CONTINUOUS DEPTH MAP RECONSTRUCTION FROM LIGHT FIELDS

Jianqiao Li and Ze-Nian Li

School of Computing Science
Simon Fraser University
Burnaby, B.C., Canada
{jianqiao_li, li} @sfu.ca

ABSTRACT

Light field analysis recently received growing interest, since its rich structure information benefits many computer vision tasks. This paper presents a novel method to reconstruct continuous depth maps from light field data. Conventional approaches usually treat depth map reconstruction as an optimization problem with discrete labels. On the contrary, our proposed method can obtain continuous depth maps by solving a linear system, which preserves richer details compared with conventional discrete approaches. Structure tensor is employed to extract raw depth information and corresponding confidence levels from the light field data. We introduce a method to reduce the adverse effect of unreliable local estimations, which helps to get rid of errors in specular areas and edges where depth values are discontinuous. Experiments on both synthetic and real light field data demonstrate the effectiveness of the proposed method.

Index Terms— Depth map reconstruction, light field, linear system

1. INTRODUCTION

Light field (LF) is a function that describes the radiance at each point in a 3D space in every direction. As the technique of capturing light fields develops, light field analysis is of great interest in recently years. Since light field data contains not only accumulated color intensity at each image point, but also some information about ray directions, many computer vision problems can be better solved by making use of this structure information, such as virtual refocusing [1], tracking through occlusions [2] and reconstructing occluded surfaces [3]. In this paper, we focus on depth map reconstruction from the light field data.

Reconstructing depth map from stereo image pairs, also known as stereo matching, is a traditional challenging computer vision task, which has been studied for more than three decades [4]. More recently, depth from moving camera [5] and depth from integral images [6] are also investigated. Studies on depth reconstruction from light fields have just started, and most of them are focused on certain plenoptic cameras [7]. More studies are needed on how the special structure of light fields can benefit depth estimation.

Most traditional approaches treat depth map reconstruction as an optimization problem with discrete labels. Markov Random Field (MRF) model [8] is widely used in this area, which supports various definitions of energy functions. However, a drawback of these multi-labelling methods is that both time complexity and memory cost increase rapidly with the image resolution and the number of labels. To solve the problem in a reasonable time, the number of discrete depth labels is usually set as a small number (32 or 64). Consequently the reconstructed depth maps usually have noticeable "stairs". Otherwise, long running time and large memory costs are needed. Now thanks to the rich structure information in light fields, continuous local estimations are available, which makes it possible to seek for a continuous final result instead of the optimal solution of a multi-labelling problem.

Our work has two main contributions. First, we introduce a refinement step to local depth estimation, which helps to reduce the effect of unreliable estimations. This is presented in Section 3. Second, we propose a continuous method to get a smooth depth map from local estimations by solving a sparse linear system. We introduce this method in Section 4. Experimental results are presented in Section 5.

2. RELATED WORK

2.1. EPIs and Structure Tensor

4D light field is first proposed in [9] and later widely used in light field analysis. We adopt the two-plane parametrization [9] of 4D LF and denote a LF as $L(x, y, s, t)$, as shown in Figure 1. Under this parametrization, a 4D LF can be seen as a 2D array of perspective views, where (s, t) can be seen as the index of different views and (x, y) are spatial coordinates within each view (see Figure 2(a)).

By fixing y and t, we can obtain a 2D (x, s) slice of a LF, as shown in Figure 2(b). Similarly, 2D (y, t) slices can be obtained if x and s are fixed. These 2D slices are called epipolar plane images (EPIs). Any point in the 3D space can be projected to a line on EPIs. And the slope of the line are

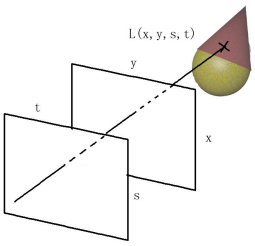
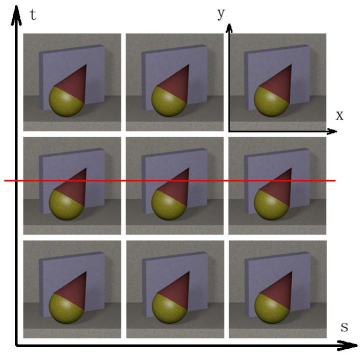
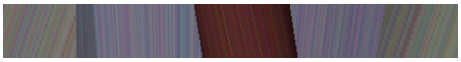


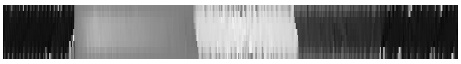
Fig. 1. Two-plane parametrization of 4D light field.



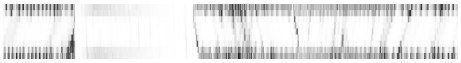
(a) A visualization of 4D light field.



(b) Epipolar plane image.



(c) Local depth estimation on EPI.



(d) Confidence map on EPI.

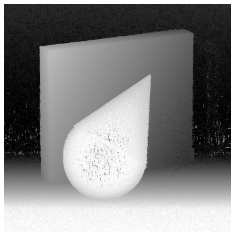
Fig. 2. Local estimation on EPIs.

shown to be related to the depth of the corresponding point in the 3D space [10].

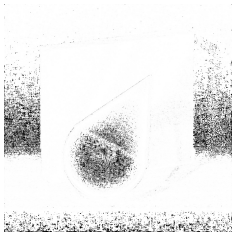
Therefore, depth values can be obtained by estimating the slope of lines in EPIs [11]. An structure tensor [12] is employed, which produces an orientation estimation at each point and the confidence level of each estimation. The depth is derived as

$$d(x,s) = -f\frac{\Delta s}{\Delta x} = -f \cdot \cot\theta(x,s), \qquad (1)$$

where θ is the estimated orientation by the structure tensor. f is the distance between the two parallel planes in Figure 1. By analyzing every 2D slice with different possible y, we can obtain the local depth estimations and their confidence levels of each view, as shown in Figure 3. Based on the local estimation we try to construct a continuous depth map instead of making use of multi-labelling methods.



(a) Local depth estimation



(b) Confidence map

Fig. 3. An example of local depth estimation.

However, the local estimation still has several limitations. It tends to give wrong estimations in areas where depth is discontinuous, but still assigns high confidence levels for these estimations, which has an adverse effect on the future step. Besides, it fails to produce reliable estimations on specular areas and texture-less areas. To fix these problems, Wanner and Goldluecke [11] employ a variational labeling method [13] to enforce visibility constraint on each EPI. However, since this optimization is on each EPI, they need to optimize hundreds of times for each LF, which usually takes several hours. An alternative way to do visibility reasoning is to construct occlusion maps, and iteratively optimize the depth maps and occlusion maps [14]. Instead of explicit visibility reasoning, Zhang et al. [5] incorporate visibility into data term of energy function using statistical information from both color and geometry. In this paper, we propose a method to refine the confidence map of the local estimation, so that wrong estimations are always assigned with low confidence.

2.2. Depth Map Reconstruction

Most conventional approaches follow the MRF model to construct depth map. An energy function is formulated, where the label costs are encoded in a data term and the spatial smoothness is enforced by a pairwise smooth term. Various methods are employed to minimize the global energy function [8], such as graph cut, loopy belief propagation. Varational labeling methods [15, 13] are used in [11] to achieve global integration. However, all those methods treat the problem as a multi-labelling problem, where depth values are quantized to certain levels. Depth map reconstruction in these frameworks is to assign the depth to a closest label. To achieve a smoother result, larger number of levels should be set. However, this will make the time and memory costs of solving the problem increase rapidly.

Given that local depth estimation is available for LF data, we discard the multi-labelling framework, but treat the problem as a continuous optimization problem. Depth values in the proposed method are never quantized except when we want to visualize the results as digital images. We rewrite the energy function in MRF model into a matrix form, and formulate a sparse linear system. A similar method is em-

ployed in [16] to propagate depth information from ground control points. However, in their work this is only an intermediate step. An multi-labelling method is later used to get quantized final results. Very few attempts to construct continuous depth maps based on linear systems have been done in this field. However, similar methods have been well practised in many other computer vision tasks, such as matting [17, 18] and colorization [19].

3. CONFIDENCE MAP REFINEMENT

Local depth estimations and their confidence levels are obtained by applying the structure tensor [12] on EPIs of LFs [11]. However, in some cases the structure tensor gives wrong estimations but assign high confidence levels for them, especially in areas where the depth is discontinuous. As shown in Figure 4(a)(b)(c), local estimation induces "fattened" boundaries along stems of the plant, and wrong confidence levels are assigned to them. In this step, we check the color consistency between different views, and give a penalty on the confidence values of wrong local estimations. As the example shows in Figure 4(d), wrong estimations along boundaries are assigned to low confidence values after this step.

In each EPI, any point (x, s) can be warped to other views, given the estimated depth $d(x, s)$. Ideally, the color at the original point and at the warped point should be identical, so are the estimated depth values. We define a matching distance to measure the difference between the original point and the one warped onto view s'.

$$\Phi(x, s, s') = ||\mathbf{I}(x, s) - \mathbf{I}(x', s')|| + c(x', s')|d(x, s) - d(x', s')|, \quad (2)$$

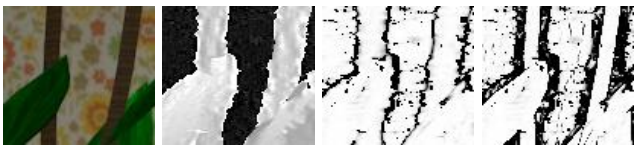
where $\mathbf{I}(x, s)$ and $\mathbf{I}(x', s')$ are three-dimensional vectors for color intensity, and $d(x, s)$ and $d(x', s')$ are estimated depth values. $c(x', s')$ is the confidence level of $d(x', s')$. If both $d(x, s)$ and $d(x', s')$ are perfectly correct, and the surface is Lambertion, the distance $\Phi(x, s, s')$ should be zero. Large distance indicates that the depth estimation at this point is not reliable.

We warp point (x, s) to different available views, and accumulate the distance. Then the accumulated distance is mapped to a penalty coefficient $p(x, s)$ for confidence level $c(x, s)$,

$$p(x, s) = 1 - \frac{1}{1 + exp(\frac{1}{\beta}(\alpha - \frac{1}{||S||}\sum_{s' \in S} \Phi(x, s, s')))} \quad (3)$$

$$c'(x, s) = p(x, s) \times c(x, s) \quad (4)$$

where S is the set of all possible views. As the accumulated distance goes larger, the penalty coefficient goes to zero, thus the refined confidence level is also closed to zero. Otherwise the estimation is considered reliable, and its confidence level is almost unchanged. Parameters α and β control the shape of the penalty function, which are empirically set as 20 and 1 respectively in the experiments.



(a) (b) (c) (d)

Fig. 4. An example of confidence refinement. (a) a close-up of the center view. (b) the local depth map, which has "fattened" boundaries. (c) the raw confidence map from structure tensor. (d) the refined confidence map, in which wrong estimation along boundaries are assigned to low confidence values. See Figure 5(a) for the full image.

Because of possible occlusions, it is better to only accumulate matching distance on visible views rather than go over all the views. Thus the temporal selection scheme in [20] is employed. Besides, to avoid long running time, we sample five views on each side of s.

Apparently, all the above analysis is also applicable to the (y, t) slices, if x and s are fixed. We actually can get two pairs of local depth maps and refined confidence maps, by analysing (x,s) slices and (y,t) slices. They are merged together by adopting the depth value from the one with higher confidence at each pixel.

4. OPTIMIZING DEPTH MAPS

4.1. Optimization by Solving a Linear System

As shown in Figure 3 and Figure 4, local depth maps are not reliable and globally consistent. At this stage, we aim at getting an optimized depth map from the local depth map and corresponding confidence map.

We write the energy function in a matrix form,

$$J(\mathbf{d}) = \mathbf{d}^T L \mathbf{d} + \lambda(\mathbf{d} - \tilde{\mathbf{d}})^T C(\mathbf{d} - \tilde{\mathbf{d}}), \quad (5)$$

where $\mathbf{d}$ and $\tilde{\mathbf{d}}$ are $N \times 1$ vectors, represent optimal depth values and local depth values respectively. N is the number of pixels in each view (i.e. $N = P \times Q$, if the resolution of the image is $P \times Q$). We want to find the optimal $\mathbf{d}$, which minimizes the energy function $J(\mathbf{d})$. In the first term, L is an affinity matrix, which enforces the points with similar colors to have similar depth values within a small neighbourhood. The second term is a data term, which makes the optimized result constrained by local depth estimations. C is a diagonal matrix, whose elements are confidence levels of corresponding pixels. Consequently, pixels with more reliable local estimations are more tightly constrained by the data term. λ controls the weight of the data term.

To optimize $\mathbf{d}$, we can take the derivative of $J(\mathbf{d})$, and try to find the optimal $\mathbf{d}$ that makes the derivative zero. As

a result, the cost function (5) can be minimized by solving a sparse linear system.

$$\frac{\partial J(\mathbf{d})}{\partial \mathbf{d}} = 2\mathbf{d}^T L + 2\lambda(\mathbf{d} - \tilde{\mathbf{d}})^T C = 0. \quad (6)$$

$$(L + \lambda C)\mathbf{d} = \lambda C\tilde{\mathbf{d}}. \quad (7)$$

By defining the affinity matrix L properly, we can make $L + \lambda C$ a symmetric positive definite matrix. Then this sparse linear system can be solved with the conjugate gradient method. Two formulations of affinity matrix are introduced, which are explained in detail in Section 4.2.

4.2. Affinity Matrix

A straightforward formulation of the affinity matrix L is

$$L = (I - W)^T(I - W). \quad (8)$$

Elements in W are defined as

$$W_{ij} = \begin{cases} \alpha_{ij}/\sum_{k\in N(i)} \alpha_{ik} & \text{if} j \in N(i) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$\alpha_{ij} = \max(\exp(\frac{-\Delta \mathbf{I_{ij}}}{\gamma}), \epsilon). \quad (10)$$

$N(i)$ is the neighbourhood of pixel i, and α_{ij} is a pair-wise weight based on color difference of neighbouring pixels. γ and ϵ control the sharpness and the lower bound of the exponential function. With this formulation, the first term in Equation (5) is identical with the typical smooth term in energy functions used in the area of stereo matching,

$$E_{smooth}(\mathbf{d}) = \sum_i (\mathbf{d_i} - \frac{\sum_{j\in N(i)} \alpha_{ij}\mathbf{d_j}}{\sum_{j\in N(i)} \alpha_{ij}}). \quad (11)$$

Although the affinity matrix is a sparse matrix, computing L and solving the linear system still take a long time with the formulation (8). If a large window size is used, which makes the matrix less sparse, even higher time and memory costs are needed to solve the system.

To make the method more efficient, we also tried another formulation, known as the matting Laplacian matrix [18]. A faster algorithm [17] with large window sizes is available to solve this system, if the matting Laplacian matrix is adopted. The (i, j) element of this matrix is defined as

$$\sum_{k|(i,j)\in\omega_k} (\delta_{ij} - \frac{1}{|\omega_k|}(1 + (\mathbf{I_i} - \mu_k)^T(\Sigma_k + \frac{\epsilon}{|\omega_k|}U)^{-1}(\mathbf{I_j} - \mu_k))), \quad (12)$$

where δ_{ij} is the Kronecker delta, μ_k and Σ_k are the mean and covariance matrix of the colors in a small local window ω_k, $|\omega_k|$ is the number of pixels in it, and U is a 3×3 identity matrix, ϵ is a regularizing parameter. More information can be found in [18]. This matrix is originally proposed to matting problem, and later widely used in haze removal, intrinsic images and colourization.

Fig. 5. The left image is the center view of a LF, and the right one is the segmentation result. The area in the black square is shown in Figure 4.

4.3. Segmentation

The matting Laplacian matrix (12) is based on an assumption that in a small local window depth is a linear transformation of color density. Apparently, the linear assumption in not always valid, especially when there is significant depth or color discontinuity. Therefore, we segment images into several pieces. For small pieces, we adopt the affinity matrix (8), which better models the relationship between the depth and color intensity. A very small window size (3×3) is used, so that the matrix can be very sparse. The costs for solving the linear system are not too high when the segments are small and the matrix is sparse. In large texture-less pieces, the matting Laplacian matrix is adopted for the sake of efficiency. Actually in large texture-less areas depth and color intensity are usually smooth, such as the light in Figure 5, which makes the linear assumption suffice. The mean shift [21] is used to segment images, which is robust and widely used in various computer vision tasks. An example segmentation result is shown in Figure 5.

5. EXPERIMENTAL RESULTS

The proposed method is tested with the HCI light field archive [11] and Stanford light field archive [22]. Our method manages to preserve rich details in the reconstructed depth maps, as shown in Figure 6. It also works very well for real light field data, as shown in Figure 7. The proposed method is compared with latest work [11], which employs the functional lifting method [15] to optimize local estimations. Their results are from the published code of [15], and depth values are quantized to 64 levels.

Quantitative results in Table 1, which is tested on the HCI light field archive demonstrate that our method effectively removes wrong estimations. Pixels whose relative estimation error is more than 3.2% are considered as wrong estimations. For the discrete functional lifting method, this threshold is equivalent to that the depth value differs from the ground truth by more than two levels. Results in the third line are produced

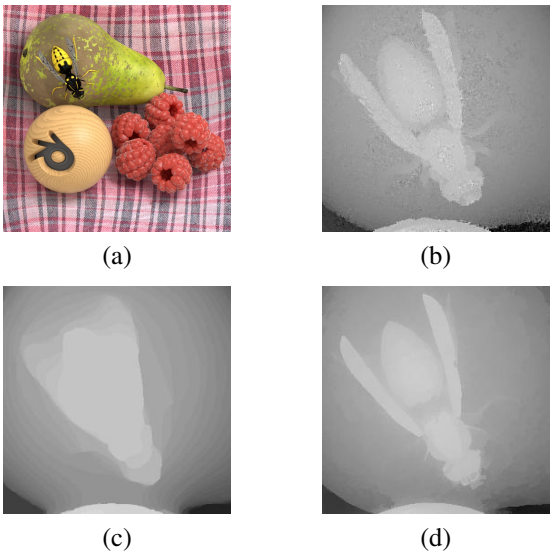


Fig. 6. A depth reconstruction result. The synthetic data is from the HCI light field archive. There are 9×9 views, and the image resolution is 768×768 each view. (a) is the center view image. (b)-(d) are close-ups of results from local depth estimation, the functional lifting [11] and the proposed method respectively. We recommend to see the electronic version of these images.

by applying our continuous method directly on local estimations from the structure tensor. If the refinement step in Section 3 is applied before the optimization step, slightly better results can be achieved, which are presented in the fourth line of Table 1.

The efficiency of the proposed method is comparative with the functional lifting method, when the latter quantizes the depth range to 64 levels. However, undesirable depth jumps between different levels are noticeable in their results as shown in the third column of Figure 6(c). If larger discrete level is set, their method will be much slower than the proposed one. Since our method is not multi-labelling based, its complexity does not change with the depth range.

6. CONCLUSION

In this paper, we propose a novel method to reconstruct continuous depth maps from 4D light fields. A refinement of local depth estimation is introduced by checking color consistency between different views. Based on the local depth estimations, we construct a sparse linear system, in which two different affinity matrices are employed. Compared with traditional multi-labelling methods, our results preserve much more details. In addition, to achieve similar level of smoothness with our results, multi-labelling methods usually take much longer time.

We made a novel attempt to reconstruct continuous depth maps with rich details, which obviously benefits many other computer vision tasks, such as 3D model reconstruction and scene understanding. More accurate and efficient models to obtain continuous depth maps are worth further investigating.

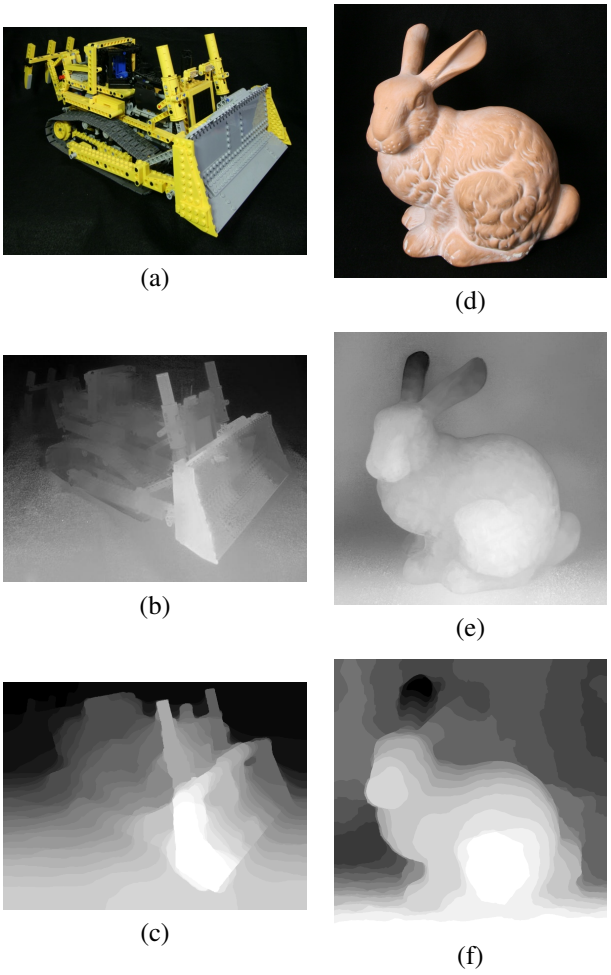


Fig. 7. Real data from Stanford light field archive. Each light field data has 17×17 views. Image resolutions of the "bulldozer" data and "bunny" data are 615×490 and 1024×1024 respectively. Images in the first row is the center views of LFs. The second and last rows are the reconstructed images by the proposed method and functional lifting [11] respectively.

7. REFERENCES

[1] Ren Ng, *Digital light field photography*, Ph.D. thesis, Stanford University, 2006.

[2] N. Joshi, S. Avidan, W. Matusik, and D.J. Kriegman, "Synthetic aperture tracking: tracking through occlusions," in *IEEE 11th International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.

[3] V. Vaish, M. Levoy, R. Szeliski, C.L. Zitnick, and S.B.

Table 1. Error rate comparison.

Method	Buddha1	Buddha2	Mona	StillLife	ConeHead
Local Estimations [11]	0.134	0.427	0.250	0.177	0.146
Functional Lifting [11, 15]	0.100	0.420	0.230	0.185	0.093
Continuous (without refinement)	0.091	0.352	0.146	0.099	0.102
Continuous (with refinement)	0.090	0.349	0.143	0.098	0.101

Kang, "Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, vol. 2, pp. 2331–2338.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, 2002.

[5] G. Zhang, J. Jia, T.T. Wong, and H. Bao, "Recovering consistent video depth maps via bundle optimization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[6] C Wu, M McCormick, A Aggoun, and SY Kung, "Depth map from unidirectional integral images using a hybrid disparity analysis algorithm," *IEEE Journal of Display Technology*, vol. 4, no. 1, pp. 101–108, 2008.

[7] T. Bishop and P. Favaro, "Full-resolution depth map estimation from an aliased plenoptic light field," *Computer Vision–ACCV*, pp. 186–200, 2011.

[8] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 1068–1080, 2008.

[9] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 31–42.

[10] R.C. Bolles, H.H. Baker, and D.H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987.

[11] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4d light fields," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 41–48.

[12] J. Bigun and G.H. Granlund, "Optimal orientation detection of linear symmetry," in *First International Conference on Computer Vision (ICCV)*, 1987, pp. 433–438.

[13] E. Strekalovskiy and D. Cremers, "Generalized ordering constraints for multilabel optimization," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2619–2626.

[14] C.K. Liang, T.H. Lin, B.Y. Wong, C. Liu, and H.H. Chen, "Programmable aperture photography: multiplexed light field acquisition," in *ACM Transactions on Graphics (TOG)*, 2008, vol. 27, p. 55.

[15] B. Goldluecke, E. Strekalovskiy, and D. Cremers, "The natural vectorial total variation which arises from geometric measure theory," *SIAM Journal on Imaging Sciences*, 2012.

[16] L. Wang and R. Yang, "Global stereo matching leveraged by sparse ground control points," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3033–3040.

[17] K. He, J. Sun, and X. Tang, "Fast matting using large kernel matting laplacian matrices," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, 2010, pp. 2165–2172.

[18] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 228–242, 2008.

[19] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," in *ACM Transactions on Graphics (TOG)*, 2004, vol. 23, pp. 689–694.

[20] S.B. Kang and R. Szeliski, "Extracting view-dependent depth maps from a collection of images," *International Journal of Computer Vision*, vol. 58, no. 2, pp. 139–163, 2004.

[21] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[22] "Stanford (new) light field archive," `http://lightfield.stanford.edu/lfs.html`.