Efficient Segmentation and Representation of Multi-View Images

Pier Luigi Dragotti and Mike Brookes
Electrical and Electronic Engineering Department, Imperial College London, Exhibition Road, London, SW7 2BT, UK

Abstract

We study the structure of multi-view images and introduce the notion of plenoptic hyper-volumes. We then present a segmentation algorithm based on the level set method to extract such hyper-volumes. .

Keywords: Plenoptic Function, Multiview Imaging, Image Segmentation

1. Introduction

Multi-view camera systems have attracted a lot of attention in recent years thanks in a large part to the dropping prices of memory and digital cameras. Multi-view systems lead to new challenging problems such as the sheer amount of data involved. Traditional algorithms do not scale properly with the number of cameras and become impracticable when the number of images acquired is large. Efficient representation and analysis methods are therefore a primordial issue. Thankfully, as we will show in the paper, the multi-view data has a very particular structure and a high degree of regularity that can be used to achieve an efficient representation of such information.

The visual information captured from any viewpoint in any direction, time and wavelength can be parameterized in a single seven dimensional function called the plenoptic function [1]. Consider the particular 3D case of the video or the space-time volume. A moving object carves out a 3D volume and the information inside it is highly regular. This is the observation reported by Ristivojevic and Konrad in [2] where they introduce object tunnels. Similarly, consider another 3D case of the plenoptic function where the time dimension is replaced by letting the viewing position move along a line. This is the case of a set of multi-baseline images or the Epipolar Plane Image (EPI) [3]. Volumes are carved out by objects at different depths in very much the same way as in the video and this was reported in [4] where Criminisi et al. introduce EPI tubes. Both cases reveal that there is a potential gain in segmentation accuracy and reliability by analyzing all the data in a single multidimensional function especially in the case of occlusions. In an effort to generalize the notion to all the dimensions of the plenoptic function, we introduce the *plenoptic hyper-volumes* and propose a hyper-volume extraction scheme based on active contours that is scalable to higher dimensions as well as takes into account the particular structure of the data.

Thanks to their ability to exploit coherence, the extraction of plenoptic hyper-volumes is a very useful step for applications such as multi-view layer based representations, MPEG-4 like object based coding and disparity compensated and shape adaptive coding of multi-view data. Image Based Rendering (IBR) in which sampling with truncated windows is used also stands to benefit from the segmentation. Finally, extracted hyper-volumes also allow for scene understanding, occlusion detection and object classification.

The paper is organized as follows: In Section 2 we analyse the structure of multi-

view data, introduce the notion of plenoptic hyper-volumes and discuss their shape constraints. Section 3 proposes a variational framework for the extraction of the volumes and derives constrained surface evolutions. Experimental results are shown in Section 4 and we conclude in Section 5.

2. Structure of Multiview Data

The plenoptic function was introduced by Adelson and Bergen in order to characterize general free-viewpoint vision. The idea is to describe the intensity of each light ray that reaches a point in space. It can therefore be characterized by seven parameters namely the visual angle, the wavelength, time and the viewing position:

$$P_7 = P(q, f, l, t, V_x, V_y, V_z)$$

Figure 1 shows the concept where a camera symbolizes the viewing point. Intuitively, we see that the camera has 3 degrees of freedom (dof) for its position in space and itself has 2 dof to address the pixels of the image. With two more parameters, namely time and wavelength, it is possible to characterize any light ray. The general function is difficult to analyze due to its high number of dimensions. However, certain valid assumptions can be made in order to reduce its complexity. First, we simplify the wavelength into three channels for red, green and blue or one channel for greyscale images. Second, we consider that air is transparent, thus intensity does not change along a light ray unless it is occluded. Third, we limit ourselves to static scenes and drop the time parameter.

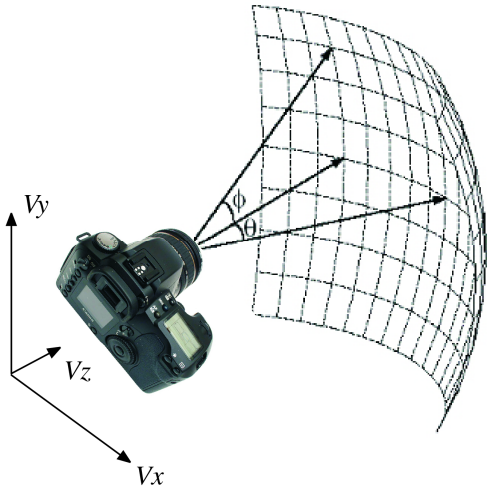

Figure 1: The plenoptic function describes the intensity of each light ray that reaches any point in space at any time. It is therefore characterized by 7 parameters, namely the viewing position, the viewing direction, time and wavelength.

Despite its apparent complexity, the plenoptic function is in fact a highly structured function especially in the case where the viewing points are constrained and can be parameterized.

Consider, for example, the case of a linear multi-camera system, that is, the case where the viewing position is constrained to be along a straight line. This setup is illustrated in Figure 3. The plenoptic function sampled by this array is characterized by three dimensions namely the two dimensions *x* and *y* of the images and the location V_x of the camera along the line. Using a projective camera model, it is straightforward to show that points in space are projected onto lines in the plenoptic function and that the slope of the line is inversely proportional to the depth of the point. Lines with higher slopes therefore always occlude lines with smaller ones. The shape carved in the plenoptic domain by the cube of Figure 2 is shown in Figure 3. In line with the work of Adelson and Bergen, we call the volume carved by the object *plenoptic hyper-volume*.

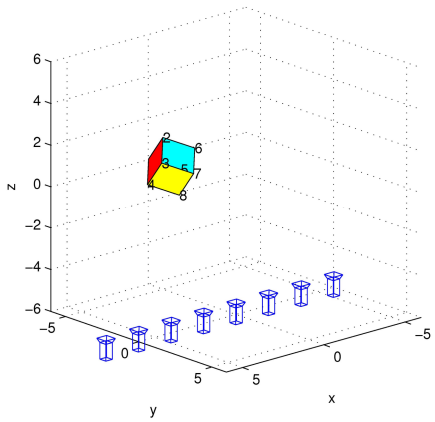

Figure 2: Linear camera array. The *X,Y* and *Z* coordinates correspond to the real world.

Similar intuitions apply to other camera setups such as the circular case illustrated in Figure 4 and Figure 5. This setup was studied in [5] where Feldmann et al. introduce Image Cube Trajectories. They show that just like in the linear array parameterization, points in space are projected on to particular trajectories in the plenoptic function and occlusion compatible orders can be defined.

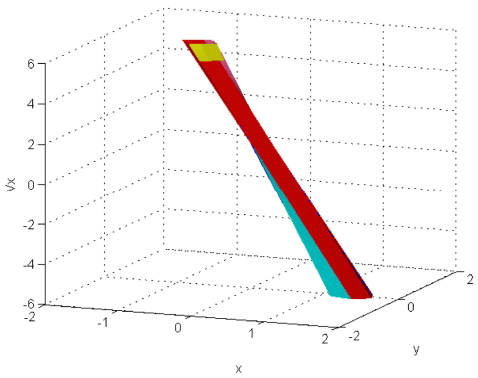

Figure 3: Structure of the plenoptic function. The shape of the plenoptic volume carved by an object or a layer is constrained by the camera setup.

In summary, in all the parameterizations of the plenoptic function, gathering a collection of lines that do not intersect generates a volume or a hypervolume n_n in which the information is highly regular. Notice that this usually corresponds to an object or a layer in the scene. The occlusion compatible order determines how occluding volumes carve through the background ones. By ordering the volumes from front to back, we can write:

$$n_n^{\perp} = n_n \cap \sum_{i=1}^{n-1} \overline{v_i^{\perp}}$$

where n_n is the hyper-volume as if there was no occlusion and $^{\perp}$ denotes that the volume has been geometrically orthogonalised with the other volumes occluding it. Higher dimensional hyper-volumes are generated in the same manner for higher dimensional plenoptic functions. In the case of the light field parameterization [6], for instance, the cameras are constrained to a plane and 4D hypervolumes are carved out by the objects.

In an attempt to extract these volumes or hypervolumes, we set ourselves in a variational framework.

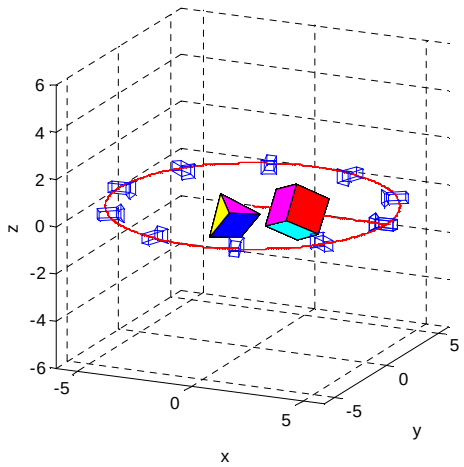

Figure 4: Circular camera array. The *X,Y* and *Z* coordinates correspond to the real world.

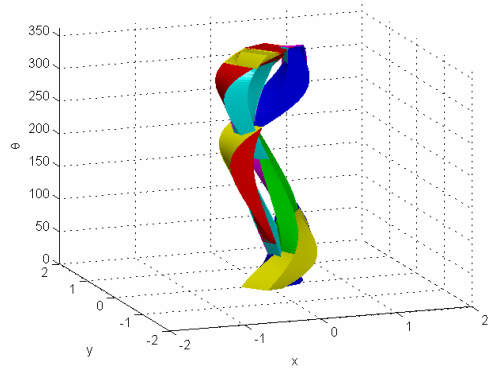

Figure 5: The plenoptic function generated by two objects in the circular setup. In this context occlusion can be predicted.

3. Extraction of Plenoptic Hyper-volumes Using a Variational Framework

Since the seminal work of Kass et al. [7], active contours have been used for numerous image and video segmentation schemes. It was rapidly noticed that the same principles can be extended to active surfaces and were used amongst other applications for space-time sequence analysis [2]. The methodology is also extendable to higher dimensions thus making it ideal for the segmentation of the plenoptic function.

In the next subsection, we briefly review the level set method which is a form of active-contour segmentation technique, we then present, in the following sub-section, our segmentation approach that takes into account the geometrical (epipolar) constraints and the occlusion ordering.

3.1 A Glimpse at the Level Set Method

There are numerous object based segmentation methods for still images and video. In the case of still images, the main criteria to segment objects are intensity gradients. One approach is to use the level set methodology [8] to grow surfaces with a speed inversely proportional to the image gradient. The level set method has been used in region-based video segmentation [2][9]. The main idea is to use the level set method in order to minimize a certain energy functional which is usually a measure of variance along a motion trajectory. The curve stops at large gradients in the motion boundary map.

The key idea of the level set method is to represent a closed curve $\Gamma(t)$ as the zero level set of a 3D surface $f(x,y,t)$, i.e. $\Gamma(t)=\{(x,y)\,|\,f(x,y,t)=0\}$. The higher dimension function f may be set to be the signed distance function of $\Gamma(t)$. In this case, we have $\|\nabla f\|=1$. The curve is grown according to the partial differential equation:

$$f_t + F|\nabla f| = 0$$

where F is the 'speed function'. This method provides several added advantages [8]. First, the curve may break or merge providing better handling in the case of topological changes. Second, the geometrical properties of the curve like curvature and normal vectors can be computed directly from the level set function. Third and most important in our case, the methodology allows for efficient numerical implementation in high dimensions.

4. A Variational Framework for Plenoptic Hyper-volumes Extraction

Without loss of generality, we derive a set of constrained evolution equations for two volumes only $n_1^{\perp}=n$ and $n_2^{\perp}=\bar{n}$ in a 3D plenoptic function where the cameras are constrained to a line (i.e. the EPI volume).

Following the variational framework, we set the extraction of plenoptic hyper-volumes as an energy minimization problem. The functional we seek to minimize can be written in the form:

$$E_{tot}(t) = \iiint_{n(t)} f(\mathbf{x})d\mathbf{x} + \iiint_{\bar{n}(t)} g(\mathbf{x})d\mathbf{x}$$

where the $f(\mathbf{x})$ and $g(\mathbf{x})$ are descriptors measuring the consistency with the front and back volumes respectively. Assuming opaque Lambertian surfaces, popular choices for the descriptors are the variance along EPI lines or cross-correlation. Notice that in order to use correspondences, these descriptors depend on depth however we assume for the moment that depth is known. Using the Euler Lagrange equations or Eulerian derivatives, it can be shown that the gradient of the energy is given by [7,9]

$$\frac{dE_{tot}(t)}{dt} = \iint_{\partial n} [g(\mathbf{x}) - f(\mathbf{x})](\mathbf{W}\cdot\mathbf{M})d\mathbf{s},$$

where ∂n is the border of the volume, $d\mathbf{s}$ is a differential surface element, $\mathbf{W}$ is the speed of the evolving interface and $\mathbf{M}$ is

its inward unit normal. Following the classical derivation, evolving the surface in a steepest descent fashion leads to the evolution equation $\vec{W} = [f(\vec{x}) - g(\vec{x})]\vec{M}$. However, this evolution does not fully take advantage of the plenoptic constraints imposed by the camera setup.

The shape of the plenoptic volume carved out by an object is constrained by the camera setup. In the EPI case as illustrated here, the volumes are constrained to tubes. It is therefore possible to write the 3D normal speed function $\vec{W} \cdot \vec{M}$ as a function of the 2D normal speed $\vec{V} \cdot \vec{N}$ of the curve at $V_x = 0$. Namely, the curve related to the first image in the stack of images. More precisely, we have that:

$$\vec{W} \cdot \vec{M} = (\vec{V} \cdot \vec{N}) a(s,t)$$

where $a(s,t)$ is a weighting factor depending on the depth map of the object or layer and the camera setup. Using this relation, we can rewrite the gradient of the energy as:

$$\frac{dE_{tot}(t)}{dt} = \int_{\partial\Omega V_x} \int [g(\vec{x}) - f(\vec{x})](\vec{W} \cdot \vec{M}) dV_x ds$$
$$= \int_{\partial\Omega} [G(s) - F(s)](\vec{V} \cdot \vec{N}) ds.$$

We now have an evolving curve in a two dimensional subspace where the speed function is essentially the original descriptor integrated over the line constituting the boarder of the volume. It is implemented as active contour instead of an active surface with evolution equation:

$$\vec{V} = [F(s) - G(s)]\vec{N}.$$

The estimation of the contours delimiting the plenoptic hyper-volume as described above requires the knowledge of the depth of the layer or the slope of the lines in the case of the EPI. We model the depth map as a linear combination of bicubic splines. The weights of the splines are determined by minimizing the energy functional where the shape of the contours is kept constant. In order to perform the minimization, we use non linear optimization methods such as the ones in Matlab's optimization toolbox. There are several advantages to this particular depth model. First, a great variety of smooth objects can be modelled. Second, only a limited amount of weights on control points need to be estimated depending on the lattice size. Finally, the depth map can be forced to have a certain shape. For instance, strictly fronto-parallel regions can be extracted by forcing all the weights to be the same for a given layer.

The overall optimization is performed by iteratively alternating depth estimation given the contour of the volume and estimation of the contour given depth until there is no significant decrease in energy. In the case of multiple volumes in a scene we perform one iteration of the evolution for each hyper-volume while keeping the other contours fixed. It is interesting to notice that by the volume construction, the plenoptic hyper-volumes only compete with the other volumes they are occluding or disoccluding. The intuition behind this property is that the evolution of an occluding layer changes the background one (i.e. the background is more or less occluded) however the rear layer just evolves behind the front one.

It is also worth mentioning that like all partial differential equation based methods, active contours require careful initialization. Classical block matching or stereo computer vision methods can be used.

5. Simulation Results

In this section, we illustrate some results for real multi-view data. In these results, the descriptor used is the variance along an EPI line and the contour evolution is performed using the level set method. There are several advantages to using this method over the classical active contour method. These include independence of

topology and numerical stability. We refer to [8] for a detailed discussion.

Figure 6: Multi-view dataset. Here we show the first and the last of the 15 multi-view images acquired.

Figures 6-8 illustrate the extracted plenoptic hyper-volumes in the case of real calibrated images. In order to segment regions at different depths, we force the depth model of the hyper-volume to be fronto-parallel therefore all the lines constituting the boarder of the volumes are parallel. Notice that the feet and the nose of the tiger are at a different depth than the body and the face. This is why separate volumes are extracted. Initialization is performed using block matching where blocks with similar motion parameters are merged.

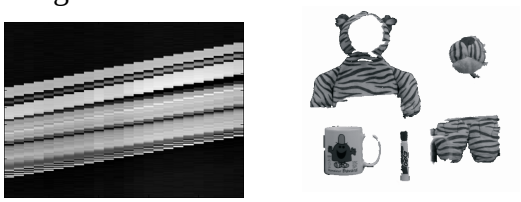

Figure 7: Left, one slice of the plenoptic function generated by the dataset of Figure 6. Right, layers extracted from the plenoptic function.

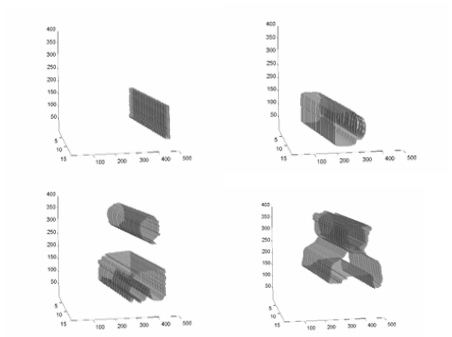

Figure 8: The four hypervolumes extracted with your segmentation method.

6. Conclusions

We have proposed a segmentation algorithm for multi-view images that is based on 3D space continuity and that takes into account occlusions explicitly. Using epipolar geometry, we reduce the 3D problem of segmenting the multi-view images into a 2D curve evolution. The speed that governs the curve evolution however is computed using the whole stack of images. The main contribution of the scheme presented lies in the competition formulation that enables a global energy minimization instead of extracting layers individually.

The extraction of the plenoptic hyper-volumes that is achieved with this algorithm is an attractive step for numerous multi-view imaging applications. In particular, we aim to study the use of such hyper-volumes for object recognition and classification.

References

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

[2] M.Ristivojevic and J.Konrad, "Space-time image sequence analysis: object tunnels and occlusion volumes," *IEEE Trans. on Image Processing*, vol. 15, no.2, pp.364-376, February 2006.

[3] 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*, vo.1, pp. 7-55, 1987.

[4] 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*, vol.97, no.1, pp. 51-85, January 2005.

[5] I. Feldmann, P. Eisert and P. Kauff, "Extension of epipolar image analysis to circular camera movements," in *Int. Conf. on Image Processing*, pp.697-700, September 2003.

[6] M. Levoy and P. Hanrahan. “Light field rendering”. In *Computer Graphics (SIGGRAPH’96)*, pages 31–42, 1996.

[7] M. Kass, A. Witkin, and D. Terzopulos, “Snakes: Active contour models,” *Int. Journal of Computer Vision*, vol.1, no.4, pp.321-331, 1988.

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

[9] S. Besson, M. Barlaud, and G. Aubert. “Detection and tracking of moving objects using a new level set based method”. In *Proc International Conference on Pattern Recognition*, Vol 3, pages 1100-1105, September 2000.

Acknowledgements

The work reported in this paper was funded by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence. This paper contains work done jointly with Jesse Berent (ICL) and Yizhou Wang (ICL).