Data-Driven Depth Map Refinement via Multi-scale Sparse Representation

Hyeokhyen Kwon[1], Yu-wing Tai[1] Stephen Lin[2]
[1]KAIST. [2]Microsoft Research.

Depth maps captured by consumer-level depth cameras such as Kinect are usually degraded by noise, missing values, and quantization. To facilitate the use of depth data, most methods have focused on the depth upsampling problem, in which a higher-resolution depth map is recovered from a lower-resolution input. These techniques can be classified as either *RGB-D based techniques* that utilize an additional RGB image to guide the depth map refinement process, or *reconstruction based techniques* that merge multiple unaligned low-quality depth maps to reconstruct a high-quality depth surface.

Our work fits into the first category and, the key idea is to transfer high-quality depth map primitives to the RAW depth map through multi-scale dictionary learning. Dictionary-based framework is adapted to address three practical issues in the depth refinement process: RGB textures uncorrelated with depth map discontinuities, and large dictionary size, and differences in geometric features at different scales.

RGB-D structure similarity measure: nAGDP To deal with the first issue of uncorrelated texture and depth discontinuities, we formulate a measure for predicting which RGB edges are most likely to coincide with depth discontinuities, refering *normalized Absolute Gradient Dot Product* (nAGDP):

$$\kappa(y_l(\mathbf{x}), y_c(\mathbf{x})) = \frac{|<g(y_l(\mathbf{x})), g(y_c(\mathbf{x}))>|}{\|g(y_l(\mathbf{x}))\|_2 \|g(y_c(\mathbf{x}))\|_2}, \tag{1}$$

where $g(y_l(\mathbf{x})) = \{\partial_x y_l(\mathbf{x}'), \partial_y y_l(\mathbf{x}')\}, \mathbf{x}' \in N(\mathbf{x})$, is a concatenation of gradients in the x (∂_x) and y (∂_y) directions within a local neighborhood $N(\mathbf{x})$ of $\mathbf{x}$, $<\cdot,\cdot>$ denotes the dot product operation, $|\cdot|$ denotes the absolute value operator, and $\|\cdot\|_2$ is the Euclidian norm of a vector. $\kappa(y_l(\mathbf{x}), y_c(\mathbf{x}))$ is set to zero if either $\|g(y_l(\mathbf{x}))\|_2$ or $\|g(y_c(\mathbf{x}))\|_2$ is smaller than a threshold.

Multi-scale solution to address degradation variation. The effects of degradation variation is significantly reduced at coarser scales, and refinement solutions at coarser scales are used as a proxy for the low-quality depth patches at finer scales.

Thus, before training our dictionary, we downsample the training data by a factor of 8 using bicubic interpolation. If missing values exist after downsampling because of large holes, they are filled by using the method in [4]. To avoid texture copying, we modify their smoothness term by setting its weight equal to $(1-\kappa)$ and use only the first order neighborhood for depth propagation within the hole regions. After hole filling, we further downsample y_l by $2\times$ to obtain the proxy solution at the coarsest resolution, $\tilde{y}_l^0$.

Scale-dependent dictionaries. Using the low-quality proxy depth map, $\tilde{y}_l^0$, and the downsampled high-quality depth map constructed by Kinect Fusion [2], y_h^0, and aligned RGB image, y_c^0, we learn dictionaries at each multi-scale to learn differences in geometric features that occur at different scales. At level i :

$$\arg\min_{D^i} \left\| \begin{bmatrix} y_h^i \\ \tilde{y}_l^i \\ y_c^i \end{bmatrix} - \begin{bmatrix} \mathbf{D}_h^i \\ \mathbf{D}_l^i \\ \mathbf{D}_c^i \end{bmatrix} \alpha^i \right\|^2 + \lambda |\alpha^i|. \tag{2}$$

where $\mathbf{D} = \{\mathbf{D}_h, \mathbf{D}_l, \mathbf{D}_c\}$ denotes their corresponding dictionaries containing basis functions, and $\alpha = \{\alpha_h, \alpha_l, \alpha_c\}$ represents sparse vectors of basis function coefficients, respectively. The learned joint dictionary is used to reconstruct and upsample the depth map progressively up to the original resolution of the RGB image.

Depth map refinement. At level i, we first estimate the sparse reconstruction coefficients of given y_l and y_c patches by considering overlapping patches to ensure consistent reconstruction between neighboring patches:

$$\begin{aligned} \arg\min_{\alpha^i(\mathbf{x})} \quad & \mathcal{P}(\mathbf{x}) \|y_h^i(\mathbf{x}) - \mathbf{D}_h^i \alpha^i(\mathbf{x})\|^2 + \|\tilde{y}_l^i(\mathbf{x}) - \mathbf{D}_l^i \alpha^i(\mathbf{x})\|^2 \\ & + \kappa(\mathbf{x}) \|y_c^i(\mathbf{x}) - \mathbf{D}_c^i \alpha^i(\mathbf{x})\|^2 + \lambda |\alpha^i(\mathbf{x})|, \end{aligned} \tag{3}$$

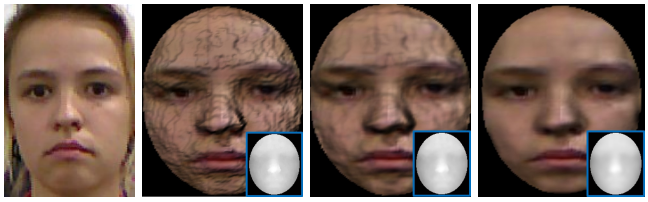
where $\mathcal{P}(\mathbf{x})$ is a binary mask that indicates parts of the reconstructed depth map that lie within patch overlap areas. By incorporating the nAGDP map, κ, into Equation (3), we avoid irrelevant edges in the RGB image which can mislead coefficient estimation.

Finally, using the estimated coefficients α_i for patches over the entire image, we reconstruct the depth map by solving the following optimization function:

$$\begin{aligned} \arg\min_{y_h^i} \quad & \sum_{\mathbf{x}} \|y_h^i(\mathbf{x}) - \sum_{\mathbf{x}'} w(\mathbf{x}') \mathbf{D}_h^i \alpha^i(\mathbf{x}')\|^2 \\ & + \mu \sum_{\mathbf{x}} (1 - \kappa(\mathbf{x})) |\nabla y_h^i(\mathbf{x})|, \end{aligned} \tag{4}$$

where $\sum_{\mathbf{x}} (1-\kappa(\mathbf{x}))|\nabla y_h^i(\mathbf{x})|$ is the total variation regularization weighted by $(1-\kappa(\mathbf{x}))$ which suppresses noise in reconstruction, $w(\mathbf{x}')$ is a blending function ($\sum_{\mathbf{x}'} w(\mathbf{x}') = 1$) for overlapping patches which is defined according to distance from the patch center.

By addressing the three important practical issues neglected in previous works, we achieved significant improvements in performance over state-of-the-art techniques through modifications of the dictionary learning and reconstruction framework to deal with these matters. Since our approach is data driven, the refinement can be targeted to a specific class of objects by employing a corresponding training set, such as human faces.



(a) RGB Image (b) Raw Image (c) Single scale (d) Multi-scale

Figure 1: Results from single and multi-scale dictionary recovery.

L_0-RGB Raw depth 1*st* level 2*nd* level 3*rd* level

Figure 2: Multi-scale refinement of the depth map. 1*st*: coarsest level; 2*nd*: middle level; 3*rd*: finest level.

[1] D.L. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

[2] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion : Real-time 3d reconstruction and interaction using a moving depth camera. In *24th annual ACM symposium on User interface software and technology, ser. UIST '11*, pages 559–568, 2011.

[3] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. *NIPS*, 19:801, 2007.

[4] J. Park, H. Kim, Y.-W. Tai, M.S. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. *ICCV*, pages 1623–1630, 2011.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.