Joint Optical Flow and Temporally Consistent Semantic Segmentation

Junhwa Hur(✉) and Stefan Roth

Department of Computer Science, TU Darmstadt, Darmstadt, Germany
junhwa.hur@visinf.tu-darmstadt.de

Abstract. The importance and demands of visual scene understanding have been steadily increasing along with the active development of autonomous systems. Consequently, there has been a large amount of research dedicated to semantic segmentation and dense motion estimation. In this paper, we propose a method for jointly estimating optical flow and temporally consistent semantic segmentation, which closely connects these two problem domains and leverages each other. Semantic segmentation provides information on plausible physical motion to its associated pixels, and accurate pixel-level temporal correspondences enhance the accuracy of semantic segmentation in the temporal domain. We demonstrate the benefits of our approach on the KITTI benchmark, where we observe performance gains for flow and segmentation. We achieve state-of-the-art optical flow results, and outperform all published algorithms by a large margin on challenging, but crucial dynamic objects.

1 Introduction

Visual scene understanding from movable platforms has been gaining increased attention due to the active development of autonomous systems and vehicles. Semantic segmentation and dense motion estimation are two core components for recognizing the surrounding environment and analyzing the motion of entities in the scene. The performance of techniques in both areas has been steadily increasing, reported and fueled by public benchmarks (e.g., KITTI [9], MPI Sintel [4], or Cityscapes [6]). Along with the increasing popularity and importance of the two areas, there has been a recent trend in the literature considering how to bridge the two themes and analyzing which benefits these tasks can additionally derive from one another.

There have been a few basic attempts to utilize optical flow to enforce temporal consistency of semantic segmentation in a video sequence [5,10,22]. Also, segmenting the scene into superpixels (without clear semantics) has been shown to help estimating more accurate optical flow, assuming that object boundaries may give rise to motion boundaries [30,36,37]. Strictly speaking, however, previous work so far simply uses the results from one task as supplementary information for the other, and there have not been many attempts to relate the two tasks

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part I, LNCS 9913, pp. 163–177, 2016.
DOI: 10.1007/978-3-319-46604-0_12

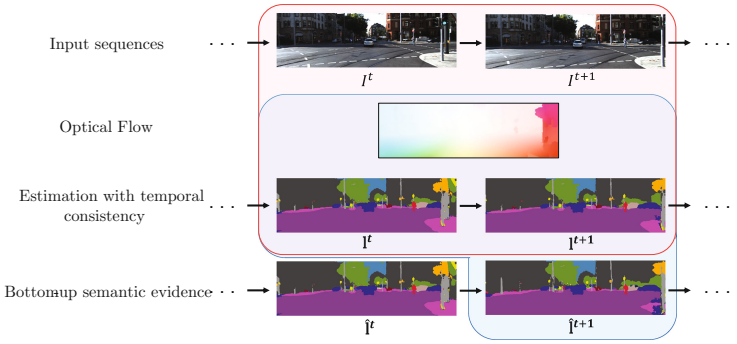


Fig. 1. Overview of our approach. The red region contributes to estimating optical flow, and the blue region ensures temporal consistency of the semantic segmentation, both given two frames. The overlapping region defines the output of our method. (Color figure online)

more closely or to solve them jointly. Yet, off-the-shelf motion estimation algorithms are not accurate enough to fully rely on [5,22]. The only exception is very recent work that uses both semantic information and segmentation to increase the accuracy of optical flow [27], however without considering the benefits of temporal correspondence for semantic labeling.

In this paper, we address this gap and present an approach for joint optical flow estimation and temporally consistent semantic segmentation from monocular video, in which both tasks leverage each other. Figure 1 shows the overview of our method. We begin by assuming that a bottom-up semantic segmentation for each frame is given. Then we estimate accurate optical flow fields by exploiting the semantic information from the given semantic segmentation. The benefit of semantic labels is that they can give us information on the likely physical motion of the associated pixels. At the same time, accurate pixel-level correspondence between consecutive frames can establish temporally consistent semantic segmentations and help refining the initial results.

We make two major contributions. First, we introduce an accurate piecewise parametric optical flow formulation, which itself already outperforms the state of the art, particularly on dynamic objects. Our formulation explicitly handles occlusions to prevent the data term from unduly influencing the results in occlusion areas. As a result, our method additionally provides occlusion information such as occlusion masks and occlusion types. Our second contribution is to jointly estimate optical flow and temporally consistent semantic segmentation in a monocular video setting. For the flow estimation, we additionally apply the epipolar constraint for pixels that should be consistent with the camera egomotion, as inferred by the semantic information. At the same time, accurately estimated flow helps to enforce temporal consistency on the semantic segmentation. We effectively realize these ideas in our joint formulation and make them feasible using inference based on patch-match belief propagation (PMBP) [2].

Our experiments on the popular KITTI dataset show that our method yields state-of-the-art results for optical flow. For estimating flows on dynamic foreground objects, which are particularly crucial from an autonomous navigation standpoint, our method outperforms all published optical flow algorithms in the benchmark by a significant margin.

2 Related Work

Piecewise Parametric Flow Estimation. Piecewise parametric approaches using a homography model have recently shown promising results on standard benchmarks [4,9] for motion estimation. Representing the scene as a set of planar surfaces significantly reduces the number of unknowns; at the same time, parametrizing the motion of surfaces by 9-DoF or 8-DoF transforms ensures sufficient diversity and generality of their motion [12,20,31–33,38]. In the stereo setting, Vogel *et al.* [31–33] proposed a scene representation consisting of piecewise 3D planes undergoing 3D rigid motion and demonstrate the most accurate results to date for estimating the 3D scene flow on the KITTI benchmark.

On the other hand, the monocular case with its limited amount of data (i.e., two consecutive images) makes the problem more challenging, hence the type of regularization becomes much more important [12,38]. Hornacek *et al.* [12] introduced a 9-DoF plane-induced model for optical flow via continuous optimization. Their method shows its strength on rigid motions, but is weaker on poorly textured regions because of the lack of global support. Yang and Li [38] instead use a 8-DoF homography motion in 2D space with adaptive size and shape of the pieces via discrete optimization.

Our approach also relies on an 8-DoF parameterization, which we found to yield accurate optical flow estimates in practice.

Epipolar Constraint-Based Flow Estimation. Several approaches have relied on the epipolar constraint for estimating motion [1]. Strictly enforcing the constraint gives the benefit of reducing the search space significantly, but causes an inherent limitation for handling independently moving objects whose motion usually violates the constraint [13,23,36,37]. Adding the constraint as a soft prior can resolve this issue, but there is still the challenge of determining where to relinquish the constraint by only depending on the data term [34,35].

Our approach explicitly resolves this ambiguity with the aid of semantic information, which provides information on the physical properties of objects (e.g., static or movable).

Temporally Consistent Semantic Segmentation. Among a broad literature on enabling temporal consistency of video segmentation, we specifically consider the case of semantic segmentation here. One common way to inject temporal consistency is to utilize motion and structure features from 3D point clouds obtained by Structure from Motion (SfM) [3,7,29]. Another way is to jointly reconstruct a scene in 3D with semantic labels through a batch process, naturally enabling temporally consistent segmentation [14,26,40]. In causal approaches that rely on

temporal correspondence, previous approaches achieve accurate temporal correspondence using sparse feature tracking [25] or dense flow maps with a similarity function in feature space [22]. A recent work [15] introduces feature space optimization for spatio-temporal regularization in partitioned batches with overlaps.

We achieve temporal consistency for semantic segmentation using a jointly estimated, accurate dense flow map, which leverages the semantic information.

Optical Flow with Semantics. The question of exploiting semantics for optical flow has only received very limited attention so far. The most related approach is the very recent work by Sevilla-Lara *et al.* [27], which treats the problem sequentially. First, the scene is segmented into 3 semantic categories, things, planes and stuff. Second, motion is estimated individually for these semantic parts and later composited. In contrast, we treat the entire problem as the minimization of a single unified energy. Moreover, motion estimation and semantic segmentation are inferred jointly instead of sequentially, hence may mutually leverage each other. Experimentally, we report significantly more accurate motion estimates for dynamic objects and demonstrate improved segmentation performance.

3 Approach

The core idea put forward in this paper is that optical flow and semantic segmentation are mutually beneficial and are best estimated jointly to simultaneously improve each other. Figure 1 shows the flow of our proposed method in the temporal domain and explains which elements contribute to achieving which task. Here, we assume that some initial bottom-up semantic evidence is already given by an off-the-shelf algorithm, such as a CNN (e.g., [16]), which is subsequently refined by having temporal consistency. In the red-shaded region, a pair of consecutive images and their refined semantic segmentation contribute to estimating optical flow more accurately. At the same time, the temporally consistent semantic labeling at time $t+1$ is inferred from its bottom-up evidence, the previously estimated semantic labeling at time t, and the estimated flow map. For longer sequences, our approach proceeds in an online manner on two frames at a time.

Similar to [38], our formulation is based on an 8-DoF piecewise-parametric model with a superpixelization of the scene. Superpixels play an important role in our formulation for connecting the two different domains: optical flow and semantic segmentation. One superpixel represents a global motion as well as a semantic label for its pixels inside, and the motion is constrained by the physical properties that the semantic label implies. For example, the motion of pixels corresponding to some physically-static objects (e.g., building or road) can only be caused by camera motion. Thus, enforcing the epipolar constraint on those pixels can effectively regularize their motion.

Another important feature of our formulation is that we explicitly formulate the occlusion relationship between superpixels [36,37] and infer the occlusion mask as well. This directly affects the data term such that it prevents occluded pixels from dominating the data term during the optimization.

3.1 Preprocessing

Superpixels. As superpixels generally tend to separate objects in images, they can be a good medium for carrying semantic labels and representative motions for their pixels. Our approach uses the recent state-of-the-art work of Yao *et al.* [39], which has shown to be well suited for estimating optical flow.

Semantic Segmentation. For the bottom-up semantic evidence, we use an off-the-shelf fully convolutional network (FCN) [16] trained on the Cityscapes dataset [6], which contains typical objects frequent in street scenes.

Fundamental Matrix Estimation. In order to apply the epipolar constraint on superpixels for which their semantic label tells us that they are surely static objects (e.g. roads, buildings, etc.), our approach requires the fundamental matrix resulting from the camera motion. We use a standard approach, i.e. matching SIFT keypoints [18] and using the 8-point algorithm [11] with RANSAC [17].

3.2 Model

Our model jointly estimates *(i)* the optical flow between reference frame I^t and the next frame I^{t+1}, and *(ii)* a temporally consistent semantic segmentation $\mathbf{l}^{t+1}$ given bottom-up semantic evidence $\hat{\mathbf{l}}^{t+1}$ and the previously estimated semantic labeling $\mathbf{l}^t$. $\mathbf{l}$ is a semantic label probability map, which has the same size as the input image and L channels, where L is the number of semantic classes. Instead of using a single label, we adopt label probabilities so that we can more naturally and continuously infer the semantic labels in the time domain. Note that we assume an online setting (i.e., no access to future information) and hence infer the segmentation at time $t+1$ rather than t. Optical flow is represented by a set of piecewise motions of superpixels in the reference frame. We define the motion of a superpixel through a homography and formulate the objective for estimating the 8-DoF homography $\mathbf{H}_s$ of each superpixel s and the temporally consistent semantic segmentation $\mathbf{l}^{t+1}$ as:

$$\begin{aligned} E(\mathbf{H}, \mathbf{l}^{t+1}, o, b) = {} & E_{\mathrm{D}}(\mathbf{H}, o) + \lambda_{\mathrm{L}} E_{\mathrm{L}}(\mathbf{H}, \mathbf{l}^{t+1}, o) \\ & + \lambda_{\mathrm{P}} E_{\mathrm{P}}(\mathbf{H}) + \lambda_{\mathrm{C}} E_{\mathrm{C}}(\mathbf{H}, o, b) + \lambda_{\mathrm{B}} E_{\mathrm{B}}(b). \end{aligned} \tag{1}$$

Here, E_{D}, E_{L}, E_{P}, E_{C}, and E_{B} denote color data term, label data term, physical constraint term, connectivity term, and boundary occlusion prior, respectively.

We adopt two kinds of occlusion variables: the boundary occlusion label b between two superpixels, and the occlusion mask o defined at the pixel level. The boundary occlusion label b regularizes the spatial relationship between two neighboring superpixels (i.e., co-planar, hinge, left occlusion, or right occlusion) [36,37]. The occlusion mask o explicitly models whether a pixel is occluded or not. One important difference to previous superpixel-based work [37] is that we additionally infer a pixelwise occlusion mask, which prevents occluded pixels from adversely affecting the data cost.

Data Terms. The data terms aggregate photometric differences

$$E_{\mathrm{D}}(\mathbf{H}, o) = \sum_{s \in S} \frac{1}{|s|} \underbrace{\sum_{\mathbf{p} \in s} (1 - o_{\mathbf{p}}) \rho_{\mathrm{D}} \Big(I^{t}(\mathbf{p}), I^{t+1}(\mathbf{p}') \Big) + o_{\mathbf{p}} \lambda_{\mathrm{o}}}_{\text{image data}} \tag{2}$$

and semantic label differences

$$E_{\mathrm{L}}(\mathbf{H}, \mathbf{l}^{t+1}, o) = \sum_{s \in S} \frac{1}{|s|} \sum_{\mathbf{p} \in s} \phi_l(\mathbf{H}, \mathbf{l}^{t+1}_{\mathbf{p}'}, o) \qquad \text{with} \tag{3}$$

$$\phi_l(\mathbf{H}, \mathbf{l}^{t+1}_{\mathbf{p}'}, o) = \frac{1}{2} \sum_{i}^{L} (1 - o_{\mathbf{p}}) \Big\| \mathbf{l}^{t+1}_{\mathbf{p}',i} - (\alpha \hat{\mathbf{l}}^{t+1}_{\mathbf{p}',i} + (1 - \alpha) \mathbf{l}^{t}_{\mathbf{p},i}) \Big\|^2 \tag{4}$$

over each pixel of each superpixel. Here, $\mathbf{p}'$ is the corresponding pixel in I^{t+1} of pixel $\mathbf{p}$ in I^t, which is determined according to the homography $\mathbf{H}_{\mathcal{S}(\mathbf{p})} \in \mathbb{R}^{3 \times 3}$ of its superpixel

$$\mathbf{p}' = \mathbf{H}_{\mathcal{S}(\mathbf{p})} \mathbf{p}, \tag{5}$$

where $\mathcal{S} : I^t \rightarrow S$ is a mapping that assigns a pixel $\mathbf{p}$ to its superpixel $s \in S$.

In the image data term in Eq. (2), the function $\rho_D(\cdot, \cdot)$ measures the photometric differences between two pixels using the ternary transform [28] and a truncated linear penalty. If a pixel $\mathbf{p}$ is occluded (i.e., $o_{\mathbf{p}} = 1$), a constant penalty λ_{o} is applied.

The label data term in Eq. (3) measures the distance between two semantic label probability distributions over each pixel: *(i)* our estimation $\mathbf{l}^{t+1}_{\mathbf{p}'}$ and *(ii)* a weighted sum of the previous estimation $\mathbf{l}^{t}_{\mathbf{p}}$, which is propagated by the optical flow, and the bottom-up evidence $\hat{\mathbf{l}}^{t+1}_{\mathbf{p}'}$, while considering its occlusion status. The motivation of the term is to penalize label differences to the bottom-up evidence and at the same time propagate label evidence over time, except when an occlusion takes place.

Physical Constraint Term. Semantic labels can provide useful cues for estimating optical flow. If pixels are labeled as physically static objects, such as building, road, or infrastructure, then they normally do not undergo any 3D motion, hence their observed 2D motion is caused by camera motion and should thus satisfy the epipolar constraint. We define the corresponding term as

$$E_{\mathrm{P}}(\mathbf{H}) = \sum_{s \in S} \min(\phi_P(s, \mathbf{H}_s), \lambda_{\text{non\_st}} + \beta[l^t_s \in L_{\text{st}}]), \tag{6}$$

$$\text{where} \quad \phi_{\text{st}}(s, \mathbf{H}_s) = \frac{1}{|s|} \sum_{\mathbf{p} \in s} \Big\| (\mathbf{p}'^{\top} \mathbf{F} \mathbf{p}) \Big\|_1 = \frac{1}{|s|} \sum_{\mathbf{p} \in s} \Big\| ((\mathbf{H}_{\mathcal{S}(\mathbf{p})} \mathbf{p})^{\top} \mathbf{F} \mathbf{p}) \Big\|_1 \tag{7}$$

measures how well the homography matrix $\mathbf{H}_s$ of a superpixel s meets the epipolar constraint from the fundamental matrix $\mathbf{F}$. For non-static objects, such as pedestrians or vehicles, we still apply the epipolar penalty, however a weak one

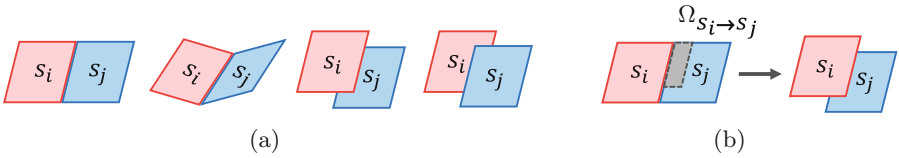


Fig. 2. *(a)* Four cases of boundary relations between two superpixels: co-planar, hinge, left occlusion, and right occlusion. *(b)* The visualization of the set of occluded pixels $\Omega_{s_\mathrm{i} \to s_\mathrm{j}}$ in the case of a left occlusion (black-colored region).

using a low truncation threshold $\lambda_{\text{non\_st}}$. This is motivated by the fact that possibly dynamic objects may in fact stand still and thus obey epipolar geometry, but we do not want to penalize them too much if they do not. For static objects, on the other hand, we augment the truncation threshold by β in order to give a stricter penalty. L_{st} is the set of semantic labels that corresponds to the physically static objects. l_s is a representative semantic label of superpixel s, which has the highest probability over the pixels in the superpixel: $l_s = \text{argmax}_i \sum_{\mathbf{p} \in s} \mathbf{l}^t_{\mathbf{p},i}$.

Connectivity Term. The connectivity term encourages the smoothness of motion between two neighboring superpixels based on their occlusion relationship:

$$E_{\mathrm{C}}(\mathbf{H}, o, b) = \sum_{s_i \sim s_j} \phi_{\mathrm{C}}(\mathbf{H}_{s_i}, \mathbf{H}_{s_j}, o, b_{ij}) \tag{8}$$

$$\text{with} \quad \phi_{\mathrm{C}}(\mathbf{H}_{s_i}, \mathbf{H}_{s_j}, o, b_{ij}) = \begin{cases} \phi_{\mathrm{co}}(\mathbf{H}_{s_i}, \mathbf{H}_{s_j}, o) & \text{if } b_{ij} = \text{co-planar}, \\ \phi_{\mathrm{h}}(\mathbf{H}_{s_i}, \mathbf{H}_{s_j}, o) & \text{if } b_{ij} = \text{hinge}, \\ \phi_{\mathrm{occ}}(s_i, s_j, o) & \text{if } b_{ij} = \text{left occlusion}, \\ \phi_{\mathrm{occ}}(s_j, s_i, o) & \text{if } b_{ij} = \text{right occlusion}. \end{cases} \tag{9}$$

As shown in Fig. 2(a), the boundary occlusion flag b_{ij} expresses the relationship between two neighboring superpixels s_i and s_j as co-planar, hinge, left-occlusion, or right-occlusion [36,37]. This categorization helps to regularize the motion of two superpixels defined by their homography matrices. We distinguish between three different potentials:

$$\phi_{\mathrm{co}}(\mathbf{H}_{s_i}, \mathbf{H}_{s_j}, o) = \frac{1}{|s_i \cup s_j|} \sum_{\mathbf{p} \in s_i \cup s_j} \|\mathbf{H}_{s_i}\mathbf{p} - \mathbf{H}_{s_j}\mathbf{p}\|_1 + \sum_{\mathbf{p} \in s_i \cup s_j} \lambda_{\mathrm{imp}}[o_p = 1] \tag{10}$$

$$\phi_{\mathrm{h}}(\mathbf{H}_{s_i}, \mathbf{H}_{s_j}, o) = \frac{1}{|\mathcal{B}_{s_i,s_j}|} \sum_{\mathbf{p} \in \mathcal{B}_{s_i,s_j}} \|\mathbf{H}_{s_i}\mathbf{p} - \mathbf{H}_{s_j}\mathbf{p}\|_1 + \sum_{\mathbf{p} \in s_i \cup s_j} \lambda_{\mathrm{imp}}[o_p = 1] \tag{11}$$

$$\begin{aligned} \phi_{\mathrm{occ}}(s_\mathrm{f}, s_\mathrm{b}, o) = &\sum_{\mathbf{p} \in s_\mathrm{f}} \lambda_{\mathrm{imp}}[o_p = 1] \\ &+ \sum_{\mathbf{p} \in s_\mathrm{b}} \Big(\lambda_{\mathrm{imp}}[\mathbf{p} \in \Omega_{s_\mathrm{f} \to s_\mathrm{b}}][o_p = 0] + \lambda_{\mathrm{imp}}[\mathbf{p} \notin \Omega_{s_\mathrm{f} \to s_\mathrm{b}}][o_p = 1] \Big) \end{aligned} \tag{12}$$

These are motivated as follows: When two superpixels are co-planar, all pixels within should follow the identical homography matrix as they are on the same plane. For a hinge relationship, only the pixels on the boundary set $\mathcal{B}_{s_i,s_j}$ can

satisfy the motion from two superpixels s_i and s_j. In both cases, there should be no occluded pixels, hence we adopt a very large 'impossible' penalty λ_{imp} to prevent occluded pixels from occurring. In case that one superpixel occludes another, their motions only affect the occlusion masks. Equation (12) expresses the case that pixels of the front superpixel s_{f} occlude some pixels of the back superpixel s_{b}. As shown in Fig. 2(b), $\Omega_{s_{\text{f}} \to s_{\text{b}}}$ is a set of pixels in s_{b} that is occluded by some pixels in s_{f} from the motion. All pixels in the front superpixel s_{f} should not be occluded, and only pixels in the set of $\Omega_{s_{\text{f}} \to s_{\text{b}}}$ in s_{b} should be occluded.

Boundary Occlusion Prior. Without an additional prior term, the boundary occlusion flag in the connectivity term would prefer to take the occlusion cases. We thus define a prior term to yield proper biases for each case:

$$E_{\text{B}}(b) = \begin{cases} \lambda_{\text{co}}[l_{s_i} \neq l_{s_j}] & \text{if } b_{ij} = \text{co-planar}, \\ \lambda_{\text{h}} & \text{if } b_{ij} = \text{hinge}, \\ \lambda_{\text{occ}} & \text{if } b_{ij} = \text{occlusion}, \end{cases} \tag{13}$$

where $\lambda_{\text{occ}} > \lambda_{\text{h}} > \lambda_{\text{co}} > 0$. Because it is less likely that two different objects are co-planar in the real world, we only apply the prior penalty for the co-planar case λ_{co} when the respective semantic labels of the superpixels differ.

3.3 Optimization

The minimization of our objective is challenging, as it combines discrete (i.e., $\{l^{t+1}, b, o\}$) and continuous (i.e., $\mathbf{H}$) variables. We use a block coordinate descent algorithm. As shown in Algorithm 1, we iteratively update each variable in the order: *(i)* homography matrices $\mathbf{H}$ for superpixels, *(ii)* occlusion variables b, o, and *(iii)* semantic label probability maps $\mathbf{l}^{t+1}$. Optimizing the homography matrices $\mathbf{H}$ is especially challenging because the matrices have 8 DoF in 2D space and their parameterization incurs a high-dimensional search space. We address this using PatchMatch Belief Propagation (PMBP) [2]; see below for details.

Once the motion $\mathbf{H}$ is updated, occlusion variables can be easily updated independently for each pair of neighboring superpixels, while other variables are held fixed. Given their motions, we first calculate the overlapping region, which can potentially be the occluded region for one of the two superpixels. Then, we calculate the energy in Eq. (1) for all four boundary occlusion cases with the candidate occlusion pixels given. The boundary occlusion case that has the minimum energy is taken, including the corresponding occlusion mask state. Finally, the semantic label probability map $\mathbf{l}^{t+1}$ can also be easily updated independently for all superpixels by minimizing label data term in Eq. (3).

Optimizing Homography Matrices Using PMBP. Our method optimizes the homography matrices in the continuous domain using PatchMatch Belief Propagation (PMBP) [12]. PMBP is a simple but powerful optimizer based on Belief Propagation. Instead of using a discrete label set, PMBP uses a set of particles that is randomly sampled and propagated in the continuous domain.

Algorithm 1. Optimization

initialization();
for $m = 1$ *to n-outer-iters* **do**
  for $n = 1$ *to n-inner-iters* **do**
    Optimizing $E(\mathbf{H}, \mathbf{l}^{t+1}, o, b)$ for $\mathbf{H}$ using PMBP
  end
  $\{b, o\} = \text{argmin}_{b,o}\, E(\mathbf{H}, \mathbf{l}^{t+1}, o, b)$
  $\mathbf{l}^{t+1} = \text{argmin}_{\mathbf{l}^{t+1}}\, E(\mathbf{H}, \mathbf{l}^{t+1}, o, b)$
end

PMBP requires an effective way of proposing the random particles; typically they are obtained from a normal distribution defined over some parameters. In our approach, however, we devise several strategies for proposing particles of the homography matrices without over-parameterization. Between two image patches, a superpixel and its corresponding region in the other frame, we estimate the homography matrix by using *(i)* LK warping, *(ii)* 3 correspondences and the fundamental matrix, *(iii)* 4 randomly perturbed correspondences, and *(iv)* sampled correspondences from neighboring superpixels. Empirically, we find that these strategies generate reasonable particles without requiring an over-parameterization, and only 5 outer-iterations are enough to be converged.

4 Experiments

We verify the effectiveness of our approach with a series of experiments on the well-established KITTI benchmark [9]. To the best of our knowledge, there is no dataset that simultaneously provides ground truth for optical flow and semantic segmentation in the same scenes; while ground truth for both is available in the KITTI benchmark, the evaluation is carried out on disjoint sequences.

We first evaluate our optical flow results on the KITTI Optical Flow 2015 benchmark and compare to the top-performing algorithms in the benchmark. In addition, we analyze the effectiveness of the semantics-related terms to understand how effectively the semantic information contributes to the estimation of optical flow. Finally, we demonstrate qualitative and quantitative results for temporally consistent semantic segmentation. We use DiscreteFlow [21] to initialize the flow estimation and utilize the FCN model [16] trained on the Cityscapes dataset [6] for bottom-up semantic segmentation evidence. We set our parameters automatically using Bayesian optimization [19] on the training portion.

4.1 KITTI 2015 Optical Flow

We compare to the top-scoring optical flow methods on the KITTI Optical Flow 2015 benchmark, which have been published at the time of submission. Note that we do not consider scene flow methods here, as they have access to multiple views. Table 1 shows the results. *Fl-bg*, *Fl-fg*, and *Fl-all* denote the flow error

Table 1. KITTI optical flow 2015: comparison to the published top-performing optical flow methods in the benchmark. Our method leads to state-of-the-art results and significantly increases the performance on challenging dynamic regions (*fg*).

Method	Non-occluded pixels			All pixels		
	Fl-bg	Fl-fg	Fl-all	Fl-bg	Fl-fg	Fl-all
MotionSLIC [36]	**6.19 %**	64.82 %	16.83 %	14.86 %	66.21 %	23.40 %
PatchBatch [8]	10.06 %	26.21 %	12.99 %	19.98 %	30.24 %	21.69 %
DiscreteFlow [21]	9.96 %	22.17 %	12.18 %	21.53 %	26.68 %	22.38 %
SOF [27]	8.11 %	23.28 %	10.86 %	**14.63 %**	27.73 %	**16.81 %**
Ours (JFS)	7.85 %	**18.66 %**	**9.81 %**	15.90 %	**22.92 %**	17.07 %

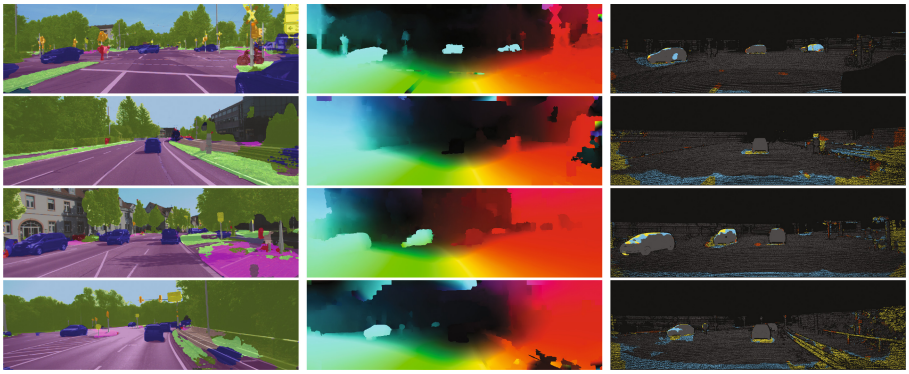


Fig. 3. Results on KITTI Optical Flow 2015. **Left**: Source images overlaid with semantic segmentation results. **Middle**: Our flow estimation results. **Right**: Qualitative comparison with DiscreteFlow: gray pixels – both methods correct, skyblue pixels – our method is correct but DiscreteFlow is not, red pixels – DiscreteFlow is correct but ours is not, and yellow pixels – both failed. (Color figure online)

evaluated for background pixels only, foreground pixels only, or for all pixels, respectively. Our method outperforms all top-scoring methods when considering all non-occluded pixels and performs very close to the leading method when considering all pixels. Especially for the flow of dynamic foreground objects, our method outperforms all published results by a large margin. This is of particular importance in the domain of autonomous navigation where understanding the motion of other traffic participants is crucial. This substantial performance gain stems from several design decisions. First, our piecewise motion representation effectively abstracts the planar surfaces of foreground vehicles, and the 8-DoF homography successfully describes the rigid motion of each surface.

The soft epipolar constraint of our model, derived from the jointly estimated semantics, contributes to the flow estimation particularly on background pixels and clear performance gains are observed for non-occluded pixels. When including occluded pixels, however, SOF [27] slightly outperforms ours.

The main reason is that their localized layer approach and planar approximation with large pieces can regularize the occluded regions better than our piecewise model based on superpixels. In future work, this gap may be addressed through an additional global support model or coarse-to-fine estimation. MotionSLIC [36] still performs better than ours on background pixels by strictly enforcing the epipolar constraint. As a trade-off, however, their strict epipolar constraint yields significant flow errors for foreground pixels and eventually increases the overall error.

Figure 3 shows visual results on the KITTI dataset (visualized as in [27]) and provides a direct comparison to DiscreteFlow, which highlights where the performance gain over the initialization originates. Our method provides more accurate flow estimates on foreground objects, but also on static objects.

4.2 Effectiveness of Semantic-Related Terms

Next we analyze the effectiveness of the semantic-related terms, the epipolar constraint term and the label data term, in order to understand how much the semantic information contributes to optical flow estimation over our basic piecewise optical flow model. We turned off each term and evaluated how each setting affects the flow estimation results on the KITTI Flow 2015 training dataset. The analysis is shown in Table 2.

We find that the label term clearly contributes to more accurate flow estimation overall, but it has a side-effect on background areas where the initial semantic segmentation may have some outliers. Using the epipolar constraint term results in more accurate flow estimates on background areas, which majorly satisfy the epipolar assumption. On foreground objects, however, the flow error slightly increases. This performance loss is coming from the trade-off of our assumption that non-static objects (e.g., vehicles) sometimes do not move, which made us apply the epipolar cost but with a small truncation threshold.

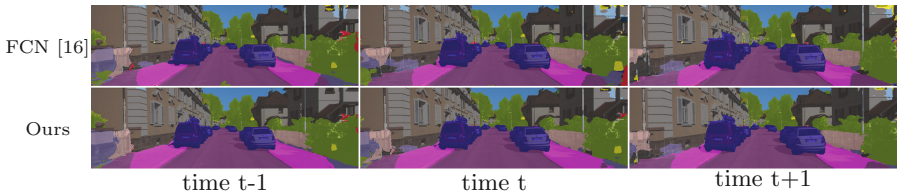
One interesting observation is that our basic piecewise flow model, without the semantic-related terms, still demonstrates competitive performance for estimating optical flow on non-occluded pixels.

Table 2. Effectiveness of semantic-related terms: the performance of our basic piecewise optical flow model is boosted further (KITTI 2015 training set).

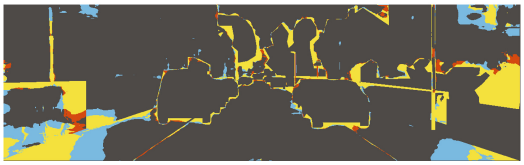
Usage of terms		Non-occluded pixels			All pixels		
Label	Epi	Fl-bg	Fl-fg	Fl-all	Fl-bg	Fl-fg	Fl-all
✓	✓	8.27 %	17.40 %	**9.83 %**	16.44 %	20.02 %	**16.98** %
✓	✗	8.45 %	**16.97 %**	9.90 %	16.73 %	**19.61 %**	17.17 %
✗	✓	**8.20 %**	17.82 %	9.84 %	**16.35 %**	20.41 ,%	16.99 %
✗	✗	8.51 %	17.21 %	10.00 %	16.84 %	19.86 %	17.31 %

Table 3. Performance of temporally consistent semantic segmentation.

IoU (%)	Sky	Building	Road	Sidewalk	Fence	Vegetation	Pole	Car	Sign	Pedestrian	Cyclist	Mean
FCN [16]	69.35	78.53	73.75	38.19	33.33	68.37	23.68	77.60	31.27	20.11	21.42	48.69
Ours	71.80	79.97	77.99	41.01	36.27	69.21	16.44	78.58	39.05	23.50	25.44	50.84



(a) Results on three consecutive frames.



(b) Performance gain/loss over bottom-up semantic segmentation.

Fig. 4. Temporally consistent semantic segmentation results.

4.3 Temporally Consistent Semantic Segmentation

We finally evaluate the performance of our temporally consistent semantic segmentation on a sequence from the KITTI dataset, which has a 3rd-party ground truth semantic annotation [24]. This, however, is a preliminary result, since the semantic segmentation model we used here is trained on the higher-resolution Cityscapes dataset [6], which possesses somewhat different statistics. Better results are expected from a custom-trained model. Table 3 shows that our joint approach increases the segmentation accuracy over the bottom-up segmentation results [16] by 2 % points in the intersection-over-union (IoU) metric. The accuracy is increased on all object classes except for the pole class, which is not well captured by our superpixels. Figure 4(a) shows our results on three consecutive frames, and Fig. 4(b) demonstrates our performance gain/loss over the bottom-up segmentation using the visualization of Fig. 3. With the aid of accurate temporal correspondences, our method revises inconsistent results and effectively reduces false positives in the time domain.

5 Conclusion

We have proposed a method for jointly estimating optical flow and temporally consistent semantic segmentation from monocular video. Our results on the challenging KITTI benchmark demonstrated that both tasks can successfully leverage each other. A piecewise optical flow model with PMBP inference builds

the basis and itself already achieves competitive results. Embedding semantic information through label consistency and epipolar constraints further boosts the performance. For dynamic objects, which are particularly important from the viewpoint of autonomous navigation, our method outperforms all published results in the benchmark by a large margin. Preliminary results on temporally consistent semantic segmentation further demonstrate the benefit of our approach by reducing false positives and flickering. We believe that a refinement of the superpixels may lead to further performance gains in the future.

Acknowledgement. We thank Marius Cordts for providing a pre-trained semantic segmentation model. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC Grant Agreement No. 307942.

References

1. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. **92**(1), 1–31 (2011)
2. Besse, F., Rother, C., Fitzgibbon, A., Kautz, J.: PMBP: PatchMatch belief propagation for correspondence field estimation. Int. J. Comput. Vis. **110**(1), 2–13 (2013)
3. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 44–57. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_5
4. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_44
5. Chen, A.Y.C., Corso, J.J.: Temporally consistent multi-class video-object segmentation with the video graph-shifts algorithm. In: WACV (2011)
6. Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
7. Floros, G., Leibe, B.: Joint 2D–3D temporally consistent semantic segmentation of street scenes. In: CVPR (2012)
8. Gadot, D., Wolf, L.: PatchBatch: a batch augmented loss for optical flow. In: CVPR (2016)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: CVPR (2012)
10. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: CVPR (2010)
11. Hartley, R.I.: In defense of the eight-point algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **19**(6), 580–593 (1997)
12. Hornáček, M., Besse, F., Kautz, J., Fitzgibbon, A., Rother, C.: Highly overparameterized optical flow using PatchMatch belief propagation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 220–234. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10578-9_15

13. Kitt, B., Lategahn, H.: Trinocular optical flow estimation for intelligent vehicle applications. In: ITSC (2012)
14. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M.: Joint semantic segmentation and 3D reconstruction from monocular video. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 703–718. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_45
15. Kundu, A., Vineet, V., Koltun, V.: Feature space optimization for semantic video segmentation. In: CVPR (2016)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
17. Lourakis, M.: Fundest: a C/C++ library for robust, non-linear fundamental matrix estimation (2011). http://www.ics.forth.gr/~lourakis/fundest/
18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
19. Martinez-Cantin, R.: BayesOpt: a Bayesian optimization library for nonlinear optimization, experimental design and bandits. J. Mach. Learn. Res. **15**, 3735–3739 (2014)
20. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015)
21. Menze, M., Heipke, C., Geiger, A.: Discrete optimization for optical flow. In: Gall, J., Gehler, P., Leibe, B. (eds.) GCPR 2015. LNCS, vol. 9358, pp. 16–28. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24947-6_2
22. Miksik, O., Munoz, D., Bagnell, J.A., Hebert, M.: Efficient temporal consistency for streaming video scene analysis. In: ICRA (2013)
23. Mohamed, M.A., Mirabdollah, M.H., Mertsching, B.: Differential optical flow estimation under monocular epipolar line constraint. In: Nalpantidis, L., Krüger, V., Eklundh, J.-O., Gasteratos, A. (eds.) ICVS 2015. LNCS, vol. 9163, pp. 354–363. Springer, Heidelberg (2015). doi:10.1007/978-3-319-20904-3_32
24. Ros, G., Ramos, S., Granados, M., Bakhtiary, A., Vazquez, D., Lopez, A.M.: Vision-based offline-online perception paradigm for autonomous driving. In: WACV (2015)
25. Scharwächter, T., Enzweiler, M., Franke, U., Roth, S.: Stixmantics: a medium-level model for real-time semantic scene understanding. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 533–548. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_35
26. Sengupta, S., Greveson, E., Shahrokni, A., Torr, P.H.S.: Urban 3D semantic modelling using stereo vision. In: ICRA (2013)
27. Sevilla-Lara, L., Sun, D., Jampani, V., Black, M.J.: Optical flow with semantic segmentation and localized layers. In: CVPR (2016)
28. Stein, F.: Efficient computation of optical flow using the census transform. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) DAGM 2004. LNCS, vol. 3175, pp. 79–86. Springer, Heidelberg (2004). doi:10.1007/978-3-540-28649-3_10
29. Sturgess, P., Alahari, K., Ladicky, L., Torr, P.H.S.: Combining appearance and structure from motion features for road scene understanding. In: BMVC (2012)
30. Sun, D., Liu, C., Pfister, H.: Local layering for joint motion estimation and occlusion detection. In: CVPR (2014)
31. Vogel, C., Roth, S., Schindler, K.: View-consistent 3D scene flow estimation over multiple frames. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 263–278. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_18
32. Vogel, C., Schindler, K., Roth, S.: Piecewise rigid scene flow. In: ICCV, pp. 1377–1384 (2013)

33. Vogel, C., Schindler, K., Roth, S.: 3D scene flow estimation with a piecewise rigid scene model. Int. J. Comput. Vis. **115**(1), 1–28 (2015)
34. Wedel, A., Pock, T., Braun, J., Franke, U., Cremers, D.: Duality TV-L1 flow with fundamental matrix prior. In: IVCNZ (2008)
35. Wedel, A., Cremers, D., Pock, T., Bischof, H.: Structure- and motion-adaptive regularization for high accuracy optic flow. In: ICCV, pp. 1663–1668 (2009)
36. Yamaguchi, K., McAllester, D., Urtasun, R.: Robust monocular epipolar flow estimation. In: CVPR (2013)
37. Yamaguchi, K., McAllester, D., Urtasun, R.: Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 756–771. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_49
38. Yang, J., Li, H.: Dense, accurate optical flow estimation with piecewise parametric model. In: CVPR (2015)
39. Yao, J., Boben, M., Fidler, S., Urtasun, R.: Real-time coarse-to-fine topologically preserving segmentation. In: CVPR (2015)
40. Zhang, C., Wang, L., Yang, R.: Semantic segmentation of urban scenes using dense depth maps. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 708–721. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_51