ISPRS Journal of Photogrammetry and Remote Sensing xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Object Scene Flow

Moritz Menze [a,*], Christian Heipke [a], Andreas Geiger [b,c]

[a] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Nienburger Str. 1, D-30167 Hannover, Germany
[b] Autonomous Vision Group, Max Planck Institute for Intelligent Systems, Spemannstr. 41, D-72076 Tübingen, Germany
[c] Computer Vision and Geometry Lab, ETH Zürich, Universitätstrasse 6, CH-8092 Zürich, Switzerland

ARTICLE INFO

Article history:
Received 26 February 2017
Received in revised form 17 September 2017
Accepted 20 September 2017
Available online xxxx

Keywords:
Scene flow
Motion estimation
Motion segmentation
3D reconstruction
Active Shape Model
Object detection

ABSTRACT

This work investigates the estimation of dense three-dimensional motion fields, commonly referred to as *scene flow*. While great progress has been made in recent years, large displacements and adverse imaging conditions as observed in natural outdoor environments are still very challenging for current approaches to reconstruction and motion estimation. In this paper, we propose a unified random field model which reasons jointly about 3D scene flow as well as the location, shape and motion of vehicles in the observed scene. We formulate the problem as the task of decomposing the scene into a small number of rigidly moving objects sharing the same motion parameters. Thus, our formulation effectively introduces long-range spatial dependencies which commonly employed local rigidity priors are lacking. Our inference algorithm then estimates the association of image segments and object hypotheses together with their three-dimensional shape and motion. We demonstrate the potential of the proposed approach by introducing a novel challenging scene flow benchmark which allows for a thorough comparison of the proposed scene flow approach with respect to various baseline models. In contrast to previous benchmarks, our evaluation is the first to provide stereo and optical flow ground truth for dynamic real-world urban scenes at large scale. Our experiments reveal that rigid motion segmentation can be utilized as an effective regularizer for the scene flow problem, improving upon existing two-frame scene flow methods. At the same time, our method yields plausible object segmentations without requiring an explicitly trained recognition model for a specific object class.

1. Introduction

Scene flow estimation provides valuable information about the dynamic nature of our three-dimensional environment. In particular, the three-dimensional scene flow field comprises all 3D motion vectors of a densely reconstructed 3D surface model, which is moving with respect to the camera. Recovering scene flow from image observations, however, is an inherently ill-posed inverse problem, requiring the development of appropriate priors for regularizing the space of solutions.

In addition to the inherent academic interest in perceiving systems, image-based scene flow estimation is relevant for a broad range of applications. While active sensors are a strong competitor in many fields, image sequences contain valuable dynamic information. Automatic navigation of autonomous platforms (Geiger et al., 2014; Zhang et al., 2013) is just one example requiring a detailed dynamic perception of the 3D environment. While warning and avoidance of moving obstacles is already part of advanced driver assistance systems, existing solutions are still restricted to certain types of objects, limited speed and ranges. The safe interaction of robots with their environment also requires up-to-date and precise information about their surroundings. Furthermore, motion cues are important for action and activity recognition for example in video surveillance applications. All of these tasks benefit from an improved perception of surrounding shapes and motions.

In this work, we propose a consistent model allowing for joint inference of both entities. In particular, we propose a unified random field model which reasons jointly about 3D scene flow as well as the location, shape and motion of vehicles in the observed scene. We formulate the problem as the task of decomposing the scene into a small number of rigidly moving objects sharing the same motion parameters. Our inference algorithm estimates the association of image segments and object hypotheses together with their three-dimensional shape and motion. We extend our model to jointly estimate the parametrized 3D shape of each vehicle in the scene. To evaluate our approach, we develop a comprehensive

* Corresponding author.
E-mail address: moritz.menze@alumni-uni-hannover.de (M. Menze).

https://doi.org/10.1016/j.isprsjprs.2017.09.013
0924-2716/

dataset and evaluation, the KITTI 2015 scene flow benchmark,[1] allowing for detailed quantitative analysis of the results and an in-depth comparison to the state-of-the-art.

1.1. Related work

Image-based methods for scene flow estimation can be categorized into variational and discrete optimization approaches. With the advent of consumer grade active sensors like the Microsoft Kinect, depth information has become readily available and is leveraged by a number of recent scene flow approaches, e.g., Herbst et al. (2013), Hornacek et al. (2014) and Quiroga et al. (2014). While active sensors work well for indoor scenes of limited extent, the focus of this paper is on the outdoor scenario with applications to autonomous driving. Therefore, we concentrate on appearance based methods in our literature review.

1.1.1. Scene flow estimation

Following the seminal approaches to optical flow (Horn and Schunck, 1981) and scene flow estimation (Vedula et al., 1999, 2005), the problem of estimating a three-dimensional displacement field is traditionally formulated in a variational setting. As depth information is needed, different ways to incorporate dense reconstruction into the variational framework have been proposed. Analogous to the 2D case, optimization has to proceed in a coarse-to-fine manner to avoid local minima of the energy functional and capture large displacements.

Pons et al. (2007) alternatingly optimize the reconstruction of a surface model and the motion field. The key contribution addresses the data term. To circumvent common assumptions of similarity measures the authors propose a global prediction error evaluating the consistency of all input images, which are warped according to the reconstructed surfaces and estimated motion. The resulting algorithm appropriately handles projective distortion and partial occlusions. To regularize the results simple smoothness constraints are imposed. The resulting energy functional is optimized in a coarse-to-fine gradient descent framework.

Huguet and Devernay (2007) generalize the variational optical flow method of Brox et al. (2004) to jointly infer geometry and motion. To this end, they propose a minimal representation of scene flow by four variables in the image domain. In particular, they compute the disparity at the first time step t_0, the optical flow with respect to the reference image and the disparity at the second time step t_1. Given a calibrated stereo camera, the three-dimensional scene flow can directly be computed from this representation. To jointly optimize stereo disparity and optical flow, Huguet and Devernay (2007) extend the respective data and smoothness terms to cover all sought entities and combine them in a unified energy functional. This formulation leads to four partial differential equations, which are optimized using the numerical scheme proposed by Brox et al. (2004) for 2D optical flow. Since stereo image matching typically has to deal with large displacements, a dedicated initialization procedure is required. To this end, pre-computed disparity and optical flow maps are employed. We leverage a similar strategy for the initialization of the proposed approach (see Section 2.3).

An important aspect of scene flow estimation is regularization. Basha et al. (2013) argue that smooth 3D motion fields can project to discontinuous 2D flow fields and thus propose a 3D model representing the scene as a point cloud with spatial motion vectors. This formulation allows to apply regularization directly to the three-dimensional motion vectors and to easily extend the method to a multi-view set-up. Vogel et al. (2011) replace the global total variation regularization by a piecewise rigid prior. Thus, sharp discontinuities in the scene flow field can be preserved more faithfully.

Valgaerts et al. (2010) discard the common assumption of a fully calibrated stereo rig and explicitly estimate the relative orientation between the stereo heads. Consequently, the results are only retrieved up to an unknown scale factor. The energy functional becomes more complex as it now comprises general stereo terms based on the unknown fundamental matrix and is minimized in a coarse-to-fine optimization scheme. Furthermore, the authors decouple regularization of shape and motion, as they do not assume respective discontinuities to coincide.

For reasons of computational efficiency, Wedel et al. (2008) completely decouple shape and motion estimation and focus on the computation of the displacement vector field, which significantly increases frame rate but also discards valuable mutual constraints between both entities. Rabe et al. (2010) parallelize the required computations on a GPU. They apply Kalman filtering to each pixel individually to smooth the resulting motion vectors over longer image sequences. Aiming at high frame rates, the recent prediction-correction approach of Derome et al. (2016) follows a similar strategy. Depth maps from stereo image matching are combined with visual odometry to predict optical flow vectors. While static parts of the scene can be recovered directly, a correction step has to be applied to account for individually moving objects.

For stereo matching and optical flow estimation, several publications have demonstrated the usefulness of slanted-plane models (Bleyer et al., 2011; Yamaguchi et al., 2013). Either the visible surface of the scene or its projected motion are assumed to vary smoothly within small regions in the reference image. Extending this idea to 3D forms the basis for the currently most successful scene flow models.

Yamaguchi et al. (2014) propose a semi-dense method, which builds on the well-known semiglobal matching described in (Hirschmüller, 2008). The stereo approach is extended to incorporate a third image from the reference camera, taken at a second time step. Based on the assumption of a static scene, this additional information increases the robustness of image matching. In addition to the disparity map in the reference image, it yields an estimate of the optical flow. To smooth and extrapolate the matching results, a slanted-plane model is optimized yielding an over-segmentation of the reference view together with dense estimates of disparity and optical flow. The combined approach is referred to as slanted plane smoothing of stereo and flow (SPS-StFl). As in previous work (Yamaguchi et al., 2013) there is a purely stereoscopic variant of the approach (SPS-St) and a dedicated version which is tailored towards optical flow estimation (SPS-Fl). In a related work, Lv et al. (2016) proposed a purely continuous factor-graph optimization using a piecewise-planar scene flow model.

Vogel et al. (2013b) propose a scene flow approach assuming piece-wise rigid surfaces (PRSF). Their formulation decomposes the 3D scene into planar regions, each undergoing a rigid motion. The reference image is decomposed into segments and for each of the segments, a parametrized representation of shape and motion is retrieved. Consequently, the number of unknowns is reduced compared to a pixel-wise representation. The smoothness assumption within each segment further implements a strong regularization. Inference in this model assigns each pixel to an image segment and each segment to one of several rigidly moving plane proposals in three-dimensional space, thus casting the task as a discrete labeling problem.

To initialize the plane proposals, 3D plane parameters and rigid body transformations are robustly fit to initial disparity and flow maps. These observations are evaluated with respect to an initial segmentation of the reference frame. During inference, the resulting planes are proposed for superpixels in the vicinity of

[1] http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php.

Please cite this article in press as: Menze, M., et al. Object Scene Flow. ISPRS J. Photogram. Remote Sensing (2017), https://doi.org/10.1016/j.isprsjprs.2017.09.013