mathscr{H}_i^\perp$ for all $i \neq n$. Region based active contour methods such as [11, 19] have shown that the gradient of the energy in this case can be written as

$$\frac{dE_{tot}(\tau)}{d\tau} = \int_{\partial \mathscr{H}_n^\perp} [d_n(\vec{\sigma}) - d_{out}(\vec{\sigma})](\vec{v}_{\Gamma_n} \cdot \vec{n}_{\Gamma_n})d\vec{\sigma},$$

where $\vec{v}_{\Gamma_n} = \frac{\partial \vec{\Gamma}_n}{\partial \tau}$ is the velocity, $\vec{n}_{\Gamma_n}$ is the outward unit normal vector, $d\vec{\sigma}$ is a differential surface element and $\cdot$ denotes the scalar product. Due to the occlusion ordering in (3), a foreground region evolving changes the background regions. That is, the evolution of $\mathscr{H}_n^\perp$ changes all the $\mathscr{H}_i^\perp$ where i goes from $n+1$ to N. Hence these occluded regions will contribute to the $\frac{dE_{tot}(\tau)}{d\tau}$ which leads to a competition. However, the other hypervolumes (i.e. $\mathscr{H}_i^\perp$ where i goes from 1 to $n-1$) are not affected by the shape changes in $\mathscr{H}_n^\perp$

and thus do not contribute to the derivative with respect to τ. Hence they will not compete. The steepest descent of the energy therefore yields the following partial differential equation

$$\vec{v}_{\Gamma_n} = [d_{out}(\vec{\sigma}) - d_n(\vec{\sigma})]\vec{n}_{\Gamma_n}$$

which will drive the evolution of the hypersurface in an unconstrained fashion. However, from Section 2, we know that the shape of the hypervolume in the light field is constrained. We can prove by doing an explicit derivation of the normal and velocity vectors using the parameterization in (2), that the normal velocity of the hypersurface $(\vec{v}_{\Gamma_n} \cdot \vec{n}_{\Gamma_n})$ projected onto the 2D subspace in $s = t = 0$ is equal to the normal velocity of the curve $(\vec{v}_{\gamma_n} \cdot \vec{n}_{\gamma_n})$ in one image[1]. The intuition behind this property is that given the fact that the region in space is approximately fronto-parallel, its contour will simply be a translated version of itself on all the images. The non fronto-parallel case is obtained by used a weighting term that compensates for the shape changes in the views. Using this constraint, the gradient of the energy can be rewritten as

$$\frac{dE_{tot}(\tau)}{d\tau} = \int_{\gamma_n} [D_n(l) - D_{out}(l)](\vec{v}_{\gamma_n} \cdot \vec{n}_{\gamma_n})dl,$$

where the $D_n(l)$ and $D_{out}(l)$ are the original descriptors integrated over the line delimiting the boarder of the hypervolume. This leads to a new evolution equation for the 2D contour

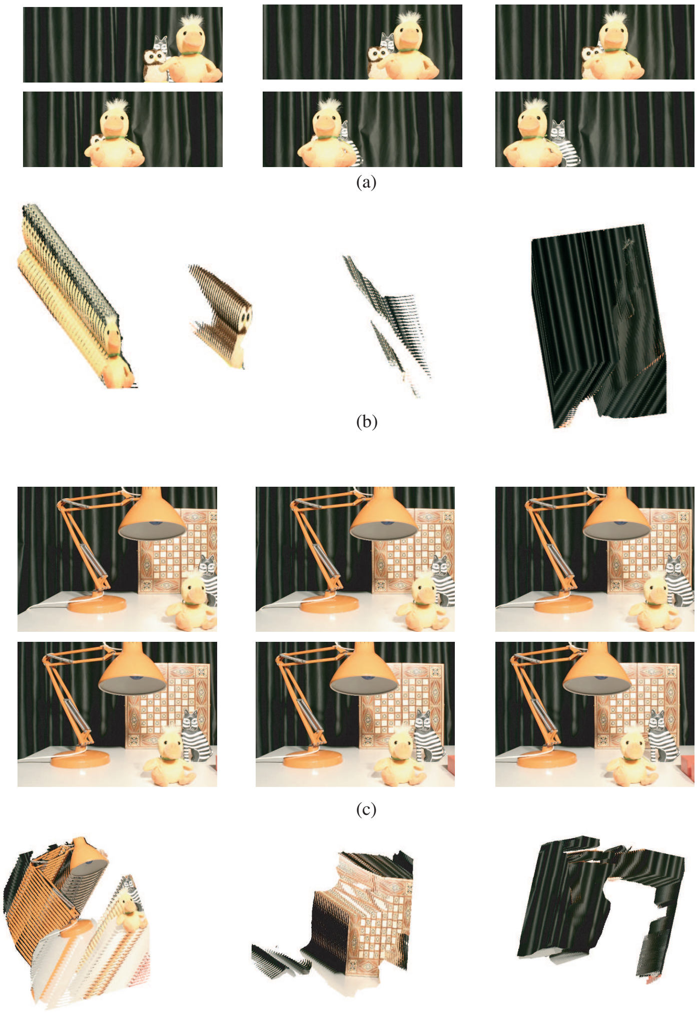
$$\vec{v}_{\gamma_n} = [D_{out}(l) - D_n(l) - \mu\kappa]\vec{n}_{\gamma_n}, \tag{4}$$

where a smoothness term proportional to the curvature κ is added in order insure regular curves. The μ is a constant weighting factor determining the influence of the smoothness term. The advantage of such a manipulation consists in the fact that the constraints are fully taken into account (i.e. points are mapped onto lines) and it can be implemented as an active contour instead of an active hypersurface which reduces the computational complexity. We perform the overall energy minimization after initialization in three iterative steps: First, estimate the slopes of the lines in each individual hypervolume $\mathscr{H}_n^{\perp}$ using classical non-linear optimization methods. Second, update the occluded regions using the occlusion constraint in (3). Third, evolve each boundary $\vec{\gamma}_n$ individually using the evolution equation in (4) and the level set method. The algorithm is stopped when there is no significant decrease in the total energy.

Some results illustrating the coherent region extraction in simplified light fields are presented. These are easier to capture and more practical to show the results while still portraying the concepts well. We use the variance along an EPI line as a descriptor to measure the consistency with a particular region. Only the visible regions of each volume are taken into account to compute this variance. Thanks to the competition formulation, we use only two parameters in the segmentation. The first one μ regulates the smoothness of the extracted regions and second one in the initialization determines how close the estimated local directions need to be in order to be merged. This will in turn determine the number of regions to extract. The algorithm takes approximately 1000 seconds using the classical level set implementation to extract the coherent regions for both natural datasets analyzed. Improvements of several orders of magnitude are expected using more efficient implementations [16]. Figure 2 illustrates some of the images[2] and the final segmentations. Due to the absence of a ground truth, we provide a qualitative assessment. The

[1] We omit this proof due to lack of space.

[2] The authors would like acknowledge the Audiovisual Communications Laboratory at the Swiss Federal Institute of Technology (EPFL) for providing the equipment to capture these datasets.



(a)

(b)

(c)

(d)

Figure 2: Automatically extracted coherent regions in simplified light fields. The first data set *animal family* consists of 32 images (235x625) some of which are illustrated in (a). The extracted regions at different depths are depicted in (b). The second data set has 25 images (400x500) some of which are illustrated in (c). The extracted volumes corresponding to the foreground, intermediate and background regions are depicted in (d).

regions are correctly captured despite the occlusions and disocclusions. The volumes also do not show the discontinuities across scan lines that may occur when individual EPIs (i.e. slices of the volume) are analyzed. All these aspects come from the fact that the data is analyzed in a global manner and coherence is enforced in all the dimensions. Some errors occur in large textureless regions due to the absence of a color based term in the energy minimization. These errors, however, will not significantly hinder the rendering quality.

4 Applications in image based rendering

The extraction of coherent regions in light fields is a useful step in numerous image based rendering applications. In this section, we demonstrate a few of these applications such as view interpolation and scene manipulations for object removal and insertion.

4.1 View interpolation

In the original light field rendering [13], the data is highly sampled and virtual viewpoints are rendered using classical quad-linear interpolation. This is equivalent to using a focal plane at infinite depth. In undersampled light fields however, this method causes blurring due to the aliasing that occurs. This sampling problem has been studied by Chai et al. [6] using a classical signal processing framework. The authors show that, due to the particular structure of the data, the band is approximately bound by the maximum and minimum depths in the scene. Moreover, it is shown that that there is an optimal rendering depth that will generate the new viewpoint with the least possible aliasing. However, the rate required often becomes very large when the scene has somewhat large depth variations. Furthermore, inherent effects such as occlusions are concentrated in space and hence contribute to spread the band of the function. Similarly to rendering with layers [15], the individual depth variations are much smaller in each of the extracted coherent volumes and an aliasing free rendering for each region can be achieved with less cameras. In order to demonstrate the reduced aliasing, we render a viewpoint using linear interpolation. Figures 3(a) and (c) illustrate rendered viewpoints using a single plane at the optimal depth in [6]. Note that some aliasing occurs in regions that are far from this depth. The images in Figures 3(b) and (d) illustrate the same rendered viewpoints generated by individually interpolating each of the extracted coherent regions. That is, the knowledge of occlusions is used and each of the regions are interpolated using their estimated constant depths. The aliasing is greatly reduced while keeping the natural aspect of the images.

4.2 Object manipulation

Similarly to layer based representations [21, 17], the extraction of the coherent regions enables to manipulate the data by recombining them in different ways. Note that a layer based representation (i.e alpha map, texture and plane or motion parameters) is not used as such. Rather, the new scenes are rendered using the full volumes. This enables to overcome some of the artifacts caused when over simplified depth models are used. The first example depicted in Figure 4 illustrates how the available coherent region $\mathscr{H}_n^{\perp}$ as shown in Figure 4(a) can be extrapolated to obtain the full region $\mathscr{H}_n$ shown in Figure 4(b). Linear extrapolation is performed using the available lines and their intensities. That is, each EPI line is simply extended along its slope.

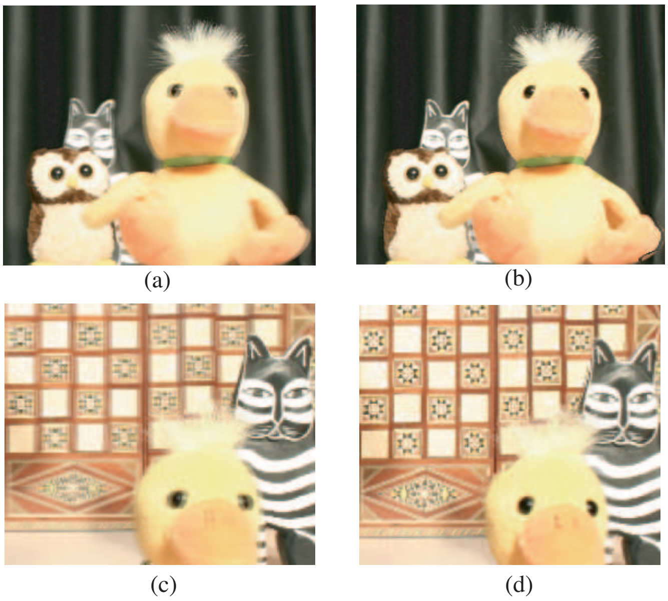


Figure 3: Image based rendering with coherent regions. The images in (a) and (c) show the rendered viewpoints using a single plane at the optimal depth. The images in (b) and (d) illustrate the same viewpoints rendered using the individual estimated constant depths for each of the extracted coherent regions.

Alternatively, the scene may be reconstructed by combining the existing coherent regions with external ones. These can be captured by other camera arrays or synthetically generated in order to perform augmented reality. As an example, we use a CAD software to generate the synthetic images of a teapot shown in Figure 5(a). The visible parts of the coherent region are determined using equation (3) which leads to the orthogonalized volume depicted in Figure 5(b). When combined with the original data, all the occlusions are naturally handled. Some of the rendered images are illustrated in Figure 5(c). Although the regions were extracted using fronto-parallel models, the objects still show their original shapes in the rendered images. This is noticeable, for instance, in the duck's beak and its shadow that are accurately rendered.

5 Conclusion and future work

In the first part of the paper, we proposed a novel multidimensional scheme based on a variational framework for the extraction of coherent regions in light fields. Since the formulation is global, coherence and consistency is enforced on all the dimensions. The method presented also takes fully advantage of the constraints namely that points in space are mapped onto particular trajectories (i.e. lines in light fields) and occlusions occur in specific order. Therefore occlusions are naturally handled and all the images are treated equally. The second part illustrated some applications in image based rendering using the extracted regions. View interpolation and scene manipulation as well as augmented reality were demonstrated in order to illustrate the benefits of extracting these regions. Since the scene is not represented as layers but as a combination of coherent hypervolumes,

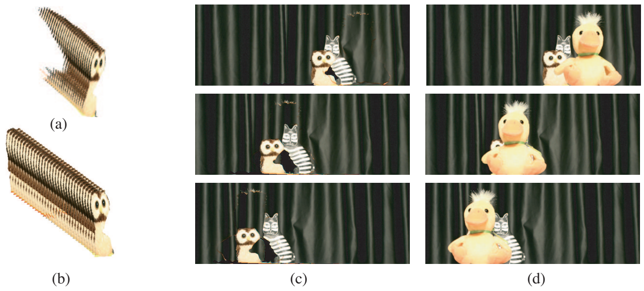


Figure 4: Coherent region extrapolation. The extracted volume illustrated in (a) is extrapolated along its EPI lines in order to reconstruct the occluded regions as shown in (b). This enables to reconstruct the scene, for example, by removing the region carved out by the duck and using the extrapolated background volumes (c). Note that there are some holes since some regions are never visible in the entire stack of images. The images in (d) show the original data for comparison.

accurate geometry is often not necessary and fronto-parallel depth models were used to extract the regions. More general depth models may also be incorporated into the energy minimization.

An interesting direction for future work is to explore the potential of the global formulation to automatically extract coherent regions in dynamic and unstructured light fields. This would imply using the constraints and imposing coherence in the four spatial dimensions as well as the time dimension.

References

[1] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. In M. Landy and J. Anthony Movshon, editors, *Computational Models of Visual Processing*, pages 3–20. MIT Press, Cambridge, MA, 1991.

[2] S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 434–441, 1998.

[3] J. Berent and P. L. Dragotti. Segmentation of Epipolar Plane Image volumes with occlusion and dissocclusion competition. In *IEEE Int. Workshop on Multimedia Signal Processing*, pages 182–185, October 2006.

[4] R.C. Bolles, H. H. Baker, and D.H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *Int. Journal of Computer Vision*, 1:7–55, 1987.

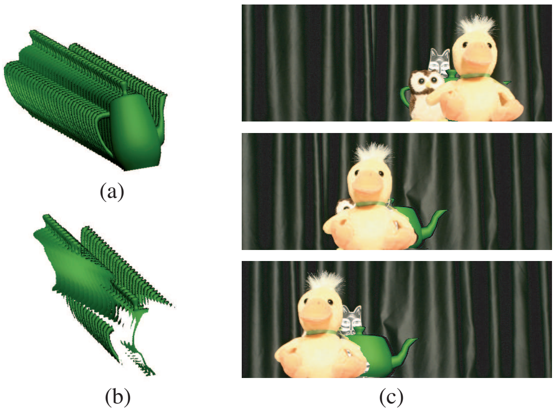
[5] C. Buehler, M. Boss, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *Computer graphics (SIGGRAPH '01)*, pages 425–432, 2001.

[6] J.-X. Chai, S.-C. Chan, H.-Y. Shum, and X. Tong. Plenoptic sampling. In *Computer graphics (SIGGRAPH '00)*, pages 307–318, 2000.

[7] S. C. Chan, K. T. Ng, Z. F. Gan, K. L. Chan, and H. Y. Shum. The plenoptic video. *IEEE Trans. on Circuits and Systems for Video Technology*, 15:1650–1659, 2005.

[8] A. Criminisi, S. B. Kang, R. Swaminathan, R. Szeliski, and P. Anandan. Extracting layers and analyzing their specular properties using epipolar-plane-image analysis. *Computer Vision and Image Understanding*, 97(1):51–85, January 2005.

[9] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *Computer graphics (SIGGRAPH '96)*, pages 43–54, 1996.



(a)

(b)

(c)

Figure 5: Image based object insertion. The synthetically generated coherent region carved out by a teapot (a) is geometrically orthogonalized with the existing extracted volumes (b). It is then straightforward to recombine the coherent regions to recreate a scene where the teapot is inserted (c).

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[11] S. Jehan-Besson, M. Barlaud, and G. Aubert. DREAMS: Deformable regions driven by an Eulerian accurate minimization method for image and video segmentation. *Int. Journal of Computer Vision*, 53(1):45–70, 2003.

[12] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. Journal of Computer Vision*, 1(4):321–331, 1988.

[13] M. Levoy and P. Hanrahan. Light field rendering. In *Computer graphics (SIGGRAPH '96)*, pages 31–42, 1996.

[14] M. Lin and C. Tomasi. Surfaces with occlusions from layered stereo. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 710–717, 2003.

[15] Z. Lin and H.-Y. Schum. A geometric analysis of light field rendering. *International journal of computer vision*, 58(2):121–138, 2004.

[16] J. Sethian. *Level Set Methods*. Cambridge University Press, 1996.

[17] J. Shade, S. Gortler, L. W. He, and R. Szeliski. Layered depth images. In *Computer graphics (SIGGRAPH '98)*, pages 231–242, 1998.

[18] H. Y. Shum, J. Sun, S. Yamazaki, Y. Li, and C. K. Tang. Pop-up light field: An interactive image-based modeling and rendering system. *ACM Trans. Graph.*, 23(2):143–162, April 2004.

[19] J.E. Solem and N.C. Overgaard. A geometric formulation of gradient descent for variational problems with moving surfaces. In *Int. Conf. on Scale Space and PDE methods in Computer Vision, Scale Space 2005*, 2005.

[20] T. Stich, A. Tevs, and M. Magnor. Global depth from epipolar volumes - a general framework for reconstructing non-lambertian surfaces. In *3DPVT*, pages 1–8, 2006.

[21] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Trans. on Image Processing Special Issue: Image Sequence Compression*, 3(5):625–638, September 1994.

[22] C. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *Computer graphics (SIGGRAPH '04)*, pages 600–608, 2004.