Fourier contrast optimization for occluded motion estimation

Ido Akov[1,2] **Roman Pflugfelder**[1,2]

Daniel Cremers[1]
[1]Technical University of Munich (TUM) [2]Austrian Institute of Technology (AIT)
{ido.akov, roman.pflugfelder, cremers}@tum.de

Abstract

Fragmented occlusion, as encountered in through-foliage observation, makes monocular motion estimation difficult because the target is visible only through sparse, discontinuous image fragments. We estimate motion by warping frames under a parametric model and maximizing the contrast of their integrated image. Although effective for 2DoF translation, this objective becomes ill-conditioned for 4DoF similarity motion. To analyze this, we derive a Fourier-domain reformulation that exposes the optimization structure and shows that static occlusion biases the objective toward zero motion. This motivates a decoupled 4DoF pipeline in which rotation and scale are estimated separately from translation. On synthetic videos with controlled fragmented occlusion, the Fourier formulation matches the spatial baseline at low-to-mid occlusion while converging faster, and the decoupled pipeline restores reliable translation recovery where joint 4DoF optimization fails.

1 Introduction

Fragmented occlusion [1] arises when an object is visible only through disconnected foreground gaps, as in observation through foliage, fences, or clutter. The visible evidence is sparse and spatially discontinuous, making correspondence-based motion estimation unreliable: classical optical flow and local registration methods [2–4] rely on brightness constancy and locally coherent support, both of which become fragile when the target never appears as a large connected region. We therefore consider monocular motion estimation from image sequences in which the object is never fully visible in any single frame, and motion must be inferred without reliable local matches.

An alternative is to estimate motion by warping and integrating observations under a low-dimensional parametric model. Variants of this principle appear in direct parametric alignment [5], synthetic-aperture reconstruction [6], and contrast-maximization methods for event cameras [7]. These approaches avoid explicit correspondences, but they do not directly address similarity motion under persistent fragmented occlusion in a monocular setting with a single global motion model.

Here we revisit contrast-based motion estimation under fragmented occlusion. The formulation works well for 2DoF translation, but becomes ill-conditioned for 4DoF similarity motion because repeated warp composition couples translation, rotation, and scale. We derive a Fourier-domain reformulation that makes the objective interpretable as a competition between moving-object alignment and static occlusion, and use this perspective to explain the instability of joint 4DoF optimization. Guided by this structure, we decouple the estimation: rotation and scale are first recovered by maximizing the same contrast objective in log-polar Fourier coordinates, and translation is then estimated separately using the same objective.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

2 Problem formulation

2.1 Spatial domain

Let $I_0, \ldots, I_{T-1}$ be a grayscale video sequence with $I_t \in \mathbb{R}^{M\times N}$, and let $W(\cdot, \theta)$ denote a parametric warp with motion parameters θ. We align each frame before temporal integration:

$$\bar{I}(\theta) = \frac{1}{T}\sum_{t=0}^{T-1} W^t(I_t, \theta), \tag{1}$$

where W^t denotes the t-fold application of the motion model. We then define

$$f_{\text{opt}}(I_{[0\ldots T-1]}, \theta) := \text{Var}\big(\bar{I}(\theta)\big), \qquad \theta^\star = \arg\max_\theta f_{\text{opt}}(I_{[0\ldots T-1]}, \theta). \tag{2}$$

Correct motion compensation sharpens the target in $\bar{I}(\theta)$, while static occluders and background become blurred by averaging. In the 2DoF setting, $\theta = [\tau_x, \tau_y]^\top$ is a coherent translation. In the 4DoF setting, we use a similarity model with translation, rotation, and isotropic scale.

2.2 Fourier domain

We first consider the translation case. By the Fourier shift theorem and linearity, the integrated image becomes

$$\mathcal{F}\{\bar{I}(\theta)\} = \frac{1}{T}\sum_{t=0}^{T-1} e^{-j\omega t\phi(\theta)}\,\mathcal{F}\{I_t\}. \tag{3}$$

where

$$\phi(\theta) := \frac{\tau_x}{M} + \frac{\tau_y}{N}. \tag{4}$$

We model each frame as the sum of a moving component and a static occlusion component,

$$I_t = W^t(I_{\text{mov}}, \theta^\star) + I_{\text{occ}}, \tag{5}$$

where I_{mov} denotes the target appearance, I_{occ} the static occluder, and $\theta^\star$ the true motion. Substituting into (3) yields

$$\mathcal{F}\{\bar{I}(\theta)\} = H(\Delta\phi)\,\mathcal{F}\{I_{\text{mov}}\} + H(\phi)\,\mathcal{F}\{I_{\text{occ}}\}, \tag{6}$$

with

$$\Delta\phi := \phi(\theta) - \phi(\theta^\star), \qquad H(\psi) := \frac{1 - e^{-j\omega T\psi}}{T(1 - e^{-j\omega\psi})}. \tag{7}$$

Since the variance objective equals non-DC Fourier energy, we obtain

$$f_{\text{opt}}(I_{[0\ldots T-1]}, \theta) \overset{\text{DFT}}{\leftrightarrow} \sum_{(m,n)\neq(0,0)} \left|H(\Delta\phi)\,\mathcal{F}\{I_{\text{mov}}\} + H(\phi)\,\mathcal{F}\{I_{\text{occ}}\}\right|^2. \tag{8}$$

This makes the optimization structure explicit: the moving-object term is maximized at the true motion, whereas the static-occlusion term is maximized at zero motion and therefore biases the objective accordingly. In higher-DoF settings, additional motion parameters further weaken the coherence of the moving term, exacerbating this competition and destabilizing joint optimization.

3 Motion decoupling for 4DoF

The Fourier formulation suggests a natural way to address this instability. In similarity motion, translation is encoded in Fourier phase, while rotation and scale are encoded in Fourier magnitude; after log-polar remapping, the latter become translations in log-polar coordinates [8, 9]. This yields a two-stage pipeline. In phase 1, each frame is mapped to Fourier magnitude, remapped to log-polar coordinates, and rotation and scale are estimated by maximizing the same contrast objective over the integrated representations. In phase 2, frames are aligned using the recovered rotation and scale, and the residual translation is estimated by maximizing f_{opt} in the image or Fourier domain. This separation prevents rotation and scale errors from being absorbed as translation drift and stabilizes optimization.

4 Experiments

4.1 Experimental setup

We evaluate the proposed methods on synthetic grayscale videos of simple high-contrast geometric shapes undergoing coherent motion. Each sequence contains $T = 8$ frames of size 128×128. This setup isolates the effect of fragmented occlusion under controlled variation in occlusion density. Fragmented occlusion is simulated by static structured masks, and occlusion density is defined as the fraction of pixels within the motion support covered by the mask. All methods are optimized with Adam, learning rate 0.1, and a fixed budget of 200 iterations for 2DoF and for each phase of the 4DoF pipeline.

In the 2DoF setting, motion is restricted to coherent translation and we compare the original spatial-domain contrast objective with its Fourier-domain reformulation. In the 4DoF setting, we compare direct joint optimization against the proposed decoupled pipeline.

4.2 Evaluation metrics

We use translation endpoint error (EPE) as the primary metric:

$$\mathrm{EPE} = \|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^\star\|_2. \tag{9}$$

We treat optimization on a single video-motion instance as successful once the recovered translation satisfies $\mathrm{EPE} < 0.5$. We additionally report median time-to-threshold (TTT), i.e. the number of optimization steps required to first satisfy this criterion. In the 4DoF setting, our primary goal is not precise estimation of rotation and scale in isolation, but successful motion decoupling: we evaluate whether the recovered rotation and scale are sufficient to remove drift and restore accurate translation recovery.

4.3 Results

Figure 1 summarizes the main empirical findings. In the 2DoF setting, the Fourier-domain objective closely matches the spatial objective at low occlusion densities, confirming that the reformulation preserves the behavior of the original contrast objective in the translation regime. Over this low-to-mid density range, the Fourier formulation reaches the EPE threshold in fewer iterations and exhibits a flatter time-to-threshold profile. At moderate and heavy occlusion, however, the spatial objective is consistently more robust in success rate, and both methods eventually break down.

The 4DoF experiment reveals a qualitatively different phenomenon. Direct joint optimization of translation, rotation, and scale fails already in the unoccluded case, indicating that the difficulty is not caused by occlusion alone but by the structure of the joint objective itself. This is consistent with the analysis in the previous sections: repeated warp composition couples the motion parameters, so that errors in rotation and scale are absorbed as translation drift. In contrast, the proposed decoupled pipeline succeeds reliably at low occlusion and degrades gradually as density increases. The gap is therefore structural rather than purely quantitative.

5 Conclusion

We introduced a Fourier-domain reformulation of a contrast-maximization objective for motion estimation under fragmented occlusion. The analysis reveals a competition between moving-object alignment and static occlusion, which increasingly biases the objective and leads to instability in joint 4DoF optimization.

This perspective leads to a decoupled estimation strategy in which rotation and scale are recovered separately from translation. Experiments confirm that, at low-to-mid occlusion densities, the Fourier formulation retains the behavior of the spatial formulation in the translation regime while improving convergence, and that decoupling resolves the failure of joint optimization in 4DoF settings.

Future work will extend the analysis to real through-foliage and other naturally occluded sequences to assess robustness beyond synthetic settings.

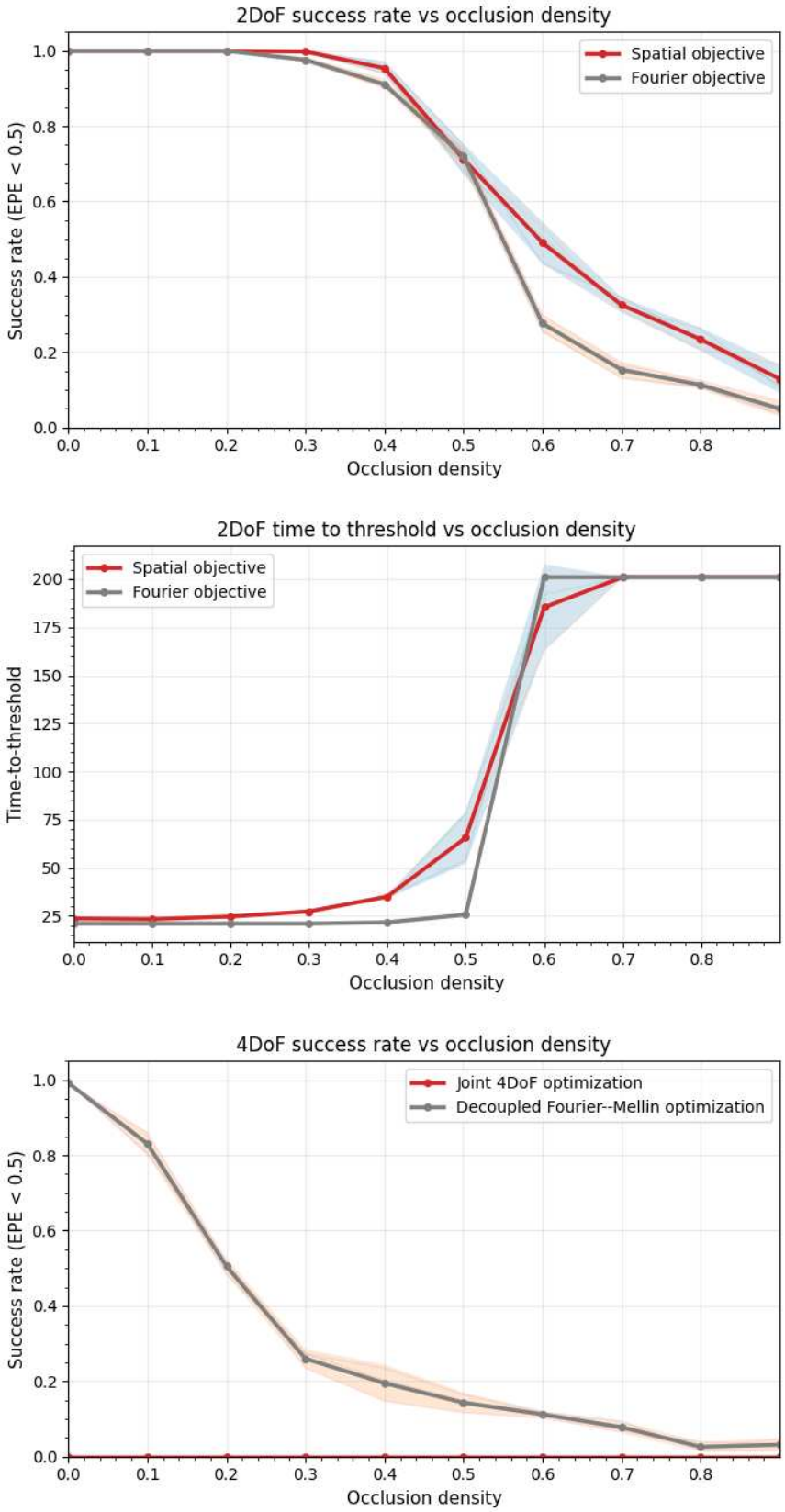

Figure 1: Optimization under fragmented occlusion. Top: 2DoF translation success rate versus occlusion density for spatial and Fourier objectives. Middle: corresponding median time-to-threshold (TTT). Bottom: 4DoF similarity-motion success rate versus occlusion density for joint optimization and the proposed decoupled pipeline.

Acknowledgements

This work received funding from the European Defence Fund under grant agreement EDF-2022-101121405-STORE.

References

[1] Julian Pegoraro and Roman Pflugfelder. The problem of fragmented occlusion in object detection, 2020. URL https://arxiv.org/abs/2004.13076.

[2] Bruce Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision (ijcai). volume 81, 04 1981.

[3] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17 (1–3):185–203, 1981.

[4] Michael J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[5] Michal Irani, B. Rousso, and Shmuel Peleg. Detecting and tracking multiple moving objects using temporal integration. *European Conference on Computer Vision*, pages 282–287, 01 1992.

[6] Vaibhav Vaish, Richard Szeliski, C. Lawrence Zitnick, Sing Bing Kang, and Marc Levoy. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[7] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 3867–3876. IEEE, June 2018. doi: 10.1109/cvpr.2018.00407. URL http://dx.doi.org/10.1109/CVPR.2018.00407.

[8] B. S. Reddy and B. N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, 1996.

[9] Santosh Thoduka, Frederik Hegger, Gerhard K. Kraetzschmar, and Paul G. Plöger. Motion detection in the presence of egomotion using the fourier-mellin transform. In *RoboCup 2017: Robot World Cup XXI*, page 252–264, Berlin, Heidelberg, 2017. Springer-Verlag. ISBN 978-3-030-00307-4. doi: 10.1007/978-3-030-00308-1_21. URL https://doi.org/10.1007/978-3-030-00308-1_21.