Supplementary Material — The Best of Both Worlds: Combining CNNs and Geometric Constraints for Hierarchical Motion Segmentation

Pia Bideau Aruni RoyChowdhury Rakesh R Menon Erik Learned-Miller
University of Massachusetts Amherst
College of Information and Computer Science
{pbideau, arunirc, rrmenon, elm}@cs.umass.edu

Abstract

Our supplementary material contains additional information about the following:

- *Modeling optical flow noise using the Sintel dataset.*
- *A proof that under projective transformations our* **set** *of rigid motion models is independent of the camera focal length* f.
- *The classical definition of motion segmentation and its connection to the related but different problem of video segmentation.*
- *Additional results on the DAVIS benchmark.*

A. Modeling the flow noise

We use the ground truth optical flow provided by the Sintel [4] dataset for modeling the characteristics of optical flow computed by the algorithm of Sun et al. [15].

We measure the variance of the observed flow noise for different magnitudes r of the ground truth flow. Figure 1 shows four histograms of the flow noise (u-component) for different ground truth flow magnitudes. The last plot shows the observed variances as blue dots and in red the exponential function that best models the relationship between flow noise variance and the motion field magnitude r. A significant relation between the variance of the flow noise and magnitude can be observed – the larger the flow magnitudes, the larger the covariance of the flow noise. For large pixel displacements the computation of optical flow becomes very noisy. To incorporate this relationship into our model, we model the variance as a function of r with an exponential function of the form $s(r) = a \cdot e^{br}$.

The least squares fit for a and b are:[1]

$$\begin{aligned}\text{Var}(n_u(r)) \quad &: \quad a = 11.45 \times 10^{-5}, b = 35.85\\ \text{Var}(n_v(r)) \quad &: \quad a = 16.35 \times 10^{-5}, b = 45.8\end{aligned}$$

Additionally we introduce a multiplier m, to add flexibility to our noise model. This is supportive for real world videos, since the measurements rely on the synthetic action movie Sintel which comes with additional challenges, like textureless regions, artificial motion blur effects and large pixel displacements. We learn the parameter m using the FBMS-59-3D motion training data set [1, 12].

B. Independence of the set of rigid motion models from the focal length f

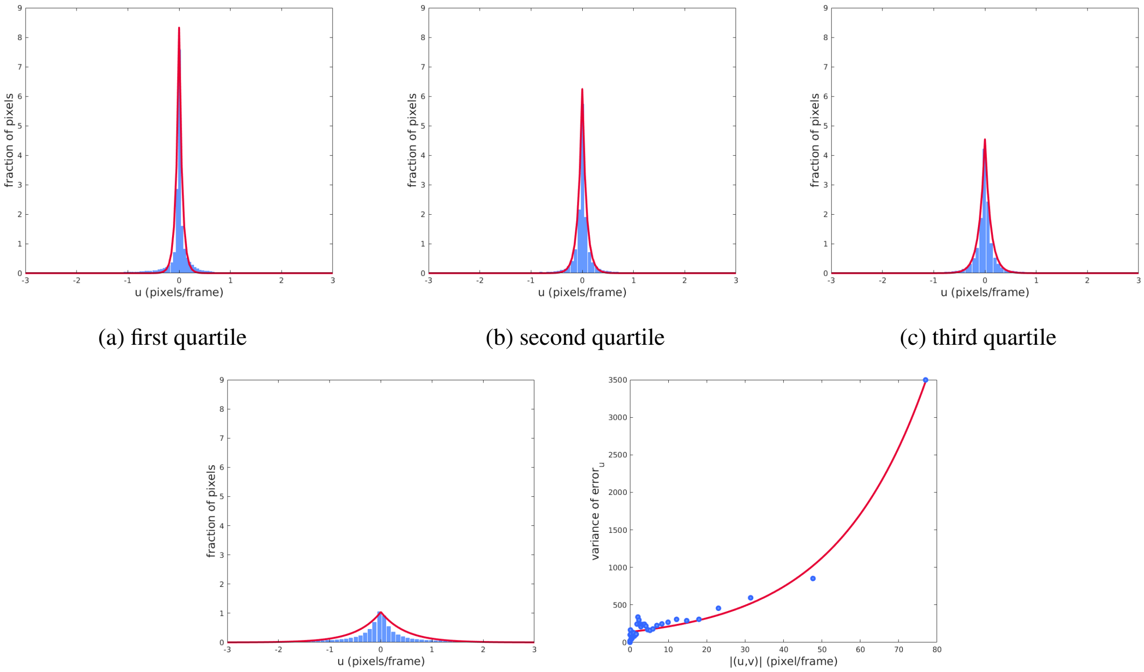
The equations that relate the translational motion of an object to the motion field (Equation 1 in the main paper),

$$u = \frac{-fU + xW}{Z}; \quad v = \frac{-fV + yW}{Z}, \tag{1}$$

show that the translational motion field (u, v) depends upon the camera focal length f. Thus, a motion field alone, without the focal length, is not enough to infer the 3D motion direction of an object. While our method segments objects based upon *different* 3D motions projected on a 2D image plane, it is not important for the method to infer the exact 3D direction. In this section, we show that for any fixed but unknown focal length, each rigid motion model maps to a unique motion direction in 3D. Thus, the rigid motion models are enough to distinguish among *different motions* even though they are not enough to distinguish the exact 3D motion. In other words, if our goal is merely to separate different types of motions, the rigid motion models are sufficient. Examples of these motion models can be seen on the right half of Figure 1 in the main paper. We present a proof that for any camera focal length f, there is a one-to-one mapping from rigid motion models to 3D motions.

Notation and Preliminaries. Let $S(f)$ be the set of all possible rigid motion models in a static environment for a camera with focal length f. $M(f, T)$ is a motion model defined by the focal length f and the motion direction $T = (U, V, W)$. Let $\mathcal{T}$ be the set of all translational directions,

[1] Parameters a and b are measured based on the *normalized flow* – the flow relative to the frame size.



(a) first quartile

(b) second quartile

(c) third quartile

(d) fourth quartile

(e) variance of flow noise

Figure 1: **Variance of flow noise.** *(a)-(d):* histograms of the optical flow noise of the first, second, third and fourth quartile of motion field magnitudes (Q_1, Q_2, Q_3, Q_4). *(e):* Visualization of the dependence of the flow noise variance and the corresponding motion field magnitude r. The blue dots show the flow noise variance for a particular motion field magnitude.

i.e., the set of points on the unit sphere. That is

$$S(f) = \{s : s = M(f, T), T \in \mathcal{T}\}, \tag{2}$$

Consider the set S^* of rigid motion models generated by the set of all possible motion directions T when the focal length f is equal to 1. We are interested in the question of how the set $S(f)$ of motion models differs from S^*, due to the difference of focal length.

Theorem 1. *Let f and f' be two different focal lengths. Let $M(f, T)$ be a canonical rigid motion model that results from the focal length f and motion direction T. The same rigid motion model can be obtained for another focal length $f' = cf$ and a different motion direction $T' = (U, V, cW)$, as $M(f', T')$. We show that*

$$M(f, T) \quad = \quad M(f', T'). \tag{3}$$

Thus the direction $\theta(x, y, f, U, V, W)$ at each pixel location (x, y) can be obtained with different focal length $f' = cf$ and a different motion direction $T' = (U, V, cW)$, or

$$\theta(x, y, f', U, V, cW) \quad = \quad \theta(x, y, f, U, V, W) \tag{4}$$

Proof.

$$\begin{aligned} & \theta(x, y, f', U, V, cW) && (5)\\ = \;& \arctan(cW \cdot y - V \cdot f', cW \cdot x - U \cdot f') && (6)\\ = \;& \arctan(c(W \cdot y - V \cdot f), c(W \cdot x - U \cdot f)) && (7)\\ = \;& \arctan(W \cdot y - V \cdot f, W \cdot x - U \cdot f) && (8)\\ = \;& \theta(x, y, f, U, V, W). && (9) \end{aligned}$$

□

Since this establishes a one-to-one mapping among rigid motion models governed by the two focal lengths, it establishes that the total set S of rigid motion models is independent of focal length. In particular, while the rigid motion model $M(f, T)$ for a particular motion direction is affected by the focal length, the *set of all possible rigid motion models S* is the same for all focal lengths.

C. Classical definition of motion segmentation

The general idea of **video segmentation** can be described as follows: given a sequence of frames the goal is to produce k regions which share one or more common properties. There are many different properties that may be

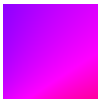
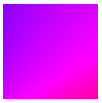
3D trans. direction $[U, V, W]$	**focal length** **in pixel**	**rigid motion model** M
$[-1, 1, 1]$	1000	
$[-1, 1, 0.001]$	1	

Table 1: **Independence of the set of rigid motion models from the focal length**. Same rigid motion model can be obtained using a different focal length and a different motion direction $[U, V, W]$.

relevant for problems addressing video segmentation. Examples of such properties are color, shape, depth or "objectness".

For **motion segmentation** algorithms, the property of interest is *3D motion*. Understanding motion is essential for understanding the world, predicting the future, understanding actions and interactions and also for understanding the definition of "objectness" itself. Thus *motion segmentation is about segmenting* ***all*** *objects which are moving independently*.

Another way to think about video segmentation is to focus on the property of objectness as primary. We can refer to this general class of video segmentation problems as **video object segmentation**, or in this context, just object segmentation, for short. For object segmentation, answering the questions *What is an object?* and *How is an object defined?* is essential. The understanding of objects might or might not incorporate knowledge about motion – a high quality object segmentation algorithm is not necessarily a good motion segmentation algorithm. An object segmentation algorithm might segment a table and chair separately regardless of whether those are moving or not, while a motion segmentation algorithm should not segment a chair or a table unless they are moving.

D. The DAVIS benchmark

DAVIS 2016 [14] is a *video object segmentation* dataset providing a densely annotated, pixel-accurate ground truth for every frame. Videos in this dataset always contain *one* ground-truth object which is considered the "most important" moving object. The authors define 15 key characteristics (such as *motion blur*, *occlusion*, *interacting objects* or *appearance change*) that describe certain aspects or challenges of a video and assign these to each video.

We note that the task of segmenting the most important object in a video differs significantly from the original motion segmentation problem which is the focus of our work. An example illustrating this difference in definition is shown in Figure 3.

For completeness, we report results of our motion segmentation algorithm on DAVIS, using the "2016 TrainVal" data split and the provided evaluation codebase.[2] Since we have multiple foreground object masks for each video frame, we consider our closest matching mask to their single ground-truth object mask, following DAVIS' prescribed "bipartite" procedure for evaluating multi-object segmentation methods [14]. Performance is reported using the mean values of $\mathcal{J}$, $\mathcal{F}$ and $\mathcal{T}$, which indicate Jaccard index (IoU), object boundary accuracy and temporal stability, respectively [14].

Results on DAVIS using some variants of our method are summarized in Table 2 and visualizations shown in Figure 2. Naively forming a binary mask by taking the union of all our predicted moving objects results in relatively low performance (*binary* matching), while finding the best match from among our multi-object masks gives better results (*bipartite* matching). Including a CRF improves performance, as expected (*Ours+CRF*).

Method	**Matching**	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{T}$
Ours	binary	0.4578	0.4417	0.5822
Ours	bipartite	0.5141	0.4712	0.5827
Ours+CRF	bipartite	0.5347	0.5005	0.6652

Table 2: **Results on DAVIS.** *Method:* We show results for our method (`ours`) and the effect of adding a CRF (`ours+CRF`). *Matching:* we can match our multiple-object predictions to the single ground-truth object by either merging all our predictions into a single foreground mask (`binary`) or by selecting the prediction that best matches the ground-truth (`bipartite`).

We compare with other "unsupervised" category methods on DAVIS "2016 TrainVal" in Table 3. There are some common methods between these and the methods evaluated on the motion segmentation datasets presented in the main paper. In terms of Jaccard-index, we are better than TRC [6] and CVOS [16], and have about one percentage difference with MSG [11]. The methods NLC [5], FST [13] and LMP [17], which showed lower performance than ours in motion segmentation on the FBMS [3], Complex Background [10] and Camouflaged Animals [2] datasets, are better on DAVIS.

On DAVIS, videos of the category Dynamic Background (DB) – usually with crowds of people and moving water, make it hard to model the camera motion at the first stage

[2] https://github.com/jponttuset/davis-matlab/tree/davis-2016

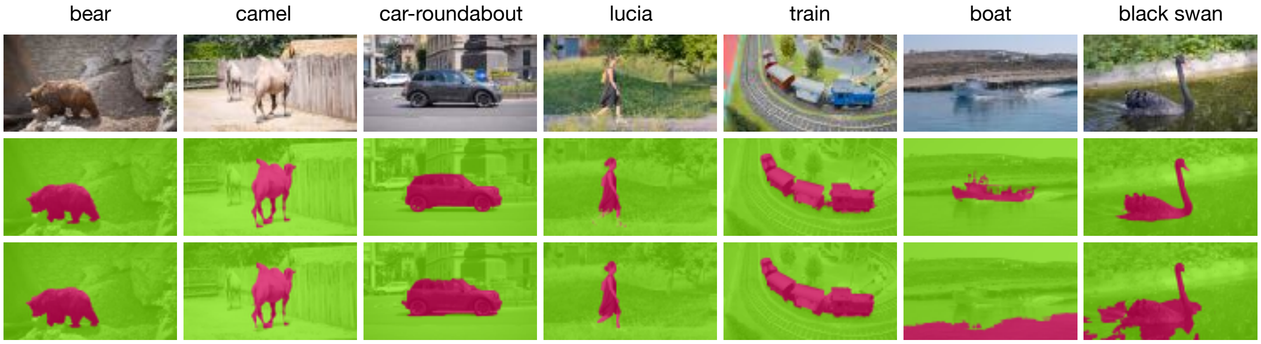


Figure 2: **Visualizations of results on DAVIS.** *Rows — top to bottom:* Original video frame, ground-truth and the output from our algorithm. Our method is not accurate in the cases of dynamic background, such as the `boat` and `black swan` videos.

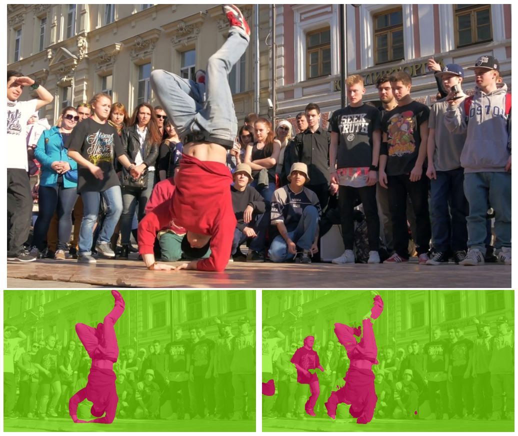


Figure 3: **Segmentation on DAVIS.** Segmentation of the `breakdance` video sequence (frame 2). *Bottom left:* ground truth, *bottom right:* motion segmentation. An evaluation on DAVIS doesn't mirror necessarily the quality of the motion segmentation method. A motion segmentation method is accurate if *all* moving objects are segmented, irrespectively whether they are are the primary object of interest or not.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{T}$
ARP [8]	0.763	0.711	0.359
FSEG [7]	0.716	0.658	0.295
LMP [17]	0.697	0.663	0.688
FST [13]	0.575	0.536	0.293
NLC [5]	0.641	0.593	0.366
MSG [11]	0.543	0.525	0.263
KEY [9]	0.569	0.503	0.210
CVOS [16]	0.514	0.490	0.256
TRC [6]	0.501	0.478	0.345
Ours+CRF	0.535	0.501	0.665

Table 3: **Comparison to state-of-the-art on DAVIS.** The performance for our final method is compared with other unsupervised methods taken from the online `DAVIS 2016 TrainVal` leaderboard.

of our method using simple rigid motion models. A rigid motion model can handle the motion of static background due to camera motion well, but not "messy" motions such as that of flowing water, which is very often present in DAVIS. These factors are mostly responsible for our performance drop on this dataset as compared to the motion segmentation results we presented in the main paper.

References

[1] P. Bideau and E. Learned-Miller. A detailed rubric for motion segmentation. *arXiv preprint arXiv:1610.10033*, 2016. 1

[2] P. Bideau and E. Learned-Miller. It's moving! A probabilistic model for causal motion segmentation in moving camera videos. In *Proc. ECCV*, pages 433–449. Springer, 2016. 3

[3] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *Proc. ECCV*, pages 282–295. Springer, 2010. 3

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. ECCV*, pages 611–625. Springer-Verlag, 2012. 1

[5] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In *Proc. BMVC*, volume 2, page 8, 2014. 3, 4

[6] K. Fragkiadaki, G. Zhang, and J. Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *Proc. CVPR*, pages 1846–1853, 2012. 3, 4

[7] S. Jain, B. Xiong, and K. Grauman. FusionSeg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. In *Proc. CVPR*, 2017. 4

[8] Y. J. Koh and C.-S. Kim. Primary object segmentation in videos based on region augmentation and reduction. In *Proc. CVPR*, pages 7417–7425. IEEE, 2017. 4

[9] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *Proc. ICCV*, pages 1995–2002. IEEE, 2011. 4

[10] M. Narayana, A. Hanson, and E. Learned-Miller. Coherent motion segmentation in moving camera videos using optical flow orientations. In *Proc. ICCV*, pages 1577–1584, 2013. 3

[11] P. Ochs and T. Brox. Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In *Proc. ICCV*, pages 1583–1590. IEEE, 2011. 3, 4

[12] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *PAMI*, 36(6):1187–1200, 2014. 1

[13] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *Proc. ICCV*, pages 1777–1784, 2013. 3, 4

[14] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proc. CVPR*, pages 724–732, 2016. 3

[15] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *Proc. CVPR*, pages 2432–2439. IEEE, 2010. 1

[16] B. Taylor, V. Karasev, and S. Soatto. Causal video object segmentation from persistence of occlusions. In *Proc. CVPR*, pages 4268–4276, 2015. 3, 4

[17] P. Tokmakov, K. Alahari, and C. Schmid. Learning motion patterns in videos. In *Proc. CVPR*, 2017. 3, 4