Computationally Efficient Background Subtraction in the Light Field Domain

Alireza Ghasemi[a] Mahdad Hosseini Kamal[b] Martin Vetterli[a]

[a]LCAV, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[b]LTS2, École Polytechnique Fédérale de Lausanne , Lausanne, Switzerland

ABSTRACT

In this paper we present a novel approach for depth estimation and background subtraction in light field images. our approach exploits the regularity and the internal structure of the light field signal in order to extract an initial depth map of the captured scene and uses the extracted depth map as the input to a final segmentation algorithm which finely isolates the background in the image.

Background subtraction is natural application of the light field information since it is highly involved with depth information and segmentation. However many of the approaches proposed so far are not optimized specifically for background subtraction and are highly computationally expensive. Here we propose an approach based on a modified version of the well-known Radon Transform and not involving massive matrix calculations. It is therefore computationally very efficient and appropriate for real-time use.

Our approach exploits the structured nature of the light field signal and the information inherent in the plenoptic space in order to extract an initial depth map and background model of the captured scene. We apply a modified the Radon transform and the gradient operator to horizontal slices of the light field signal to infer the initial depth map. The initial depth estimated are further refined to a precise background using a series of depth thresholding and segmentation in ambiguous areas.

We test on method on various types real and synthetic of light field images. Scenes with different levels of clutter and also various foreground object depth have been considered in the experiments. The results of our experiments show much better computational complexity while retaining comparable performance to similar more complex methods.

Keywords: Plenoptic Function, Light Field Imaging, Segementation, Background Substraction

1. INTRODUCTION

Light field or plenoptic imaging has attracted increased attention in both research and consumer community in recent years. This exponentially growing popularity stems mostly from the introduction of the consumer-targeted light field cameras Lytro[1] and Raytrix. In light field imaging, multiple views of the scene of interest are captured using the directional information of the light rays incident on the sensing device. In color imaging in contrast, only intensities of light rays are captured on the sensor cells.

The advanced sensing capability of a light field camera enables capturing much more information from a scene than traditional pinhole ones. For example information such as depth map and occlusion ordering can be extracted using the information inherent in the light field signal.[2] Therefore, we can think of light field imaging as a method to recover the information which are lost during the color image formation process in pinhole cameras which involves projecting the real-world three-dimensional scene into a two-dimensional image plane.

The extra information content of a light field image can be very efficiently exploited in order to improve many of the current computer vision algorithms. Computer vision tasks such as segmentation,[3] depth map estimation[4] and three-dimensional reconstruction can highly benefit light field information for performance improvement. This requires adaptation of current algorithms to work in the light field space or develop new methods from scratch.[5]

One of the major applications of computer vision is background subtraction. In background subtraction, we assume that there is a dominant, clearly distinguishable object in the scene and we want to separate that object (or probably objects) from its background.

Computational Imaging XII, edited by Charles A. Bouman, Ken D. Sauer, Proc. of SPIE-IS&T Electronic Imaging,
SPIE Vol. 9020, 90200U · © 2014 SPIE-IS&T · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2041161

Object recognition can be used as a very important and useful preprocessing step in computer vision tasks such as object detection and image retrieval. Moreover, it can be considered and independent task by itself, mostly for entertainment purposes.

A lot of approaches have been proposed for background subtraction. Most of them rely on motions of the scene and try to model the background using statistical approaches.[6] Because of the need to motion information, these approaches require sequential frames captured from a scene, rather than a single image. For detecting background pixels, these approaches mostly rely on color histogram information as well as motion of the objects.

On the other hand, there are approaches which utilize light-field information to infer a depth map from the scene.[4] When the depth map is available, background subtraction can be simply reduced to a thresholding on the depth values. these approaches usually try to infer the depth map as precise as possible and use variational approaches.[7] therefore, they are computationally expensive.[8]

In his paper, we use the facts about background subtraction to improve the speed of light-field based methods. First, depth estimation is a much more difficult task than background subtraction. Therefore, we may not need to do full depth estimation in order to just separate background from foreground. actually, one would argue that a single-bit precision is enough for depth estimation if we want it only for background subtraction.

Moreover, we notice that background pixels usually form large background patches. Therefore, depth estimation is not required for every pixel.And finally, noisy estimations can be removed using spatial information. We use these facts in derivation of our algorithm.

2. THE PROPOSED APPROACH

2.1 The Plenoptic Function and Formation of Light-Fields

When we talk about images, we usually mean a two-dimensional function of spatial axes x and y. Now consider a video sequence, i.e a time-ordered sequence of images. We can consider the overall data can as a 3D shape in the $x-y-t$ space. This can be thought of as a generalization of the traditional intensity image.

In the next step, we also relax the assumption of having a single-sensor camera and add a fourth parameter for wavelength of the light at any position (x_0, y_0) on any time (t_0). This way we have a four dimensional function or hypervolume corresponding to the aggregation of video sequences taken at different wavelengths. Each $x-y$ slice of this volume (assuming other parameters are kept constant) refers to a 2D image captured at a specific time under a specific wavelength. Similar argument applies to $x-y-t$ (video signals) and other slices.

A Plenoptic function is a generalization of a two-dimensional image, in which we have as well as spatial coordinates of the image plane, five more dimensions for the time(t), wavelength (λ) and the position of the camera.[9] This leads to:

$$P = P_7(x, y, V_x, V_y, V_z, t, \lambda) \tag{1}$$

in which (V_x, V_y, V_z) are the spatial coordinates of the camera position.

The plenoptic function is highly structured. It contains a significant amount of information about the captured scene, camera setup, and other acquisition parameters.

However, to make use of this information we should first filter out irrelevant information and make the data simpler to analyze by reducing its dimensions.

We usually reduce the number of parameters by introducing constraints on the capture process.

A commonly known 3D restriction of the plenoptic function is the $x-y-V_x$ slice. This is known as the Epipolar Plane Image[11] or the light-field.[8] Light-fields have attracted a lot of attention in recent years.[12–15]

We can capture EPIs using either a linear camera array or equivalently (and more practically) by assuming that the scene is static and linearly moving a single camera (Figure 1). This latter setup is more practical since a plenoptic function (more precisely a slice of it) can be captured easily and efficiently using a handheld device's internal camera.

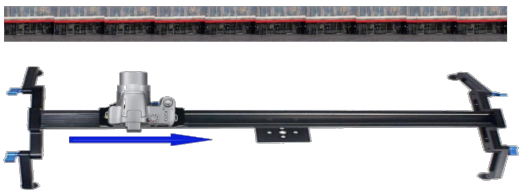

Figure 1: capturing an EPI volume.

2.2 Regularity of the Plenoptic Function

Some conditions are explicitly and implicitly assumed when studying plenoptic function and its properties. Firstly, we assume the pinhole camera model for each individual camera and Cartesian coordinates in the plenoptic function. Secondly, we neglect the wavelength parameter λ by considering grayscale images or different color channels. Finally, surfaces in the scene are assumed to be opaque Lambertian surfaces to avoid reflection issues.

The simplest setup for plenoptic function is what we already know as a still (two-dimensional) image. It can be captured using a single camera and is therefore a sample of the plenoptic function where all parameters except image coordinates (x and y) are kept constant. In the pinhole camera model, each point $\vec{X} = (X, Y, Z)$ in the scene is mapped to a point (x, y) in the image plane with the following rule:[2]

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \mapsto \begin{pmatrix} \frac{X}{Z} \\ \frac{Y}{Z} \end{pmatrix}, \tag{2}$$

In the above model, it is assumed that the focal length is 1 and there is match between camera and world coordinates. A significant amount of information is lost in this mapping. This is why current methods of processing images rely mostly on color and texture.

Now, assuming the pinhole camera model and Lambertian surfaces,[3] consider an image sequence, taken by moving the camera V_x units along the horizontal axis for each image (i.e the light-field case). Adding a third dimension for the camera location (set to V_x), the mapping from a scene point $P = (X, Y, Z)^T$ to its projection into each image in the sequence can be described as:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \mapsto \begin{pmatrix} f\frac{X}{Z} - f\frac{V_x}{Z} \\ f\frac{Y}{Z} \\ V_x \end{pmatrix}. \tag{3}$$

This is how a light-field is formed. We infer from (3) two important facts:

1. Each scene point corresponds to a line (a single-parameter curve) in one of the $x - V_x$ slices of the epipolar volume (the slice corresponding to $y = \frac{Y}{Z}$).
2. The gradient (slope) of this line is proportional to the depth (Z value) of the scene point.

Figure 2 depicts these properties in a typical EPI volume.

Please note that the above facts are valid only when all our initial assumptions have been met. For example, if the motion is non-linear or speed varies, then the feature paths are no longer lines. An $x - V_x$ slice of an epipolar volume is called an "epipolar plane".

We will use these properties in deriving our scale-invariant representation for light-field images.

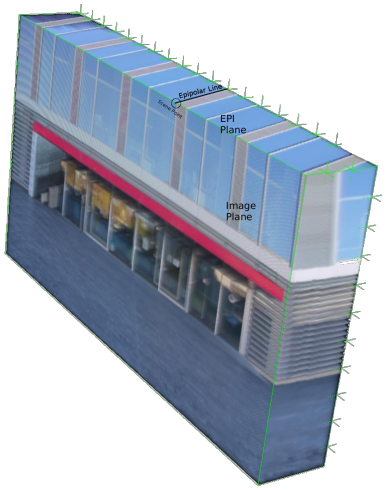

Figure 2: A Typical EPI volume

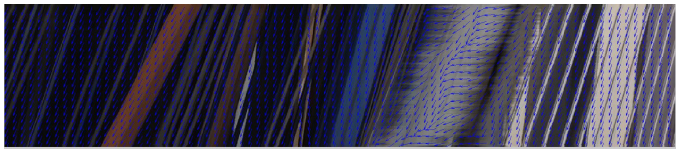
Figure 3: Line Segment Detection (LSD) by gradient computation.

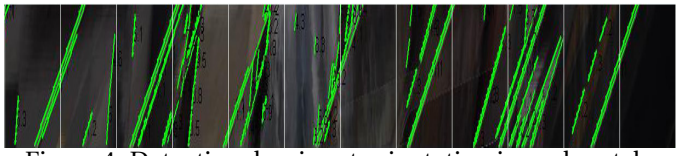
Figure 4: Detecting dominant orientation in each patch.

2.3 Detecting Backgrounds in the Plenoptic Space

To avoid intolerable time complexity, we attack the problem of background detection by detecting the depth of image patches and putting a threshold in order to separate background points. We can then improve the background detection by applying a simple segmentation algorithm.

To detect the depth of image patches, we need to find the gradients of the dominant epipolar lines in each patch. To achieve this goal we need a line segment detector which can operate in a very fast manner.

The LSD approach proposed in[16] is a promising approach which can detect line segments in an image rapidly and robustly. The algorithm is based on gradient computation. First gradient in every pixel is computed. Then, all pixels are clustered based on angles of gradient values (Figure 3). Finally, statistically non-random (significant) clusters are selected as line segments.

After line segment detection, we have to estimate the relative depth of different parts of the light-field. To do so, the whole light-field volume is partitioned into light-field patches. The size of patches can be selected according the complexity of the background and foreground. It may further be adapted, as we will mention.

After partitioning the light-field, the slope and length (strength) of lines in each patch are accumulated. Finally, a weighted average and variance is computed in each patch. This indicates the average depth of a patch and how confident the depth computation has been for that patch. Figure 4 shows this.

In the process of patch depth estimation, we may require to further adapt and change the size of the patches. We may require further divide a light-field patch for some reasons such as:

- If the variance of lines' slopes is higher than a threshold.

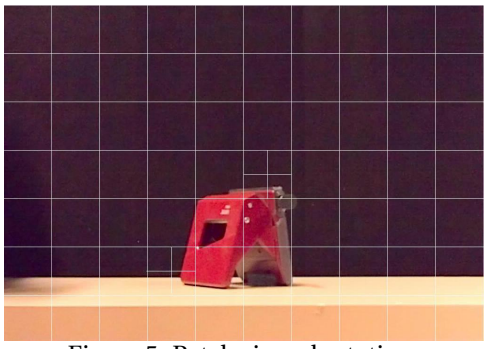

Figure 5: Patch size adaptation.

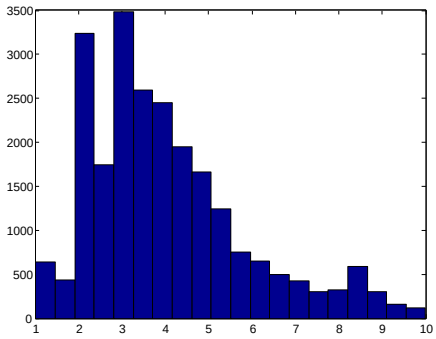

Figure 6: Threshold estimation.

- If neighbors are patches of different type (BG/FG)

Or we may require further to combine neighboring light-field patches, for example if:

- the overall strength of detected line segments is low.
- no line is detected.

We continue these processes until the patch sizes are in a predefined range (Figure 5).

In the last remaining step of our algorithm, we have to threshold the detected slopes, in order to discriminate between background and foreground. To do so, we first form a histogram of computed depth values, such as the one in the Figure 6. We smooth the histogram to reduce the effect of noisy peaks.

Then, we locate two peaks from the beginning and the end of the histogram, to count for dominant foreground and background depth, respectively. Finally, we place the threshold on a depth (i.e. slope) between the two peaks. this way we make sure that there is heuristically enough marging between foreground objects and the background.

3. EXPERIMENTS AND RESULTS

To test the background subtraction algorithm, we need a dataset of light-fields. We used a set of light-field images captured from the EPFL campus (Figure 7). In each light-field image there is a total of 300 sub-view images in a one meter baseline. sub-view images are in HD (1280×720) resolution. In the captured images, background is the sky or far away buildings. Depth resolution of the used configuration is up to 50 meters. We compared our approach against a frame based GMM modeling.[17]

Two sample background separation results are shown in Figures 8 and 9. We can see the improved performance of our approach compared to the GMM-based approach.

Figure 7: Three Sample Light-Fields in the EPFL campus dataset.

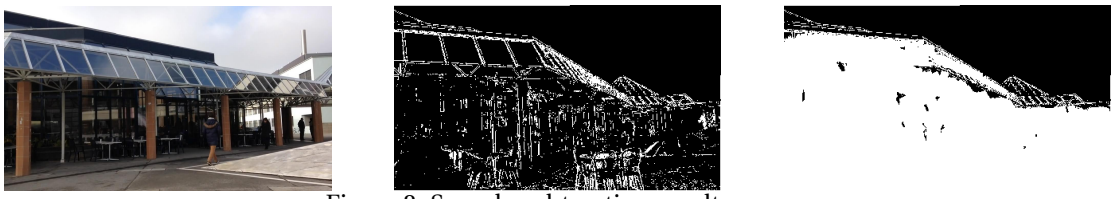

Figure 8: Sample subtraction result.

To further analyze the effect of the baseline and the number of sub-aperture views, we reduced the number of sub-aperture views from 300 to 30 which causes the baseline to drop to 10 centimeters. Results are shown in the Figures 10 and 11. We can see that this time less points have been classified as background (i.e. a higher proportion of the binary result image is white). This is mostly because of the noise in depth estimation which causes he threshold to be determined a bit incorrectly.

Regarding the time complexity of our approach, we notice that light-field processing approaches take about 10 minutes computing a depth map. On the other hand, the frame-based traditional approaches spend about 120 seconds processing 300 frames. Our approach takes about the same time as the traditional approaches, but produces better results which are comparable to the former.

4. SUMMARY

Since the invention of first camera, people have always thought of new ways to capture moments. With the advent of digital photography, not only imaging systems have progressed, but also technology provides new user experience. The recently introduced light field photography enables users with new imaging modalities such as digital refocusing and looking through occlusions, for example.

The specific structure of light field signals highlights the importance of adapting different routines of traditional imaging to exploit the power of light field photography. In recent years, the consumer market has expressed a huge interest in analyze and editing of light field images which necessitates adaptation of traditional background subtraction methods, as the first step of most editing tools, to operate in the light field space. The goal of this work is to develop a computationally efficient algorithm making use of the properties of light fields for background subtraction. Our approach is based on a novel transform, which maps each light ray into a new space where background subtraction is achieved by a simple thresholding.

ACKNOWLEDGMENTS

This project has been co-funded by the Commission for Technology and Innovation (CTI) of Switzerland.

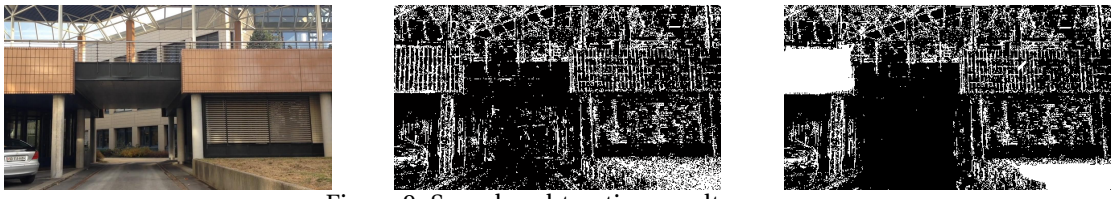

Figure 9: Sample subtraction result.

Figure 10: Sample subtraction result, Reduced baseline.

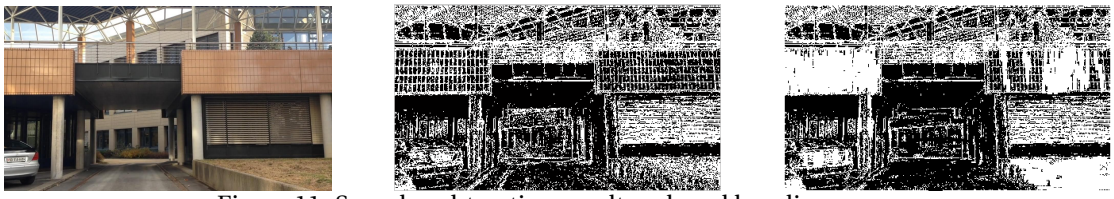

Figure 11: Sample subtraction result, reduced baseline.

References

[1] Georgiev, T., Yu, Z., Lumsdaine, A., and Goma, S., "Lytro camera technology: theory, algorithms, performance analysis," in [*IS&T/SPIE Electronic Imaging*], 86671J–86671J, International Society for Optics and Photonics (2013).

[2] Berent, J. and Dragotti, P., "Plenoptic manifolds," *Signal Processing Magazine, IEEE* **24**(6), 34–44 (2007).

[3] Berent, J. and Dragotti, P., "Segmentation of epipolar-plane image volumes with occlusion and disocclusion competition," in [*Multimedia Signal Processing, 2006 IEEE 8th Workshop on*], 182–185, IEEE (2006).

[4] Wanner, S., Straehle, C. N., and Goldluecke, B., "Globally consistent multi-label assignment on the ray space of 4d light fields," in [*CVPR*], 1011–1018, IEEE (2013).

[5] Wang, Y., Brookes, M., and Dragotti, P., "Object recognition using multi-view imaging," in [*Signal Processing, 2008. ICSP 2008. 9th International Conference on*], 810–813, IEEE (2008).

[6] Bouwmans, T., El Baf, F., Vachon, B., et al., "Statistical background modeling for foreground detection: A survey," *Handbook of Pattern Recognition and Computer Vision* , 181–199 (2010).

[7] Wanner, S. and Goldluecke, B., "Variational light field analysis for disparity estimation and super-resolution," (2013).

[8] Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., and Gross, M., "Scene reconstruction from high spatio-angular resolution light fields," *To appear ACM Trans. Graph.(Proc. SIGGRAPH)* (2013).

[9] Adelson, E. H. and Wang, J. Y. A., "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.* **14**(2), 99–106 (1992).

[10] Venkataraman, K., Lelescu, D., DuparrÃl', J., McMahon, A., Molina, G., Chatterjee, P., Mullis, R., and Nayar, S., "Picam: An ultra-thin high performance monolithic camera array," in [*ACM SIGGRAPH 2013 Asia*], ACM (2013).

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

[12] Tosic, I., Shroff, S. A., and Berkner, K., "Dictionary learning for incoherent sampling with application to plenoptic imaging," in [*ICASSP*], 1821–1825, IEEE (2013).

[13] Do, M. N., Marchand-Maillet, D., and Vetterli, M., "On the bandwidth of the plenoptic function," *IEEE Transactions on Image Processing* **21**(2), 708–717 (2012).

[14] Bando, Y., Holtzman, H., and Raskar, R., "Near-invariant blur for depth and 2d motion via time-varying light field analysis," *ACM Trans. Graph.* **32**(2), 13 (2013).

[15] Marwah, K., Wetzstein, G., Bando, Y., and Raskar, R., "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Trans. Graph.* **32**(4), 46 (2013).

[16] Von Gioi, R. G., Jakubowicz, J., Morel, J.-M., and Randall, G., "Lsd: A fast line segment detector with a false detection control," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **32**(4), 722–732 (2010).

[17] Bouwmans, T., El Baf, F., Vachon, B., et al., "Background modeling using mixture of gaussians for foreground detection-a survey," *Recent Patents on Computer Science* **1**(3), 219–237 (2008).