22^{nd} *Computer Vision Winter Workshop*
Nicole M. Artner, Ines Janusch, Walter G. Kropatsch (eds.)
Retz, Austria, February 6–8, 2017

Object Scene Flow with Temporal Consistency

Michal Neoral and Jan Šochman
Center for Machine Perception, Department of Cybernetics, FEE CTU in Prague
Zikova 1903/4, 166 36 Prague 6, Czech Republic
{neoramic,jan.sochman}@cmp.felk.cvut.cz

Abstract. *In this paper, we propose several improvements of the Object Scene Flow (OSF) algorithm [14]. The OSF does not use the scene flow estimated in previous frame nor the object labels and their corresponding object motion information. The goal of this paper is to use this information in order to produce temporarily consistent output throughout the whole video sequence. We evaluate the progress on the KITTI'15 multiframe dataset. We show that propagating the labels and the corresponding motion information using the estimated flow reduces the false negative rate (missed cars). Together with two further proposed improvements the overall reduction of false negative is 42%. The proposed improvements also reduce EPE on the KITTI'15 scene flow from 10.63% to 9.65%.*

1. Introduction

In this paper, we focus on the temporal consistency in the scene flow estimation. Most of the current methods do not use temporal consistency or data from previously computed frames. We propose improvements, which add temporal consistency to the Object Scene Flow (OSF) algorithm [14]. The OSF estimates independently moving objects as part of scene flow estimation. We show that adding temporal consistency leads to a more accurate scene flow estimation as well as more precise estimation of independently moving objects.

Accurate and efficient estimation of the scene flow is still an unresolved problem. Figure 1 shows examples of estimated optical flow by state-of-the-art scene flow and optical flow estimation algorithms on our own sequences. Even the best methods often fail when the conditions differ from the ones of KITTI [14]. In accordance with the official KITTI results, it could be seen that stereo methods work better than monocular. We have observed OSF produce more stable results consistent over various conditions than the KITTI first ranked PRSM [24]. It also provides independent motion segmentation.

The original OSF is not using any temporal consistency or any information from the previous image frames of the same sequence. The algorithm uses only two consecutive stereo frames. The main contribution of this paper is the addition of temporal consistency to the OSF algorithm to stabilise scene flow estimation so that the same independently moving objects are detected more often through the sequence of images. We experimentally verify the proposed improvements on the KITTI'15 multi-frame vision benchmark. These improvements reduce the number of missed vehicles by 42%. Moreover, they also reduce the erroneous pixel percentage from 10.63% to 9.65% of estimated scene flow.

2. Related Work

The three-dimensional scene flow aims to recover dense or semi-dense 3D geometry and 3D motion. Vedula et al. introduced the concept of scene flow [21] as a three-dimensional vector field describing the motion of each three-dimensional point on each surface in a scene. We can look at scene flow as a combination of dense stereo reconstruction and optical flow estimation, which are both challenging problems themselves. The scene flow can be used as input for a lot of high-level application as e.g. obstacle detection or prediction collision.

Scene flow is computed using consecutive video frames from calibrated stereoscopic cameras. Similarly to the optical flow methods, approaches for scene flow estimation are often based on variational methods [21, 1, 8, 22, 26], feature matching [13], or scene flow representation using stixels [17].

The algorithms for scene flow must cope with the

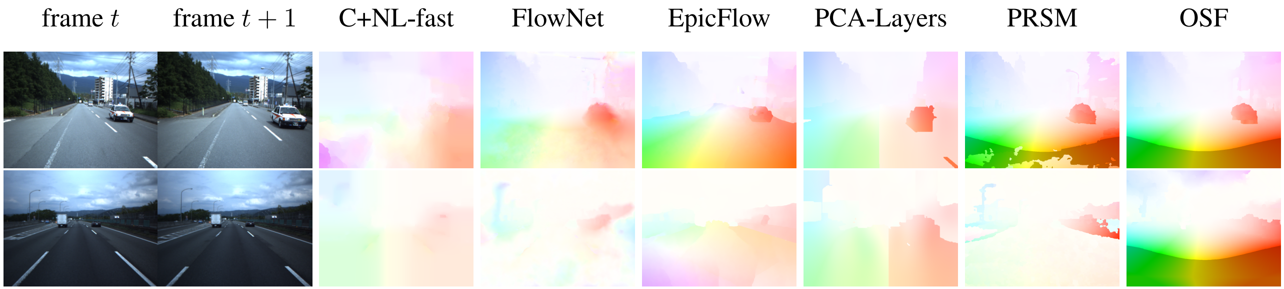


Figure 1. State of the art optical flow estimation results on our internal dataset. Compared optical flow methods: C+NL-fast [20], FlowNet [5], EpicFlow [19], PCA-Layers [27] and scene flow methods PRSM [24], OSF [14]

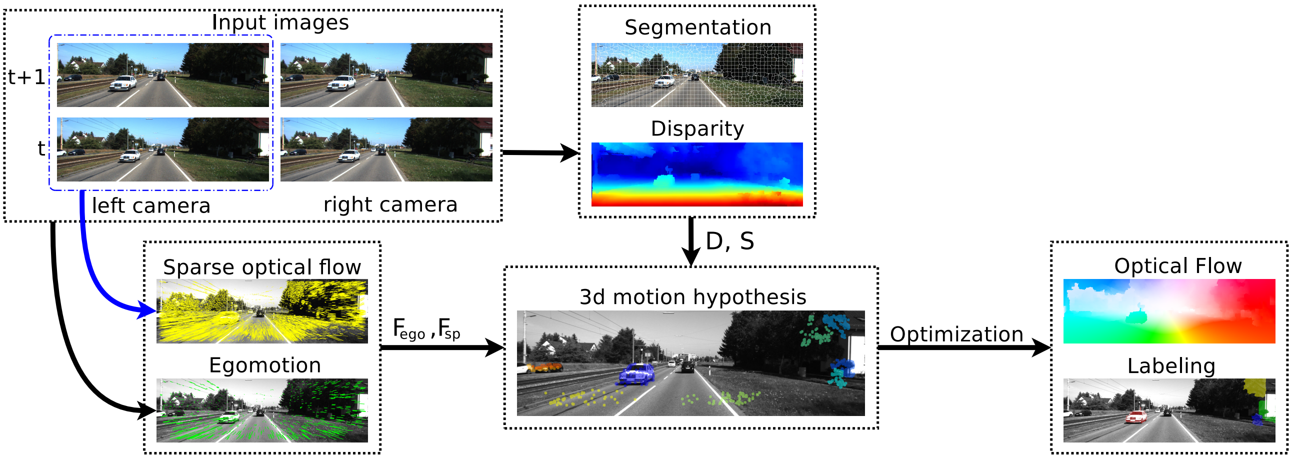


Figure 2. Overview of the Object Scene Flow algorithm [14]. Superpixel segmentation S, disparity D, sparse-flow F_{sp} and ego-motion F_{ego} of the camera are estimated from input stereo images. Then the independent motion candidates are estimated. Labels of motion candidates to proper segments are assigned during optimization. Finally, the output in the form of scene flow and label map are computed from motion candidates and segments parameters using MP-PBP [16] and TRW-S [10].

same problems as algorithms for optical flow and disparity estimation i.e. occlusions, large displacements or radiometric challenges [11, 14]. Although more information is avaible for the scene flow estimation (more cameras), there are also more parameters to estimate. Recently, many successful methods [23, 28, 24, 14, 13] started using small planar patches for a description of the scene instead of simple pixel-wise representations [8, 21]. Segmentation of the scene into rigid planar regions increases robustness and decreases the number of parameters which must be estimated [23].

The Object Scene Flow (OSF) [14] further introduces an idea that scene flow can be composed from a small number of independent motions. This assumption leads to a strong regularisation for scene flow computation and results into a more accurate estimation. Each independent motion is further restricted spatially, allowing independently moving object segmentation.

Since Murray and Buxton [15], various approaches using temporal consistency have been proposed for optical and scene flow. Some of them rely on smoothness assumption of trajectory over multiple frames. A spatio-temporal smoothness term for the optical flow was proposed in [15]. However, the algorithm does not work well for large displacements. Volz at al. [25] proposed adaptive trajectory regularisation over five consecutive frames. Motion fields of all frames are parametrised with respect to the central reference frame.

The others use some kind of tracking. Devernay at al. [4] used tracking of 3D points and surfels (small planar square regions) for temporally consist scene flow estimation also proposed extension of [12] using multiple cameras. Rabe at al. [18] used extended Kalman filter for tracking, but instead of tracking matched features they tracked dense scene flow computed by [26]. Although the algorithm is real-time, its use is rather limited, since it is not able to handle with fast motions. Using robustly linked frames, Hung at al. [9] proposed optical flow and stereo estimation from long-temporal motion trajectories but algorithm needs the whole sequence for the computation, thereby it is inappropriate for online estimation.

Recently, Vogel at al. [23] achieved temporal coherence using sliding temporal windows for their both viewpoints and multi-frames consistent model. However, the method does not produce independent motion segmentation like OSF, which is not only desirable as an output but functions also as a strong regularisation for the scene flow estimation.

3. Object Scene Flow Algorithm

The OSF decomposes each dynamic scene using a small number (hundreds) of 3D planar patches. The algorithm assumes that each patch belongs to one of a few indentedly moving objects, each with its own rigid motion (6 degrees of freedom). Each of the planar patch is parametrised by four variables: Three of them for the plane parameters and one for a label index. Each label corresponds to an object motion. Further, it is assumed that the set of independently moving objects is small (up to ten). Scene flow estimation is solved as a labelling problem, where each of the planar patches is assigned to one of the rigid body motions using a discrete-continuous CRF. The CRFs objective is defined as a weighted sum of unary and pairwise terms computed from disparity, superpixels, sparse optical flow and motion candidates.

The structure of the OSF algorithm is shown in Figure 2. The input to the algorithm are two consecutive stereo frames. The left image at time t is used as the reference image. The superpixels and the initial disparity is computed by StereoSLIC [28] and SGM [7]. All reference view pixels are segmented into superpixels. The superpixels and the initial disparity are computed by StereoSLIC [28] and SGM [7]. Each superpixel is assigned with its patch. The planar patch plane is computed by fitting a plane to the corresponding disparity measurements. The camera ego-motion is computed [6] since is assumed that dominant motion of the scene is induced by the motion of the camera.

The rigid body motion hypotheses of independently moving objects are computed next. First, the ego-motion outliers are found used as an input to a sequential RANSAC which greedily produces hypotheses. Finally, the CRF is solved, and the planar patches are assigned with the motion hypotheses as mentioned above. The OSF uses max-product particle belief propagation [16] and tree-reweighted message passing [10] for the optimization. Details can be found in the original paper [14]. The estimated dense scene flow is computed from the planar patches parameters.

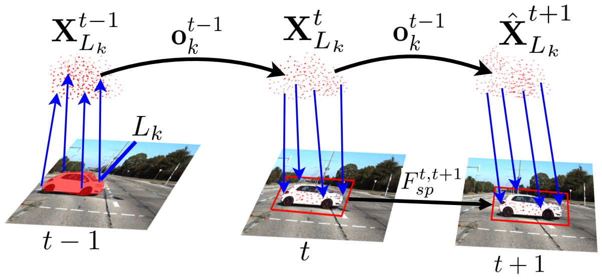


Figure 3. Scheme of object motion labels propagation. Dense 3D point cloud $\mathbf{X}^{t-1}_{L_k}$ is computed from segments assigned with motion $\mathbf{o}^{t-1}_k$. 3D points $\mathbf{X}^{t-1}_{L_k}$ are transformed by $\mathbf{o}^{t-1}_k$ to the frame t and $t+1$ as $\mathbf{X}^{t}_{L_k}$ and $\hat{\mathbf{X}}^{t+1}_{L_k}$ respectively. Sparse correspondences $F^{t,t+1}_{sp}$ with larger density for motion estimation are computed in the appropriated areas where $\hat{\mathbf{X}}^{t}_{L_k}$ and $\hat{\mathbf{X}}^{t+1}_{L_k}$ are reprojected.

4. Object Scene Flow with Temporal Consistency

In this section, we present our extensions to the OSF algorithm. We focus on the temporal consistency of the independently moving objects. We will discuss benefits of individual improvements in Section 5.

4.1. Notation

The OSF algorithm decomposes a dynamic scene into a set of 3D planar patches $\mathbf{s}_i = (\mathbf{n}_i, l_i)$, where $\mathbf{n}_i$ is a normal of the plane, l_i is a label of 3D motion $l_i \in \{1, \ldots, |O|\}$ and O is a set of a few independent motions. Each 3D motion $\mathbf{o}_k \in O$ is parametrised by rotation $\mathbf{R}_k \in \mathrm{SO}(3)$ and translation $\mathbf{t}_k \in \mathbb{R}^3$. Each plane normal $\mathbf{n}_i$ is computed from a superpixel $i \in S$, where S is a set of superpixels in the reference frame, by robustly fitting a plane to the depth values obtained from the disparity. The plane parameters provide the mapping between 3D points $\mathbf{X}_i = [x_i, y_i, z_i]^{\mathrm{T}}$ and its corresponding 2D points $\mathbf{x}_i = [u_i, v_i]^{\mathrm{T}}$. The frame t is considered to be the reference frame, $t+1$ is the next frame and $t-1$ is the previous frame etc.

4.2. Object Motion Labels Propagation

As noted above, the OSF does not use any information from the previous stereo image pairs. We expect that adding temporal consistency will lead to a more accurate scene flow estimation. We also expect that some objects that are missed by standard

OSF will be detected thanks to the temporal consistency. Finally, the object labels should become stable throughout the sequence.

Assuming constant velocity, we propagate estimated motion parameters $\mathbf{o}_k^{t-1}$ from the frame $t-1$ and use them for estimation of $\mathbf{o}_k^t$ in frame t. However, we do not simply use $\mathbf{o}_k^{t-1}$ as a motion candidate at frame t. Instead, we use disparity and sparse correspondences between the frames t and $t+1$ to find a good candidate motion $\mathbf{o}_k^t$ as it is shown in Figure 3.

Let $L_k = \{i; l_i = k\}$ be a set of indices of all planar patches $\mathbf{s}_i$ sharing the same motion $\mathbf{o}_k^{t-1}$ and let $\mathbf{X}_{L_k}^t$ be a set of all 3D points associated to all segments from L_k. We get a set of 3D points $\hat{\mathbf{X}}_{L_k}^{t+1}$ using a constant motion assumption.

$$\hat{\mathbf{X}}_{L_k}^{t+1} \simeq \mathbf{R}_k^{t-1}\mathbf{X}_{L_k}^t + \mathbf{t}_k^{t-1} \qquad (1)$$

Since the constant motion assumption is only approximately valid, we use 3D points $\mathbf{X}_{L_k}^t$ and $\hat{\mathbf{X}}_{L_k}^{t+1}$ only to estimate the bounding box of expected object location. The actual positions of 3D points are then estimated from disparity and sparse correspondences in the following way. We re-project the 3D points $\mathbf{X}_{L_k}^t$ and $\hat{\mathbf{X}}_{L_k}^{t+1}$ back to the image plane as 2D points $\hat{\mathbf{x}}_{L_k}^t$ and $\hat{\mathbf{x}}_{L_k}^{t+1}$, respectively. We compute sparse flow $F_{sp}^{t,t+1}$ correspondences [6] between frames t and $t+1$, with larger density than in the original OSF (five times in our case). These correspondences are computed in the image area bordered with the smallest rectangular bounding box containing all reprojected points $\hat{\mathbf{x}}_{L_k}^t$ and $\hat{\mathbf{x}}_{L_k}^{t+1}$, respectively. We enlarge the bounding box by 20 pixels at each side to increase robustness. For all computed correspondences, we estimate their corresponding 3D points $\mathbf{X}_{F_{sp}}^t, \mathbf{X}_{F_{sp}}^{t+1}$ using stereo camera calibration and estimated disparity.

To remove obvious outliers, we remove all points $\mathbf{X}_{F_{sp}}^{t+1}$ (with their $\mathbf{X}_{F_{sp}}^t$ correspondences) which are further away from the $median(\hat{\mathbf{X}}_{L_k}^{t+1})$ than a threshold θ_{sp}[1]. We also remove all correspondences which have similar motion as the camera ego-motion. Motion hypothesis candidate $\mathbf{o}_k^t = \left(\mathbf{R}_k^t, \mathbf{t}_k^t\right)$ is estimated on the remaining correspondences by RANSAC. We propagate every object motion $\mathbf{o}_k^{t-1}$ except the ego-motion.

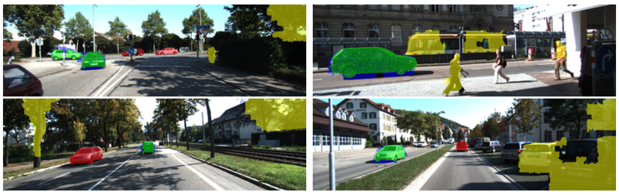


Figure 4. Evaluation of independently moving objects labelling from Object Scene Flow algorithm [14] on the KITTI'15 dataset. Missed vehicles are coloured in red; correctly detected vehicles are in green, and falsely detected vehicles are coloured in yellow. As false positive detection are considered also moving objects e.g. cyclists, trucks, persons, etc. since foreground ground-truth contains only moving cars.

4.3. Ego-motion Outlier Redefinition

This motion hypothesis propagation has a positive effect on the error of the estimated scene flow and decreases the number of missed vehicles. However, the label propagation also increases false positive detection of moving vehicles (Table 2). This is caused mostly by propagating additional false positive detection from previous frames as could be seen in Figure 4. The most of them are at the sides of the images as is shown in the Figure 5.

The OSF algorithm finds 3D motion hypotheses as ego-motion outliers in sparse flow correspondences. A correspondence is considered as ego-motion outlier when its end-point-error $E_{epe}(u,v)$ is greater than a fixed threshold (2 px) for all (u,v) where F_{sp} is defined.

Figure 5 shows the ego-motion outliers of the original approach (labelled with red colour). It could be observed that the fixed threshold works well at medium flow magnitudes but worse at the boundary of the images where the optical flow is larger and a small disparity error causes significant EPE. To eliminate this effect, we propose to use a dynamic threshold which depends on the motion magnitude. Correspondence in the image point (u,v) is labelled as ego-motion outlier if

$$(E_{epe}(u,v))^2 \geq \max\left(\left\|F_{ego}^{u,v}\right\|_2, \theta_{min}\right), \qquad (2)$$

where $\theta_{min} = \sqrt{2}$ is a minimal optical flow threshold to increase robustness.

Application of this change is shown in Figure 5. The false ego-motion outliers at image edges and the true estimated outliers are found disappear on the distant vehicles.

[1] θ_{sp} = 3 m; Similar process as is used in [14]

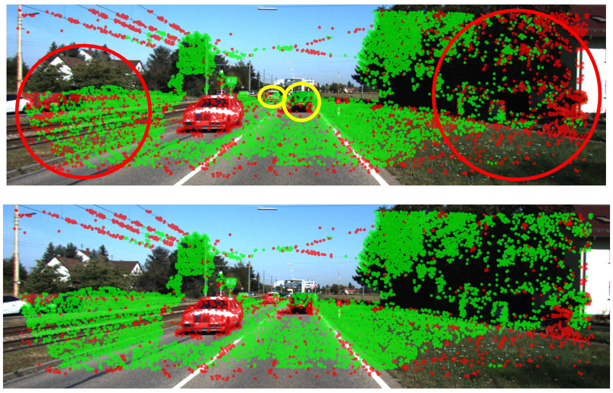


Figure 5. Demonstration of ego-motion outlier redefinition. Green color marks ego-motion inliers and red color ego-motion outliers. (top) original approach and (bottom) proposed approach. The most significant difference is on the sides of the images, where lots of false-positive ego-motion outliers disappeared. Red ellipses mark areas with significant number of false-positives ego-motion outliers and yellow ellipses mark false-negative ego-motion outliers.

4.4. Robust Motion Hypotheses Generation

Examining further the results, most of the remaining errors are caused by the random nature of the algorithm. Depending on the initialisation, we observed a high variance in its output. Due to inaccurate matches, this approach of multi-instance model fitting could produce imprecise models. Inaccurate models hypotheses are then discarded during labelling and optimization step of the OSF algorithm. Particle filtering in the optimization loop should fix the inaccuracy of the models, nevertheless this works only for small deviations.

Figure 6 compares the best and the worst case from 10 randomly initialised runs of the algorithm on the same input data. Possible reason of the poor hypotheses could be a small number of correspondences and inaccurate estimated disparity of which are computed 3D points as an input to the RANSAC. A small error at disparity leads to a significant error of estimated model.

To alleviate these problems we decided to use LO-RANSAC [3] to generate motion hypotheses. Because of its local optimization step it tends to produce more precise motion hypotheses. As we will demonstrate in Section 5, the number of missing cars is the lowest from all tested combinations. The variance of the algorithm is also reduced.

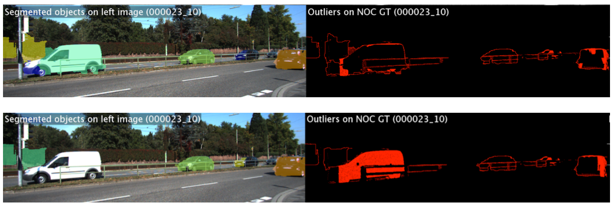

Segmented objects on left image (000023_10)
Outliers on NOC GT (000023_10)
Segmented objects on left image (000023_10)
Outliers on NOC GT (000023_10)

Figure 6. Demonstration of OSF variance. (top) shows the best result and (bottom) shows the worst result on a random KITTI'15 sequence with ten randomly initialised computations. Left images show the final labels of independently moving objects (background not shown) and right images represent EPE of the found optical flow (red color for EPE $\geq$ 3px).

	$\overline{\text{FP}}$	σ_{FP}	$\overline{\text{FN}}$	σ_{FN}
OSF [14]	236.6	14.7	170.4	3.3
+ label propagation	276.1	11.7	153.5	8.4
+ dynamic outliers	**219.6**	8.3	142.7	3.0
+ LO-RANSAC (3 stereo pairs)	244.4	12.4	125.3	2.9
+ LO-RANSAC (5 stereo pairs)	235.3	9.3	**121.0**	3.6
+ LO-RANSAC (12 stereo pairs)	236.3	17.0	123.2	5.4

Table 2. Comparison of detection results of moving vehicles. Tested on the KITTI'15 training multiview dataset. We run listed algorithms algoritm 5 time for each sequence and each extension. $\overline{\text{FP}}$ and $\overline{\text{FN}}$ denote mean of false positive (wrong detection) and mean of false negative (missed detection) of moving vehicles respectively. In addition the standard deviations σ_{FP} and σ_{FN} are shown for better comparison.

5. Experiments

In this section, we demonstrate the results of the proposed extensions. For evaluation, we used the standard KITTI'15 benchmark [14]. The benchmark contains stereo camera sequences with large displacements and nontrivial environment conditions.

5.1. Evaluation protocol

To evaluate the scene flow, optical flow, and disparity, we use the standard KITTI'15 metric – a percentage of erroneous pixels. Pixels are considered erroneous when the end-point-error exceeds 3 pixel and 5% according to ground truth. We also report the number of undetected moving vehicles – false negatives (FN), and the number of falsely detected vehicles – false positives (FP). We label object O_k as true positive if

$$\frac{|L^{\mathrm{GT}_m} \cap L^{O_k}|}{|L^{\mathrm{GT}_m} \cup L^{O_k}|} \geq 0.5, \tag{3}$$

where L^{GT_m} is the set of pixels of the mth moving vehicle marked in the ground truth and L^{O_k} is a set of pixels labelled as k-th object by the proposed al-

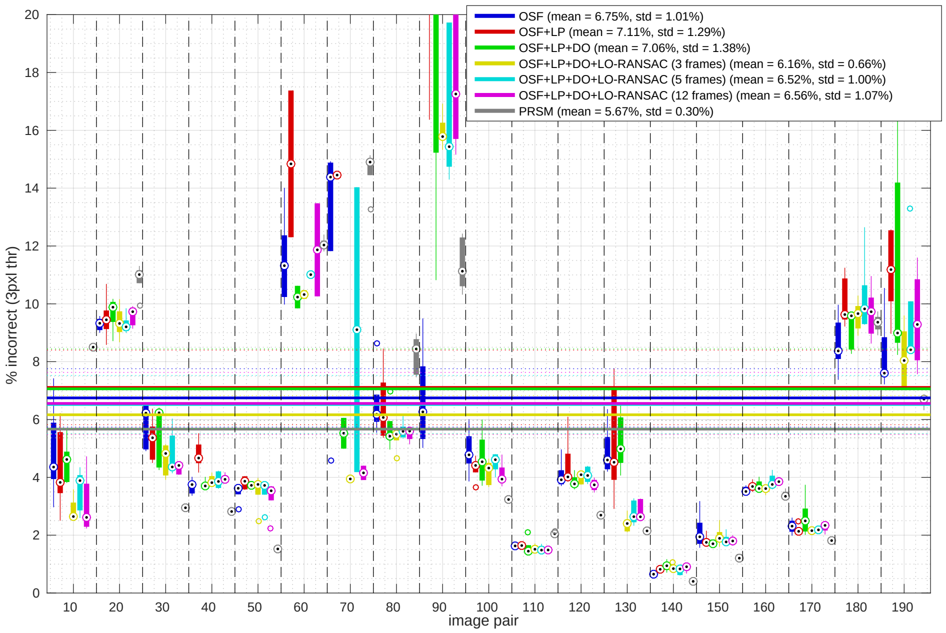


Figure 7. Evaluation across combination of proposed improvements and OSF on all sequences from KITTI'15 training dataset (bargraphs only for every 10th, mean and variance for all 200 sequences). Results show huge variace of OSF (and OSF with proposed improvements) during 5 computations on the same data.

Alg.\Error	D1-bg	D1-fg	D1-all	D2-bg	D2-fg	D2-all	Fl-bg	Fl-fg	Fl-all	SF-bg	SF-fg	SF-all
PRSF [23]	4.74 %	13.74 %	6.24 %	11.14 %	20.47 %	12.69 %	11.73 %	27.73 %	14.39 %	13.49 %	33.72 %	16.85 %
CSF [13]	4.57 %	13.04 %	5.98 %	7.92 %	20.76 %	10.06 %	10.40 %	30.33 %	13.71 %	12.21 %	36.97 %	16.33 %
OSF [14]	4.54 %	12.03 %	5.79 %	5.45 %	19.41 %	7.77 %	5.62 %	22.17 %	8.37 %	7.01 %	28.76 %	10.63 %
PRSM [24]	**3.02** %	10.52 %	**4.27** %	**5.13** %	**15.11** %	**6.79** %	**5.33** %	17.02 %	**7.28** %	**6.61** %	23.60 %	**9.44** %
OSF+TC (ours)	4.11 %	**9.64** %	5.03 %	5.18 %	15.12 %	6.84 %	5.76 %	**16.61** %	7.57 %	7.08 %	**22.55** %	9.65 %

Table 1. Quantitave comparison with the state-of-the-art results (all pixels). Columns mark categories of evalution: scene flow as **SF**, optical flow as **F**, disparity **D1** for the first frame of the test pair and **D2** for the second; and subcategories: evaluation over **bg** background regions, evaluation over **fg** foreground regions and **all** evaluation over all ground-truth. Categories where OSF+TC performs better than the original OSF are highlighted in grey and the best results are in bold.

gorithm.

5.2. Object Scene Flow Variance Analysis

As was mentioned above, we noticed that the OSF results vary significantly depending on the random seed initialization. To investigate this variance, we removed all fixed random generator seeds in all parts of the OSF algorithm and instead initialised all the seeds randomly for each computation. We then run the OSF algorithm 30 times for each sequence. Figure 8 shows variance of the OSF results. We noticed that sequences with large variance have difficult radiometric conditions or large displacements. We thus report mean and standard deviation of all our result for better comparison.

5.3. Evaluation of the Proposed Object Scene Flow Extensions

We compare results of our improvements according to various quantitative criteria. Results from the evaluation of vehicle detection are shown in Table 2. In Table 3 we show the results of erroneous pixels percentage for different scene flow estimation variants. Figure 7 shows results (mean and standard deviation) of the OSF algorithm and its extensions.

Object Motion Label Propagation. The motion hypotheses propagation influence is shown in Table 2. The number of undetected vehicles decreases. On the other hand, the number of false positive detections increases. Besides that we observe the increase of scene flow error as shown in Table 3. As discussed above, this is connected with the propagation of false positives in time.

Ego-motion Outlier Redefinition. Next, we evaluate the ego-motion outlier re-definition. It helps to decrease the number of false positive detections as shown in Table 2. Also the number of false detections decreases. The scene flow is still worse than the original OSF but gets slightly better results compared to the previous experiment (Figure 7).

Alg.\Error	D1-bg	D1-fg	D1-all	D2-bg	D2-fg	D2-all	Fl-bg	Fl-fg	Fl-all	SF-bg	SF-fg	SF-all
OSF [14]	3.77 %	7.30 %	4.45 %	**4.22** %	12.41 %	5.60 %	**4.45** %	20.36 %	7.04 %	**5.48** %	22.95 %	8.20 %
+ label propagation	3.85 %	6.78 %	4.52 %	4.38 %	10.34 %	5.65 %	4.68 %	18.90 %	7.44 %	5.73 %	21.41 %	8.51 %
+ dynamic outliers	**3.74** %	6.50 %	4.39 %	4.37 %	9.57 %	5.56 %	4.75 %	17.77 %	7.34 %	5.78 %	20.43 %	8.45 %
+ LO-RANSAC (3 stereo pairs)	**3.78** %	**6.11** %	**4.33** %	4.41 %	**8.49** %	**5.30** %	4.79 %	**12.74** %	**6.37** %	5.82 %	**15.21** %	**7.52** %
+ LO-RANSAC (5 stereo pairs)	3.81 %	6.33 %	4.42 %	4.43 %	9.04 %	5.51 %	4.81 %	14.05 %	6.79 %	5.86 %	16.54 %	7.91 %
+ LO-RANSAC (12 stereo pairs)	3.81 %	6.28 %	4.42 %	4.44 %	8.76 %	5.47 %	4.78 %	14.01 %	6.80 %	5.82 %	16.52 %	7.92 %

Table 3. Scene flow evaluation of proposed improvements. Tested on the KITTI'15 training multiview dataset. Columns marks categories of evalution: scene flow as **SF**, optical flow as **F**, disparity **D1** for the first frame of the test pair and **D2** for the second; and subcategories: evaluation over **bg** background regions, evaluation over **fg** foreground regions and **all** evaluation over all ground-truth.

Alg.\Error	D1-bg	D1-fg	D1-all	D2-bg	D2-fg	D2-all	Fl-bg	Fl-fg	Fl-all	SF-bg	SF-fg	SF-all
PRSF [23]	4.41 %	13.09 %	5.84 %	6.35 %	16.12 %	8.10 %	6.94 %	23.64 %	9.97 %	8.35 %	28.45 %	11.95 %
CSF [13]	4.03 %	11.82 %	5.32 %	6.39 %	16.75 %	8.25 %	8.72 %	26.98 %	12.03 %	10.26 %	32.58 %	14.26 %
OSF [14]	4.14 %	11.12 %	5.29 %	4.49 %	16.33 %	6.61 %	**4.21** %	18.65 %	6.83 %	**5.52** %	24.58 %	8.93 %
PRSM [24]	**2.93** %	10.00 %	**4.10** %	**4.13** %	12.85 %	5.69 %	4.33 %	14.15 %	6.11 %	5.54 %	20.16 %	8.16 %
OSF+TC (ours)	3.79 %	**8.66** %	4.59 %	4.18 %	**12.06** %	**5.59** %	4.34 %	**12.86** %	**5.89** %	**5.52** %	**18.02** %	**7.76** %

Table 4. Quantitave comparison with state-of-the-art results (non-occluded pixels). Columns marks categories of evalution: scene flow as **SF**, optical flow as **F**, disparity **D1** for the first frame of the test pair and **D2** for the second; and subcategories: evaluation over **bg** background regions, evaluation over **fg** foreground regions and **all** evaluation over all ground-truth. Categories where OSF+TC performs better than the original OSF are highlighted in grey and the best results are in bold.

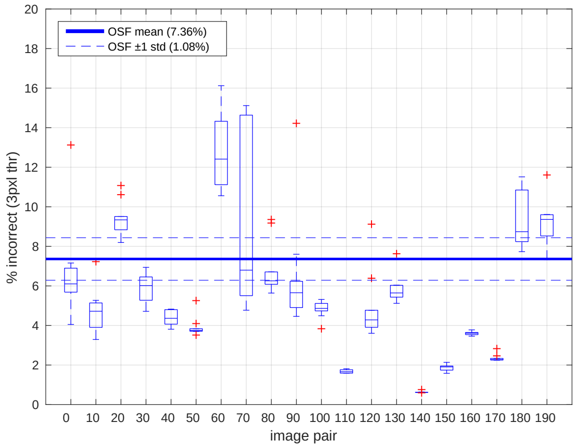


Figure 8. Variance of the OSF algorithm (bargraphs only for every 10th, mean and variance over all 200 sequences). The central line inside each box indicates the median. The bottom of the box refers the 25th percentile and the top of the box refers to the 75th percentiles, respectively. Outliers are marked with the red symbol '+'.

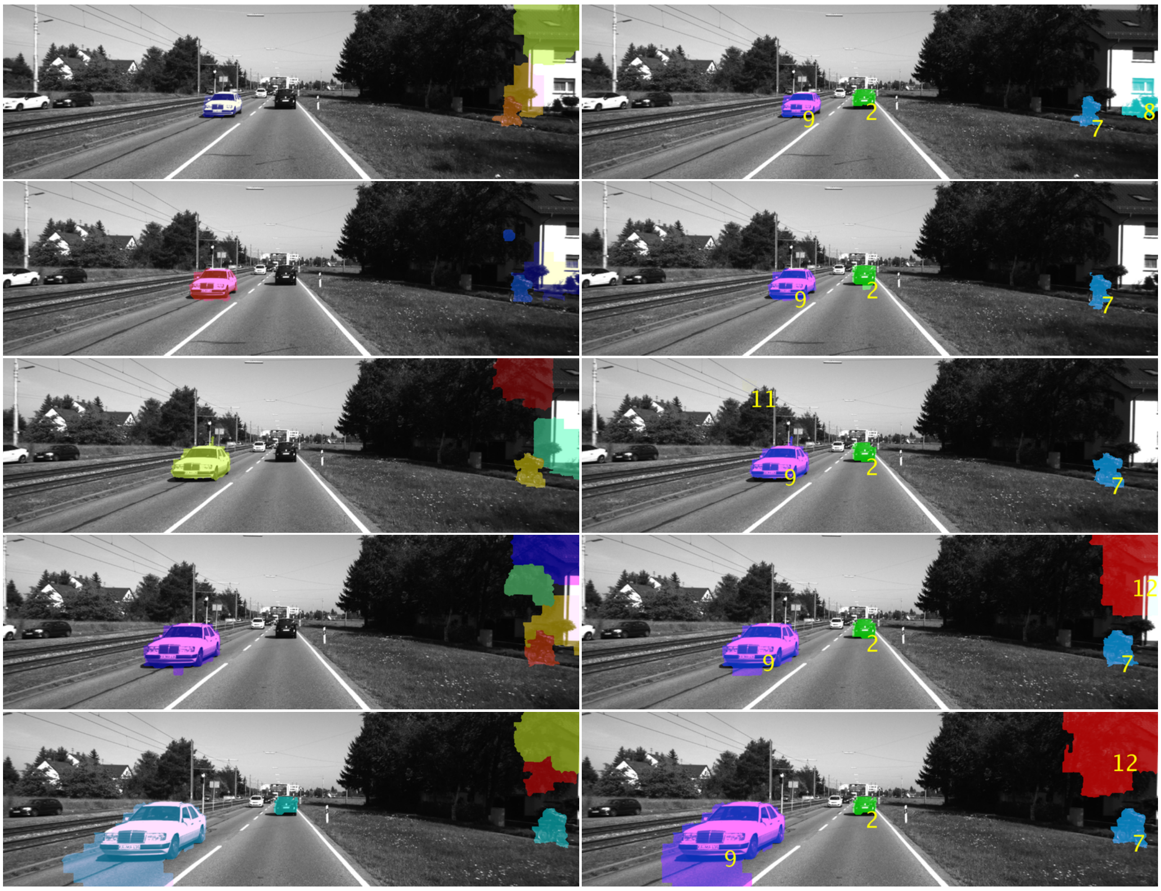
Robust Motion Hypotheses. The additional application of LO-RANSAC in the motion hypotheses estimation leads to a significant decrease of the scene flow error from 8.2% to 7.52% (Tab. 3) compared to the original OSF algorithm. Besides, the number of false negatives also decreases. However, the number of the false positive increases (Tab. 2). In the case of application of all extensions, we try a different number of frames as input to the temporally consistent OSF.

We run experiments for 3, 5 and 12 frames. Using more frames reduces FPs and FNs as shown in Table 2, however the scene flow results degrade a bit (Table 3, Figure 7). This effect is most likely caused by the different density of ground-truth in the KITTI dataset (foreground is about 4× denser). Every super-pixel falling on both fg and bg is more likely to be removed from the motion hypothesis when propagated longer. This nibbling of the car borders however causes higher fg scene-flow errors as shown in Table 3. We believe that this effect could be mitigated by adding temporal consistency also to the super-pixels [2], but leave it as a future work.

Summary. Based on the results, the propagation through three frames was chosen for further comparison with the state of the art. The level of false positives and false negatives is similar to other variants, but the scene flow errors are significantly lower. The method is termed OSF+TC in the comparisons. Figure 8 shows the mean and standard deviation of the obtained results over all sequences.

5.4. Comparison with the State of the Art

We compare the best combination of all proposed extensions (using three stereoscopical frames temporal consistency) with the best-ranked KITTI'15 submissions in the scene flow category. Table 1 shows the results for evaluation on all pixels from ground-truth in the image frame. OSF+TC decreases EPE of the original OSF from 10.63% to 9.65% and achieves the second position in the scene flow estimation total. The loss to the first place (PRSM [24]) is less



(a) Original (b) Proposed

Figure 9. Propagation of moving object label through time. Three moving objects (id=2,7,9) in (b) have stable label over the whole sequence as opposed to the original approach in (a). Car 2 is detected earlier due to the stronger LO-RANSAC model estimation and is then correctly propagated. Object 7 is a man on a bicycle. Also many false positives are reduce due to the ego-motion outlier redefinition.

than a quarter percent. Moreover, OSF+TC ranked first for scene flow evaluation on non-occluded pixels as shown in Table 4, with an improvement to the original algorithm by 1%. Finally, OSF+TC achieves the first position for scene flow and optical flow over foreground regions for both, non-occluded pixel evaluation and all pixel evaluation.

6. Conclusion

We presented improvement of Object Scene Flow algorithm aiming to increase accuracy and robustness of individually moving objects hypotheses estimation. We demonstrated that the original algorithm was overcome with our improvements in the challenging KITTI’15 flow benchmark by 1% in the flow category. Moreover, we stabilised the labelling of moving objects and reduced the number of non found objects to half compared to [14].

Acknowledgements

The research was supported by Toyota Motor Europe HS and by CTU student grant SGS15/155/OHK3/2T/13

References

[1] T. Basha, Y. Moses, and N. Kiryati. Multi-view scene flow estimation: A view centered variational approach. *International journal of computer vision*, 101(1):6–21, 2013. 1

[2] J. Chang, D. Wei, and J. W. Fisher. A video representation using temporal superpixels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2051–2058, 2013. 7

[3] O. Chum, J. Matas, and J. Kittler. *Locally Optimized RANSAC*, pages 236–243. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. 5

[4] F. Devernay, D. Mateus, and M. Guilbert. Multi-camera scene flow by tracking 3-D points and surfels. In *2006 IEEE Computer Society Confer-*

ence on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 2203–2212, 2006. 2

[5] A. Dosovitskiy, P. Fischery, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, Dec. 2015. 2

[6] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 963–968. IEEE, 2011. 3, 4

[7] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005. 3

[8] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *2007 IEEE 11^{th} International Conference on Computer Vision*, pages 1–7, Oct. 2007. 1, 2

[9] C. H. Hung, L. Xu, and J. Jia. Consistent binocular depth and scene flow with chained temporal profiles. *International Journal of Computer Vision*, 102(1):271–292, 2013. 2

[10] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1568–1583, 2006. 2, 3

[11] D. Kondermann, R. Nair, K. Honauer, K. Krispin, J. Andrulis, A. Brock, B. Gussefeld, M. Rahimimoghaddam, S. Hofmann, C. Brenner, and B. Jahne. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016. 2

[12] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981. 2

[13] Z. Lv, C. Beall, P. F. Alcantarilla, F. Li, Z. Kira, and F. Dellaert. A continuous optimization approach for efficient and accurate scene flow. *CoRR*, abs/1607.07983, 2016. 1, 2, 6, 7

[14] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[15] D. W. Murray and B. F. Buxton. Scene segmentation from visual motion using global optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(2):220–228, Mar. 1987. 2

[16] J. Pacheco, S. Zuffi, M. J. Black, and E. B. Sudderth. Preserving modes and messages via diverse particle selection. In *ICML*, pages 1152–1160, 2014. 2, 3

[17] D. Pfeiffer and U. Franke. Efficient representation of traffic scenes by means of dynamic stixels. In *2010 IEEE Intelligent Vehicles Symposium*, pages 217–224, June 2010. 1

[18] C. Rabe, T. Müller, A. Wedel, and U. Franke. *Dense, Robust, and Accurate Motion Field Estimation from Stereo Image Sequences in Real-Time*, pages 582–595. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. 2

[19] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1164–1172, 2015. 2

[20] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014. 2

[21] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 722–7292, 1999. 1, 2

[22] C. Vogel, K. Schindler, and S. Roth. 3D scene flow estimation with a rigid motion prior. In *2011 International Conference on Computer Vision*, pages 1291–1298. IEEE, 2011. 1

[23] C. Vogel, K. Schindler, and S. Roth. Piecewise rigid scene flow. In *The IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013. 2, 3, 6, 7

[24] C. Vogel, K. Schindler, and S. Roth. 3D scene flow estimation with a piecewise rigid scene model. *International Journal of Computer Vision*, 115(1):1–28, 2015. 1, 2, 6, 7

[25] S. Volz, A. Bruhn, L. Valgaerts, and H. Zimmer. Modeling temporal coherence for optical flow. In *2011 International Conference on Computer Vision*, pages 1116–1123, Nov. 2011. 2

[26] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In *European conference on computer vision*, pages 739–751. Springer, 2008. 1, 2

[27] J. Wulff and M. J. Black. Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 120–130. IEEE, 2015. 2

[28] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *European Conference on Computer Vision*, pages 756–771. Springer, 2014. 2, 3