Deep Convolutional Neural Fields for Depth Estimation from a Single Image

Fayao Liu, Chunhua Shen, Guosheng Lin
University of Adelaide, Australia; Australian Centre for Robotic Vision.

Estimating depths from a single monocular image is challenging as no reliable depth cues are available, e.g., stereo correspondences, motions etc. Previous methods either exploit geometric assumptions [3] or employ non-parametric methods [1]. The former is constrained to model particular scene structures, while the latter is prone to propagate errors through different decouple stages. Recent efforts have been focusing on exploiting additional sources of information, e.g., semantic labels [2], which are generally not available. We in this paper present a deep convolutional neural field model for estimating depths from a single image, without relying on any geometric assumptions nor extra information. Specifically, we propose a deep structured learning scheme which learns the unary and pairwise potentials of continuous conditional random field (CRF) [4] in a unified deep convolutional neural network (CNN) framework.

In our method, the integral of the partition function can be analytically calculated, thus we can exactly solve the log-likelihood optimization. Moreover, solving the MAP problem for predicting depths of a new image is highly efficient as closed-form solutions exist. We experimentally demonstrate that the proposed method outperforms state-of-the-art depth estimation methods on both indoor and outdoor scene datasets.

Let $\mathbf{x}$ be an image and $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ be a vector of continuous depth values of all n superpixels in $\mathbf{x}$. We model the conditional probability distribution of the data with the following density function:

$$\Pr(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathrm{Z}(\mathbf{x})} \exp(-E(\mathbf{y},\mathbf{x})), \tag{1}$$

where Z is the partition function: $\mathrm{Z}(\mathbf{x}) = \int_{\mathbf{y}} \exp\{-E(\mathbf{y},\mathbf{x})\}\mathrm{d}\mathbf{y}$; E is the energy function. Here, because $\mathbf{y}$ is continuous, the integral in $\mathrm{Z}(\mathbf{x})$ can be analytically calculated under certain circumstances (refer to paper for details). This is different from the discrete case, in which approximation methods need to be applied. To predict the depths of a new image, we solve the maximum a posteriori (MAP) inference problem:

$$\mathbf{y}^\star = \operatorname*{argmax}_{\mathbf{y}} \Pr(\mathbf{y}|\mathbf{x}). \tag{2}$$

We formulate the energy function as a typical combination of unary potentials U and pairwise potentials V over the nodes (superpixels) $\mathcal{N}$ and edges $\mathcal{S}$ of the image $\mathbf{x}$:

$$E(\mathbf{y},\mathbf{x}) = \sum_{p\in\mathcal{N}} U(y_p,\mathbf{x}) + \sum_{(p,q)\in\mathcal{S}} V(y_p,y_q,\mathbf{x}). \tag{3}$$

The unary term U aims to regress the depth value from a single superpixel. The pairwise term V encourages neighbouring superpixels with similar appearances to take similar depths. We aim to jointly learn U and V in a unified CNN framework.

Unary potential The unary potential is constructed from the output of a CNN by considering the least square loss:

$$U(y_p,\mathbf{x};\theta) = (y_p - z_p(\theta))^2, \ \forall p = 1,\ldots,n. \tag{4}$$

Here z_p is the regressed depth of the superpixel p parametrized by the CNN parameters θ.

Pairwise potential We construct the pairwise potential from K types of similarity observations, each of which enforces smoothness by exploiting consistency information of neighbouring superpixels:

$$V(y_p,y_q,\mathbf{x};\beta) = \frac{1}{2} R_{pq}(y_p - y_q)^2, \ \forall p,q = 1,\ldots,n. \tag{5}$$

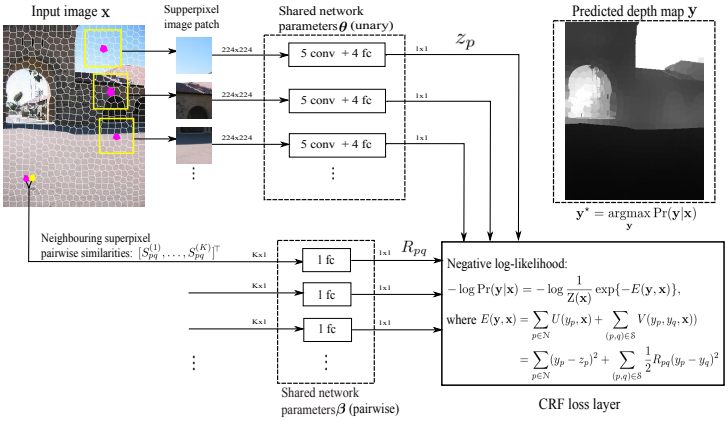

Figure 1: An illustration of our deep convolutional neural field model for depth estimation. The input image is first over-segmented into superpixels. In the unary part, for a superpixel p, we crop the image patch centred around its centroid, then resize and feed it to a CNN composed of 5 convolutional and 4 fully-connected layers. In the pairwise part, for a pair of neighbouring superpixels (p,q), we consider K types of similarities, and feed them into a fully-connected layer. The outputs of unary part and the pairwise part are then fed to the CRF structured loss layer, which minimizes the negative log-likelihood. Predicting the depths of a new image $\mathbf{x}$ is to maximize the conditional probability $\Pr(\mathbf{y}|\mathbf{x})$, which has closed-form solutions.

Here R_{pq} is the output of the network in the pairwise part (see Fig. 1) from a neighbouring superpixel pair (p,q). We use a fully-connected layer here:

$$R_{pq} = \beta^\top [S_{pq}^{(1)}, \ldots, S_{pq}^{(K)}]^\top = \sum_{k=1}^{K} \beta_k S_{pq}^{(k)}, \tag{6}$$

where $\mathbf{S}^{(k)}$ is the k-th similarity matrix whose elements are $S_{pq}^{(k)}$ ($\mathbf{S}^{(k)}$ is symmetric); $\beta = [\beta_1, \ldots, \beta_k]^\top$ are the network parameters.

Learning We minimizes the negative conditional log-likelihood of the training data:

$$\min_{\theta,\beta\geq 0} -\sum_{i=1}^{N} \log \Pr(\mathbf{y}^{(i)}|\mathbf{x}^{(i)};\theta,\beta) + \frac{\lambda_1}{2}\|\theta\|_2^2 + \frac{\lambda_2}{2}\|\beta\|_2^2, \tag{7}$$

where $\mathbf{x}^{(i)}$, $\mathbf{y}^{(i)}$ denote the i-th training image and the i-th depth map; N is the number of training images; λ_1, λ_2 are weight decay parameters.

Implementation of the method using MatConvNet [5] and details of network architectures are described in the paper. We conclude that the proposed method provide a general framework for joint learning of deep CNN and continuous CRF, which can be used for depth estimations of general scenes.

[1] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depthtransfer: Depth extraction from video using non-parametric sampling. *PAMI*, 2014.

[2] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *CVPR*, 2014.

[3] David C. Lee, Abhinav Gupta, Martial Hebert, and Takeo Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010.

[4] Tao Qin, Tie-Yan Liu, Xu-Dong Zhang, De-Sheng Wang, and Hang Li. Global ranking using continuous conditional random fields. In *NIPS*, 2008.

[5] Andrea Vedaldi. MatConvNet. http://www.vlfeat.org/matconvnet/, 2013.

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