DEPTH MAP ESTIMATION IN LIGHT FIELDS USING AN STEREO-LIKE TAXONOMY

ESTIMACIÓN DE MAPA DE PROFUNDIDAD EN CAMPOS DE LUZ MEDIANTE EL USO DE UNA TAXONOMÍA TIPO ESTÉREO

Francisco C. Calderon,[1] Carlos A. Parra[1], Cesar L. Niño[1]

[1.] This work was partially founded by Colciencias Grant #2011-528 Grupo de Investigación en Sistemas Inteligentes, Robótica y Percepción.
Pontificia Universidad Javeriana, Bogotá DC, Colombia
Facultad de Ingeniería Electrónica

Recibido: 10 Febrero de 2015
Aceptado: 2 Febrero de 2016

*Correspondencia del autor: Francisco C. Calderon. E-mail: calderonf@javeriana.edu.co

ABSTRACT

The light field or LF is a function that describes the amount of light traveling in every direction (angular) through every point (spatial) in a scene, this LF can be captured in several ways, using arrays of cameras, or more recently using a single camera with an special lens, that allows the capture of angular and spatial information of light rays of a scene (LF). This recent camera implementation gives a different approach to find the dept of a scene using only a single camera. In order to estimate the depth, we describe a taxonomy, similar to the one used in stereo Depth-map algorithms. That consist in the creation of a cost tensor to represent the matching cost between different disparities, then, using a support weight window, aggregate the cost tensor, finally, using a winner-takes-all optimization algorithm, search for the best disparities. This paper explains in detail the several changes made to an stereo-like taxonomy, to be applied in a light field, and evaluate this algorithm using a recent database that for the first time, provides several ground-truth light fields, with a respective ground-truth depth map.

Palabras claves: Stereo, Light field, smoothing filter, Depth Map, Stereo Taxonomy

RESUMEN

El campo de luz (CL) es una función que describe la cantidad de luz que viaja en toda dirección (angular) a través de cada punto (espacial) en una escena; este CL se puede capturar de varias maneras, utilizando matrices de cámaras, o más recientemente utilizando una sola cámara con un lente especial que permite la captura de información angular y espacial de los rayos de luz de una escena. Esta reciente implementación de cámara brinda un acercamiento diferente para hallar la profundidad de una escena mediante el uso de una sola cámara. Para estimar la profundidad, describimos una taxonomía, similar a la utilizada en algoritmos en estéreo de mapa de profundidad, que consisten en la creación de un tensor de coste para representar el costo coincidente entre diferentes disparidades, luego, con el uso de una ventana de apoyo de peso, agregamos el tensor de coste, finalmente, mediante el uso de un algoritmo de optimización del 'ganador toma todo', buscamos las mejores disparidades. Este documento explica en detalle varios de los cambios realizados a una taxonomía tipo estéreo, para ser aplicada en un campo de luz y evaluar este algoritmo mediante el uso una base de datos reciente que por primera vez, provee varios campos de luz de realidad sobre el terreno, con un respectivo mapa de profundidad de realidad sobre el terreno.

Keywords: estéreo, campo de luz, filtro de suavizado, mapa de profundidad, taxonomía en estéreo

1. INTRODUCTION

Thanks to different devices that capture the light field (LF), new applications have emerged to use this representation of the light in the scene.

The light field, as stated, is a 4D function, that describes the light that comes from the scene, using a parametric representation, the most commonly used, consist in the parameterization using the intersection points of the light rays with two parallel planes, separated by a distance f, usually $f = 1$. The intersection points are two, denoted by the 4 coordinates *L(u, v, s, t)* the first two coordinates (*u, v*) are called the pixel or spatial indexes, and (*s, t*) the camera or angular indexes. This model was first proposed by Levoy and Gortler in (1), and later improved and well explained in (2). As shown in figure 1.

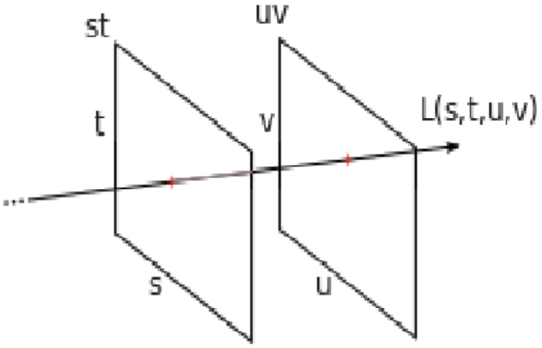

Figure 1 Light field model *L(u, v, s, t)* of Levoy, Gortler(2)

This LF can be used to display different views of the scene, or to create a synthetic aperture focusing. Another area of recent interest is the light field rendering, which deals with angular interpolation of novel views, starting from the originally captured views. Also, the depth estimation can be applied in the LF, to improve the virtual refocusing or to characterize the scene. (2) The structure of the light field, makes it very similar to an array of multiple cameras, so, it is possible to extract the depth of the scene from a LF, using an algorithm based in a stereo or multi-camera a taxonomy. Like in (3, 4).

The taxonomy of these multi-camera algorithms is well known, and is divided in four classes:

- The first class of algorithms operates by computing a 3D cost function, then, extracting a surface from this volume. The cost function may differ, and also the surface extraction method. These algorithms use a local or global optimization method, to extract an optimal surface, for example in (5), The Optimization method solve a sparse lineal system to obtain a LF depth map.
- The second class of algorithms, works by iteratively evolving an initial depth estimation to minimize a cost function, for example (4).
- The third class of algorithms, are the image-space methods, these class, first computes a set of depth maps, to then merge the set of depth maps into a 3D scene as a post process (4).
- Finally, the fourth class, consists of algorithms who first extract and match a set of feature points, to then fit a surface to the reconstructed features points, as is done in (6), or more recently to LF in (7).

The Taxonomy that is used in this paper, follows the first class of algorithms, it also is based in the taxonomy of stereo like algorithms, treated in (3). It is possible to find a distance between the cameras and an object, from images taken by two cameras with known locations, using just a simple trigonometric relationship,

$$(1)\ Z = f\frac{D}{(x_L - x_R)} = f\frac{D}{d}$$

where Z is the distance to an interest point in space, D is the rectified distance between the optical camera centers, f the focal length of the cameras and $\{x_L, x_R\}$ are the distances reflected in the image plane, measured from the beginning of each image $\{O_L, O_R\}$. Where $(x_L - x_R)$ is called the disparity and is denoted by the letter d, from this, it is possible to find the distance at which there is a point in the image space common to both cameras, using equation 1, which is known a priori, the focal length of the calibrated cameras, and the distance between the cameras, which is an easily measurable parameter (8).

The LF is a slightly more complicated structure than stereo, however, the depth information is also encoded in the data present in the LF. To analyze the depth in the LF, it is first necessary to discuss the epipolar images (EPI), an EPI is a simplification in 2D, of two mixed coordinates in the LF, one Angular, and one spatial, the most used one is E(*u, s*), follow by E(*v, t*). the scene elements are represented in the EPI, as lines, the depth of these objects are proportional to the slope of these lines, for example, an object in the focal plane will create a completely vertical line in the EPI.

As another object is moving towards or away from the focal plane, their slope in the EPI will change proportionally. An example of a EPI is shown in Figure 2, a visual explanation of depth in the LF is shown in Figure 3.

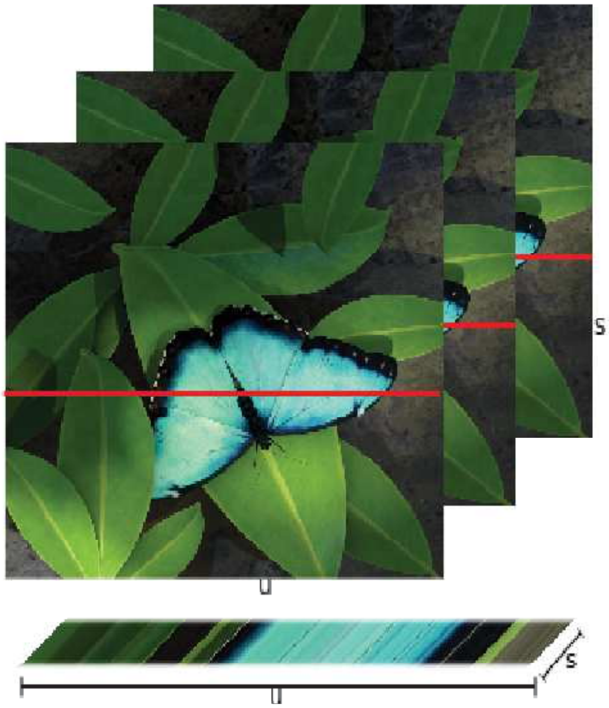

Figure 2 A sample of LF, three spatial samples along the s dimension, and an EPI. From the database of (9).

1.1 Stereo Algorithms

This approach makes that the process of determining the distance to a point, seen by a pair of stereo cameras is equivalent to find the disparity between the images of two rectified cameras.

Most of stereo matching algorithms can be divided into a general taxonomy, proposed in (3) consisting of four basic steps:

1. Matching cost computation.
2. Cost (support) aggregation.
3. Disparity computation / optimization.
4. Disparity refinement.

In this taxonomy stereo matching algorithms can be classified into two major groups, local and global, depending on how they operate the image. In a local algorithm the disparity search is done in small regions using windows to aggregate locally on the neighborhood of a pixel, this have the advantage of be quick and easy to implement, these algorithms assume that all pixels within the matching region have the same disparity, this is known as the fronto-parallel assumption, and is not valid around disparity discontinuities, and leads commonly to the fattening of these discontinuous regions when large windows are used.

Contrary to local methods, global algorithms perform a search in the whole image, their goal as is shown in (3), is to find an optimal disparity *d* per pixel $p_{(x,y)}$ which minimizes a cost function.

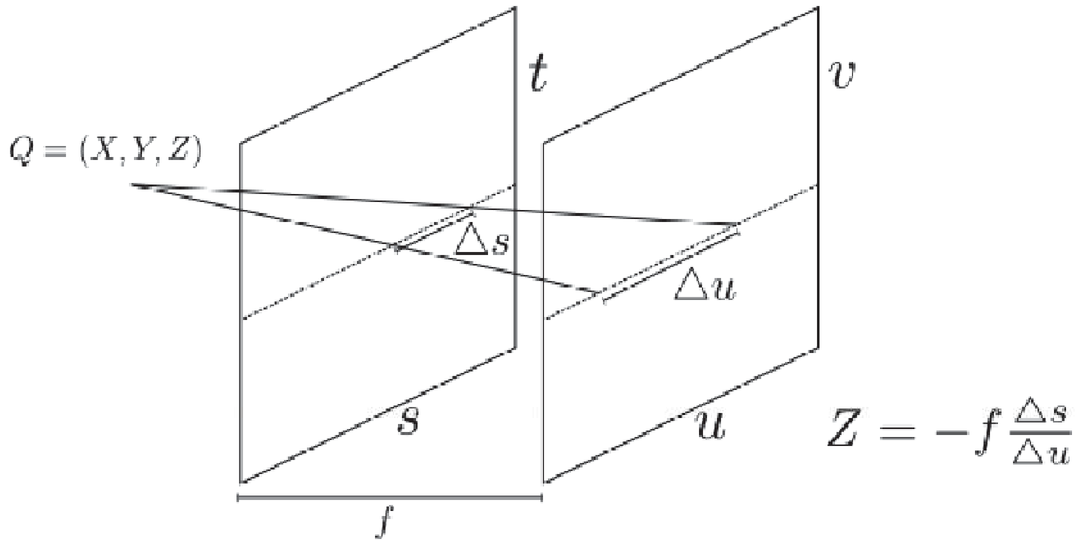

Figure 3 A scene point *P*(*X, Y, Z*) in space, and its projection in the LF, with planes separated by a distance *f* using Light field model *L*(*u, v, s, t*) of (2). It is possible to find the depth *Z* by using the differences in the epipolar projection of the point. This is equivalent to find the slope 0 *s*/0 *u* .

2 Depth Evaluation in the LF

Based in the taxonomy of stereo algorithms, is possible to create a depth estimation algorithm to be used in light fields, the changes of an standard stereo algorithm are discussed in detail in the next sub section:

2.1 Cost computation

A Cost Volume $C_s(p_{(u,v)}\ , d_s)$, is created from the LF, given by equation 2, all elements of this volume, stores the color difference (or dissimilarity) at several disparities between pixels. Each pixel in the light $L(s, t, u, v)$, is compared with all pixels in $L(s, t, u - sd_s, v)$, for $ds = \{\mathrm{d}_{\mathrm{smin}}, ..., \mathrm{d}_{\mathrm{smax}}\}$, where dsmin and dsmax are the minimum and maximum expected disparities, the values of $\mathrm{d}_{\mathrm{smin}}$ and $\mathrm{d}_{\mathrm{smax}}$ are determined by the capture conditions, for example, the camera intrinsics, and the characteristics of the scene. Normally camera intrinsics are previously known.

The cost then is given by:

(2) $C_s(p_{(u,v)}, d_s) = / _{i=1}^{i=c} N_{ds} \ (E_s[L_c \{s,t,u\text{-}sd_s,v\}]\, I_s\text{-}L_c \{s,t,u\text{-}sd_s,v\})$

where c is an index of color channels in the radiance of the light rays, this model, only takes in to account color spaces where the values of channels are equivalent, for example RGB color space. The variable E_s, is an expected value estimator along one of the two angular coordinates. For example, the coordinate s in Equation 2. In this document, the mean and the median are used as E_s. The function N_{ds} is the cost measure distance of a pixel $p_{(u,v)}$, for an specific disparity d_s. Finally, I_s is an identity matrix, placed to repeat $E_s[L_c\{s, t, u - sd_s, v\}]$ along the s dimension and allow the dimensionality of the difference.

2.2 Cost aggregation

The cost function $C_s(p_{(u,v)}, \mathrm{d}_s)$ to find the disparity, usually is object of many manipulations to increase the performance in the final step. A common operation consist in the aggregation of cost using windows of constant disparity d, over a three-dimensional cost Csp_((u,v) , ds). An extensive comparison of mean, Gaussian, shiftable and others square windows, can be found in (3). Recently, adaptive support windows have been proposed, and are well studied in (10). In this work, as an evaluation procedure, are used squared Gaussian windows of size N and standard deviation v_{N} , following the three-sigma rule.

The cost aggregation can be represented by the Equation 3

(3) $C_{ws}(p_{(u,v)}, d) = W(q_{w(m,n)}\ , d)*C_s(p_{(u,v)}, ds)$,

where W represents the window filter of size m × n, in the evaluation, only squared kernel windows are used, of size $n \times n$, to calculate $C_{ws}(\mathrm{p}_{(u,v)}, d)$.

2.3 Disparity computation / optimization

Disparity is selected searching in Equation 3, for the minimum cost per pixel along the disparity dimension $c_{min}(u, v)$, which occurs at disparity $d_{mins}(u, v)$. No other optimization algorithm was used in this step. The same occurs with the last step of the taxonomy, no disparity refinement algorithm was implemented, in order to have smoothing operations only in the aggregation step of the taxonomy.

3 Evaluation Methodology

The basic structure of our Evaluation methodology, follows completely the first three steps of the taxonomy proposed by (3), which consists in applying on a pair of rectified images, first, a matching cost computation (2.1), followed by the aggregation of cost using windows (2.2), then a disparity selection via Winner-Takes-All the final step or the post-processing is not taken into consideration in this document, there are many techniques but this step do not provide more information to the other three steps of the taxonomy and is heuristic. We decided to follow a taxonomy in order to make the algorithm more understandable, so that the results presented here can be used and reproduced. Each step of the evaluation is discussed below.

3.1 Cost computation

As the the cost measure distance function N_{ds} , are used the L_1 and L_2 norms. As an equivalent of the Sum of Absolute Differences (SAD), and the Sum of Squared Differences (SSD). Also as the expected value estimator variable E_s, are used the Median, and Mean.

3.2 Cost Aggregation

The performance of local stereo methods depends entirely on the election of a support weight, used in this step, [3, 11]. In this paper are used Gaussian windows, at 12 different squared sizes, of: 1(No window), and [3, 5, 7, ..., 29]. With larger than 29 × 29 windows, the algorithm lost performance due to the well known fattening effect(10).

3.2.1 Windows

The mean window represent only an average of the square vicinity of a central pixel, is the simplest window and do not require any parameter other than its size *n*.

Gaussian window requires Besides its size *n*, the value of the variance *σ*, this is calculated using an commonly used heuristic value in the literature of implementation(8) presented in equation 4.

$$(4)\ \sigma = 0.3(\frac{n}{2} - 1) + 0.8$$

3.3 Evaluation

The results were tested on the recent ground truth light fields, of the Heidelberg Collaboratory for Image Processing HCI, (9). which consist of 13 LF datasets the mean values over the 13 datasets are taken into consideration.

The results shown in the next section are the mean values of the following two error measures.

1. E_{RMS} Represent the *RMS* or Root Mean Squared error given by the equation 5.

$$(5)\ E_{RMS} = \sqrt{\frac{1}{T} /_{(x,y)} \{d_O\,(p_{(x,y)})\text{-}d_{GA}\,(p_{(x,y)})\}^2}$$

Where *T* is the number of pixels in the image, $d_{GA}\,(p_{(x,y)})$ the ground truth disparity map in all valid pixels.

B or percentage of of bad pixels given by the equation 6.

$$(6)\ B = \frac{1}{T}\ \sum_{(x,y)} \{|d_O\,(p_{(x,y)}) - d_{GA}\,(p_{(x,y)})| > \delta_d\}$$

Where δ_d is a disparity error tolerance. For the experiments we use 1.0 since this coincides with previously published studies (3, 6, 12, 13, 14, 15).

4 Results and conclusions

The results are plotted in Figures 4 through 7. The Figure 4 shows the mean *RMS* error, the lower, the better, in this case, the Median- L_1 estimator achieves the best results, with a minimum for a 5 × 5 window, The Figure 5 shows the mean percentage of bad pixels for a $\delta_d = 1$, the lower, the better, in this case, again the Median- L_1 estimator achieves the best results, with a minimum also with a 5 × 5 window.

In the Figure 6, is shown a ground-truth depth map for the sample (papillon) of the database of (9). A synthetic color-map was applied to the depth map, to increase the depth perception. The best depth map obtained in the *B* metric is shown on Figure 7. Finally in the Figure 8 is shown a resulting depth map without any aggregation window applied to the cost volume.

In the 4 cases studied, the best couple was Median-*L*1, which resembles other studies done in stereo (3, 16), however, it is noteworthy, that contrary to stereo, the size of the windows with the best solution is lower, this means that the inflection point in where the fattening effect that affects the metric, is lower than in stereo. as may be compared to previous studies (3), we believe this is due to the greater amount of redundant information obtained from the multiple views in the LF, that makes it require less filtering, and also the lack of noise in the synthetic LF .

An analysis of the curves that make use of the L^1 norm and the L^2 norm, it can be seen that L^1 gives a better estimate of the depth when is used in conjunction with an aggregation window, which is the usual case, only was overwhelmed by a short range by the *L*1 norm, when no window is used.

It is hoped that this work will contribute to the analysis of the depth of the light fields, providing a well detailed taxonomy, for future researchers to improve existing algorithms and provide a tool to guide future depth map estimation algorithms using light field, all metrics in this document can be compared with the state-of-the-art of similar works in (3, 9). A sample code to support this article can be found in matlab-central.

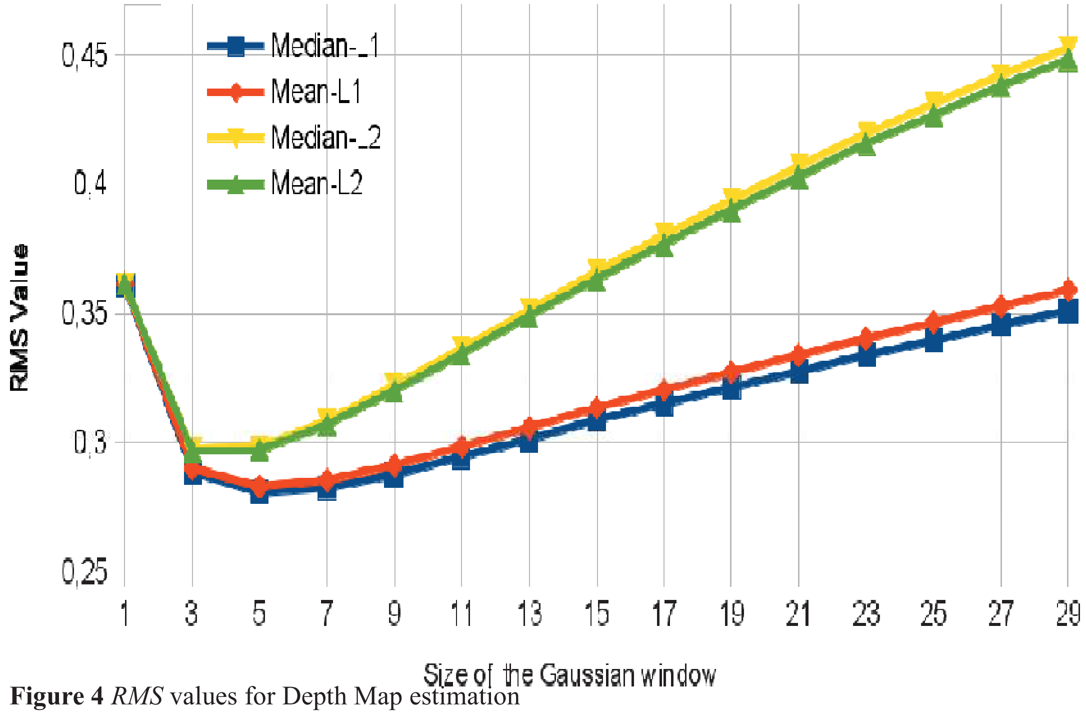

Figure 4 *RMS* values for Depth Map estimation

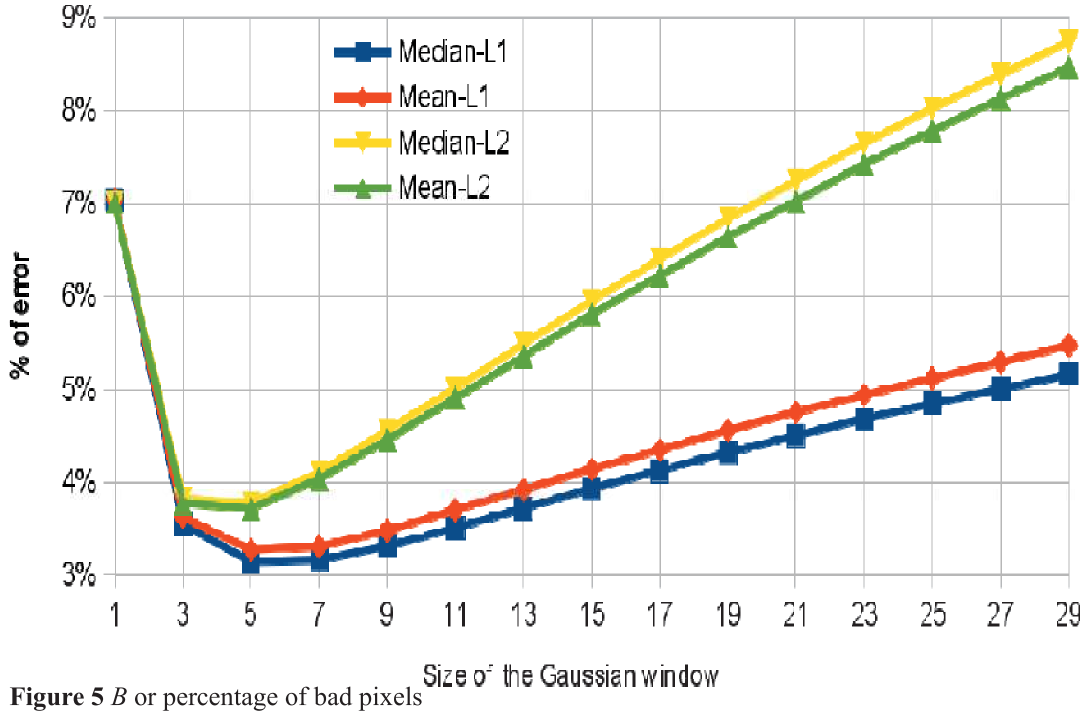

Figure 5 *B* or percentage of bad pixels

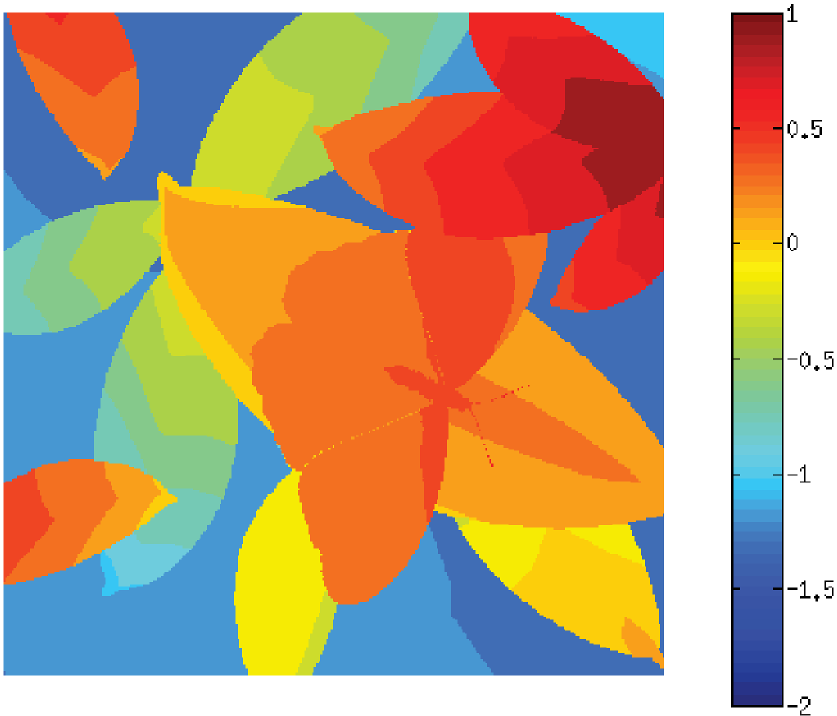

Figure 6 Ground truth for the LF database of (9), for the sample called papillon.

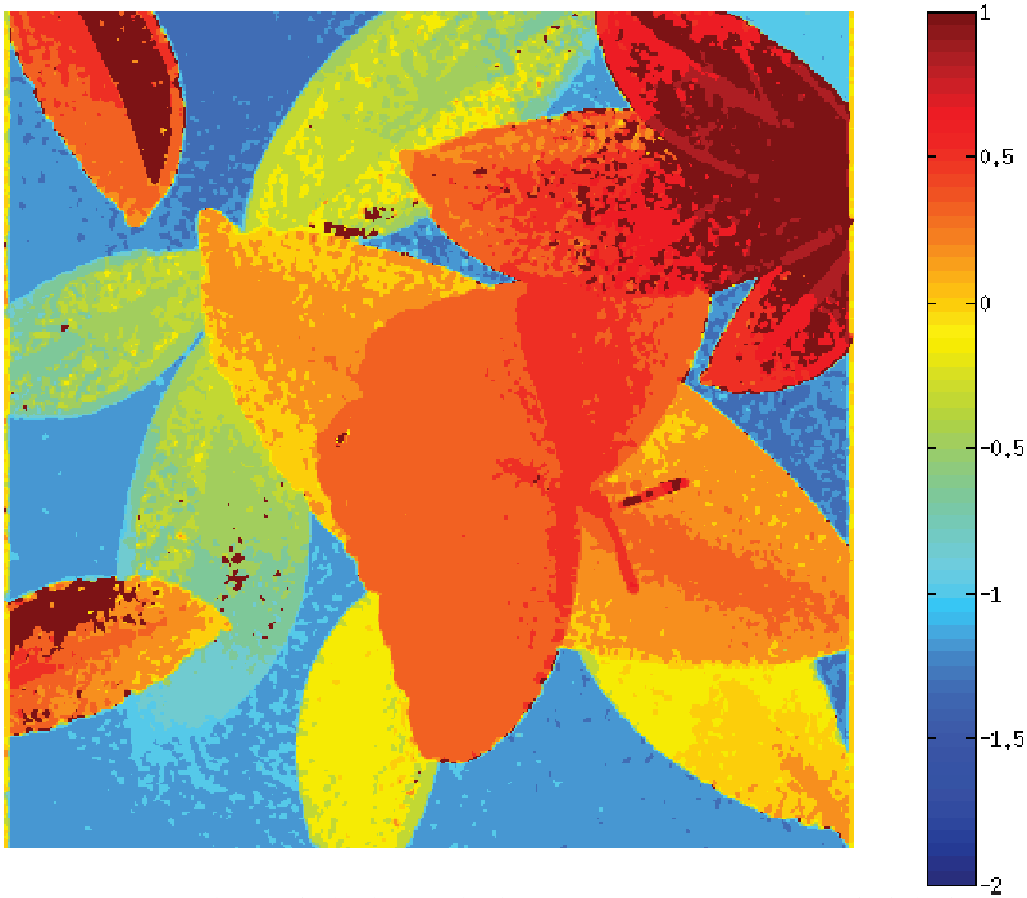

Figure 7 Best Depth map obtained for evaluation parameter *B* or percentage of bad pixels.

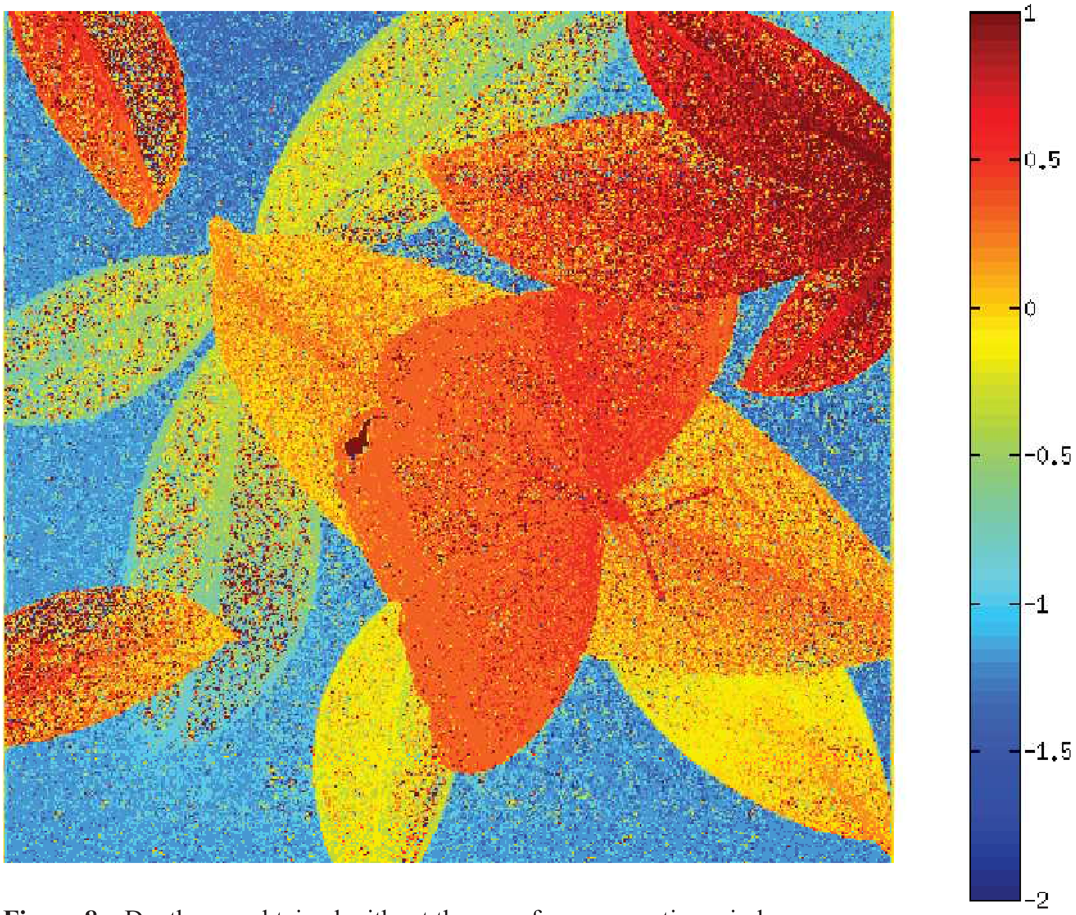

Figure 8 Depth map obtained without the use of an aggregation window.

BIBLIOGRAPHY

1. Levoy M, Hanrahan P. Light Field Rendering. In: SIGGRAPH; 1996.p. 1–12.
2. Levoy M. Light fields and computational imaging. IEEE ComputerSociety. 2006.
3. Scharstein D, Szeliski R. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal ofComputer Vision. 2002;47(1-3):7–42.
4. Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Volume 1 CVPR06. vol. 1. IEEE. Ieee; 2006. p. 519–528.
5. Li J, Li Z. Continuous depth map reconstruction from light fields. In: Multimedia and Expo (ICME), 2013 IEEE. vol. 2; 2013. .
6. Geiger A, Roser M, Urtasun R. Efficient large-scale stereo matching. Computer Vision ACCV 2010 Springer Berlin. 2010;6492:25–38.
7. Liang Ck, Lin Th, Wong By, Liu C, Chen HH. Programmable Aperture Photography: Multiplexed Light Field Acquisition. In: SIGGRAPH; 2014. p. 1–10.
8. Bradski GR, Kaehler A. Learning OpenCV: Computer Vision with the OpenCV Library. 1st ed. Sebastopol, Canada: O'Reilly; 2008.

9. Wanner S, Meister S, Goldluecke B. Datasets and Benchmarks for Densely Sampled 4D Light Fields. In: Vision, Modelling and Visualization (VMV); 2013. .
10. Hosni A, Bleyer M, Gelautz M. Secrets of Adaptive Support Weight Techniques for Local Stereo Matching. Computer Vision and Image Understanding. 2013 Feb;.
11. Hirschmuller H, Scharstein D. Evaluation of stereo matching costs on images with radiometric differences. IEEE transactions on pattern analysis and machine intelligence. 2009 Oct;31(9):1582–99.
12. Damjanovi ́ c S, van der Heijden F, Spreeuwers LJ. Local Stereo Matching Using Adaptive Local Segmentation. ISRN Machine Vision. 2012;2012:1–11.
13. Heijden FVD, Spreeuwers LJ, Group S. Local Stereo Matching Using Adaptive Local Segmentation. 2012;p. 1–15.
14. Klaus A, Sormann M, Karner K. Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. In: 18th International Conference on Pattern Recognition (ICPR’06). IEEE; 2006. p. 15–18.
15. Vladimir Kolmogorov RZ. Multi-camera Scene Reconstruction via Graph Cuts. In: European Conference on Computer Vision; 2002. p. 82 – 96.
16. Nalpantidis L, Sirakoulis G, Gasteratos A. Review of Stereo Vision Algorithms: From Software to Hardware. International Journal of Optomechatronics. 2008;2(4):435 – 462.