In The Name of God.
The Merciful, The Compassionate.

Extracting 3D Scene-consistent Object Proposals and Depth from Stereo Images

By Michael Bleyer, Christoph Rhemann, and Carsten Rother

1 Abstract and Introduction

- The goal is to jointly extract objects and estimate depths from stereo images
- Main contribution is to introduce the concept of 3D scene consistency in stereo matching
- Few works on 3D reasoning with respect to stereo images
- Object stereo [1]: the goal was to improve depth estimation by object extraction.
- This work: main focus is on object extraction.
- Inspired by the work of [12]. Proposed the following 3-step pipeline for object extraction:
    1. generate large pool of object proposals
    2. rank object proposals by learning objectness score
    3. perform object recognition on top ranked proposals
- This work differs in the case that it takes an stereo image as input and generates a pool of scene proposals which consist:
    1. disparity map
    2. object map: each pixel $\longrightarrow$ an object
- Object stereo [1]: did not introduce the concept of computing a pool of object maps.
- Key difference is objects in [1] were approximated by flat 2D planes. We enclose them by using a 3D bounding box $\Longrightarrow$ we can exploit physical constraints.

2 Model

Each pixel $p \in \mathcal{I}$ is assigned to a 3D plane. Computes a mapping $F : \mathcal{I} \to \mathcal{F}$ where $\mathcal{F}$ denotes the set of all possible 3D planes. Disparity d_p is defined using its plane f_p as $d_p := a_{f_p} p_x + b_{f_p} p_y + c_{f_p}$.

Second mapping for objects: $O : \mathcal{I} \to \mathcal{O}$ where O is object map and $\mathcal{O}$ is set of all objects. An object is defined by 2 parameters:

1. Oriented 3D bounding box
2. Color model

$\langle F, O \rangle$ forms the scene proposal.

The quality of a scene proposal is measured by an energy, as:

$$E(F,O) = E_{pc}(F) + E_{col}(O) + E_{ol}(O,F) + E_{tight}(O) + E_{is}(O) + E_{gravity}(O) + E_{mdl}(O) \tag{1}$$

The individual terms are explained informally next:

- $E_{pc}(F)$: photo consistency; penalizes difference between left and right image given f_p with local smoothing.
- $E_{col}(O)$: color; prefers objects that are compact in color. Color of an object is modelled by GMM.
- $E_{ol}(O, F)$: Bounding box(BB) outlier; penalizes count of 3D points P outside object O_p's BB. (How BB is computed?)
- $E_{tight}(O)$: tightness $\to \sum_{o \in O} volume(o)$; penalizes BBs from unnecessarily fill free space.
- $E_{is}(O)$: intersection.
- $E_{gravity}(O)$: gravity; encourages objects to stand on top of each other
- $E_{mdl}(O)$: mdl; encourages small number of objects as possible.

3 Optimization

Steps of the optimization procedure are described below:

1. Initial Disparity Map Computation: Optimizes E_{pc} and returns F' which is initial mapping of pixels to 3D pixels.
2. Object Map Generation: depth segmentation on left view

   (a) first, divides the image into color segments using Mean Shift
   (b) A disparity plane is fitted to each color segment using F'
   (c) Group segments with similar disparity planes

   Varying the parameters of Mean Shift generates different object maps.

3. Object Map Refinement: process each object map O_i separately. Optimization is based on hypotheses testing. Starts with objects that have lowest BB volume. They are aligned to on top of another BB that is spatially close. If it leads to lower energy, this move is accepted. A random shift of border shift is also done if it leads to lower energy.

4. Object Map Fusion: The goal is to select objects from object maps $O_1, \ldots, O_n$. Selected objects form a new object map O'. Simulated annealing is used to find optimal selection of objects.

5. Joint Object/Disparity Map Refinement: The goal is to find refined O, F. Joint optimization is done by assigning each pixel to a plane and object label. Patch Match based joint optimization is as follows:

   (a) Initialization: with assigning each pixel to to a random plane and random object (values of F', O'). Labelling cost of each pixel is computed by evaluation of pc, col, ol.
   (b) Propagation: plane and object label of a pixel are propagated to its neighbours. Accept this propagation if it leads to lower energy.