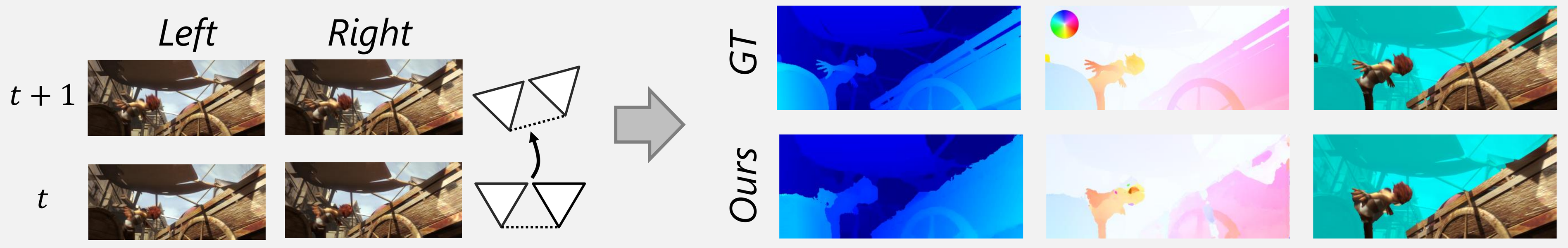
Fast Multi-frame Stereo Scene Flow with Motion Segmentation

Tatsunori Taniai (RIKEN AIP/Univ. of Tokyo) **Sudipta N. Sinha** (Microsoft Research) **Yoichi Sato** (Univ. of Tokyo)

CVPR July 21-26 Honolulu 2017

Introduction

Input — Output



A sequence of stereo image pairs recorded by a moving stereo camera rig.

Estimate stereo disparity, optical flow, and motion segmentation indicating moving object mask with 6 DoF camera ego-motion.

- We propose a stereo scene flow method that simultaneously recovers moving-object mask (motion segmentation) and camera ego-motion as well as disparity and optical flow maps.
- Our method takes 2 – 3 seconds to process each frame in the KITTI dataset using only CPU, which is 1 – 3 orders of magnitude faster than state-of-the-art methods.

Contributions

Unified framework where multiple tasks benefit from each other

- Optical flow: 2D flow motion for rigid background (rigid flow) is recovered parametrically using known depth and camera motion, reducing computational burden of general (non-rigid) optical flow.
- Stereo: Given camera motion, disparity at left-right occluded regions is improved via multi-view stereo on consecutive frames.
- Motion segmentation: The segmentation mask is a byproduct of our flow estimation that fuses non-rigid and rigid flow maps.
- Visual odometry: Camera motion estimates are recovered more robustly by utilizing the moving object mask information.

In contrast to existing joint methods

- We decompose the task into several simple optimization problems, rather than directly optimizing a single complex function.

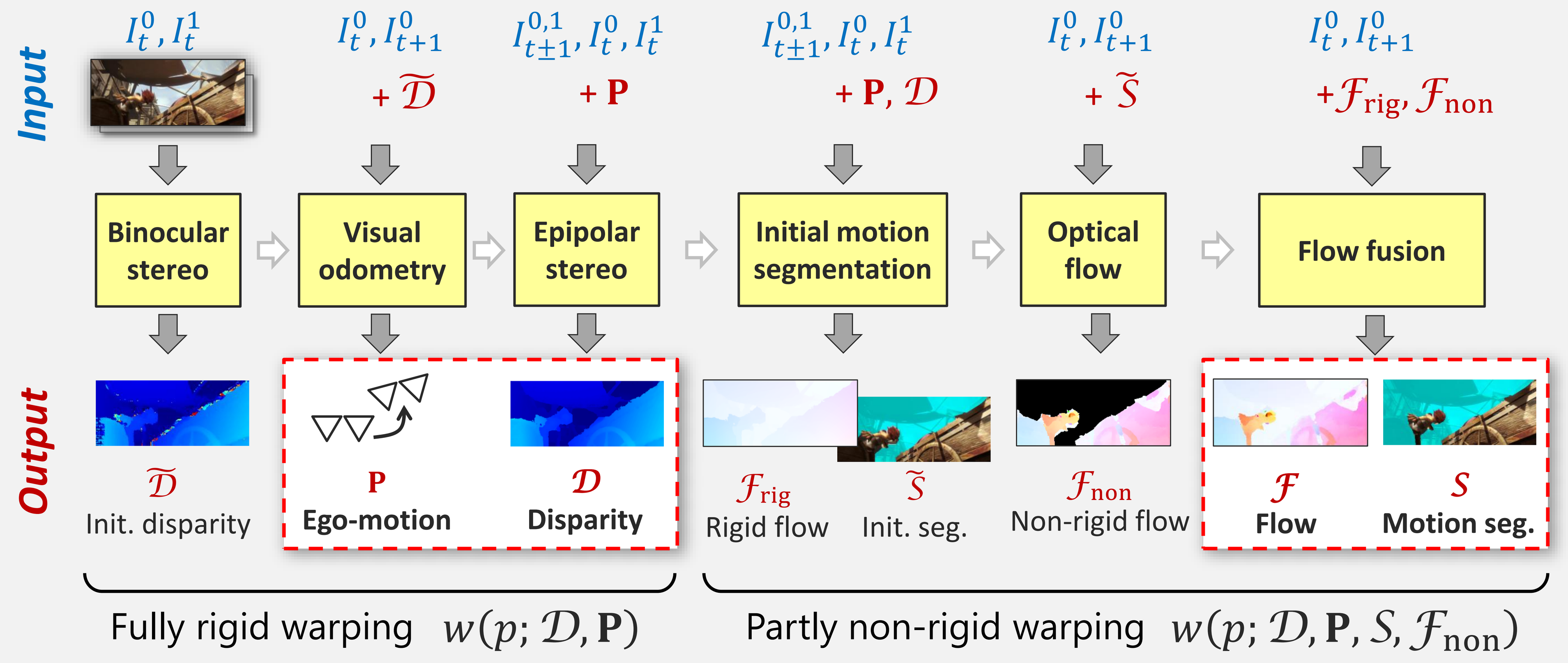
Multi-staged pipeline framework

We estimate disparity $\mathcal{D}$, camera motion $\mathbf{P}$, moving-object mask S, and moving-object flow $\mathcal{F}_{\mathrm{non}}$ (non-rigid flow) by implicitly minimizing image residual

$$E(\mathcal{D},\mathbf{P},S,\mathcal{F}_{\mathrm{non}}) = \sum_{\mathbf{p}} \left\| I_t^0(\mathbf{p}) - I_{t+1}^0(w(\mathbf{p};\mathcal{D},\mathbf{P},S,\mathcal{F}_{\mathrm{non}})) \right\|$$

using bimodal warping

$$w(\mathbf{p};\mathcal{D},\mathbf{P},S,\mathcal{F}_{\mathrm{non}}) = \begin{cases} \mathbf{p} + \mathcal{F}_{\mathrm{non}}(\mathbf{p}) & \text{if } S(\mathbf{p}) = \text{foreground} \\ \mathbf{p} + \mathcal{F}_{\mathrm{rig}}(\mathbf{p};\mathcal{D},\mathbf{P}) & \text{if } S(\mathbf{p}) = \text{background} \end{cases}$$



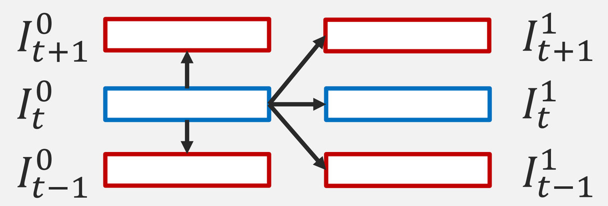
Binocular stereo uses SGM to get an initial disparity map.

Visual odometry estimates camera motion by minimizing

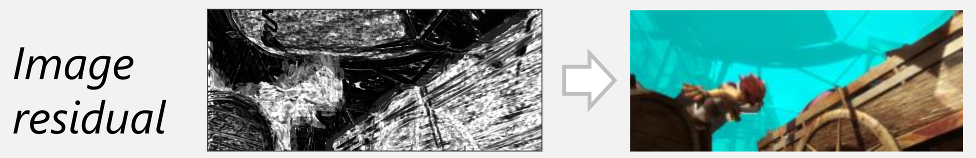
$$\min_{\mathbf{P}} \sum_{\mathbf{p}} w_{\mathbf{p}} \left\| I_t^0(\mathbf{p}) - I_{t+1}^0(w(\mathbf{p};\mathcal{D},\mathbf{P})) \right\|$$

We downweight moving object regions by $w_{\mathbf{p}}$ predicted by previous $\{S, \mathcal{F}_{\mathrm{non}}\}$.

Epipolar stereo refines disparity using temporarily adjacent frames. We blend left-right matching costs with matching costs for four adjacent frames.



Initial segmentation finds moving object regions. We use GrabCut with image residual as soft seeds for moving foreground.



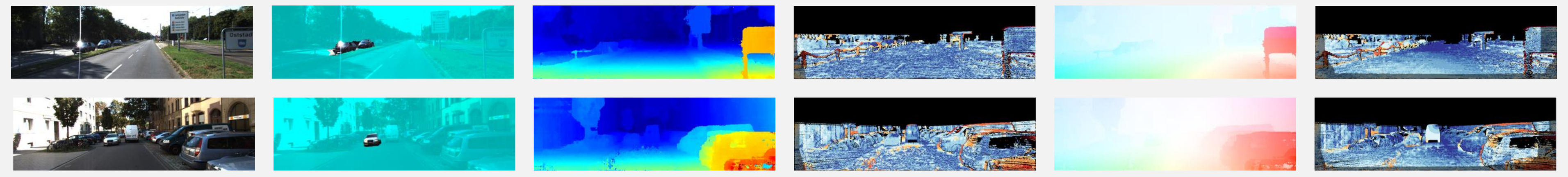
Optical flow estimates 2D flow map for only the predicted moving object regions. We use the SGM algorithm.

Flow fusion combines rigid and non-rigid flow proposals by a fusion move.

Experiments

KITTI 2015 stereo scene flow benchmark (in November 2016)

Rank	Method	D1-bg	D1-fg	D1-all	D2-bg	D2-fg	D2-all	Fl-bg	Fl-fg	Fl-all	SF-bg	SF-fg	SF-all	Time
1	PRSM [43]	**3.02**	**10.52**	**4.27**	**5.13**	**15.11**	**6.79**	**5.33**	**17.02**	**7.28**	**6.61**	**23.60**	**9.44**	300 s
2	OSF [30]	4.54	12.03	5.79	5.45	19.41	7.77	5.62	22.17	8.37	7.01	28.76	10.63	50 min
3	**FSF+MS** (ours)	5.72	11.84	6.74	7.57	21.28	9.85	8.48	29.62	12.00	11.17	37.40	15.54	2.7 s
4	CSF [28]	4.57	13.04	5.98	7.92	20.76	10.06	10.40	30.33	13.71	12.21	36.97	16.33	80 s
5	PR-Sceneflow [42]	4.74	13.74	6.24	11.14	20.47	12.69	11.73	27.73	14.39	13.49	33.72	16.85	150 s
8	PCOF + ACTF [10]	6.31	19.24	8.46	19.15	36.27	22.00	14.89	62.42	22.80	25.77	69.35	33.02	0.08 s (GPU)
12	GCSF [8]	11.64	27.11	14.21	32.94	35.77	33.41	47.38	45.08	47.00	52.92	59.11	53.95	2.4 s

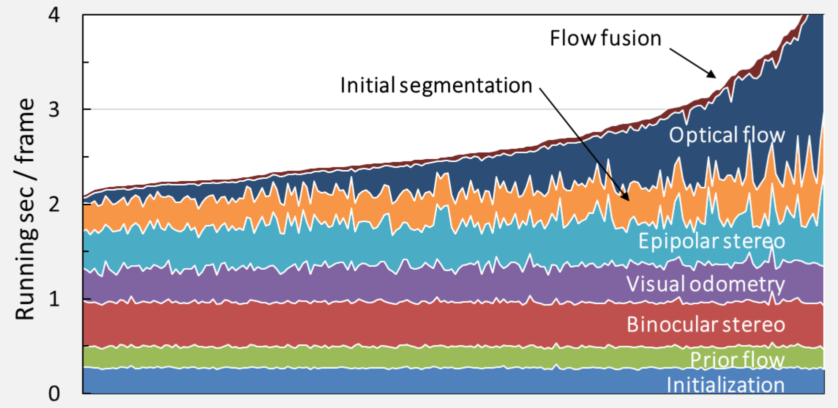


Reference — Motion segmentation — Disparity — Disparity error — Optical flow — Flow error

Improvements by epipolar stereo

	all pixels			non-occluded pixels		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
Binocular ($\tilde{\mathcal{D}}$)	7.96	12.61	8.68	7.09	10.57	7.61
Epipolar ($\mathcal{D}$)	**5.82**	**10.34**	**6.51**	**5.57**	**8.84**	**6.06**

Per-stage running times



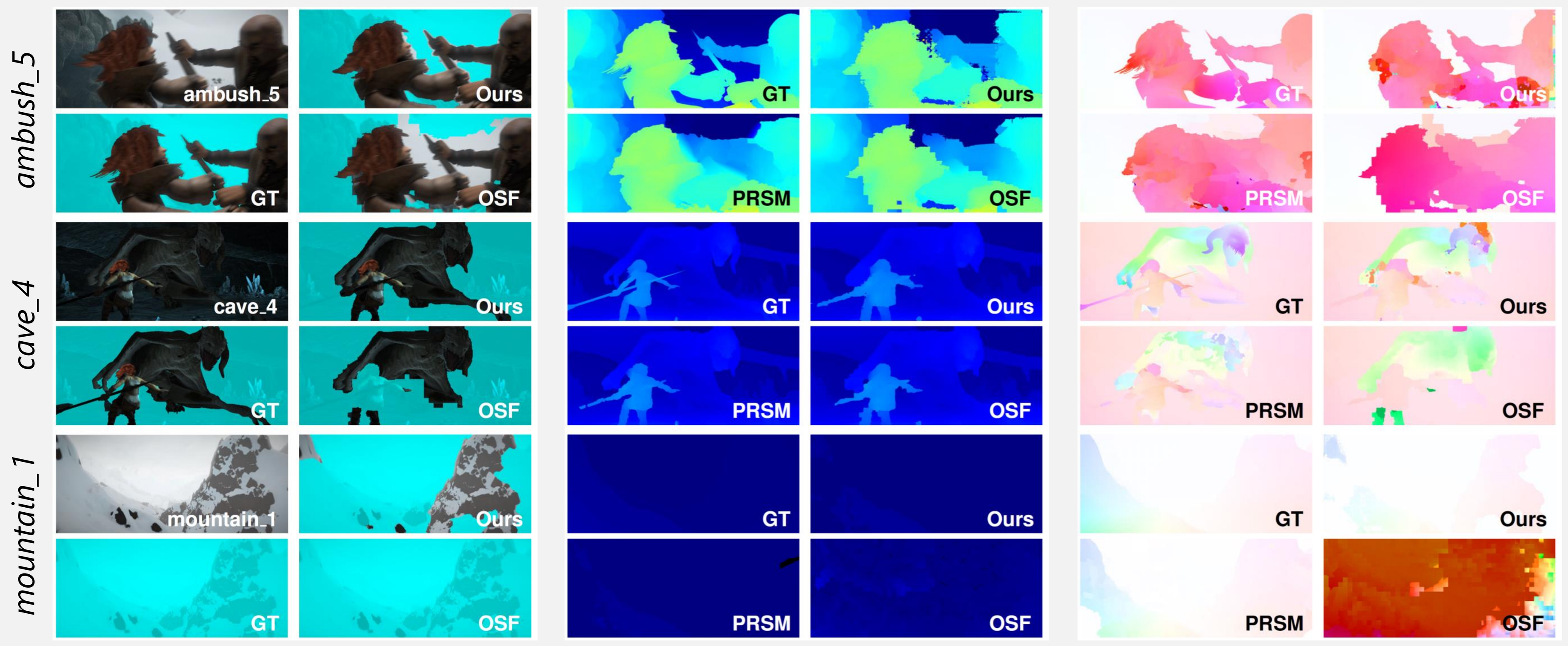
200 scenes from the KITTI benchmark

Evaluation on Sintel dataset

	D1			Fl			SF			SG	
	Ours	OSF	PRSM	Ours	OSF	PRSM	Ours	OSF	PRSM	Ours	OSF
alley_1	5.92	0.64	-1.50	2.11	-5.22	0.53	6.91	-3.13	-0.98	5.40	-12.06
alley_2	2.08	0.77	1.29	1.20	-0.23	0.12	2.99	0.50	1.36	1.94	0.63
ambush_2	36.93	-18.20	-4.83	72.68	-14.69	21.35	80.33	-10.63	18.41	1.72	-31.04
ambush_4	23.30	-0.74	-0.79	45.23	-3.93	3.25	49.81	-3.45	3.67	20.98	1.16
ambush_5	18.54	-1.00	0.82	24.82	-19.88	-0.42	35.15	-17.11	1.04	2.50	-16.89
ambush_6	30.33	4.14	0.92	44.05	-10.70	2.07	49.93	-8.53	2.85	53.95	28.97
ambush_7	23.47	-48.11	-11.60	27.87	5.40	24.52	44.51	-33.44	7.58	26.77	-9.31
bamboo_1	9.67	-0.05	2.33	4.11	0.07	1.70	11.05	0.23	2.70	4.43	0.26
bamboo_2	19.27	1.18	2.20	3.65	-1.21	0.07	21.39	0.16	2.16	4.08	-0.46
bandage_1	20.93	1.56	-0.30	4.00	-14.41	0.70	23.72	-12.84	0.36	33.32	-13.34
bandage_2	22.69	-0.84	0.26	4.76	-8.36	0.70	24.19	-8.13	0.57	16.37	-24.77
cave_4	6.22	0.37	1.95	14.62	-19.32	-1.70	17.53	-18.51	-0.18	16.13	-0.80
market_2	6.81	0.20	1.53	5.17	-4.91	0.40	10.38	-4.15	1.84	8.97	-4.93
market_5	13.25	-0.42	-2.12	26.31	-3.27	-2.07	29.93	-1.67	-2.07	15.26	-0.07
market_6	10.63	0.34	1.64	13.13	-3.26	2.41	18.07	-2.11	2.98	3.59	-34.04
mountain_1	0.23	-0.54	-0.18	17.05	-71.55	13.34	17.05	-71.56	13.20	31.63	31.63
shaman_2	24.77	-3.50	-0.72	0.56	-1.10	0.10	25.07	-4.35	-0.68	30.98	3.95
shaman_3	27.09	-25.13	-6.83	1.31	-10.14	-0.44	27.61	-27.91	-6.82	3.81	-25.83
sleeping_2	3.52	0.55	1.78	0.02	0.00	0.01	3.52	0.55	1.78	0.00	-0.54
temple_2	5.96	0.42	1.04	9.66	-0.86	0.15	9.82	-0.73	-0.05	1.32	-2.81
temple_3	10.65	-5.97	-0.40	62.34	-19.05	30.24	63.56	-18.30	28.96	4.20	-21.22
AVERAGE	15.35	-4.49	-0.64	18.32	-9.84	4.62	27.26	-11.67	3.75	13.68	-6.26

■ Our method is better ■ Our method is worse

Comparison with state-of-the-art methods (PRSM, OSF) on Sintel dataset



Reference / Motion segmentation — Disparity maps — Flow maps