This paper has been accepted for CVPR 2012. The paper will be updated after the camera ready copy is available.

Joint 2D-3D Temporally Consistent Semantic Segmentation of Street Scenes

Georgios Floros Bastian Leibe
UMIC Research Centre
RWTH Aachen University
{floros,leibe}@umic.rwth-aachen.de

Abstract

In this paper we propose a novel Conditional Random Field (CRF) formulation for the semantic scene labeling problem which is able to enforce temporal consistency between consecutive video frames and take advantage of the 3D scene geometry to improve segmentation quality. The main contribution of this work lies in the novel use of a 3D scene reconstruction as a means to temporally couple the individual image segmentations, allowing information flow from 3D geometry to the 2D image space. As our results show, the proposed framework outperforms state-of-the-art methods and opens a new perspective towards a tighter interplay of 2D and 3D information in the scene understanding problem.

1. Introduction

Visual scene understanding from moving platforms has become an area of very active research, and many approaches have been proposed towards this goal in recent years [16, 27, 4, 25, 5, 7]. Multi-class image segmentation has become a core component for many of those approaches, as it can provide semantic context information to support the higher-level scene interpretation tasks [27, 4, 15]. This development has been assisted by significant improvements in state-of-the-art segmentation frameworks [24, 14, 15]. In this paper, we build upon this recent progress in order to address the problem of segmenting urban street scenes into semantically meaningful classes, such as *road*, *building*, *street marking*, *car*, *etc*. (see Fig. 1).

Despite their motivation by mobile applications, most previous semantic scene segmentation approaches operate on individual 2D images (*e.g.*, [4, 14, 9]) or single stereo pairs [15], ignoring temporal continuity information. We believe that such single-frame semantic segmentation is fundamentally limited, since at any point in time, large parts of the scene will not be visible at sufficient resolution to make confident decisions. As mobile platforms often have the capability to move through the scene and observe it from several viewpoints, we argue that scene understanding systems should make use of this temporal information to enforce temporal consistency between the semantic labelings acquired at different time steps.

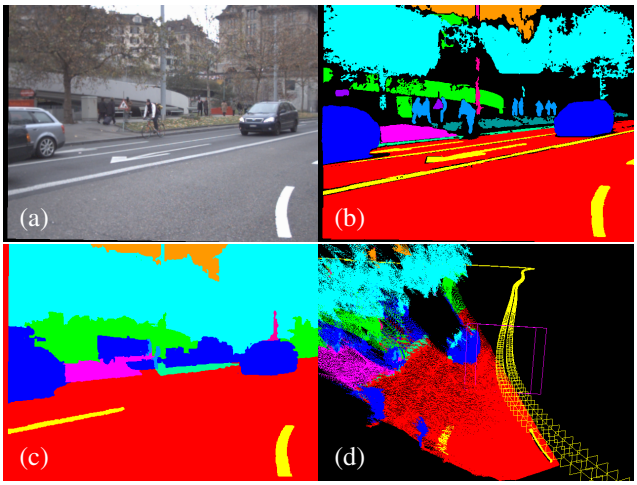

Figure 1. We propose a novel approach to integrate temporal consistency and local 3D geometry information into CRF image segmentation formulations. (**(a)** input image; **(b)** ground truth labels; **(c)** sematic segmentation results; **(d)** semantic 3D reconstruction).

It has also been argued that 3D information plays an important role for scene understanding, and several approaches have been developed to estimate 3D information from single 2D images as part of the semantic segmentation process [11, 9, 26, 7]. In parallel, a line of interesting research has emerged on segmentation of 3D point clouds acquired from highly accurate 3D laser range sensors (*e.g.*, Velodyne) [18, 19]. So far, however, there has been no connection between those two research areas. We aim at bridging this gap by incorporating directly estimated 3D information into the 2D image segmentation process. In contrast to [18, 19, 23], we however derive our 3D information from dense stereo depth, which is far noisier than laser data and which therefore requires special provisions.

In this paper, we propose a unified framework which incorporates both of those motivations. We introduce a novel CRF framework which forms temporal consistency constraints over consecutive frames through higher-order potentials defined on the points of a local 3D point cloud reconstruction (see Fig. 1(d)). Our framework offers a principled way to incorporate local 3D geometry information