LaTeX Author Guidelines for WACV Proceedings

Anonymous WACV submission

Paper ID ****

Abstract

In this paper, we propose a new method for single monocular image depth estimation through user interactions requiring size-related information from user instead of traditional geometric or depth-related information.

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 unreal 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 a effective solution, but with two major shortcomings. Firstly, to train a deep neural network, plenty of training datas are required, while RGB-D datasets generated by devices like Kinect and Lidar are limited in both type of scene and number of images. Popular RGB-D datasets such as NYU Depth [**?**, **?**] and SUN RGB-D [**?**] 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 methods try to exploit human prior through user interaction. They require users to directly label geometric information or depth for several significant pixels or label relative relationships of depth for some pairs or groups of pixels.

Our method belongs to the latter family. Compared to other methods in this family, we adopt a new concept to deal with depth estimation.

2. Related works

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

Some traditional learning-based methods formulate depth estimation as a Markov Random Field (MRF) learning problem. Saxena *et al*. [**?**] using linear regression and a MRF to predict depth from a set of image feature. Liu *et al*. [**?**] 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*. [**?**] make use of predicted semantic labels to guide 3D reconstruction by enforcing depth related to class and geometry prior.

Most of the 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*. [**?**] design a global coarse-scale deep CNN to regress a coarse depth map directly from an input image. They then train a local fine-scale network to make local refinements. Liu *et al*. [**?**] 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*. [**?**] 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 they are not able to generalize to unseen images which are not of trained image types. Some of the methods [**?**, **?**] also need result of segmentation algorithm.

User interactive methods exploit human ability to interpret 2D images, using input information from user. Some previous works make use of geometric elements (lines or plains) in images to predict depth. Criminisi *et al*. [**?**] describe a way to compute 3D affine measurements given user

inputs providing geometric information determined from the image. Later works [**?**] 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 inferring. Lopez *et al*. [**?**] formulate the problem as an optimizing 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 user-defined constraints. Our method also belongs to this family. In contrast, our method requires only size information, which is more trivial for user 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 evoluted 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

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

implys 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 divison 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, we model the conditional probability distribution of RGB-D data with

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

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

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

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

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 of annotation

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 convolutional neural networks. Different convnets are trained for each granularity of patch division. [***Need more description***]

3.4. Conditional random field

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

$$E(\mathbf{x},\mathbf{y}) = \sum_{p\in P} E_{unary}(y_p,\mathbf{x}) + \lambda \sum_{p_x,p_y\in A} E_{binary}(p_x,p_y,\mathbf{x}) \tag{4}$$

where P is the set of pixels in the image, and A is set of adjacent pair of pixels. E_{unary} and E_{binary} are relatively unary and binary loss terms. E_{unary} restricts the depth of pixel to align to the annotated depth of its belonging patches. E_{binary} tend to assign similar depth for neighboring pixels which encourage local continuity of depth.[***Need more description***]

3.5. Learning

3.6. Implementation details