SIMULTANEOUS ESTIMATION OF SUPER-RESOLVED DEPTH AND ALL-IN-FOCUS IMAGES FROM A PLENOPTIC CAMERA.

F. Pérez Nava[+]*, J. P. Lüke*

[+]Departamento de Estadística, Investigación Operativa y Computación
Departamento de Física Fundamental y Experimental, Electrónica y Sistemas
Universidad de La Laguna. 38271. Canary Islands. Spain.

E-mail:{ fdoperez@ull.es[+]*, jpluke@ull.es}*

ABSTRACT

This paper presents a new technique to simultaneously estimate the depth map and the all-in-focus image of a scene, both at super-resolution, from a plenoptic camera.

A plenoptic camera uses a microlens array to measure the radiance and direction of all the light rays in a scene. It is composed of $n \times n$ microlenses and each of them generates a $m \times m$ image. Previous approaches to the depth and all-in-focus estimation problem processed the plenoptic image, generated a $n \times n \times m$ focal stack, and were able to obtain a $n \times n$ depth map and all-in-focus image of the scene. This is a major drawback of the plenoptic camera approach to 3DTV since the total resolution of the camera n^2m^2 is divided by m^2 to obtain a final resolution of n^2 pixels.

In our approach we propose a new super-resolution focal stack that is combined with multiview depth estimation. This technique allows a theoretical resolution of approximately $n^2m^2/4$ pixels. This is an $o(m^2)$ increment over previous approaches.

From a practical point of view, in typical scenes we are able to increase 25 times the resolution of previous techniques. The time complexity of the algorithm makes possible to obtain real-time processing for 3DTV using appropriate hardware (GPU's or FPGA's) so it could be used in plenoptic video-cameras.

Index Terms— 3DTV, plenoptic camera, lightfield, depth, all-in-focus image, focal stack, super-resolution.

1. INTRODUCTION

Three-dimensional television (3DTV) [1] is regarded as the next step in the evolution of television. The 3DTV offers a three-dimensional (3D) depth impression of the observed scene and provides a more natural and life-like visual home entertainment experience to the user. Together with the advent of Digital Video Broadcasting (DVB) and the progress in the area of autostereoscopic 3D displays, results seem now to form a critical mass for the successful introduction of 3DTV in the consumer market.

A 3DTV system is composed of several modules: image capture, 3D scene reconstruction and representation, coding, transmission, rendering and display. In this paper we focus on the image capture and 3D reconstruction modules. A 3DTV video camera that integrates the scene image capture and 3D reconstruction modules would be of great importance in 3DTV systems where the techniques for 3D capture and 3D display are decoupled from each other and where the capture device should provide the computerized representation of the 3D scene.

To obtain a 3D scene reconstruction, it is necessary to acquire multiview information of the scene that can be obtained from a configuration of several cameras, generally two or an array of them. Another way to achieve this 3D reconstruction is to use a plenoptic video-camera [2] that captures multiview information using micro-lenses. This multiview information can be used to solve the 3D scene reconstruction problem building a focal stack from the plenoptic image [3]. Plenoptic cameras offer several advantages over a configuration of cameras: there is no need for geometric or color calibration and no frame synchronization problem. Also the camera is fully portable. They also have several drawbacks: less depth resolution due to a smaller baseline and less spatial resolution due to their spatio-angular lightfield sampling.

In this paper we present a new technique to overcome the spatial resolution problem. This problem has also been addressed in [4] for the "plenoptic 2.0" camera. However our approach is the first to obtain super-resolved depth and all-in-focus images from plenoptic cameras. It also obtains the maximum attainable spatial resolution.

This paper is divided in five sections. In Section 2 we introduce the super-resolution focal stack. Section 3 shows how to solve the 3D reconstruction problem with this new focal stack. Section 4 contains some experimental results and Section 5 includes conclusions and future work.

2. THE SUPER-RESOLUTION FOCAL STACK

In this section we present the Focal Stack transform, which is based on the Photography transform, and the super-resolution extension that we propose.

2.1. The Focal Stack Transform

To introduce the Photography transform we parameterize the lightfield defined by all the light rays inside a camera. We will use the two-plane parameterization and write $L_F(\boldsymbol{x},\boldsymbol{u})$ as the radiance travelling from position $\boldsymbol{u}=(u_1, u_2)'$ (apostrophe means transpose) on the lens plane to position $\boldsymbol{x}=(x_1, x_2)'$ on the sensor plane. F is the distance between the lens and the sensor (see Fig. 1 adapted from [5]).

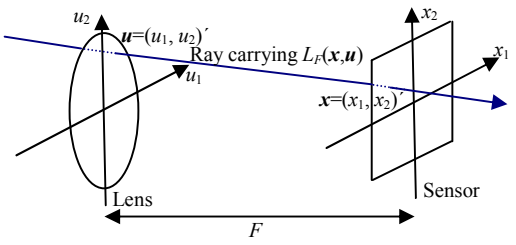

Figure 1 - Two plane parameterization of the lightfield.

The lightfield L_F can be used to compute conventional photographs at any depth αF. Let $\mathcal{P}_\alpha$ be the operator that transforms a lightfield L_F at a sensor depth F into a photograph at sensor depth αF, then we have [5]:

$$\mathcal{P}_\alpha[L_F](\boldsymbol{x}) = \frac{1}{\alpha^2 F^2}\int L_F\left(\boldsymbol{u}\left(1-\frac{1}{\alpha}\right)+\frac{\boldsymbol{x}}{\alpha},\boldsymbol{u}\right)d\boldsymbol{u}. \qquad (1)$$

This equation shows how to compute $\mathcal{P}_\alpha[L_F]$ at different depths from the lightfield L_F. When we compute the photographs for every sensor depth αF we obtain the focal stack transform $\mathcal{S}$ of the lightfield defined as [3]:

$$\mathcal{S}[L_F](\boldsymbol{x},\alpha) = \mathcal{P}_\alpha[L_F](\alpha\boldsymbol{x}), \qquad (2)$$

A plenoptic camera only captures discrete samples of the lightfield. We will assume that the plenoptic camera is composed of $n\times n$ microlenses and each of them generates a $m\times m$ image. A discrete lightfield captured from a plenoptic camera is shown on Figure 2 [5]. To obtain a discretized version of the focal transform we need to interpolate the lightfield L_F and to discretize the integral in (1). There are several techniques to perform this interpolation: nearest neighbour, Kaisser-Bessel filter [5] or the Dirichlet filter [3]. A common result from these techniques is that the focal stack is composed of images with $n\times n$ pixels. Several focal stacks can be built depending on the application. We can use the radiance in the lightfield L_F to obtain the usual focal stack, measures of focus to obtain the laplacian focal stack [3] or measures of photoconsistency to obtain variance focal stack.

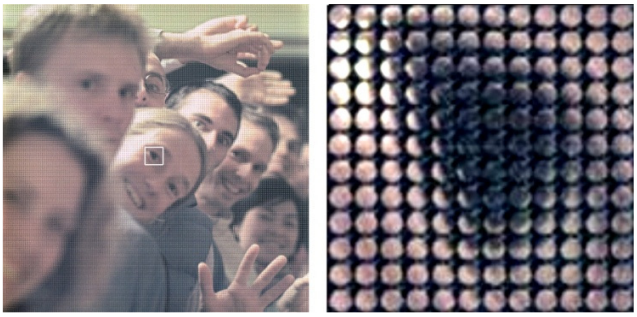

Figure 2 – Left: Lightfield captured from a plenoptic camera. Right: Detail from the white square on the left image.

To introduce the super-resolution focal stack we first show how to compute the Photography transform on a discretized lightfield. To simplify the exposition we work in 2D instead of 3D. The extension of the results to 3D is straightforward. In the 2D case to compute the Photography transform (1) at sensor depth αF we need to compute a line integral on the lightfield. For each of the n pixels in the x axis, a sensor depth αF generates a line through x with determined slope that depends on α (see Figure 3). Since we have a continuous line, to compute the integral we have to interpolate the values that are not available [3], [5].

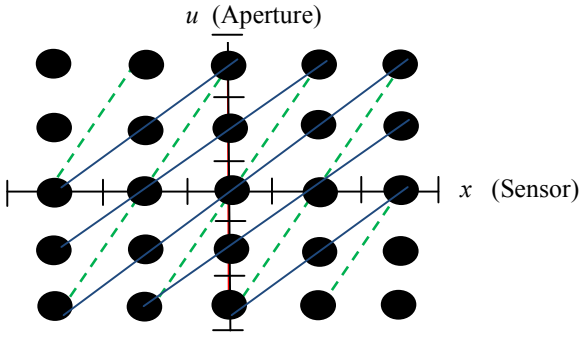

Figure 3 – Photography transform for two particular α values.

In the focal stack transform we check m different slopes for each pixel in the x axis so the size of the focal stack in 2D is $n\times m$.

2.2. The Super-resolution Discrete Focal Stack Transform (SDFST)

The super-resolution extension to the Focal Transform (SDFST) is based on performing the Focal Stack Transform for a different set of lines. The selection of this new set of lines is explained on Figure 4.

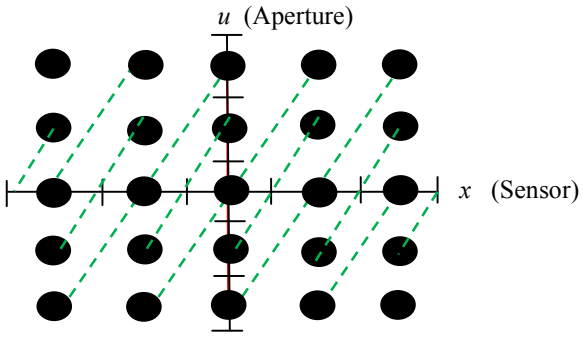

Figure 4 – The lines used in the SDFST for a particular distance

As we see in Figure 4, for a fixed slope we use all the lines with that slope passing through all points in the discretized lightfield. Intersecting these lines with the x axis we see that we have now twice the resolution than we had before (see Figures 3 and 4). Since our interest is real-time plenoptic video processing we do not interpolate the lightfield and we only use the samples that lie exactly on each line. Formally we state the following theorem and corollary:

Theorem

Given a lightfield $L_F(x,u)$, $|x|\leq r$ $|u|\leq s$, $n=2r+1$, $m=2s+1$, $0<|\Delta x|\leq r$, $0<|\Delta u|\leq s$ $\Delta u,\Delta x\in Z$. If we pad $L_F(x,u)$ with $|\Delta x_g|s$ zero-columns on each side of the x-axis, the super-resolution focal stack image for slope $\Delta u/\Delta x$ has $|\Delta u_g|n+|\Delta x_g|m-|\Delta u_g|-|\Delta x_g|+1$ points with $\Delta u_g=\Delta u/g$ and $\Delta x_g=\Delta x/g$ where g is the greatest common divisor of $|\Delta u|,|\Delta x|$.

Corollary

Under the conditions of the preceding theorem the collection of slopes $\{\Delta u/\Delta x$ $0<|\Delta x|<s$, $\Delta u=s$, Δu, $\Delta x\in Z$ with s a prime number$\}$ can generate a super-resolution focal stack with Δu n $-\Delta u$ $+1$ pixels in each image.

Therefore, the final resolution is approximately $\Delta u n \approx mn/2$ that is, half the full resolution. The extension of the above results to 3D using squared aperture/microlenses is trivial. The final resolution in that case is approximately the full resolution divided by 4. In real plenoptic images the resolution is somewhat slower due to the circular shape of microlenses and their border effects. The algorithm to compute an image of the super-resolution focal stack is:

Input: Lightfield $L_F(\boldsymbol{x},\boldsymbol{u})$, $\boldsymbol{x}=(x_1, x_2)'$, $\boldsymbol{u}=(u_1, u_2)'$, $|x|\leq r$ $|u|\leq s$.
Slope $\Delta u/\Delta x$ with $\Delta u,\Delta x\in Z$ as in theorem.
Output: Super-resolution focal stack image $F_{\Delta u/\Delta x}(\boldsymbol{k})$

Compute g=g.c.d.$(\Delta u,\Delta x)$, $\Delta u_g=\Delta u/g$ and $\Delta x_g=\Delta x/g$
For each microlens $\boldsymbol{x}=(x_1, x_2)'$
  For each pixel in that microlens $\boldsymbol{u}=(u_1, u_2)'$
    Compute $\boldsymbol{k}=(k_1, k_2)'$, $\boldsymbol{k}=\Delta u_g\boldsymbol{x}-\Delta x_g\boldsymbol{u}$
    Update the mean value in $F_{\Delta u/\Delta x}(\boldsymbol{k})$ using $L_F(\boldsymbol{x},\boldsymbol{u})$.

The result of the algorithm is shown on Figure 5 for a region of the image in Figure 2.

Figure 5 – Results of the super-resolution focal stack with a 25× resolution magnification

3. MULTIVIEW STEREO

The multiview stereo algorithms can operate in the super-resolution laplacian [3] or variance focal stack and are able to obtain a $t\times t$ estimation of depth ($t=\Delta u$ n $-\Delta u$ $+1$) with m-3 different depth values and a $t\times t$ all-in-focus image.
We will examine in detail the use of the variance focal stack. The variance focal stack is based on the photoconsistency assumption (PA) which states that the radiance of rays from a 3D point over all directions is the same (the lambertian model). To test the PA assumption we simply estimate the variance of all points along every line (see Figure 4). A multiview stereo algorithm can use this variance measure to choose the optimal depth. A straightforward approach to solve the multiview stereo problem would be to choose for each pixel the line with less variance. However it is well known that to obtain good results it is necessary to add more assumptions. If we assume that surfaces are smooth we can use the variance focal stack and the Markov Random Field (MRF) approach to obtain the optimal depth minimizing an energy function [6]. This energy function is composed of a data energy E_d and smoothness energy E_s, $E=E_d+\lambda E_s$, where the parameter λ measures the relative importance of each term. The data energy is simply the sum of the per-pixel data costs $E_d=\Sigma_{\boldsymbol{p}}$ $c_{\boldsymbol{p}}(d)$ where $c_{\boldsymbol{p}}(d)$ is the variance measure for pixel $\boldsymbol{p}$ in the super-resolved image and $d=\Delta u/\Delta x$. To define the smoothness energy we use the standard 4-connected neighborhood system, so the smoothness energy E_s can be written $E_s=\Sigma_{p,q}$ $V_{pq}(d_p, d_q)$ where $\boldsymbol{p}$ and $\boldsymbol{q}$ are 4-connected pixels and $V_{pq}(d_p, d_q)=\min(\mu,|d_p-d_q|)$. Since there are occasions (autostereoscopic displays) where the desired final resolution may be lower than the maximum attainable resolution we optimize the energy function E using hierarchical belief propagation [7]. The belief propagation local updating rule also makes easier to port the whole technique to GPU's and FPGA's.

4. EXPERIMENTAL RESULTS

In this section, we describe the experiments used to evaluate the super-resolution algorithm.

4.1 Test data

To evaluate the technique we have used the image shown on Figure 2 [5]. The image has 4380×4380 pixels with 292×292 microlenses and 15×15 pixels for each microlens. We compensate the intensity fall-off effects on the edges of microlens using a histogram-based non-linear transform.

4.2 Experimental results

Using the 292×292×15×15 lightfield, previous estimation algorithms would give a 292×292 all-in focus and depth map. This is insufficient for current autostereoscopic displays. We use an autostereoscopic Philips 20-2D2W04/00 display with 800×600 3D resolution [8] so we have employed a 3× magnification factor to use the full display resolution. In figure 6 we can see the all-in-focus image and depth estimation from the super-resolution focal stack and multiview stereo algorithms. In Figures 7 and 8 we have selected an area from the plenoptic image to compare super-resolution (right) with an bicubic interpolation-based magnification (left) of the all-in focus 292×292 image computed with previous approaches. Results show a substantial improvement in detail and noise-reduction.

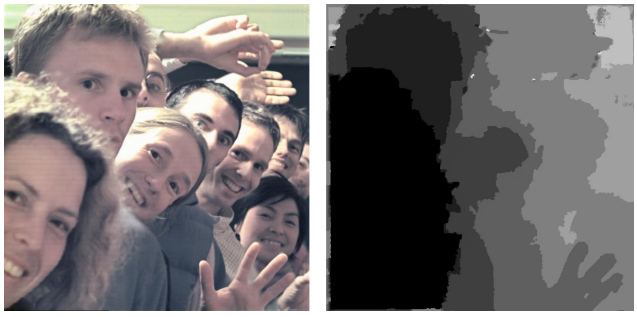

Figure 6 – Super-resolved depth an all-in-focus image

Figure 7 – Super-resolved (right) and bicubic interpolation based magnification (left)

Figure 7 –Super-resolved (right) and bicubic interpolation based magnification (left)

5. CONCLUSIONS AND FUTURE WORK

We have implemented a new technique for the simultaneous estimation of super-resolved depth maps and all-in-focus images from a plenoptic camera. This technique enables the use of current plenoptic photo and video cameras in present autostereoscopic displays. Future improvements will consist on porting the technique to GPU's (Graphic Processing Units) and FPGA (Field Programmable Gates Arrays) to obtain real-time processing. Another future extension will be the integration of our technique with conventional motion based super-resolution.

6. ACKNOWLEDGMENTS

This work has been funded by "Programa Nacional I+D+i" (Project DPI 2006-09726) of the "Ministerio de Ciencia y Tecnología", and by the "European Regional Development Fund" (ERDF).

7. REFERENCES

[1] L. Onural: "Television in 3-D: What Are the Prospects?", Proc. of the IEEE vol 95, nº 6, pp. 1143 – 1145, 2007.

[2] The Cafadis Camera: International Patent number PCT/ES2007/000046 (WIPO WO/2007/082975)

[3] F. Pérez, J.G. Marichal, and J.M. Rodríguez-Ramos, "The Discrete Focal Stack Transform", Proc. of the Eusipco, 2008.

[4] A. Lumsdaine, T. Georgiev, Full Resolution Lightfield Rendering, Adobe Tech Report, January 2008.

[5] R. Ng, "Fourier Slice Photography". Proc. of SIGGRAPH, 2005.

[6] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/2/3):7-42, 2002

[7] P.F. Felzenszwalb; D.R., Huttenlocher, "Efficient belief propagation for early vision," *Computer Vision and Pattern Recognition, 2004. CVPR 2004.* vol.1, pp. I-261-I-268, 2004

[8] Philips Wowx Display Website: http://www.business-sites.philips.com/3dsolutions/Products/3DScreens/Index.html