Size to Depth: A New Perspective for Single Image Estination

Anonymous WACV submission

Paper ID ****

Abstract

In this paper, we propose a new method for single monocular image depth estimation through human interactions exploiting size information labeled by human instead of traditional geometric or depth information. We divide the image to be processed into equally sized patches, and then label real-world size for the dominant component in each patch manually, or automatically through deep convolutional neural network which is a faster alternative for manual labeling. With size information for each patch, we design an algorithm to generate a coarse depth map. At last we make refinements on the depth map by applying conditional random field. Experimental evaluation shows feasibility and superiority of our method.

1. Introduction

Single image depth estimation is a essential problem in computer vision which has found various applications in tasks like generating depth for a dataset without depth label and estimate depth for imaginary images such as cartoon. As an important component of understanding geometric relations within a scene, it also serves as a basis for plenty of advanced problems. Single image depth estimation is a quite challenging task for it's ill-posedness due to the inherent ambiguity of single images. Learning-based methods are recently proved to be an effective solution, but with two major shortcomings. Firstly, state-of-art methods based on deep neural network require plenty of training data, while RGB-D datasets generated by devices like Kinect and Lidar are limited in both number of images and type of scene. Popular RGB-D datasets such as NYU Depth [10, 8] and SUN RGB-D [11] typically contain images of indoor scenes less than or on the order of million. Secondly, for unseen images that are not of the same type with training data (*e.g.*, using a neural network trained by indoor scenes to predict depth for images of outdoor scenes), learning-based methods may perform poorly, as neural network will learn particular bias from training data as well, which shows its lack of generalization. Another kind of popular method try to exploit human prior through interaction. They require human to directly label geometric information or depth for several pixels or label relative relationships of depth for some pairs or groups of pixels. Our method belongs to the latter family, which requires only real-world size label from human, instead of depth or geometric information in traditional algorithms. Because labeling size is an much easier task for human compared to labeling depth or geometric information, our method takes advantage. What's more, we try to use machine to label size, which makes labeling extremely efficient.

2. Related works

Increasing number of methods are trying to estimate depth for single monocular image, which can be roughly classified into two families: learning-based method and human interactive method.

Some traditional learning-based methods formulate depth estimation as a markov random field (MRF) learning problem. Saxena *et al*. [9] use linear regression and MRF to predict depth from a set of image feature. Liu *et al*. [6] propose a discrete-continuous conditional random field (CRF) model to take relations between adjacent superpixels into consideration. Realizing the strong correlation between depth estimation and semantic segmentation, Liu *et al*. [4] make use of predicted semantic labels to guide 3D reconstruction by enforcing depth related to class and geometry prior.

Most of recent learning-based approaches rely on the application of deep learning, among which deep convolutional neural network (CNN) is used most commonly. Eigen *et al*. [2] design a global coarse-scale deep CNN to regress a rough depth map directly from an input image. They then train a local fine-scale network to make local refinements. Liu *et al*. [5] propose a deep convolutional neural field model for depth estimation by exploring CNN and CRF. They jointly learn the unary and pairwise potentials of CRF in a unified deep CNN framework. Like in traditional learning-based method, Wang *et al*. [12] use a deep CNN to jointly predict a global layout composed of pixel-wise depth values as well as semantic labels, and improved

performance by allowing interactions between depth and semantic information.

Compared to our method, learning-based methods require plenty of ground truth data and are not able to generalize to unseen images which are not of trained image types. Some of the methods [4, 12] need result of segmentation algorithm.

Human interactive methods exploit human ability to interpret 2D images, using human annotation. Some previous works make use of geometric elements (lines or plains) in images to predict depth. Criminisi *et al*. [1] describe a way to compute 3D affine measurements given human inputs providing geometric information determined from the image. Later works [3] by Lee *et al*. try to generate plausible interpretations of a scene from a collection of line segments automatically extracted from a single indoor image. These methods are limited for requiring a large amount of straight lines or plains in the image to provide enough evidences for 3D structure inference. Lopez *et al*. [7] formulate the problem as an optimization process by assuming that image regions with low gradients will have similar depth values. In their method, depth values are propagated between pixels with small image gradients under a number of human-defined constraints. Our method, In contrast, requires only size information, which is more trivial for human to label than geometric or depth information, and even for machine to label. We will show this in the following sections.

3. Our method

// Notations, conventions

3.1. Overview

Our goal is to estimate pixelwise depth for single image of general cases. We develop an algorithm that takes advantage of human annotation to estimate depth. Note that, directly labeling depth of pixel in numeric value is impractical for humans. Senses of human are envolved to be sensitive to comparing relative depth but not estimating raw value of depth. One can readily distinguish the nearer object between two, but struggles to name that distance to an object is 10m, 12m or 15m. The basic equation in computer vision

$$\text{depth} \propto \text{focal length} * \frac{\text{real-world size}}{\text{size in photo}} \tag{1}$$

implies that depth of object is proportional to its real-world size given focal length and its size in photo is fixed, which is satisfied in our problem setting. And labeling real-world size of objects is a much better formulation of annotation since size of objects is common knowledge for human.

In order to simplify the labeling process while get global layout of depth by annotation, we propose a patch-to-size way of annotation. In this formulation, image is divided into grids of equally sized patches. And human are asked to label real-world size of dominant component in each patch. To involve both local and global horizon, this is done several times for coarse-to-fine patch division of image.

After patch-to-size step, we obtain size for each patch (which is equivalent to depth). But the result is rather coarse with rigid boundary. Another depth refinement step is introduced to smooth depth gaps and interpolate between pixels under the constraint of depth annotation.

The depth refinement step is formulated as an energy function optimization problem in conditional random field (CRF). Mathematically, let $\mathbf{x}$ be the RGB-D image, $\mathbf{y}$ be the corresponding depth, θ be the model parameter (if any, *e.g*. parameters of patch-to-size CNN), we model the conditional probability distribution of RGB-D data with

$$\Pr(\mathbf{y}|\mathbf{x},\theta) = \frac{1}{Z(\mathbf{x}|\theta)}\exp(-E(x,y|\theta)) \tag{2}$$

where $E(\mathbf{x},\mathbf{y}|\theta)$ is the energy function and $Z(\mathbf{x}|\theta)$ is a normalization term given by

$$Z(\mathbf{x}) = \int_y \exp(-E(\mathbf{x},\mathbf{y}))dy. \tag{3}$$

Model parameter θ^* that best fits the training data is found by maximum likelihood estimation with regularization

$$\theta^* = \text{argmax}_\theta \log \Pr(\mathbf{y}|\mathbf{x},\theta) + \text{reg}(\theta). \tag{4}$$

And maximum a posteriori (MAP) solution $\mathbf{y}^*$ gives the depth $\mathbf{y}$ of maximum probability for observed image $\mathbf{x}$ which is the best estimation of depth

$$\mathbf{y}^* = \text{argmax}_\mathbf{y} \log \Pr(\mathbf{y}|\mathbf{x},\theta). \tag{5}$$

In the following, we will give a detailed discussion about components of our algorithm.

3.2. Annotation formulation

As discussed, knowledge of object shape and spatial relation is fundamental for depth estimation. Realizing the difficulty to model them by learning, we introduce human knowledge to ease the urge. And we must decide whose size to label. Having the annotator draw the contour of labelled object would be too complex and time-consuming. We come up with the patch dividing idea to avoid this problem. In our patch-to-size formulation, specification of labelled object is substituted by a flexible concept of dominant component. Patch-to-size makes the general assumption that depth of dominant component in the patch is representative of the entire patch(this does not mean we will assign same depth for every pixel in the patch, refer to Section 3.4). The assumption is definitely inapplicable in some cases such as images depicting a person in the background of sky. But it can be fixed by introducing CRF loss terms as shown in section 3.4.

3.3. Faster alternative for manual labeling

Although patch-to-size greatly reduces annotation effort, one still have to label up to 10 by 10 patches which takes 3-5 minutes. We make the attempt to speedup this process by learning to label size by CNNs. Different CNNs are trained for each granularity of patch division. [Description of CNN architecture, need more experiment]

3.4. Conditional random field

As is typically formulated, energy function of CRF consists of unary and binary potential terms.

$$E(\mathbf{x},\mathbf{y}) = \sum_{p\in P} E_{unary}(y_p,\mathbf{x}) \tag{6}$$

$$+\lambda \sum_{x,y\in A} \frac{1}{2} E_{binary}(y_x,y_y,\mathbf{x}) \tag{7}$$

where P is the set of pixels in the image, A is set of adjacent pair of pixels. E_{unary} and E_{binary} are relatively unary and binary loss terms.

The unary term is given by

$$E_{unary}(y_p,\mathbf{x}) = (y_p - \sum_i w_i d_i)^2 \tag{8}$$

$$= (y_p - w^T\mathbf{d})^2 \tag{9}$$

where d_i is the depth annotation. This term basically requires the depth of image to match with depth of its belonging patches.

The binary term is given by

$$E_{binary}(y_x,y_y,\mathbf{x}) = \text{sim}(x,y)(y_x - y_y)^2 \tag{10}$$

where $\text{sim}(x,y)$ is a similarity function of pixel. This term penalizes gap of depth between neighboring pixels which encourages local continuity of the depth field. Similarity between pixels can be represented by image gradient

$$\text{sim}(x,y) = |I_x - I_y| \tag{11}$$

where I denotes image intensity.

3.5. MAP in CRF

In the case of manual annotation, depth d_i is predefined by user, enabling us to pre-process the $w^T\mathbf{d}$ term to be a single variable d_p for every pixel. And if we measure similarity by image gradient, energy function can be written as

$$E(\mathbf{x},\mathbf{y}) = \sum_{p\in P}(y_p - d_p)^2 + \lambda \sum_{x,y\in A}(I_x - I_y)(y_x - y_y)^2. \tag{12}$$

Optimization of E can be performed by taking gradient w.r.t $\mathbf{y}$ and doing gradient descent. Rewrite energy function E as a function of $\mathbf{y}$:

$$E = \mathbf{y}^T(\,(1 + \sum_{y,(x,y)\in A} |I_x - I_y|\,)I + W\,)\mathbf{y} - 2\mathbf{y}^T d + d^T d \tag{13}$$

$$= \mathbf{y}^T W'\mathbf{y} - 2\mathbf{y}^T d + d^T d \tag{14}$$

where I is the identity matrix, d is the vector of d_p, $W_{ij} = |I_x - I_y|$ and

$$W' = 1 + \sum_{y,(x,y)\in A} |I_x - I_y|. \tag{15}$$

Normalization term Z can be analytically integrated:

$$Z(\mathbf{x}) = \int_y \exp(-E(\mathbf{x},\mathbf{y}))dy \tag{16}$$

$$= \frac{\pi^{\frac{n}{2}}}{|W'|^{\frac{1}{2}}} \exp(d^T W'^{-1} d - d^T d). \tag{17}$$

and the log-likelihood of posterior is given by

$$\log \Pr(\mathbf{y}|\mathbf{x}) = \frac{n}{2}\log\pi - \frac{1}{2}|W'| + d^T W'^{-1} d - d^T d \tag{18}$$

With l2-loss, the objective of CRF can be written as

$$\min \sum \log \Pr(\mathbf{y}|\mathbf{x}) \tag{19}$$

The optimization problem is efficiently solved by gradient descent.[More detailed derivation of gradient]

3.6. Maximum likelihood estimation in CRF

Our CRF also fit in parameterized learning scheme. Parameters might come from the patch-to-size CNN to speed up patch-to-size. In that case,

$$E_{unary} = (y_p - w^T\mathbf{d}(\theta))^2 \tag{20}$$

and the final gradient w.r.t θ is given by

$$\frac{\partial E_{unary}}{\partial \theta} = \frac{\partial E_{unary}}{\partial \mathbf{d}} \frac{\partial \mathbf{d}}{\partial \theta} \tag{21}$$

$$= \sum \mathbf{y}\text{grad}(\theta) \tag{22}$$

where $\text{grad}(\theta)$ is calculated by back propagation in CNN.

Another possible parameterization is by applying CNN as distance metric. Note that inferring depth on pixelwise scale in CRF is computationally extensive. Thus we can infer in larger scale patches in CRF while generate finer-grain depth map by upsample. Designing a reasonable mathematical representation of distance metric between patches is nontrivial. Thus we introduce CNN to perform a projection from high-dimension image space to low-dimension

distance space where euclidean distance of feature vector is representative of distance between patches. In this setting,

$$E_{binary}(y_x, y_y, \mathbf{x}) = \text{sim}(x, y, \theta)(y_x - y_y)^2 \quad (23)$$

$$= (f(x, \theta) - f(y, \theta))^2 (y_x - y_y)^2 \quad (24)$$

where $f(\mathbf{x}, \theta)$ is the transform function defined by the CNN. The final gradient w.r.t θ is given by

$$\frac{\partial E_{binary}}{\partial \theta} = \sum 2(f(x, \theta) - f(y, \theta))(\text{grad}(\mathbf{x}, \theta) - \text{grad}(\mathbf{y}, \theta)) \quad (25)$$

$$(y_x - y_y)^2 \quad (26)$$

The projection CNN can be optimized by performing back propagation.

References

[1] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *International Journal of Computer Vision*, 40(2):123–148, Nov 2000.

[2] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *CoRR*, abs/1406.2283, 2014.

[3] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2136–2143, 2009.

[4] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[6] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[7] A. Lopez, E. Garces, and D. Gutierrez. Depth from a Single Image Through User Interaction. In A. Munoz and P.-P. Vazquez, editors, *Spanish Computer Graphics Conference (CEIG)*. The Eurographics Association, 2014.

[8] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[9] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1161–1168. MIT Press, 2006.

[10] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*, 2011.

[11] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. pages 567–576, 06 2015.

[12] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.