Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Joint Depth and Semantic Inference from a Single Image via Elastic Conditional Random Field ☆

Rongrong Ji [a], Liujuan Cao [a,*], Yan Wang [b]

[a] Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, Xiamen, Fujian 361005, China

[b] Department of Electrical Engineering, Columbia University, New York, NY 10027, USA

ARTICLE INFO

Article history:
Received 28 July 2015
Received in revised form
20 February 2016
Accepted 9 March 2016

Keywords:
Depth estimation
Semantic labeling
Conditional random field
Structured Support Vector Machine
Content analysis
Scene understanding

ABSTRACT

The estimations of depth and regional semantics from a single image have traditionally been considered as two separated problems. In this paper, we argue that these two tasks provide complementary information, which therefore can be performed jointly to reinforce individual tasks in terms of both accuracy and speed. In particular, we propose an Elastic Conditional Random Field (E-CRF) deployed upon superpixel segmentations, which models the interdependency between depth and semantics to refine each other in an iterative manner. Differing from the traditional CRFs, E-CRF makes edges elastically hidden/emergent during inference to conduct fast Loopy Belief Propagation, while explicitly modeling the depth-label interdependency to achieve high inference accuracy. Moreover, the Structured Support Vector Machine (SSVM) is further introduced to drastically speed up the inference. We have conducted extensive evaluations on both Make3D and NYU benchmark datasets, which demonstrated that our E-CRF method significantly outperforms state-of-the-art techniques in terms of precision, while significantly accelerating the inference speed (2–3 orders of magnitude).

1. Introduction

Given a single image, it is easy for humans to understand the depth layouts and semantic labels of different regions. However, the tasks of depth layout estimation and semantic labeling from a single image retain as open problems, despite the great efforts made in the literature, respectively [1–4]. While both tasks are traditionally treated as separate research problems, the human visual system typically understands an image by analyzing both its 3D layout and semantic information jointly. For instance as shown in Fig. 1(a), by given only the semantic information with simple shapes, the human visual system is still possible to infer the depth layout, i.e., one tree is in front of the other trees. As another instance in Fig. 1(b), by given only the rough depth patterns, the human visual system is also above to identify semantic labels from different regions, i.e., a road/ground/grass below the sky. In this paper, we explore the intrinsic interdependency between semantic labeling and depth estimation in a single image. We targets at showing how both tasks can be mutually reinforced each other, in terms of both inference accuracy as well as the inference speed.

1.1. Motivation

Depth estimation and semantic labeling from a single image are both fundamental problems in computer vision, which have broad application prospects in various areas including stereo vision, robotics, scene understanding, as well as image retrieval. Depth estimation from a single image typically involves estimating local depth at individual pixels or superpixels [1,2,4], followed by contextual inference like MRF to model the spatial relations, i.e., connected structures, co-planarity and co-linearity [4]. However, it retains as an open problem in terms of both accuracy and speed, the latter of which typically costs tens of seconds to process each image on regular PCs [3]. Region labeling aims to decompose an image into semantically meaningful regions, which also modeling contextual consistency to smooth labels of individual regions, i.e., the appearance and the structure of the scene [3,5,6]. Similar to depth estimation, region labeling is also challenging in terms of both accuracy and speed, also due to the complex models and inference processes.

From another perspective, it is also helpful to incorporate depth

☆This work is supported by the Special Fund for Earthquake Research in the Public Interest No. 201508025, the Nature Science Foundation of China (No. 61402388, No. 61422210 and No. 61373076), the Fundamental Research Funds for the Central Universities (No. 20720150080 and No. 2013121026), and the CCF-Tencent Open Research Fund.

* Corresponding author.
E-mail address: caoliujuan@xmu.edu.cn (L. Cao).

http://dx.doi.org/10.1016/j.patcog.2016.03.016
0031-3203/

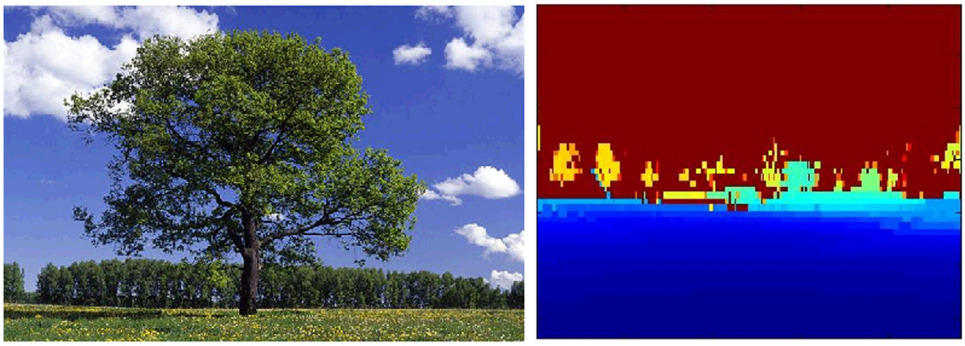

Fig. 1. Human vision system may use semantic labels to help 3D layout estimation, as shown in (a), and use depth to infer semantic info, as shown in the case of *road*/*ground*/*grass* under *sky* in (b).

into the estimation of semantic labels. For example, by distinguishing two regions as far away or close from each other, strong context can be added into the inference of region labels, a.k.a., a sky is always far away from car (for outdoor scene), while screen is more likely to be close to keyboard (for indoor scene).

It is motivated that both depth estimation and region labeling share a similar paradigm, which actually provide complementary information to assist respective tasks in the human visual system. In this paper, we targets at doing these two inferences together with the following motivation:

- To perform region labeling in a more intrinsic domain: Without knowing the depth, only image-level spatial consistency can be incorporated, which is less accurate comparing to using the 3D geometric layouts.
- To integrate the object/scene labels to improve the depth estimation precision, as indicated in [2,4]. For instance, the *sky* always has large depth compared to other objects in a scene. For another instance, knowing a pedestrian's height in advance can improve the depth estimation of the pedestrian's region.
- To accelerate inferences on both tasks: Cues from both sides can be exploited to design more efficient inference algorithms while maintaining or improving the inference accuracy.

In this paper, by treating depth or labeling as the hidden/latent cues to reinforce each other, we demonstrate significant gains in terms of inference accuracy, as well as a fundamental speedup via a very efficient (almost real-time) discriminative solution. Such accurate yet on-the-fly inference of both depth and semantics have high impact on various practical applications. For instance, various scene structure recovery algorithms also benefit from the estimated depth information, which helps to mitigate the difficulty in separating foreground and background objects. For another instance, some live interactive augmented reality applications also need to parse the depth map and semantic labels in real time.

Typically, depth information is recoverable only when stereo or multi-view observations are available. The proposed work is not intended to prove otherwise. Instead, it proposed to explore an emerging approach reported in recent works [33,2,7] that utilize two sources of information to alleviate the difficulty: first, the data to depth correlation can be learned from training datasets and applied to estimate depth in new data assuming the learnt patterns hold in the test domain; and second, iterative inferencing processes exploring the dependency between the depth and semantics can be used to further improve the accuracy. This is essentially a data driven learning based approach, distinct from the typical physical geometrical model seen in the computer vision literature. Like other learning based solutions, the proposed method will work well only if the test and training sets share similar characteristics. Cross-domain generalization remains an important open issue that is outside the scope of the current manuscript.

The learning based approach assumes there are consistent patterns that can be learned in predicting depth from visual features and semantic labels. For example, different semantic objects (*road*, *building*, etc.) render consistent yet distinct data features at different depths. Such distinct relations can be used to estimate depth from visual features and semantic labels. Conversely, the distinct correlations allow inferring semantic labels from visual features and depth estimation.

A problem may arise in the initial step, when no semantic labels are available and the initial depth estimation is made with the visual features only. We address this issue by learning the feature-to-depth relation from the training data set, and equally importantly, by incorporating the visual features of the surrounding neighbors in the initial depth estimation. After the initial step, the semantic labels (estimated from the visual features, depth and semantic labels of the surrounding neighborhood) can be incorporated to refine the depth estimation. Our empirical results confirm the feasibility of such approaches at least when the test and training domains are similar.

The authors are aware that there were already some inspiring work [33,2,7] touching this interesting topic. But as will be analyzed in details in the following sections, the proposed approach has some substantial differences from the existing works, and may help build new perspectives toward the problems, i.e. a more elegant model for the problem, a large-margin formulation providing efficient optimization, and more cross-dataset generality because of the existence of a "latent" layer.

1.2. Challenges

Semantic labels have been leveraged as additional cues to improve the estimation accuracy of depth in [2]. However, an integrated bi-directional inference for both depth and label has not been studied, mainly due to the difficulty in designing robust yet efficient inference model. To model the interdependency between depth and labeling, simply extending the traditional generative models like Conditional Random Field (CRF) or Markov Random Field (MRF) would dramatically increase the model complexity, despite that the separate estimation process is already quite slow. For example, the depth estimation approach proposed in [2] needs 30 seconds for each image of a small resolution (240×320 pixels).

Please cite this article as: R. Ji, et al., Joint Depth and Semantic Inference from a Single Image via Elastic Conditional Random Field, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.03.016