Unsupervised Representation and Understanding of Multi-View Images

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

Abstract

We study the structure of multi-view images and review the notion of plenoptic hyper-volumes. We then present a segmentation algorithm based on the level set method to extract such hyper-volumes and show how this segmentation can be used to dis-occlude hidden objects and to facilitate the understanding of the monitored scene. Finally, we highlight preliminary results on classification and clustering of plenoptic hyper-volumes based on a multi-dimensional version of the SIFT algorithm.

Keywords : Plenoptic Function, Multiview Imaging, Image Segmentation

1. Introduction

The data acquired by densely sampled multi-view camera systems is highly regular with the exception of object boundaries. It is therefore beneficial in numerous multi-view imaging applications to extract these boundaries in order to exploit the regularity of the data inside and outside them. Traditional algorithms for the analysis of multi-view data 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 paramount issue. In this paper, we present a segmentation scheme based on a variational framework to extract these boundaries in a way that takes into account the nature of the data, the camera setup and occlusions. We then propose a compact representation of such extracted boundaries based on the Scale Invariant Feature Transform (SIFT) presented in [1].

The visual information captured from any viewpoint in any direction, time and wavelength can be parameterised in a single seven-dimensional function called the plenoptic function [2]. 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 [3] 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) [4]. Volumes are carved out by objects at different depths in very much the same way as in the video and this was reported in [5] 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 generalise the notion to all the dimensions of the plenoptic function, we have introduced in [6] the notion of *plenoptic hyper-volumes* and proposed a hyper-volume extraction scheme based on active contours that are scalable to higher dimensions as well as taking 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

scene understanding, occlusion detection and object classification.

The paper is organised as follows: In section 2 we review the structure of multi-view data and review the notion of plenoptic hyper-volumes. Section 3 presents a variational framework for the extraction of the volumes and derives constrained surface evolutions. Experimental results are shown in section 4, in particular, we present some results on the extraction of hyper-volumes from a lightfield data-set (i.e., from 4-D data). This is done in order to highlight the ability of our algorithm to operate on high dimensional data. We also show that it is possible to make partially occluded object visible. In section 5, we present a multidimensional version of SIFT for compact representation of the plenoptic hyper-volumes and we conclude in section 6.

2. Structure of Multiview Data

The plenoptic function was introduced by Adelson and Bergen in order to characterise general free-viewpoint vision [2]. The plenoptic function corresponds to the function representing the intensity and chromaticity of the light observed from every position and direction in the 3-D space, and can therefore be parameterised as a 7-D function:

$$P_7 = P(\theta, \phi, \lambda, t, V_x, V_y, V_z).$$

The three coordinates V_x, V_y, V_z correspond to the position of the camera, θ and ϕ give its orientation, t is the time and λ corresponds to the frequency considered. The high dimensionality of this function, however, makes it difficult to handle.

If there is no restriction on the position of the cameras, but we fix the time t and the frequency λ (i.e. grayscale images or separate RGB channels), we obtain a 5-D representation of the plenoptic function. Moreover, camera positions can be constrained to a plane, a line or a point to remove one, two or three additional dimensions respectively.

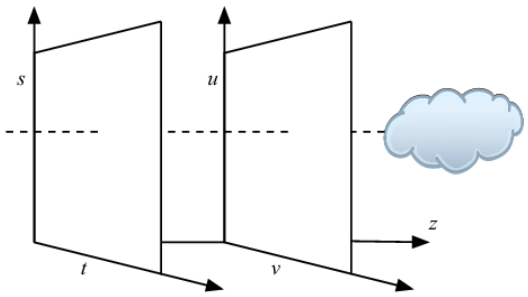

Figure 1: 4D light field (or light slab) parameterization. A light ray is uniquely characterised by its intersection between the camera plane (s,t) and the image plane (u,v). Each light ray is therefore addressed by its coordinate (u,v,s,t)

Figure 2: Planar camera array. (Note that this multi-camera system is supported by SEAS-DTC and The Royal Society)

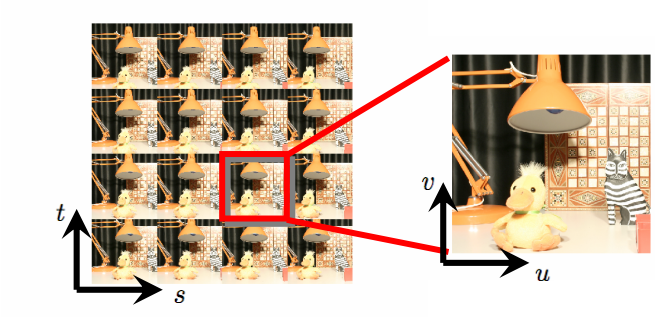

Figure 3: The lightfield parameterisation. The planar camera array shown in Figure 2 leads to the 4-D lightfield shown above. Here, the *t,s* coordinates correspond to the camera locations, while the u,v coordinates corresponds to the image plane

For example, the case when cameras are on a plane leads to the 4-D lumigraph or lightfield [7] parameterization. This parameterization is obtained by using two parallel planes: the focal plane (or camera plane, coordinate *t,s*) and the retinal plane (or image plane, coordinate *u,v*). A ray of light is therefore parameterised by its intersection with these two planes. The coordinates in the focal plane give the position of the pinhole camera, while the coordinates in the retinal plane give the point in the corresponding image, this is schematically shown in Figure 1. An example of a camera configuration leading to a lightfield data is shown in Figure 2 and the corresponding data set is shown in Figure 3.

If we now assume that cameras are placed along a straight line, we obtain the Epipolar Plane Image (EPI)} [4] which is a 3-D plenoptic function. The EPI has a structure that is similar to a video sequence, but the motion of the objects can be fully characterised by their positions in the scene. Notice that, in this set-up, points in the world scene are converted into lines in the plenoptic domain. Alternative 3-D plenoptic function can be obtained, for example, by placing cameras on a circle. In this case, however, a point in the world scene is not converted into a straight line anymore. Finally, if we constrain the camera position to a single point, we have a 2-D function which is in fact a conventional still image.

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 parameterised. For example, in the case of a linear multi-camera system, that is, in the case where the viewing position is constrained to be along a straight line, the plenoptic function is characterised 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 4 is shown in Figure 5. In line with the work of Adelson and Bergen, we call the volume carved by the object *plenoptic hyper-volume*.

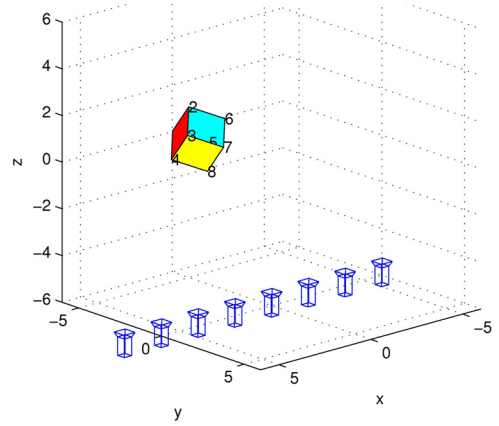

Figure 4: Linear camera array. The *X,Y* and *Z* coordinates correspond to the real world

Similar intuitions apply to other camera setups such as the circular case or the lightfield case.

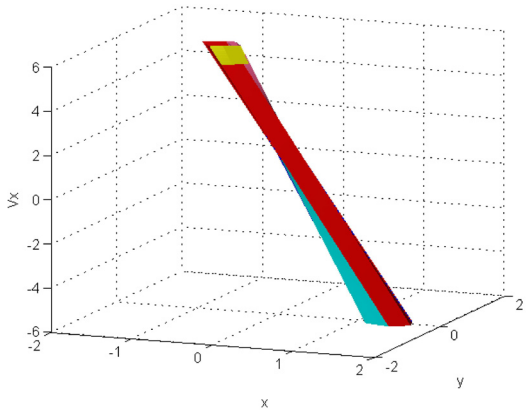

Figure 5: Structure of the plenoptic function. The shape of the plenoptic volume carved by an object or a layer is constrained by the camera setup

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 ν_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

$$\nu_n^{\perp} = \nu_n \cap \overline{\sum_{i=1}^{n-1} \nu_i^{\perp}}$$

where ν_n is the hyper-volume as if there was no occlusion, $^{\perp}$ denotes that the volume has been geometrically orthogonalised with all other volumes that occlude it, and $^{-}$ denotes the complement. This is graphically shown in Figure 6.

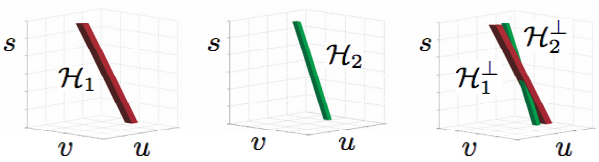

Figure 6: Orthogonalisation of plenoptic hyper-volumes

Higher dimensional hyper-volumes are generated in the same manner for higher dimensional plenoptic functions. In the case of the light field parameterization [7], for instance, the cameras are constrained to lie on a plane and 4D hypervolumes are carved out by the objects.

In order to extract these volumes or hypervolumes, we resort to a variational framework and the level-set method.

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

In this section, for the sake of clarity, we review some of the material already presented in [6].

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 [3]. 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 used to segment objects involve 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 [3][9]. The main idea is to use the level set method in order to minimise 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 $\phi(x,y,t)$, i.e. $\Gamma(t) = \{(x,y) \mid \phi(x,y,t) = 0\}$. The higher dimension function ϕ may be set to be the signed distance function of $\Gamma(t)$. In this case, we have $\|\nabla\phi\| = 1$. The curve is grown according to the partial differential equation

$$\phi_t + F|\nabla\phi| = 0$$

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 such as 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 $\nu_1^{\perp} = \nu$ and $\nu_2^{\perp} = \bar{\nu}$ 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 minimise can be written in the form:

$$E_{tot}(\tau) = \iiint_{\nu(\tau)} f(\vec{x})d\vec{x} + \iiint_{\bar{\nu}(\tau)} g(\vec{x})d\vec{x}$$

where the $f(\vec{x})$ and $g(\vec{x})$ 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]

$$\frac{dE_{tot}(\tau)}{d\tau} = \iint_{\partial\nu} [g(\vec{x}) - f(\vec{x})](\vec{W}\cdot\vec{M})d\vec{\sigma},$$

where $\partial\nu$ is the border of the volume, $d\vec{\sigma}$ is a differential surface element, $\vec{W}$ is the speed of the evolving interface and $\vec{M}$ 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

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

where $\alpha(s,t)$ 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

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

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 dis-occluding. 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.

4. 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. However, the main advantage of the level set method over other existing active contours techniques, is that it naturally scales to higher dimensions. This is of central importance in our context where the plenoptic function can have up to 7-D. The experiments in Figure 6 make this point more evident. In this example, the data is acquired using a planar camera set-up and this leads to the 4-D lightfield parameterization discussed in Section 3. In order to operate on the entire dataset jointly, the level-set method has to evolve hyper-surfaces in a 4-D space. This would have not been possible with different segmentation strategies. Moreover the constrained evolution introduced in [6] allows the algorithm to handle occlusions systematically. The extract hyper-layers are shown in Figure 6. These are related to the dataset shown in Figure 3.

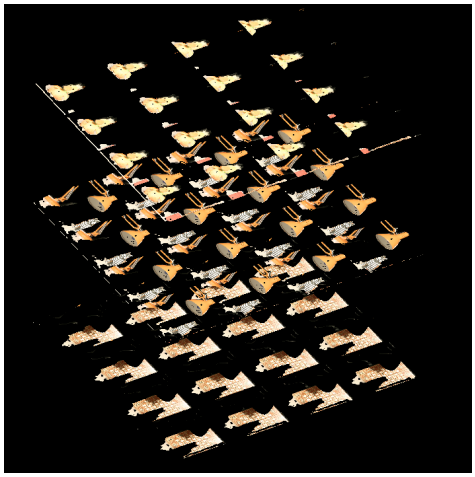

Figure 6: The hyper-volumes extracted with the level-set method from the lightfield shown in Figure 3

A second important advantage of using a segmentation method that operates on the whole stack of images is that occlusions can be highlighted and partially occluded object can be recognised and made visible. This is shown in the following

demonstration where a tank is behind a wall containing some windows. Three of the original 15 input images are shown in Figure 7. In this case, the camera positions lie on a straight line. Figure 8 shows three different dis-occlusion techniques. The proposed method, shown at the bottom of Figure 8, clearly outperforms the existing methodologies.

Figure 7: Three of the original 15 input images

5. Ongoing Work: Compact Representation of Plenoptic Hyper-volumes

Each extracted hyper-volume is fundamentally a collection of segmented images, where each segmented region is related to the same object or layer. The pending issue now is to develop an efficient and possibly compact way to describe each hyper-volume. Numerous methods have been devised for extracting salient local features from images in order to provide an efficient description of an object in an image. However, such methodologies have rarely been extended to the case of multi-view images.

Figure 8: Disocclusion of the tank. Top: Dis-occlusion using synthetic aperture [10]. Middle: Dis-occlusion using state-of-the-art stereo algorithm [11]. Bottom: Dis-occlusion using the plenoptic hyper-volume extraction algorithm

Among all the possible image feature descriptors, we have chosen the Scale Invariant Feature Transform (SIFT) presented in [1]. SIFT was chosen because of its good invariance properties to affine transformation and change of luminance, and for its low complexity. Each SIFT descriptor is a vector of 128 components with a particular location and orientation.

In our algorithm, the SIFT features are extracted from each segmented image and then the corresponding feature points in every image are tracked along the plenoptic lines. Dynamic Programming (DP) is used to find the global track with the minimum distortion. The distortion is, in our case, given by the Euclidean distance between

the feature vectors of two adjacent images - the rational being that two feature points describing the same region of the same object should be very similar- and the epipolar constraint, namely two feature points should be located on the same plenoptic line. Once tracks have been found, we have studied the variation of the SIFT vectors along each track and tried to fit this variation with a polynomial of maximum degree two. In most cases, a constant polynomial is sufficient to model such variation. In other words it is sufficient to describe all the SIFT points along a single track by their average vector. The collection of such average vectors for each hyper-volume, therefore, represents a compact representation of each object in the multi-view data sets and will be used in future for object classification.

Some examples of the tracking of SIFT features are shown in Figures 9 and 10.

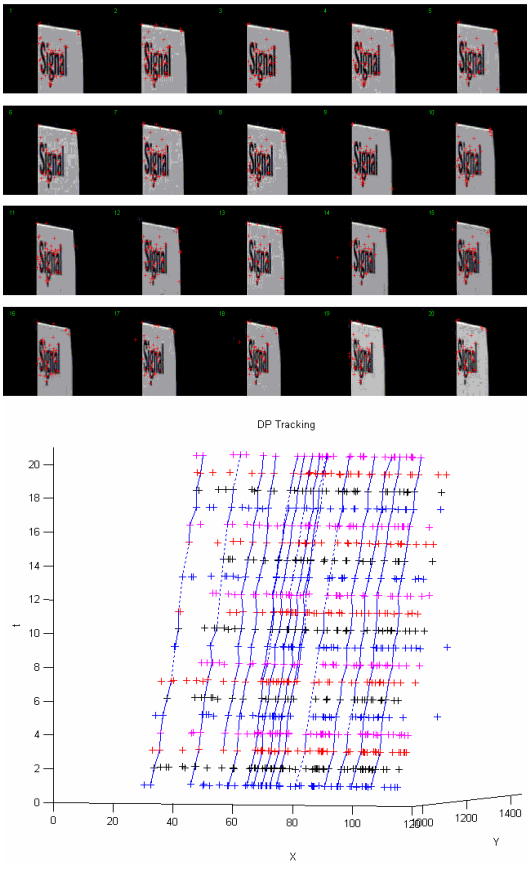

Figure 9: Tracking of SIFT feature for a linear camera set-up. *Top*: Original dataset, the red crosses on the images are related to the SIFT feature points. *Bottom*: The corresponding tracks

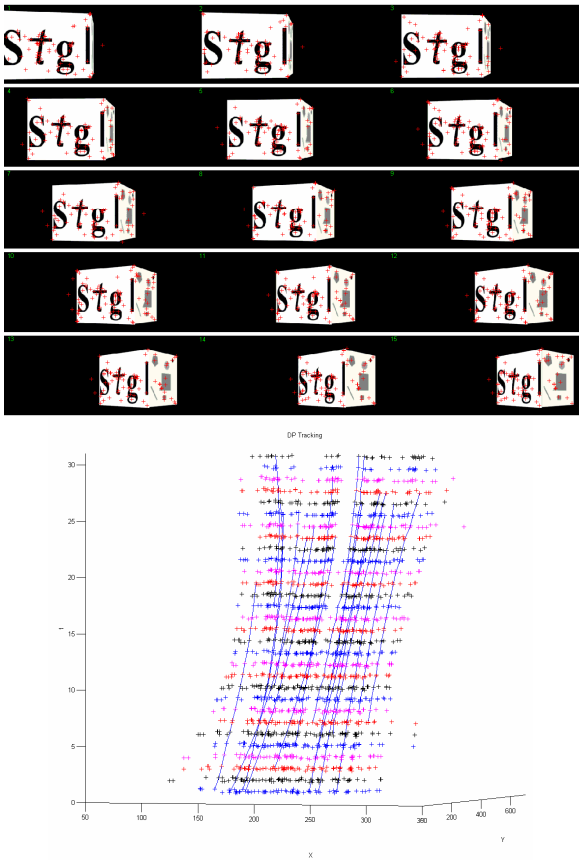

Figure 10: Tracking of SIFT feature for a circular camera set-up (rotation from 0 to 90 degrees). Top: Original dataset, the red crosses on the right are related to the SIFT feature points. Bottom: The corresponding tracks

6. Conclusions

We have proposed a segmentation algorithm for multi-view images that is based on space continuity and that takes into account occlusions explicitly. We have shown that the proposed algorithm can scale to any dimension and that hyper-volumes of high dimension can be extracted. 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 have shown that this makes it possible to make occluded objects visible. We are now exploring the use of such

hyper-volumes for object recognition and classification.

References

[1] D. Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, vol. 60, no.2, 91-110, 2003.

[2] 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.

[3] 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.

[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, vo.1, pp. 7-55, 1987.

[5] A. Criminisi, S.B. Kang, R. Swaminathan, R. Szeliski and P. Anandan, *Extracting layers and analyzing their specular properties using epipolar-plane-image analysis*, Computer Vision and Image Understanding, vol.97, no.1, pp. 51-85, January 2005.

[6] P.L. Dragotti and M. Brookes, *Efficient Segmentation and Representation of Multi-View Images*, SEAS-DTC workshop, Edinburgh, July 2007.

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

[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.

[10] A. Isaksen, L. McMillan and S. J. Gortler, *Dynamically reparameterized light fields*, In Proc. SIGGRAPH, pages 297-306, 2000.

[11] A. S. Ogale and Y. Aloimonos, *Shape and the stereo correspondence problem*, Int. Journal of Computer Vision, Vo. 65(3), pages 147-162, December 2005.

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).