TWO-STAGE ADAPTIVE OBJECT SCENE FLOW USING HYBRID CNN-CRF MODEL

*Congcong Li, Haoyu Ma, Qingmin Liao**

Tsinghua University

licc18, hy-ma17@mails.tsinghua.edu.cn liaoqm@tsinghua.edu.cn

Abstract

Scene flow estimation based on stereo sequences is a comprehensive task relevant to disparity and optical flow. Some existing methods are time-consuming and often fail in the presence of reflective surfaces. In this paper, we propose a two-stage adaptive object scene flow estimation method using a hybrid CNN-CRF model (ACOSF), which benefits from high-quality features and the structured modelling capability. Meanwhile, in order to balance the computational efficiency and accuracy, we employ adaptive iteration for energy function optimization, which is flexible and efficient for various scenes. Besides, we utilize high-quality pixel selection to reduce the computation time with only a slight decrease in accuracy. Our method achieves competitive results with the state-of-the-art, which ranks second on the challenging KITTI 2015 scene flow benchmark.

Method

Our ACOSF mainly consists of two stages based on the hybrid CNN-CRF model. In the first stage, we use CNNs to obtain initial disparity and optical flow estimation. Then we integrate the initial results into a CRF-based model.

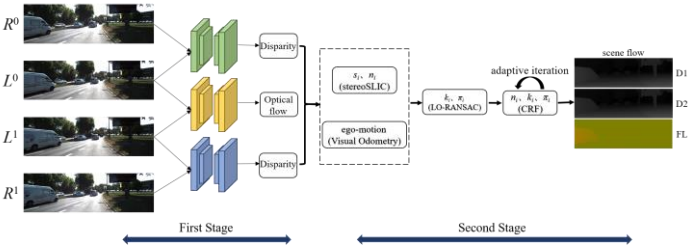

Fig. 1. The diagram of the proposed ACOSF scene flow estimation method.

➢ 3D Geometry and 2D Optical Flow Estimation

In the first stage, the convolutional neural networks are used to obtain the initial disparity and optical flow estimation, which are powerful to extract high-quality features for matching and searching correspondences.

➢ CRF Model for Scene Flow

In the second stage, we over-segment the reference image L^0. And we follow the assumption in [1] that there are a finite number of traffic participants moving rigidly. Each planar region B_i in the image is allocated to superpixel $s_i \in S$, which is described by a random variable $v_i = (n_i, k_i)^{\mathrm{T}}$. Each object O_k is associated with a variable $\pi_k \in SE(3)$ describing its rigid motion.

$$E(v,\pi) = \sum_{s_i \in S} \underbrace{\varphi_i(v_i,\pi)}_{\text{data}} + \sum_{s_i \sim s_j} \underbrace{\psi_{ij}(v_i,v_j)}_{\text{smoothness}}$$

➢ Efficiency

1) To balance the accuracy and computational efficiency, we make use of high-confidence matching obtained in the first stage. The algorithm samples a small number of pixels to construct the cost volume instead of all pixels in the region B_i.

2) Our ACOSF can dynamically adjust the number of iterations n in the MP-PBP to suit different scenarios by comparing the continuous variation of the energy function with the pre-set threshold T.

Experimental Results

➢ Comparison with OSF

Method	D1	D2	Fl	SF	Time
OSF [11]	4.97	6.34	6.54	7.73	56.18s
CNN-based	3.01	3.98	4.32	5.75	46.75s
+LO-RANSAC	2.92	3.63	4.13	5.48	48.13s
+adaptive iteration	2.98	3.72	4.22	5.63	39.41s

Table 1. Results on the validation portion of KITTI training set. It validates the improvements of the two-stage model over the original OSF algorithm.

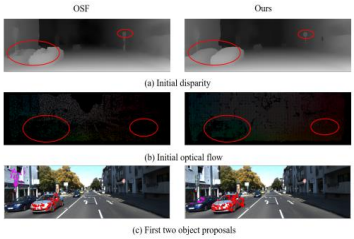

Fig. 2. The right row indicates better initial predictions for optimization. In row (c), we display the first two proposals of moving objects to reflect the improvements clearly.

➢ Comparison with state-of-the-art

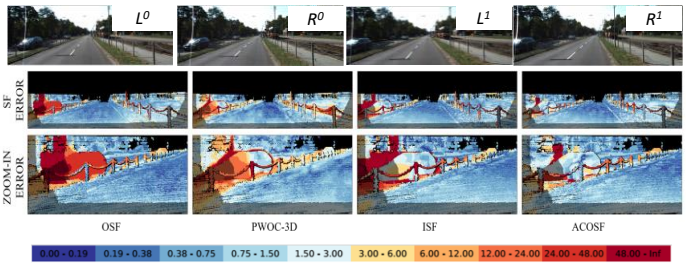

Fig. 3. Comparison of visual results on the KITTI test set .

Method	D1			D2			Fl			SF			Run Time
	bg	fg	all	bg	fg	all	bg	fg	all	bg	fg	all	
DRISF [33]	**2.16**	**4.49**	**2.55**	**2.90**	**9.73**	**4.04**	**3.59**	**10.40**	**4.73**	**4.39**	**15.94**	**6.31**	0.75s(G)
ACOSF(ours)	**2.79**	7.56	3.58	**3.82**	12.74	**5.31**	**4.56**	12.00	**5.79**	**5.61**	19.38	**7.90**	5min(C)
ISF [12]	4.12	6.17	4.46	4.88	11.34	5.95	5.40	**10.29**	6.22	6.58	**15.63**	8.08	10min(C)
PRSM* [29]	3.02	10.52	4.27	5.13	15.11	6.79	5.33	13.40	6.68	6.61	20.79	8.97	5min(C)
OSF+TC* [31]	4.11	9.64	5.03	5.18	15.12	6.84	5.76	13.31	7.02	7.08	20.03	9.23	50min(C)
SSF [32]	3.55	8.75	4.42	4.94	17.48	7.02	5.63	14.71	7.14	7.18	24.58	10.07	5min(C)
OSF [11]	4.54	12.03	5.79	5.45	19.41	7.77	5.62	18.92	7.83	7.01	26.34	10.23	50min(C)
DWARF [18]	3.20	**3.94**	**3.33**	6.21	**9.38**	6.73	9.80	13.37	10.39	11.72	18.06	12.78	0.14-1.43s(G)
PWOC-3D [34]	4.19	9.82	5.13	7.21	14.73	8.46	12.40	15.78	12.96	14.30	22.66	15.69	0.13s(G)
CSF [43]	4.57	13.04	5.98	7.92	20.76	10.06	10.40	25.78	12.96	12.21	33.21	15.71	80s(C)

Table 2. Results on KITTI 2015 scene flow dataset at the time of submission.

References

[1] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3061 3070.

[2] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5410–5418.

[3] J. Xu, R. Ranftl, and V. Koltun, "Accurate optical flow via direct cost volume processing," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1289–1297.

[4] W.-C. Ma, S. Wang, R. Hu, Y. Xiong, and R. Urtasun, "Deep rigid instance scene flow," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3614–3622.

[5] F. Aleotti, M. Poggi, F. Tosi, and S. Mattoccia, "Learning end-to-end scene flow by distilling single tasks knowledge," in Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.