Dense Wide-Baseline Scene Flow From Two Handheld Video Cameras — Supplemental Material —

Christian Richardt [1, 2, 3] Hyeongwoo Kim [1] Levi Valgaerts [1] Christian Theobalt [1]

[1] Max Planck Institute for Informatics [2] Intel Visual Computing Institute [3] University of Bath

1. Variational scene flow computation

The method of Valgaerts et al. [2] estimates the scene flow between two successive time steps by minimising an energy functional of the form

$$E = \int_\Omega \Bigg(\underbrace{\sum_{i=1}^{4} E_{\mathrm{D}}^i}_{\text{data}} + \underbrace{\sum_{i=1}^{2} \alpha_i \cdot E_{\mathrm{E}}^i}_{\text{epipolar}} + \underbrace{\sum_{i=1}^{3} \beta_i \cdot E_{\mathrm{S}}^i}_{\text{smoothness}} \Bigg) \,\mathrm{d}\mathbf{x}. \quad (1)$$

The first part of this energy collects four data terms that measure the difference in brightness between corresponding points in the four-frame configuration of Figure 1:

$$E_{\mathrm{D}}^1 = \Psi\big(\|I_1^{t+1}(\mathbf{x}+\mathbf{u}_1) - I_1^t(\mathbf{x})\|_2^2\big), \quad (2)$$
$$E_{\mathrm{D}}^2 = \Psi\big(\|I_2^{t+1}(\mathbf{x}+\mathbf{u}_1+\mathbf{u}_2+\mathbf{u}_3) - I_2^t(\mathbf{x}+\mathbf{u}_2)\|_2^2\big), \quad (3)$$
$$E_{\mathrm{D}}^3 = \Psi\big(\|I_2^t(\mathbf{x}+\mathbf{u}_2) - I_1^t(\mathbf{x})\|_2^2\big), \quad (4)$$
$$E_{\mathrm{D}}^4 = \Psi\big(\|I_2^{t+1}(\mathbf{x}+\mathbf{u}_1+\mathbf{u}_2+\mathbf{u}_3) - I_1^{t+1}(\mathbf{x}+\mathbf{u}_1)\|_2^2\big). \quad (5)$$

Here, $\mathbf{u}_1$ and $\mathbf{u}_2$ denote the optical flow in the first view and the stereo flow at time t, respectively, while $\mathbf{u}_3$ closes the correspondence loop from I_1^t to I_2^{t+1}. The images I_1^t are colour-corrected to match I_2^t using the transform $[\mathbf{A}\ \mathbf{a}]$ estimated for Equation 3 in the main paper – without this appearance normalisation, matching would be much harder. To handle the remaining appearance differences, we also include the gradient difference for improved matching in the presence of noise and lighting changes over time. We also disable the data terms for pixels that are marked as occluded in the occlusion mask, so that their flow value is chiefly determined by the epipolar and smoothness terms. For all terms, $\Psi(s^2) = \sqrt{s^2 + 10^{-6}}$ is the regularised ℓ_1 penaliser. We use $(\alpha_i, \beta_1, \beta_2, \beta_3) = (10, 31, 60, 200)$ for all results.

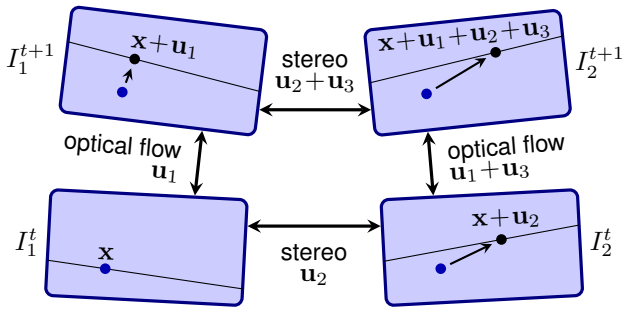

Figure 1. Four-frame configuration used in scene flow computation.

The second term of the energy favours correspondences that satisfy the epipolar constraint between I_1 and I_2:

$$E_{\mathrm{E}}^1 = \Psi\Big(\big((\mathbf{x}+\mathbf{u}_2)^\top \mathbf{F}_t\, \mathbf{x}\big)^2\Big), \quad (6)$$
$$E_{\mathrm{E}}^2 = \Psi\Big(\big((\mathbf{x}+\mathbf{u}_1+\mathbf{u}_2+\mathbf{u}_3)^\top \mathbf{F}_{t+1}(\mathbf{x}+\mathbf{u}_1)\big)^2\Big), \quad (7)$$

where $\mathbf{F}_t$ and $\mathbf{F}_{t+1}$ are the fundamental matrices at times t and $t+1$. Note that the variational formulation uses different data and epipolar terms than our matching cost (Equation 1 in main paper), as the terms used in our variational formulation are sufficient when provided with a good initialisation, as in our case.

The last term imposes regularized total-variation smoothness – the standard TV norm is defined as $\|\nabla \mathbf{u}\|_2$ [1] – on the estimated flows by penalising their spatial derivatives:

$$E_{\mathrm{S}}^i = \Psi\Big(\|\nabla \mathbf{u}_i\|_2^2\Big), \quad \text{for } i = 1, 2, 3. \quad (8)$$

2. Camera motion in the used datasets

Most of the datasets we use in our paper (BEAR, BOAR, BOY, DEER) were captured with independently moving, handheld cameras. This is clearly visible when looking at the camera baselines and angles between cameras over time, which are shown in Figure 2. The camera baselines vary by more than 50 percent, and up to 250 percent (DEER), while the angle between cameras varies over a range of 4 degrees (BOAR) to 36 degrees (DEER). The ODZEMOK dataset has a constant camera baseline, but the angle between cameras varies between about 10 and 20 degrees. The TRAFFIC2 dataset (now shown in Figure 2) uses a fixed stereo calibration with constant baseline and parallel cameras for all video frames.

References

[1] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992. 1

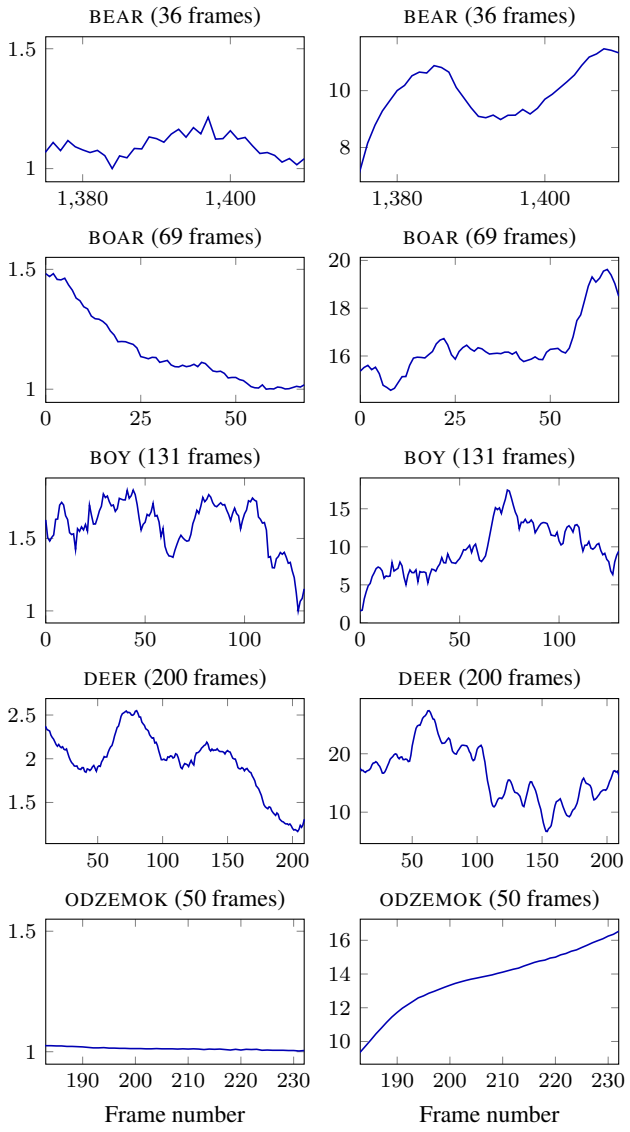

Figure 2. Visualisation of independent camera motion. **Left:** The baseline between cameras over time, normalised so that the minimum baseline is equal to one. **Right:** Angle between cameras over time (in degrees), specifically the angle between the principal axes of both cameras.

[2] L. Valgaerts, A. Bruhn, H. Zimmer, J. Weickert, C. Stoll, and C. Theobalt. Joint estimation of motion, structure and geometry from stereo sequences. In *ECCV*, 2010. 1