Volume 4, Issue 8, August 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Automatic 2D-to-3D Image and Video Conversion by Learning Examples and Dual Edge-Confined Inpainting

Jean Maria Dominic
CSE Department, VJCET, MG University
India

Arsha J K
CSE Department, VJCET, MG University
India

Abstract— Most successful 2D-to-3D image and video conversion methods involve human operators, which is time-consuming and costly. Fully-automatic methods may work well in some cases and make strong assumptions about the 3D scene. But, a deterministic scene model that covers all possible background and foreground combinations is very difficult to construct. More over those have not achieved the same level quality of semi-automatic methods. Here a new 2D to 3D conversion method is proposed which is based on globally estimating the entire depth field of a query directly from a repository of image+depth pairs. The method uses a nearest neighbour-based regression. A key observation and an assumption are made for the proposal of this work. The key observation is that millions of 3D images are available on-line and for each 2D input query there exist many 3D content matches. The key assumption is that two photo metrically similar 3D images will also have similar depth fields. Also a new method to enhance the consistency between a colour image and its depth map is introduced. Here, first finds the edges in both 2D colour image and depth map. In the areas where the edges do not match, our approach uses a dual edge-confined inpainting technique to fill the gap.

Keywords— 3D images, depth estimation, depth maps, edge detection, 3D conversion.

I. INTRODUCTION

Stereoscopy, also called stereoscopic or 3D imaging is a technique for creating or enhancing the illusion of depth in an image by means of stereopsis for binocular vision. Most stereoscopic methods present two offset images separately to the left and right eye of the viewer. These two-dimensional images are then combined in the brain to give the perception of 3D depth. This technique is distinguished from 3D displays that display an image in three full dimensions, allowing the observer to increase information about the 3-dimensional objects being displayed by head and eye movements.

2D-to-3D conversion adds the binocular disparity depth cue to digital images perceived by the brain, thus, if done properly, greatly improving the immersive effect while viewing stereo video in comparison to 2D video. However, in order to be successful, the conversion should be done with sufficient accuracy and correctness: the quality of the original 2D images should not deteriorate, and the introduced disparity cue should not contradict to other cues used by the brain for depth perception. If done properly and thoroughly, the conversion produces stereo video of similar quality to "native" stereo video which is shot in stereo and accurately adjusted and aligned in post-production. The process of 2D to 3D conversion consists of two steps: depth estimation for a given 2D image and depth-based rendering of a new image in order to form a stereo pair.

3D capable hardware like 3D TVs, Blu-Ray players, handheld gaming consoles, cell phones, still and video cameras are widely available in the market. But this hardware availability is not yet matched 3D content production. Current available 2D to 3D conversion methods have not achieved a high quality level. The most successful approaches are interactive. That means it involve human operators. Therefore these methods are time consuming and costly.

Early versions of learning-based approach to 2D-to-3D image conversion either suffered from high computational complexity or were tested on only a single dataset. The aim of the project is to develop learning based approach to automatically convert 2D image to 3D image. Also a method to enhance the consistency between a colour image and its depth map is proposed. With 2D colour image's edges and depth map's edges, the proposed method finds the areas where the edges do not match, and uses a dual edge-confined inpainting technique to inpaint the edge mismatched areas..

II. STATE OF THE ART

Two approaches to 2D to 3D conversion can be loosely defined: quality semiautomatic conversion for cinema and high quality 3DTV, and low-quality automatic conversion for cheap 3DTV, VOD and similar applications. In semiautomatic conversion a skilled operator assigns depth to various parts of an image or video. Based on this sparse depth assignment, a computer algorithm estimates dense depth over the entire image or video sequence. In the case of automatic methods, no operator intervention is needed and a computer algorithm automatically estimates the depth for a single image or video. Automatic methods estimates shape from shading, structure from motion or depth from defocus. Electronics manufacturers use stronger assumptions to develop real-time 2D-to-3D converters. Such methods may work well in specific scenarios. But generally it is very difficult to construct heuristic assumptions that cover all possible background and foreground combinations.

In order to reduce operator involvement in the semiautomatic conversion process and therefore, lower the cost while speeding up the conversion, research effort has recently focused on the most labour-intensive steps of the manual involvement, namely spatial depth assignment. Guttmann et al. [4] have proposed a dense depth recovery via diffusion from sparse depth assigned by the operator. The focus of the method proposed by Agnot et al. [7] is the application of cross-bilateral filtering to an initial depth map. The authors propose to use a library of initial depth maps from which an operator can choose one that best corresponds to the image being converted. They also suggest estimation of the initial depth map based on image blur but show only one very simple example; this initialization is unlikely to work well in more complex cases. Phan et al. [12] propose a simplified and more efficient version of the Guttmann et al. [4] method using scale-space random walks that they solve with the help of graph cuts. Liao et al.[9] further simplify operator involvement by first computing optical flow, then applying structure-from-motion estimation and finally extracting moving object boundaries. The role of an operator is to correct errors in the automatically computed depth of moving objects and assign depth in undefined areas.

The problem of depth estimation from a single 2D image, which is the main step in 2D-to-3D conversion, can be formulated in various ways, for example as a shape-from shading problem [16]. However, this problem is severely under-constrained; quality depth estimates can be found only for special cases. Other methods, often called multi-view stereo, attempt to recover depth by estimating scene geometry from multiple images not taken simultaneously. For example, a moving camera permits structure-from-motion estimation [18] while a fixed camera with varying focal length permits depth from-defocus estimation [19]. Both are examples of the use of multiple images of the same scene captured at different times or under different exposure conditions. Although such methods are similar in spirit to the methods proposed here, the main difference is that while these methods use images known to depict the same scene as the query image, the proposed method uses all images available in a large repository and automatically select suitable ones for depth recovery.

Recently, machine-learning-inspired techniques employing image parsing have been used to estimate the depth map of a single monocular image [5], [3]. Such methods have the potential to automatically generate depth maps, but currently work only on few types of images using carefully-selected training data. The data-driven approaches to 2D-to-3D conversion is inspired by the recent trend to use large image databases for various computer vision tasks, such as object recognition [14] and image saliency detection [15].

III. SYSTEM STRUCTURE

The proposed method is based on globally estimating the entire depth field of a query directly from a repository of image+depth pairs using nearest neighbour-based regression. This approach is built upon a key observation and an assumption. The key observation is that among millions of 3D images available on-line, there likely exist many whose 3D content matches that of the 2D input query. The key assumption is that two 3D images whose left images are photometrically similar are likely to have similar depth fields. Also a new method to enhance the consistency between a colour image and its depth map is also introduced. With 2D colour image's edges and depth map's edges, the proposed method finds the areas where the edges do not match, and uses a dual edge-confined inpainting technique to inpaint the edge mismatched areas.

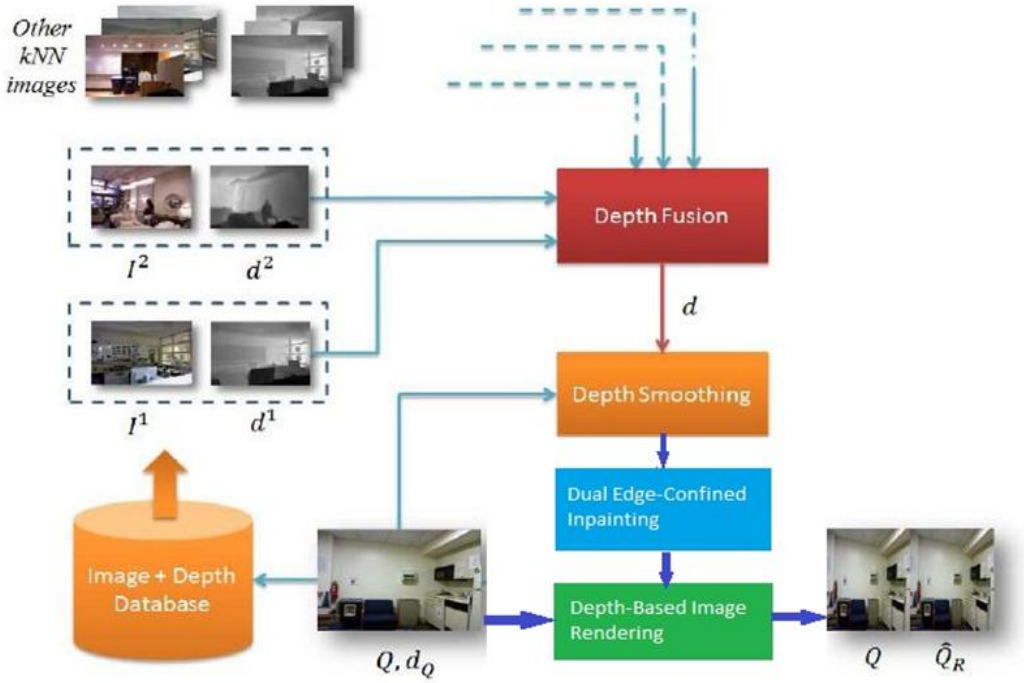


Fig. 1. Block diagram of the overall algorithm

The proposed method converts both 2D image and video into 3D. In case of video, first find key frames and frame rate. Then apply image conversion method for each frame. The system includes five main components:

A. kNN Search

There exist two types of images in a large 3D image repository: those that are relevant for determining depth in a 2D query image, and those that are irrelevant. As per the assumption images that are not photometrically similar to the 2D query need to be rejected. Because they are not useful for estimating depth and could potentially be more harmful to the 2D-to-3D conversion process. The selection of a smaller subset of images provides the added practical benefit of computational tractability when the size of the repository is very large.

One method for selecting a useful subset of depth-relevant images from a large repository is to select only the k images that are closest to the query where closeness is measured by some distance function capturing global image properties such as colour, texture, edges, etc. As this distance function, we use the Euclidean norm of the difference between histograms of oriented gradients (HOGs) computed from two images. Each HOG consists of 144 real values (4 × 4 blocks with 9 gradient direction bins) that can be efficiently computed. Perform a search for top matches to our monocular query Q among all images , k = 1...K in the 3D database I. The search returns an ordered list of image+depth pairs, from the most to the least photometrically similar images in the query. Then discard all but the top k matches (kNNs) from this list.

B. Depth Fusion

After kNN search generally none of the NN image+depth pair $(I^i, d^i), i \in$ K match the query Q accurately. However, the location of some objects and parts of the background is quite consistent with those in the respective query. If a similar object appears at a similar location in several kNN images, it is likely that such an object also appears in the query, and the depth field being sought should reflect this. This depth field can be computed by applying the median operator across the kNN depths at each spatial location x as follows:

$$d[x] = median\{d^i[x] \forall i \in K\} \quad (1)$$

The median-based fusion helps make depth more consistent globally. The fused depth is overly smooth and locally inconsistent with the query image due to edge misalignment between the depth fields of the kNNs and the query image. This results in the lack of edges in the fused depth where sharp object boundaries should occur and/or the lack of fused-depth smoothness where smooth depth is expected.

C. Depth Smoothing

After depth fusion the cross-bilateral filtering (CBF) is applied to overcome its limitations. CBF is a variant of bilateral filtering. It is an edge-preserving image smoothing method that applies anisotropic diffusion controlled by the local content of the image itself. In CBF the diffusion is controlled by an external input. CBF is applied to the fused depth d using the query image Q to control diffusion. This allows us to achieve two goals simultaneously: alignment of the depth edges with those of the luminance Y in the query image Q and local noise/granularity suppression in the fused depth d. This is implemented as follows:

$$\hat{d}[x] = \frac{1}{\gamma[x]} \sum_y d[y] h_{\sigma_s}(x-y) h_{\sigma_e}(Y[x]-Y[y]),$$
$$\gamma[x] = \sum_y h_{\sigma_s}(x-y) h_{\sigma_e}(Y[x]-Y[y]), \quad (2)$$

where $\hat{d}[x]$ is the filtered depth field and h_{σ_s} is a Gaussian weighting function.

D. Dual Edge-Confined Inpainting

The idea of this module is to find and inpaint the unmatched areas between a depth image and its corresponding colour image. Firstly, edge detection will be conducted on the colour image and the depth image. Then, locate the areas where edges in the two images are unmatched. Next, determine if the areas require inpainting, determine the direction of inpainting, and finally perform inpainting. A flowchart of the proposed inpainting method is shown in Fig. 2.

1) **Edge Detection**: The method utilizes edge information from both depth image and color image to find unmatched edge locations. Therefore, edge detection is required to find the edge information of interest. The Canny edge detector is selected. Because it has good detection capability, good edge positioning and good response values.

2) **Fast Edge Interpolation:** Fast Edge Interpolation method divides an image into smooth pixel areas and edge pixel areas. Then linear interpolation is applied to smooth pixels and non-linear interpolation is applied to edge pixels. The inpainting process includes three major steps: (1) Edge removal and combination, (2) selecting inpainting areas, and (3) dual edge inpainting. An unmatched edge is inpainted with these three steps.

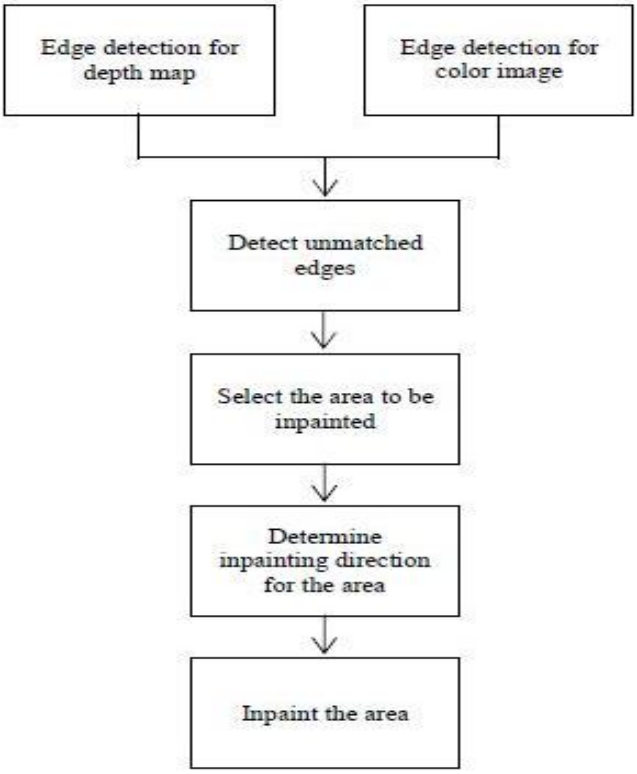
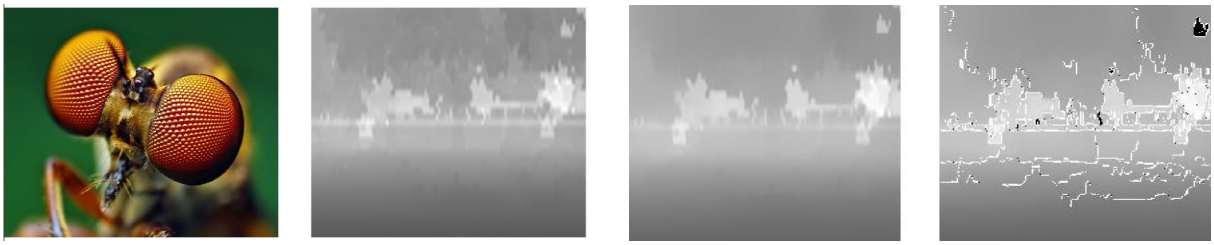


Fig. 2. A flowchart of the inpainting method



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

Fig. 3. (a) Input image (b) Fused depth map (c) Smoothened depth map (d) Inpainted depth map

E. Stereo Rendering

Final step is the rendering of stereopair. In order to generate an estimate of the right image $\hat{Q}_R$ from the monocular query Q to compute a disparity from the estimated depth d. Assume the fictitious image pair (Q, $\hat{Q}_R$) was captured by parallel cameras with baseline B and focal length f. The disparity $\delta[x,y] = Bf / \hat{d}[x]$. The right image can be produced from the 2D query Q using

$$\hat{Q}_R[x+\delta[x,y],y] = Q[x,y] \quad (3)$$

Fig. 4. (a) Input image (b) Global method 3D output (c) Proposed method 3D output

An example of 2D to 3D image conversion is shown in Fig.3. The proposed system output is validated against the existing global method [1]. In Fig.4 the comparison of the proposed and global

methods 3D outputs is shown. In case of video, first find key frames and frame rate. For each key frame create its depth map and then apply image conversion method for all other frames based on that depth map.

IV. EXPERIMENTAL EVALUATION

The proposed 2D to 3D conversion method has been evaluated and compared with an existing compared with an existing conversion method in the same scenario. The objective is to evaluate edge consistency between depth images and the corresponding colour images. It also evaluated the depth closeness of each depth map to its corresponding colour image. All the algorithms were implemented using MATLAB. Make 3d dataset is used for depth calculation.

A. Comparison of Edge Consistency

To evaluate edge consistency various depth images are generated by applying proposed algorithm and global algorithm in a dataset of eight images. Then edge detection is performed on all the depth images and ground depth images using canny edge detector. For each image, the number of common edges between the derived depth image and the original colour image are recorded for the calculation. The number of common edges are detected by performing AND operation.

$$\text{Number of common edges} = Count_{image} \cap Count_{Depthmap}$$

A bar graph(Fig.5) is plotted with the calculated number of common edges. There is an increase in number of common edges for the proposed system, compared to the ground and global depth maps.

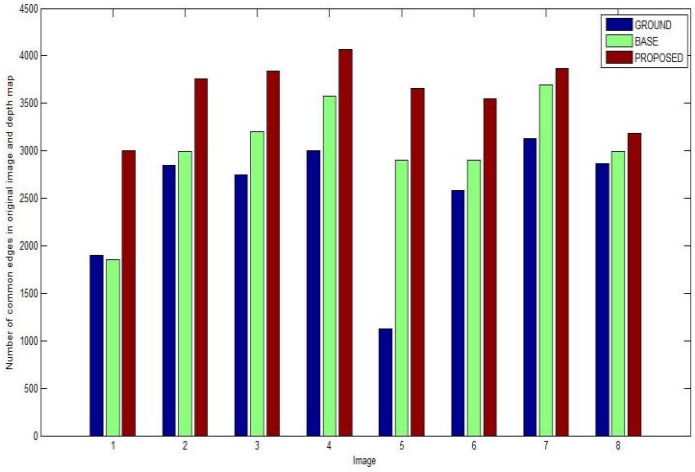


Fig. 5. Comparison of edge consistency

B. Depth Accuracy

The accuracy of depth can be calculated by some distance function. Here a Euclidean difference between histograms of oriented gradients (HOGs) is computed from original image and derived depth maps. The lower value will have more accurate depth image. A graph is plotted for all images against different depth maps, which is shown in Fig.6. The depth map generated by the proposed method has lower value. That means it is more similar to the original input image.

$$\text{Depth closeness} = \sqrt{HOG(image) \wedge 2 - HOG(depth\_image) \wedge 2}$$

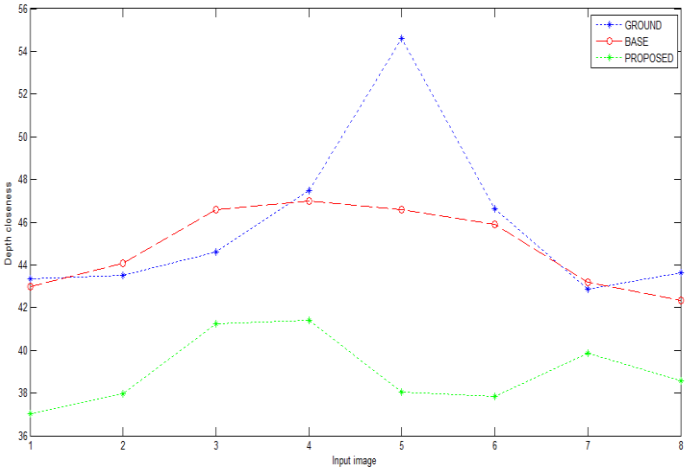


Fig. 6. Comparison of depth accuracy

V. CONCLUSIONS

A new class of method aimed at 2D -to- 3D image conversion is implemented. The proposed 2D to 3D conversion method estimates the entire depth field of a query image directly from a repository of image+depth pairs within a fraction of time. Also dual edge-confined inpainting technique enhanced the consistency between a color image and its depth map. The proposed method improved the number of common edges between the original image and depth image. Also the accuracy of the depth map is improved. The proposed method validated against existing global method and it performed well than global method.

ACKNOWLEDGMENT

The authors express their sincere thanks to HOD, group tutor and staff in Computer Science department, Viswa Jyothi College of Engineering and Technology for many fruitful discussions and constructive suggestions during the implementation of this paper.

REFERENCES

[1] Janusz Konrad, Fellow, Meng Wang, Prakash Ishwar,Chen Wu, and Debargha Mukherjee "Learning-Based, Automatic 2D-to-3D Image and Video Conversion",IEEE Trans. Image Process., vol. 22, no. 9, pp. 3485_3496, Sep. 2013.

[2] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in Advances in Neural Information Processing Systems. Cambridge, MA,USA: MIT Press, 2005.

[3] A. Saxena, M. Sun, and A. Ng, _Make3D: Learning 3D scene structure from a single still image,_ IEEE Trans. Pattern Anal. Mach. Intell.,vol. 31, no. 5, pp. 824_840, May 2009.

[4] M. Guttmann, L. Wolf, and D. Cohen-Or, "Semi-automatic stereo extraction from video footage", in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2009, pp. 136_142.

[5] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 1253_1260.

[6] R. Phan, R. Rzeszutek, and D. Androutsos, "Semi-automatic 2D to 3D image conversion using scale-space random walks and a graph cuts based depth prior,"in Proc. 18th IEEE Int. Conf. Image Process., Sep. 2011, pp. 865_868.

[7] M. Grundmann, V. Kwatra, and I. Essa, "Auto-directed video stabilization with robust L1 optimal camera paths," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,Jun. 2011, pp. 225_232.

[8] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non- parametric sampling," in Proc. Eur. Conf. Comput. Vis., 2012, pp. 775_788.

[9] M. Liao, J. Gao, R. Yang, and M. Gong, "Video stereolization: Combining motion analysis with user interaction," IEEE Trans. Visualizat. Comput. Graph., vol. 18, no. 7, pp. 1079_1088, Jul. 2012.

[10] J. Konrad, M. Wang, and P. Ishwar, "2D-to-3D image conversion by learning depth from examples," in Proc. IEEE Comput. Soc. CVPRW, Jun. 2012, pp. 16_22.

[11] Ming-Fu Hung, Shaou-Gang Miaou, and Chih-Yuan Chiang "Dual Edge-Confined Inpainting of 3D Depth Map Using Color Image's Edges and Depth Image's Edges" Signal and Information Processing Association Annual Summit and Conference (AP-SIPA), 2013 Asia-Pacific Oct. 29 2013-Nov. 1 2013.

[12] J. Konrad, G. Brown, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Automatic 2D-to-3D image conversion using 3D examples from the Internet,"Proc. SPIE, vol. 8288, p. 82880F, Jan. 2012.

[13] L. Angot, W.-J. Huang, and K.-C. Liu, "A 2D to 3D video and image conversion technique based on a bilateral" lter,_ Proc. SPIE, vol. 7526, p. 75260D, Feb. 2010.

[14] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 11, pp. 1958_1970, Nov. 2008.

[15] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley, "Image saliency: From intrinsic to extrinsic context," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,Jun. 2011, pp. 417_424.

[16] R. Zhang, P. S. Tsai, J. Cryer, and M. Shah, "Shape-from-shading: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 8, pp. 690_706, Aug. 1999.

[17] L. Angot, W.-J. Huang, and K.-C. Liu, "A 2D to 3D video and image conversion technique based on a bilateral" lter,_ Proc. SPIE, vol. 7526, p. 75260D, Feb. 2010.

[18] R. Szeliski and P. H. S. Torr, "Geometrically constrained structure from motion: Points on planes," in Proc. Eur. Workshop 3D Struct. Multiple Images Large-Scale Environ., 1998, pp. 171_186.

[19] M. Subbarao and G. Surya, "Depth from defocus: A spatial domain approach," Int.J. Comput. Vis., vol. 13, no. 3, pp. 271_294.

