Depth Estimation and Inpainting with an Unconstrained Camera

Arnav V. Bhavsar
arnav.bhavsar@gmail.com

A. N. Rajagopalan
raju@ee.iitm.ac.in

Image Processing and Computer Vision Lab
Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai, India

We formulate a general framework for depth estimation using multiple shifted and blurred images, by elegantly coupling the blur and motion via the common unknown of depth. Indeed, stereo [3], depth from defocus (DFD) [2] and related methods are restrictive special cases of a framework such as ours. Our framework relaxes the restrictions on motion and internal parameter variations and offers more freedom in operating the camera. We extend our approach for inpainting images as well as depth, using observations with missing areas. We exploit the motion cue which allows the correspondence / color information, missing in some images, to be present in others. The inpainting also considers the blurring process.

The observed images g_is can be modeled to be warped and blurred manifestations of an ideally non-blurred image f as

$$g_i(n_1,n_2) = \sum_{l_1,l_2} h_i(n_1,n_2,\sigma_i,\theta_{1i}(l_1),\theta_{2i}(l_2)) \cdot f(\theta_{1i}(l_1),\theta_{2i}(l_2)) + \eta_i(n_1,n_2) \quad (1)$$

where $\theta_{1i}(l_1)$ and $\theta_{2i}(l_2)$ denote the warped pixel coordinate for the reference pixel (l_1,l_2), and the kernel $h_i(n_1,n_2,\sigma_i,\theta_{1i}(l_1),\theta_{2i}(l_2))$ blurs a pixel at $(\theta_{1i}(l_1),\theta_{2i}(l_2))$ in the i^{th} image $f(\theta_{1i}(l_1),\theta_{2i}(l_2))$.

The geometric transformation between the reference and the i^{th} view can be expressed in terms of the depth Z as,

$$\theta_{1i} = \frac{v_i a_{i11} l_1 + v_i a_{i12} l_2 + v_1 v_i a_{i13} + v_1 v_i \frac{t_{xi}}{Z}}{a_{i31} l_1 + a_{i32} l_2 + v_1 a_{i33} + v_1 \frac{t_{zi}}{Z}} \quad (2)$$

$$\theta_{2i} = \frac{v_i a_{i21} l_1 + v_i a_{i22} l_2 + v_1 v_i a_{i23} + v_1 v_i \frac{t_{yi}}{Z}}{a_{i31} l_1 + a_{i32} l_2 + v_1 a_{i33} + v_1 \frac{t_{zi}}{Z}} \quad (3)$$

where the lens-image plane distance in the i^{th} view is denoted by v_i, the camera translations along the 3 axes by t_{xi}, t_{yi}, t_{zi}, and elements of the camera rotation matrix by a_{pq}, $1 \le p,q \le 3$.

The blur kernel can be modeled by a 2D Gaussian function whose variance σ, defined as the blur parameter, is related to the absolute depth from the lens [2]. Denoting the depth of a point from the reference camera as Z, its depth from the i^{th} view can be expressed as $Z_i = a_{i31}X + a_{i32}Y + a_{i33}Z + t_{zi}$. The blur parameter σ_i, for the 3D point in the i^{th} view is related to Z_i through the aperture r_i, focal length f and v_i as

$$\sigma_i = \rho r_i v_i \left(\frac{1}{f} - \frac{1}{v_i} - \frac{1}{Z_i}\right) \quad (4)$$

Also, the blur kernel is centered at $(\theta_{1i}(l_1),\theta_{2i}(l_2))$ for the i^{th} view.

Our depth estimation uses the belief propagation (BP) method [1], which involves computing messages and beliefs on an image-sized grid, which are, in turn, functions of data costs and the prior costs.

Expressing the reference image g_1 as result of local convolutions of the blur kernels and the i^{th} image, the data cost at a node is defined as

$$E_{di}(n_1,n_2) = |g_1(n_1,n_2) - h_{ri}(\sigma_i,n_1,n_2) * g_i(n_1,n_2)| \quad (5)$$

where, h_{ri} signifies the relative blur kernel corresponding to blur parameter $\sqrt{\sigma_1^2 - \sigma_i^2}$; a function of Z. The symbol $*$ denotes convolution.

The above data cost assumes that g_1 is more blurred than the g_i at all points, which need not be always true. To resolve this, we use the sign of $\sigma_1^2 - \sigma_i^2$. If for a depth label, $\sigma_1^2 - \sigma_i^2 < 0$, we modify the data cost as,

$$E_{di}(n_1,n_2) = |g_i(\theta_{1i},\theta_{2i}) - h_{ri}(\sigma_i,n_1,n_2) * g_1(n_1,n_2)| \quad (6)$$

We incorporate the notion of visibility in the above data term using a binary visibility function V which modulates the datacost as

$$E_{vi}(n_1,n_2) = V_i(n_1,n_2) \cdot E_{di}(n_1,n_2) \quad (7)$$

The total data cost E_d is the average of the individual data costs E_{vi} $(i > 1)$.

We also use the color image segmentation cue [4] to improve the depth estimate. We compute the color segmented image and a binary map that specifies the reliability of the depth estimate at each pixel. After the first iteration, we compute a plane-fitted depth map using the current depth estimate, segmented image and the reliability map. This plane-fitted depth map regularizes the data cost in subsequent iterations as

$$E_{ds}(n_1,n_2) = E_d(n_1,n_2) + w \cdot |Z(n_1,n_2) - Z_p(n_1,n_2)| \quad (8)$$

where Z_p denotes the plane-fitted depth map and the binary weight w is 0 if the pixel is reliable and 1 if it is not.

The prior cost constrains the neighbouring nodes to have similar labels. We define the smoothness prior as a truncated absolute function

$$E_p(n_1,n_2,m_1,m_2) = \min(|Z(n_1,n_2) - Z(m_1,m_2)|, T) \quad (9)$$

where, (n_1,n_2) and (m_1,m_2) are neighbouring nodes in a 4-connected neighbourhood. The truncation allows discontinuities in the solution.

For estimating the inpainted depth, we note that camera motion may allow the correspondences/color information missing in the reference image g_1 to be found in other images. We denote the set of missing pixels as M. If a pixel $g_1(l_1,l_2) \notin M$ and $g_i(\theta_{1i},\theta_{2i}) \in M$, $(i > 1)$ then the data cost between the $g_1(l_1,l_2)$ and $g_i(\theta_{1i},\theta_{2i})$ is not computed. If $g_1(l_1,l_2) \in M$, we look at $(\theta_{1i},\theta_{2i})$ and $(\theta_{1j},\theta_{2j})$ for a depth label. If both $g_i(\theta_{1i},\theta_{2i})$ and $g_j(\theta_{1j},\theta_{2j}) \notin M$, the matching cost between them is defined as $(1 < i < j)$

$$E_{vi}(n_1,n_2) = V_{ij}(n_1,n_2) \cdot |g_i(\theta_{1i},\theta_{2i}) - h_{rij}(\sigma_{ij},n_1,n_2) * g_j(n_1,n_2)| \quad (10)$$

$$h_{rij}(\sigma_i,n_1,n_2) * g_j(n_1,n_2) = \sum_{l_1,l_2} h_{rij}(\sigma_i,n_1-\theta_{1j},n_2-\theta_{1j}) \cdot g_j(\theta_{1j},\theta_{2j}) \quad (11)$$

$$V_{ij}(n_1,n_2) = V_i(n_1,n_2)V_j(n_1,n_2) \quad (12)$$

Here, h_{rij} denotes the blur kernel corresponding to $\sqrt{\sigma_i^2 - \sigma_j^2}$. The *compound* visibility V_{ij} signifies that the matching cost is not computed if a pixel is not observed in either the i^{th} or the j^{th} view. The total data cost is the average of matching costs over image pairs with visible pixels. The above process can yield some errors in pixel labeling. To mitigate these we invoke the segmentation cue as earlier, albeit with some modifications to account for the erroneous segments corresponding to missing regions.

Given the inpainted depth map, the data cost for image inpainting compares $g_i(\theta_{1i},\theta_{2i})$, $i > 1$ with intensity labels L if $g_i(\theta_{1i},\theta_{2i}) \notin M$

$$E_{di}(n_1,n_2) = V(n_1,n_2) \cdot |L - h_{ri}^p(\sigma_i,n_1,n_2) * g_i(n_1,n_2)| \quad (13)$$

The kernel superscript p denotes that, during the convolution, h_{ri}^p carries out a partial sum for only those pixels in its support which $\notin M$.

(a) (b) (c)

Figure 1: Real result: (a) One out of four observations. (b) Estimated depth map (c) Inpainted image.

[1] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, pages 1: 261–268, 2004.

[2] A. N. Rajagopalan and S. Chaudhuri. *Depth from defocus: A real aperture imaging approach.* Springer-Verlag New York, Inc., New York, 1999.

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

[4] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister. Stereo matching with color-weighted correlation, heiarchical belief propagation, and occlusion handling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(3):492–504, 2009.