ACCURATE NON-ITERATIVE DEPTH LAYER EXTRACTION ALGORITHM FOR IMAGE BASED RENDERING

James Pearson, Pier-Luigi Dragotti, Mike Brookes

Dept. of Electronic and Electrical Engineering
Imperial College

ABSTRACT

Image based rendering is an attractive alternative for generating novel views compared to model based rendering due to its lower complexity and potential for photo-realistic results. We present a fast unsupervised method for synthesising arbitrary viewpoints of a scene from a set of existing views. Our novel improvements include optimising the placement of depth layers to take advantage of the composition of real world scenes and hierarchically building our simple geometric model to maximise its accuracy.

Index Terms— View synthesis, Plenoptic function, depth layer

1. INTRODUCTION

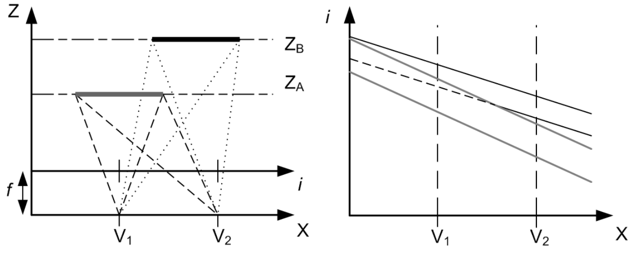
There are several methods of generating an arbitrary new view of a scene from a set of existing views [1]. One approach is to create a textured 3D model of the entire scene and to use this for synthesising new views. Alternatively, in image based rendering (IBR), new views are generated by combining individual pixels from a densely sampled set of input images. Between these two extremes lies a range of methods with varying proportions of geometric and image input information that has been explored, for example in [2]. Using a complete 3D model allows freedom in the final rendering but requires more computation and often creates noticeably artificial output images. In contrast the IBR approach requires little geometric information and can give potentially photo-realistic results but requires many more input images. A layer-based representation of the scene geometry [3, 4] represents a compromise that has low geometric complexity while allowing view synthesis from a moderate number of input images. In this paper we present an unsupervised non-iterative procedure for extracting depth layers and synthesizing new views from multiple input images that is accurate, robust and is able to generate new photo-realistic output images. In Section 2 we discuss the plenoptic function and its relation to the depth layer model. Section 3 describes our new algorithm and in Section 4 its performance is evaluated. Finally, Section 5 concludes the paper.

2. THE PLENOPTIC FUNCTION AND LAYER APPROXIMATION

A convenient interpretation of a multiview image set is to consider the collection of light rays emanating from the scene. The complete parameterisation of the rays at any position and time requires the seven dimensional Plenoptic Function, P_7, which was introduced by Adelson and Bergen [5]. Simplifying assumptions about the sensing setup leads to the Light Field [6] and the Lumigraph [7]. In this paper we are addressing the specific case of an image sequence in which the camera moves along a horizontal line and we also assume the images have been rectified so that disparities are all horizontal. These assumptions result in a reduced version of the full plenoptic function,

$$P = P_3(i, j, \tau), \tag{1}$$

where i and j are the coordinates of the intersection of a ray with the image plane and τ is the camera position along the x-axis, (V_x). A layered scene is illustrated in Fig. 1(a) where the ray intersections with the image plane for two camera positions and two layers are shown. Fig. 1(b) demonstrates how the intersection point of a ray with the image plane changes as the camera position moves. This locus is known as the



(a) Layered view

(b) EPI lines

Fig. 1. Two layers observed by a camera in two positions τ = 1 and 2, i is the image plane and f is the focal length.

We would like to acknowledge the Rabin Ezra Scholarship Trust for providing partial funding of this work.

EPI line, and its gradient is inversely proportional to the layer depth z. Lines with a steeper gradient occlude lines with a shallower gradient when they intersect. Each camera view can be regarded as a planar sample sliced through a three-dimensional EPI volume. We use a depth layer based geometric model because it is robust, computationally efficient and offers a good description of many real scenes. Each layer is a plane perpendicular to the optical axis and is therefore at a constant depth. Associated with each layer, l, is a unique disparity gradient, d_l, which is the ratio of the disparity to the camera motion, V_x, and is inversely proportional to the layer depth, z_l. The number of layers, L, represents a trade-off between fidelity and computation complexity. Chai et al. [2] used plenoptic theory to estimate the number of layers, L_{min}, necessary for successful rendering without aliasing, known as the minimum sampling criterion (MSC).

3. LAYER EXTRACTION AND VIEW SYNTHESIS

For the plenoptic based sampling to be achieved, each pixel of the input images and of the virtual view needs to have an assigned disparity gradient. We use a segment based approach and assign disparities on a segment by segment basis rather than pixel by pixel. The advantage of this is that most layer changes will occur at segment boundaries which normally coincide with the object boundaries; additionally this method is more robust to individual pixel matching errors. This gives an efficient algorithm resulting in sharp and consistent edges. The layer extraction is achieved by segmenting one or more key images, using for example [8], and assigning each segment, S_n, to a layer with disparity gradient d_l by matching the segment in other images. The layer assignment disparity gradient map can then be projected onto the other input images and the virtual view. Hirschmüller and Scharstein [9] found that for most sample sets the most effective S_n matching error score (MES) metric to use, when estimating disparities, is the absolute intensity difference (SAD). A key innovation by Berent [10] was to do the layer assignment in two passes, the first treats each segment in isolation and the second takes into account the predicted occlusions from surrounding segments. We calculate the matching error ϵ using,

$$\epsilon\left(S_n\right) = \frac{\sum_{k=0}^{K_n-1} |I_1(i_{n,k}, j_{n,k}) - I_2(i_{n,k}+d_n, j_{n,k})|O_{n,k}}{\left(\sum_{k=0}^{K_n-1} O_{n,k}\right)\log_{10}\left(\sum_{k=0}^{K_n-1} O_{n,k}\right)} \tag{2}$$

where K_n is the total number of pixels within the segment S_n, I_1 and I_2 are input images, $(i_{n,k}, j_{n,k})$ are pixel coordinates for segment n and index k and d_n is the segment disparity, for the pixel index k within the segment. O is a visibility mask where $O_{n,k}$ = 1 is true for the first pass

$$O_{n,k} = \begin{cases} 1 & \text{if } I_2(i_{n,k}+d_n, j_{n,k}) \text{ is visible;} \\ 0 & \text{if } I_2(i_{n,k}+d_n, j_{n,k}) \text{ is occluded.} \end{cases}$$

is true for the second pass.

3.1. Non uniformally spaced layers

In our algorithm, we first set the number of layers, L, to a high value and determine the disparity gradient histogram, shown in Fig. 2. Rather than using the uniformally spaced values of layer disparity gradient, d_l, shown in the upper set of vertical bars, we use the Lloyd-Max algorithm [11] to find the values of d_l for our reduced output L, that minimize the mean-squared error, shown in the lower set of vertical bars. It can be seen that these cluster around the salient feature of the scene.

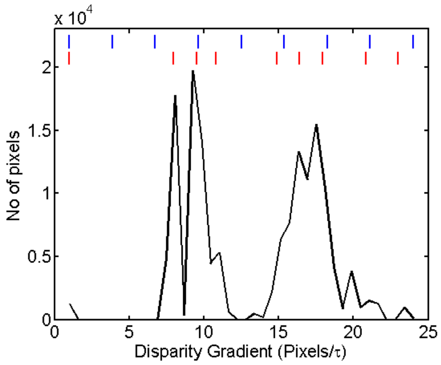


Fig. 2. Disparity gradient distribution for Teddy sequence 'leave 1 image out' with uniform and variably assigned layers, the two types of layer positions are shown by vertical lines on the top of the graph, L is 9

3.2. Hierarchical Layer extraction

A weakness of the two pass d assignment procedure is that any errors in the first pass propagate through into the second pass. It was found that an incomplete but accurate occlusion map, restricted to segments with a low MES, reduced the number of misassigned segments. We know from the plenoptic theory that occlusions are hierarchical and predictable and assuming that the first pass is relatively accurate all S_n assigned to the top level, d_{max}, should be well assigned as they will not have any occlusions. If their MES is sufficiently low then they are used to form the occlusion map for the subsequent layer. If the matching score is too high then it is likely that it has been been misassigned and so it is ommited from the occlusion map. This process is repeated for each layer until d_{min} is reached. Segments with a poor matching score are ignored until the very end at which point they are then assigned using the most recent occlusion map. The benefits of this hierarchical procedure are that occlusions are estimated from the second-pass layer assignments rather than the first-pass assignments and that unreliably assigned segments are ignored when estimating occlusions. We note that the hierarchical approach does not increase the complexity of the previous method in that it only changes the order in which segments are tested. The weighting in the MES (2) prevented segments ‘hiding’ behind other segments to improve their match.

3.3. Synthesising View

All input images are samples lying on a plane through the EPI volume. Novel views can be generated by interpolating new points from the existing EPI volume of input images, Fig. 3 shows how the new sample on an EPI line, at position $V_{1.7}$, can be interpolated from input images, V_1 and V_2, either side. For points (P, R, S) the EPI line is un-occluded on both sides so the new sample can be a blended distance-dependant mixture of the two input images. In the case of (Q) only one side of the EPI line is un-occluded so only the sample from V_2 will be used.

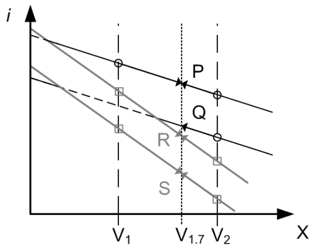


Fig. 3. Synthesising new view $V_{1.7}$ by interpolating samples along EPI lines for existing views V_1 and V_2

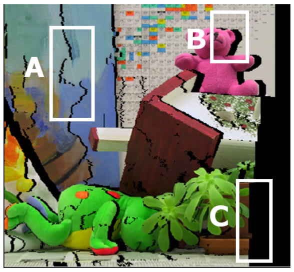


Fig. 4. Projection from image 0 to the image 8 of the teddy sequence, with resulting disocclusions marked with black pixels

3.4. Multiple key images

When a virtual camera view is synthesised, regions of the scene become dissocluded leaving gaps in the output. The three types of disocclusion possible are illustrated in Fig. 4; (A) shows tearing, where a missing region appears in a oblique surface which is assigned to multiple depth layers; (B) shows a region of true disocclusion; (C) demonstrates where dissoccluded areas can also occur when regions that were outside the field of view become visible. If only a single key image is used the correct disparity of these regions can not be estimated robustly. Using two or more key images increases the coverage of the scene and allows reliable layer assignment of these regions. For a linear image sequence, using the two end images gives the largest parallax and maximises coverage. The procedure is shown in Fig. 5 in which the histograms of Fig. 2 are estimated for each key image independently but the Lloyd-Max algorithm is applied jointly to both in order to estimate a common set of layer disparity gradients. For most scenes increasing the number of key frames beyond two provides little additional benefit.

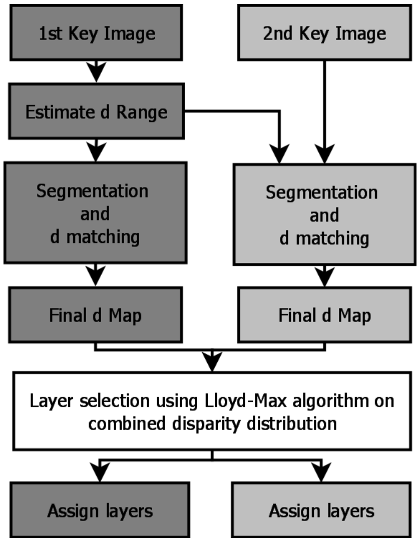
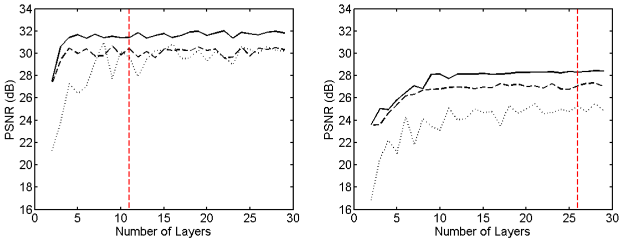

1st Key Image
2nd Key Image
Estimate d Range
Segmentation and d matching
Segmentation and d matching
Final d Map
Final d Map
Layer selection using Lloyd-Max algorithm on combined disparity distribution
Assign layers
Assign layers

Fig. 5. Dual key image d generation process, showing joint initialization of key parameters

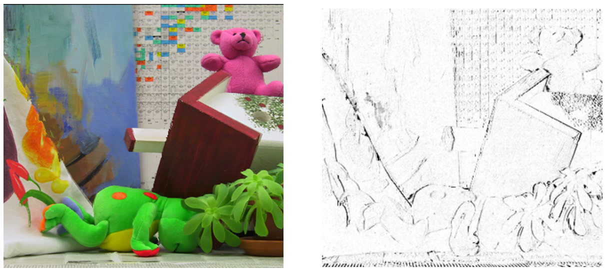
4. EVALUATION

For evaluation we used the Teddy dataset of nine images [12, 13] and the key images were segmented using the mean shift algorithm [14, 8]. We used the ‘leave m out’ method of evaluation in which only every $(m + 1)^{th}$ image is included in the input image set. These are used to synthesize one of the omitted images for which the ground truth is known. In Fig. 6 we compare the performance of [10] with our new algorithm using one key image (image 0) and using two key images (images 0 and 8). In all cases an infilling algorithm was used to fill any holes with the lowest adjacent disparity. The vertical dashed line on each graph indicated the predicted minimum sampling criterion. Fig. 6(a) shows that the original algorithm plateaus at around L_{min}, however our new algorithm plateaus significantly earlier, showing the advantage of adaptive layer spacing. Additionally using two key images improves the PSNR by about 1.8 dB due to its ability to assign dissocluded regions accurately. Fig. 6(b) shows a more challenging case and the improvements made are even more apparent. As well as plateauing at about 12 layers rather than the L_{min} of 26 there is over 2 dB worth of improvement in the PSNR. In both cases our algorithm gives smoother results because it can pick the best layer assignment and the hierarchical procedure results in fewer miss-assignments. Fig. 7(a) shows a typical rendered results for frame 2 in the ‘leave 3 out' case. Fig. 7(b) shows the luminance error map which only has a median of 1 and mean of 1.64. Virtually all the errors are at the edges of objects and are mostly due to aliasing effects.



(a) Leave 1 image out (b) Leave 3 images out

Fig. 6. Comparing the interpolation ability of the orignal (dot), single key image (dash) and dual key image (line) approaches on the Teddy sequence.



(a) Rendered result (b) Error map

Fig. 7. Example rendered output from worst case, 'leave 3 out', where this example is furthest from an input image, Output Frame = 2, Key images = 0 and 8, L = 22, PSNR = 28.3 dB

5. CONCLUSION

In this paper we have presented three novel improvements on existing methods to take advantage of our knowledge of the typical structure of multiview data and the strictly hierarchical nature of occlusions. In contrast to previous work, the layers are assigned with non uniform spacing, a hierarchical approach is taken to assign depths and one or two key images were used. All this leads to a significant performance improvement at low extra cost. We have shown that the minimum sampling criterion that emerges from plenoptic theory can be relaxed when layers are non uniformly distributed. In future work, we wish to use Light Field plane input sources and multiple key images to produce accurate interpolations of viewpoints that do not lie on the input camera plane. We also aim to investigate using post-processing and alpha blending techniques to improve the view synthesis.

6. REFERENCES

[1] H.-Y. Shum, S.-C. Chan, and S. B. Kang, *Image Based Rendering*. Springer, 2007.

[2] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *International Conference on Computer Graphics and Interactive Techniques*. ACM Press, 2000, pp. 307 – 318.

[3] 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.*, vol. 23, no. 2, pp. 143–162, 2004.

[4] J. Shade, S. Gortler, L.-w. He, and R. Szeliski, "Layered depth images," in *Proc. SIGGRAPH*, New York, 1998, pp. 231–242.

[5] E. H. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," *Computational Models of Visual Processing*, pp. 3 – 20, 1991.

[6] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. SIGGRAPH*, New York, 1996, pp. 31–42.

[7] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. SIGGRAPH*, New York, 1996, pp. 43–54.

[8] C. Christoudias, B. Georgescu, and P. Meer, "Synergism in low level vision," in *Proc. 16th International Conference on Pattern Recognition*, vol. 4, 2002, pp. 150–155.

[9] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1582 –1599, sept. 2009.

[10] J. Berent, P. L. Dragotti, and M. Brookes, "Adaptive layer extraction for image based rendering," in *International Workshop on Multimedia Signal Processing*, 2009.

[11] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Information Theory*, vol. 28, no. 2, pp. 129 – 137, mar. 1982.

[12] Middlebury. (2003) Teddy stereo dataset. Website. Middlebury. [Online]. Available: http://vision.middlebury.edu/stereo/data/

[13] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *IEEE Computer Vision and Pattern Recognition*, vol. 1, Jun. 2003, pp. 195–202.

[14] C. M. Christoudias and B. Georgescu. (2002) Website. Rutgers University. [Online]. Available: http://coewww.rutgers.edu/riul/research/code/EDISON/index.html