Optical Flow Revisited: how good is dense deep learning based optical flow?

Jeongmin Kang, Zoran Sjanic and Gustaf Hendeby

DOI: https://doi.org/10.1109/SDF-MFI59545.2023.10360502

IEEE Symposium Sensor Data Fusion and International Conference on Multisensor Fusion and Integration (SDF-MFI), 2835-947X, No.

Optical Flow Revisited: how good is dense deep learning based optical flow?

Jeong Min Kang*, Zoran Sjanic*†, and Gustaf Hendeby*
* Dept. of Electrical Engineering, Linköping University, Linköping, Sweden
e-mail: {jeongmin.kang,zoran.sjanic,gustaf.hendeby}@liu.se
† Saab AB, Linköping, Sweden

***Abstract*—Accurate localization is a part of most autonomous systems. GNSS is today the go to solution for localization but is unreliable due to jamming and is not available indoors. Inertial navigation aided by visual measurements, *e.g.*, optical flow, offers an alternative. Traditional feature-based optical flow is limited to scenes with good features, current development of deep neural network derived dense optical flow is an interesting alternative. This paper proposes a method to evaluate the result of dense optical flow on real image sequences using traditional feature-based optical flow and uses this to compare six different dense optical flow methods. The results of the dense methods are promising, and no clear winner amongst the methods can be determined. The results are discussed in the context of how they can be used to support localization.**

I. Introduction

Reliable and robust localization is an essential capability for autonomous navigation in indoor and outdoor environments. Global navigation satellite systems (GNSS), such as GPS, have turned into the *de facto* standard for localization in outdoor environments. However, GNSS is sensitive to disturbances and jamming, suffers from degraded performance in urban canyons, and does not work at all in indoor environments where the GNSS signals are blocked. To handle these situations alternatives to GNSS navigation are needed.

One common way to obtain a stand-alone navigation system is to use inertial sensors. However, all inertial navigation systems (INS) suffer from drift and need an additional source of information to provide a viable long term navigation solution. Image sensors can provide this extra information, and their relative light weight and low price makes them an attractive sensor alternative. Current research direction in visual navigation which has shown promising results is visual-simultaneous localization and mapping (V-SLAM), which is an example of a SLAM solution [1]. Another method to extract information about motion from images is visual odometry (VO) [2]. Both V-SLAM and VO aim to estimate the camera's pose using structure for motion concept from computer vision. They extract feature points from images and estimate camera pose by matching feature points within consecutive frames. The state or poses of the camera is formulated in such a way that it can be solved online using a Kalman filter or as a batch problem to which standard optimization solvers can be applied.

Conventional visual navigation tracks the movement of a sparse set of image features between consecutive frame [3]. The sparse set provides reliable visual measurements to navigation systems in visual information rich environment, but it can lead to performance degradation in featureless or dynamic environment in which the reliability of feature tracking decreases between successive image frames. The optical flow, which can be obtained from the sparse feature set, is the movement of a point between two frames. Assuming that the feature points represent the same point in the 3D world in two frames, the optical flow is then the change of the 2D projection into the image plane of this 3D point. This, in turn, carries information about the motion of the camera. Therefore, extracting robust optical flow plays a key role in these systems. If the optical flow is computed from a sparse set of feature points only a sparse optical flow is available in the tracked points. The features can be extracted using standard feature detectors, *e.g.*, the scale-invariant feature transform (SIFT) [4] or ORB [5], and tracked using some feature tracker, *e.g.*, the Lucas-Kanade tracker (KLT) [6].

Recently, as the result of the rapid development of deep learning, methods have been developed that provide dense optical flow. These methods are fed a pair of images and produce an optical flow estimate in each pixel, regardless of if there are any feature points or not. Deep learning-based optical flow estimation methods implement convolutional neural networks (CNNs), which derive optical flow for all image pixels through stacking networks, image, or feature warping. The first end-to-end optical flow network was FlowNet [7], which uses CNNs with an encoder-decoder architectures. Later, FlowNet2 [8] stacked multiple FlowNetC and FlowNetS, and it remedied the low accuracy achieved with FlowNet. However, this stacking increases the number of parameters and the computational complexity. PWC-net [9] combined the coarse-to-fine strategy with warping features and computing a cost volume. It outperformed both classical and previous learning-based methods. LiteFlowNet [10] was proposed to achieve competitive accuracy while having fewer parameters than FlowNet2. The estimation performance of the deep learning-based optical flow estimation was further improved by applying recurrent neural network (RNN) techniques. Recurrent all-pairs field transforms [11], which is well known as RAFT, was proposed. It builds 4D correlation volumes between all pairs of pixels and uses that information to iteratively refine the flow field. Based on the RAFT framework, a global motion aggregation (GMA) module was proposed [12]. It adds an attention mechanism of transformer networks to RAFT, which has a multi-scale 4D correlation layer and a recurrent update component. The module of GMA improves occluded pixels estimation by adding a 4D attention matrix to the layered feature structure.

Since the proposal of the first end-to-end optical flow method [7], deep neural network (DNN)-based dense optical flow methods have achieved significant improvement in their accuracy. However, due to the evaluation metric that requires the ground truth in all pixels, the development of the algorithm often depends on the availability of synthetic datasets, and it usually focuses on obtaining clean flow fields represented by average end-point error (EPE) loss from open synthetic dataset. This results in a lack of comparisons of dense flow methods based on real-world scenes. Therefore, given the potential importance of optical flow in navigation applications, comparative studies of optical flow estimation methods based on both classical and DNN methods are needed. This paper aims to address this shortcoming and provides a comparison of classical feature-based and DNN-based optical flow estimation methods.

The main contributions of this paper are:

- a method to compare dense optical flow methods, using sparse feature-based optical flow as baseline;
- a comparison of dense optical flow methods, and a feature-based sparse optical flow method using standard datasets; and
- a discussion about the suitability of using dense optical flow computed as a supporting sensor in inertial navigation.

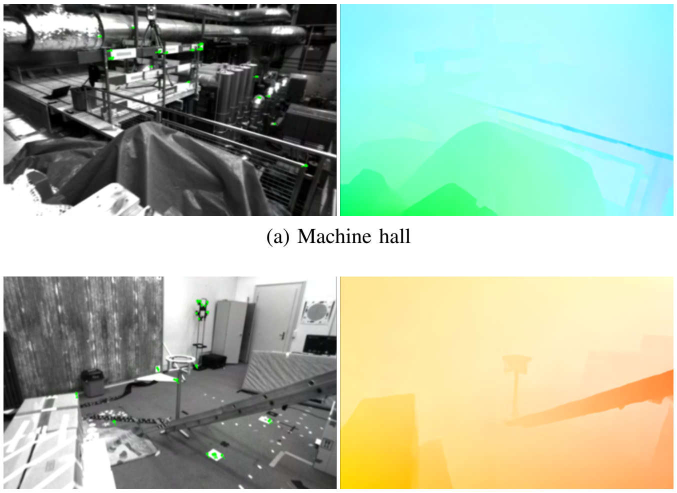
The remainder of this paper is organized the following way. First, Sec. II, it is described how to compute sparse and dense optical flow, then the result from the experimental comparison of the different optical flow methods are presented and discussed in Sec. III, and finally conclusions are presented in Sec. IV.

II. Optical Flow

Optical flow is defined as the movement of a pixel in an image stream between consecutive frames. The underlying assumption is that the pixel should represent the same point in 3D, which then makes it possible to derive properties of the motion of the camera without knowledge of the underlying 3D world. Traditionally optical flow has been computed using feature points, which results in a sparse optical flow. Advances in machine learning have resulted in the ability to produce dense optical flow. The differences between these two approaches are illustrated Fig. 1.

A. Feature-Based Optical Flow

When computing feature-based optical flow, the first step is to obtain stable features to track between frames. Any suitable features can be used, which one to use depends on the ability to find features in the scene at hand, and the ability to find appropriate feature correspondences between frames. The second step is to compute feature correspondences between the images. These correspondences are then used to compute the optical flow.



(a) Machine hall

(b) Vicon Room

Fig. 1: Examples of sparse and dense optical flow, computed for two images from two EuRoc datasets [13]. The GMA [12] method is used to obtain the dense optical flow images. The dense optical flow uses the flow field color encoding [14].

Assume a set of correspondences between image I^k and I^{k+1}, $\left\{\left((x_i^k, y_i^k), (x_i^{k+1}, y_i^{k+1})\right)\right\}_i$ obtained as described above. The feature-based optical flow is now given for feature pair i by

$$\boldsymbol{\delta}_i^{k,\text{feature}} = \begin{pmatrix} \Delta u_i^k \\ \Delta v_i^k \end{pmatrix} = \begin{pmatrix} x_i^{k+1} - x_i^k \\ y_i^{k+1} - y_i^k \end{pmatrix}. \tag{1}$$

The optical flow vector $\boldsymbol{\delta}_i^{k,\text{feature}}$ describes the displacement of feature point i in the reference image I^k in the frame I^{k+1}. The feature-based optical flow in the given image I^k is then $\left\{\boldsymbol{\delta}_i^{k,\text{feature}}\right\}$. This is typically much fewer points than points in the image, which motivates the name sparse optical flow.

B. DNN-Based Optical Flow

Dense optical flow refers to optical flow computed methods where every point in the reference image I_k is assigned an optical flow. Due to the advancements in machine learning, in particular in the area of deep neural networks (DNN), this can now be achieved. As described in the introduction, this can be done in many different ways. Basic approach of the DNN architecture is constructing CNNs which can solve the optical flow estimation problem as a supervised learning task. In order to improve quality and speed, the methods have evolved into applying various training data protocols and adopting a layer that correlates feature vectors at different image locations. To directly compare with the feature-based optical flow, we allocate estimated dense flow at the same pixel location as the feature-based flow. Therefore, DNN-based optical flow in the given image I^k is defined as $\left\{\boldsymbol{\delta}_i^{k,\text{DNN}}\right\}$, indicates the displacement given by the neural networks.

C. Comparison metrics

In order to compare different optical flow methods, we define metrics to compare the outputs. Given a specific point (x_i^k, y_i^k) in the image I^k with two optical flows $\boldsymbol{\delta}_i^{k,A}$ and $\boldsymbol{\delta}_i^{k,B}$ of methods A and B, define the angle error α_i^k as

$$\alpha_i^k = \arccos\left(\frac{\boldsymbol{\delta}_i^{k,A} \cdot \boldsymbol{\delta}_i^{k,B}}{\|\boldsymbol{\delta}_i^{k,A}\|\|\boldsymbol{\delta}_i^{k,B}\|}\right) [\text{rad}]. \tag{2}$$

In a similar way, the relative difference in magnitude can be defined

$$\mathcal{M}_i^k = \left(\frac{\|\boldsymbol{\delta}_i^{k,A}\| - \|\boldsymbol{\delta}_i^{k,B}\|}{\|\boldsymbol{\delta}_i^{k,A}\|}\right) \times 100\,[\%]. \tag{3}$$

The relative magnitude is used not to be too influenced by the magnitude of the optical flow.

In general, ground truth data for optical flow in real images is scarce, in particular for dense optical flow. As a result, many comparative studies of optical flow methods are only conducted using artificial image sequences. Though, providing some insights of the methods compared, using artificial images always runs the risk of introducing artifacts related to the creation of the artificial images. Hence, in this paper an alternative method of comparison is suggested and used to evaluate the different dense optical flow methods considered. The idea is to use optical flow from a stable feature-based sparse optical flow method as ground truth. The methods are then compared in the points where the sparse optical flow is computed. This way authentic image material can be used, which is an important benefit.

III. Experimental Method Comparison

In this section, six different DNN-based methods for dense optical flow are evaluated and compared using the publicly available EuRoc datasets [13]. As baseline, as true dense optical flow ground truth data is unavailable, sparse feature-based optical flow from ORB features [5] are used.

A. DNN Optical Flow Methods

In this comparison the following network models are used to obtain DNN dense optical flow: GMA, RAFT, LiteFlowNet2, PWC, FlowNet2CSS of FlowNet2, FlowNetC of FlowNet. All six networks are pre-trained on FlyingChairs [7] for all methods. The models are then fine-tuned on a combination of FlyingThings3D [15] (excluded for the FlowNetC), Sintel [16] (excluded for FlowNet2), and KITTI [17] (only LiteFlowNet2), considering each training method on public benchmarks [16]. The methods are all implemented using the PyTorch [18] with the mixed precision strategy. The modules are initialized from scratch with random weights using Tesla T4 GPU. Furthermore, standard optical flow training conventions [8, 11, 16] are followed. For example, in case of GMA, pre-train is performed using FlyingThings for 120 000 iterations with a batch size of 8, followed by another 120 000 iterations on FlyingThings with batch size of 6. Finally, the model is fine-tuned on Sintel for 120 000 iterations with batch size 6.

B. Experimental setup

The trained DNN optical flow methods are evaluated on image sequences from the EuRoC dataset [13], in particular the “Machine Hall 01” (here denoted “MH 01 easy”), “Machine Hall 03” (“MH 03 medium”), “Vicon Room 02” (“V2 02 easy”), and “Vicon Room 03” (“V2 03 difficult”) sequences. The comparisons are all performed against the sparse optical flow computed using ORB features. The results of these comparisons are presented and discussed below.

C. Results

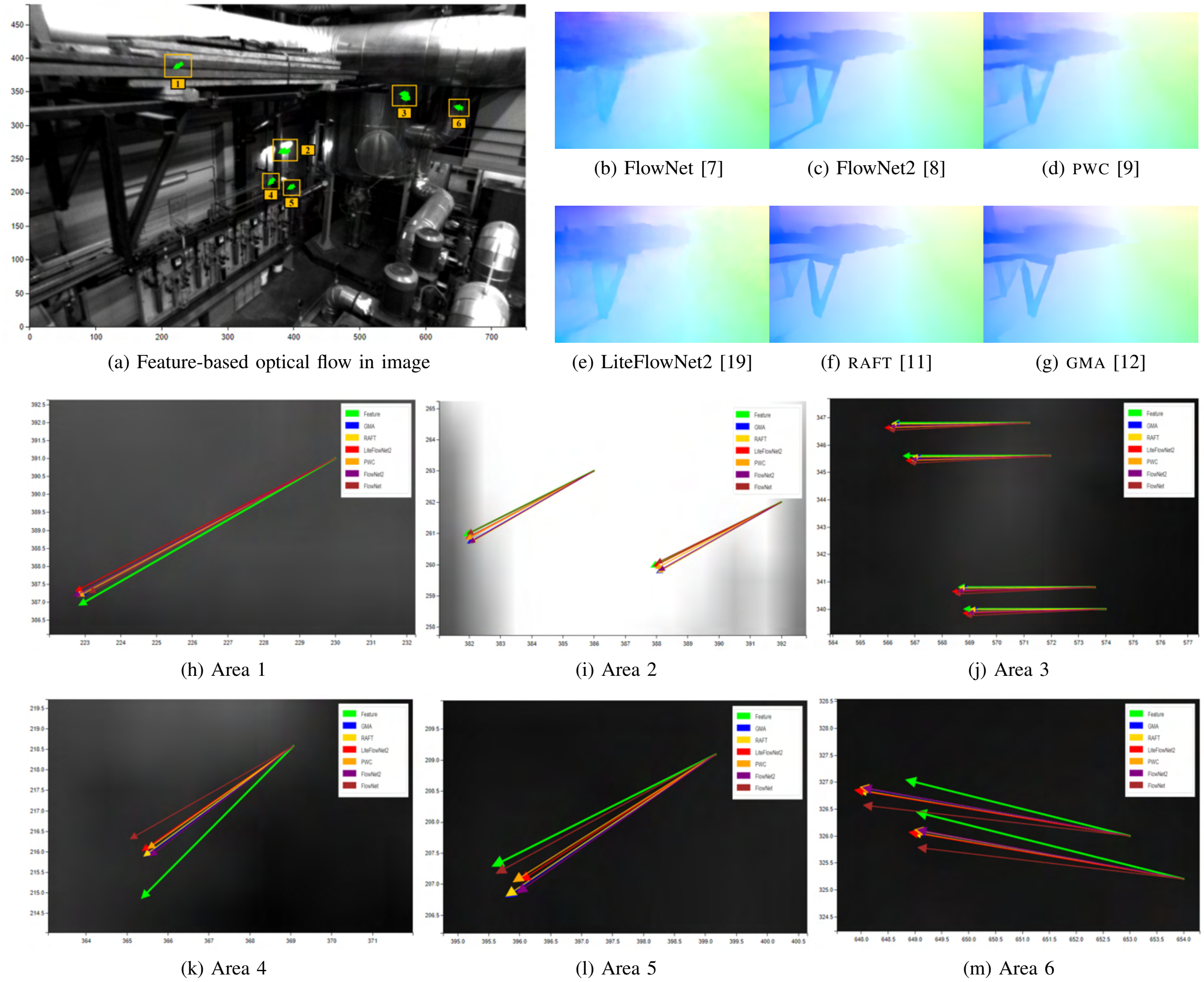
The results of all the comparisons made are summarized in Table I, which contains average errors for all the tests conducted. The table indicates that all the evaluated methods perform well on average, with only minor differences between the methods.

To complement the table, Fig. 2 and 3 provide qualitative comparisons of the methods for two frames with distinctive characteristics. For each frame, the evaluated frame is depicted, and the computed sparse optical flow is visualized. Next to it, the dense optical flow estimated with the different DNN methods are visualized. The estimated dense flows are all similar, but differ, *e.g.*, in the level of detail captured. That means, outside of the evaluated points, there are differences in the computed optical flow. Which one to prefer depends on the application, it can be observed that FlowNet and LiteFlow Net2 provides a smoother (locally average) than the other methods. Furthermore, the figure zooms in on 6 different regions with a sparse optical flow and shows the sparse optical flow in comparison to the optical flow computed by the DNN methods in the same point. The estimated optical flows match well in most cases, but there are exceptions such as Fig. 3(i) and (m), where in both cases it is questionable if the right feature match has been made between the images given the fairly homogeneous image patch.

A quantitative comparison of the angle and relative magnitude difference are provided in the Fig. 4 and 5, respectively, using box plots. Small angle differences implies that the orientation of the optical flow estimates from the compared methods matches and indicates a camera movement in the same direction. The means and distributions of the angle and magnitude differences of DNN methods in each sequence show similar distribution pattern. On thing to note though is that the “V2 03 difficult sequence” (Fig. 4d and 5d) has a smaller variance in the distribution compared to other sequences for both the angle and magnitude is present. Overall, the models for DNN-based dense optical flow estimation have evolved to return flows directly comparable to feature-based optical flow. Also, there are in general small differences between neural network estimates in comparison to the sparse method. The comparison results show that the overall dense methods provide similar vector fields compared to the sparse method, which shows that it is suitable for utilizing dense optical flow in the visual measurements to navigation system provided by sparse optical flow.

TABLE I: Summary of comparison results.

		GMA	RAFT	LiteFlowNet2	PWC	FlowNet2	FlowNet
MH 01 easy	Avg. angle diff. [rad]	0.0897	0.0894	0.0942	0.0885	0.0858	0.0991
(14 300 flows in 3 682 frames)	Avg. rel. magnitude diff. [%]	2.4509	2.4031	0.9282	1.8520	1.5576	0.5058
MH 03 medium	Avg. angle diff. [rad]	0.0831	0.0814	0.0854	0.0801	0.0777	0.0938
(8 671 flows in 2 700 frames)	Avg. rel. magnitude diff.	3.2090	3.1496	1.7667	2.7522	2.2259	2.7305
V2 01 easy	Avg. angle diff. [rad]	0.0872	0.0850	0.0951	0.0826	0.0801	0.1031
(8 546 flows in 2 280 frames)	Avg. rel. magnitude diff.	2.4257	2.2809	0.6701	1.8993	1.5551	1.9715
V2 03 difficult	Avg. angle diff. [rad]	0.0465	0.0466	0.0546	0.0466	0.0443	0.0587
(4 357 flows in 1 922 frames)	Avg. rel. magnitude diff.	0.8022	0.6527	0.6257	0.5817	0.5012	0.9879



(a) Feature-based optical flow in image

(b) FlowNet [7]

(c) FlowNet2 [8]

(d) PWC [9]

(e) LiteFlowNet2 [19]

(f) RAFT [11]

(g) GMA [12]

(h) Area 1

(i) Area 2

(j) Area 3

(k) Area 4

(l) Area 5

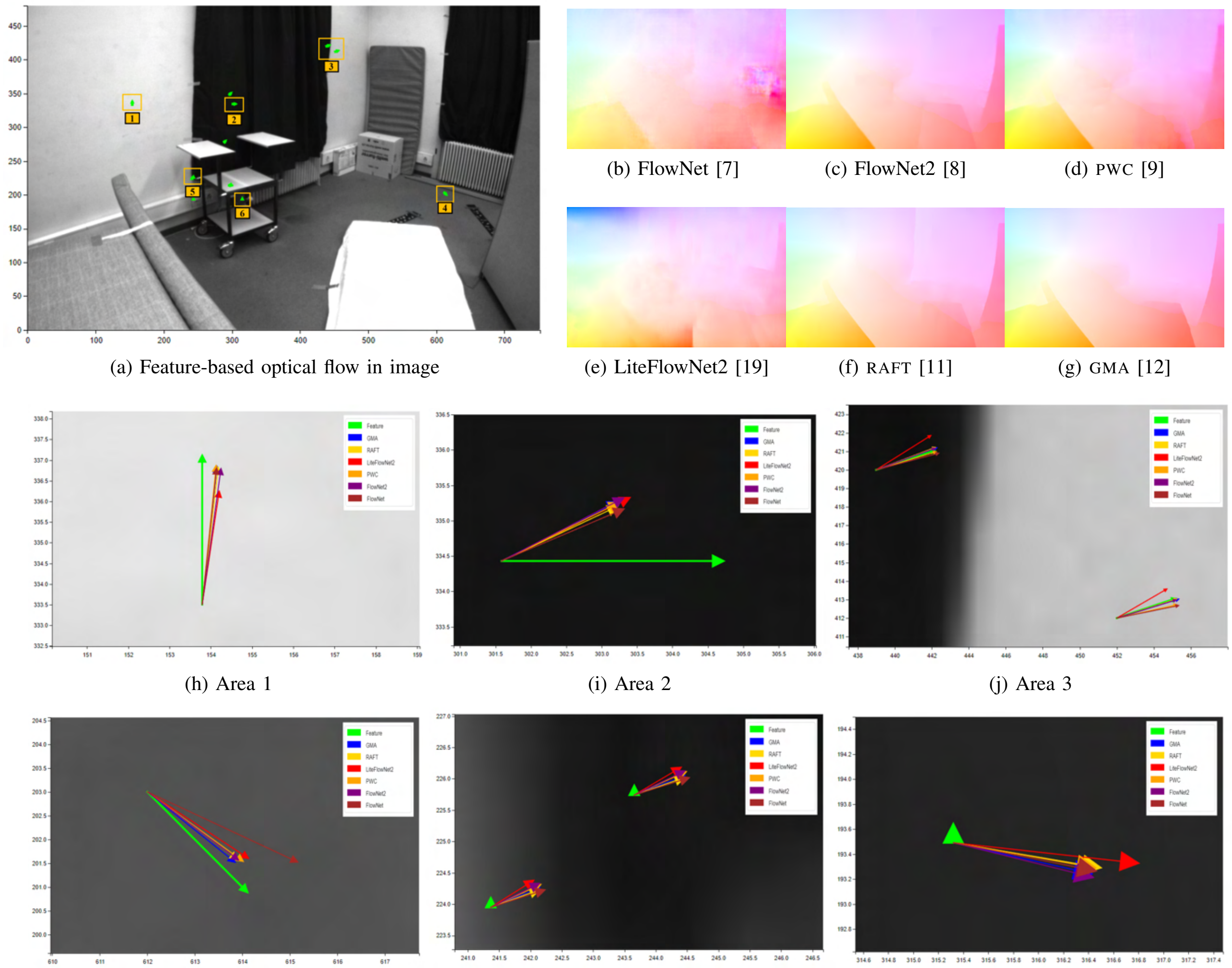
(m) Area 6

Fig. 2: Flow predictions and comparison results on Machine Hall of EuRoc dataset. The dense optical flow uses the flow field color encoding [14].

D. Discussion

Overall, the methods for dense optical flow evaluated perform similarly compared to the sparse feature-based optical method used as ground truth. The results are good, and in these points, for most usages, the DNN-based methods could be used as substitute for the feature-based optical flow. As can be seen in Fig. 2 and 3, in between these points, the overall behavior is similar, but difference can be observed in the level of “detail” captured. Whether to prefer the higher resolution result or the more averaged result depends on the application. Hence, it is important to consider this aspect before choosing to use one or the other method. One can also conclude that it could be important which points are used when extracting the optical flow in the end. It would be interesting to study if there are indicators of regions in an image that will provide reliable optical flow estimates.

As indicated in the introduction, the authors’ intention is to



(a) Feature-based optical flow in image

(b) FlowNet [7]

(c) FlowNet2 [8]

(d) PWC [9]

(e) LiteFlowNet2 [19]

(f) RAFT [11]

(g) GMA [12]

(h) Area 1

(i) Area 2

(j) Area 3

(k) Area 4

(l) Area 5

(m) Area 6

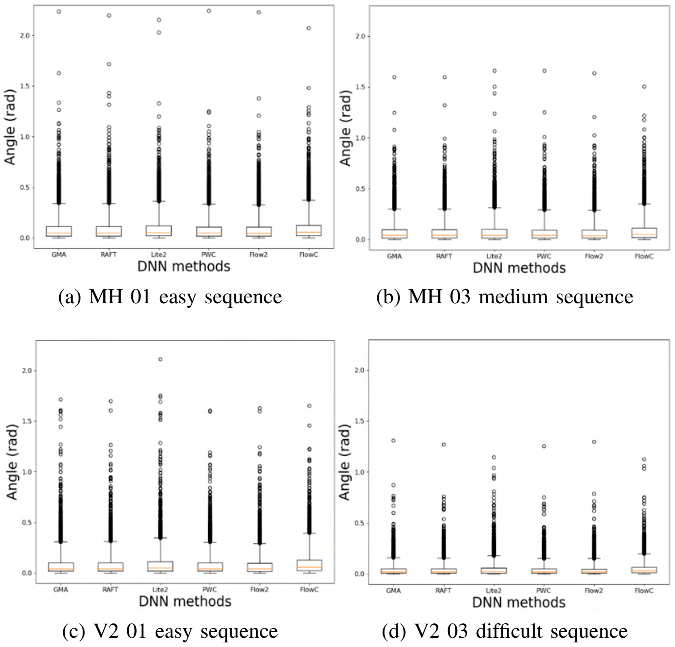
Fig. 3: Flow predictions and comparison results on Vicon Room of EuRoc dataset. The dense optical flow uses the flow field color encoding [14].

use the dense optical flow as a supporting sensor for visual inertial odometry. The straightforward application of optical flow in this application does only use the direction of the optical flow. This means that the angle error is much more important than the relative magnitude. Overall, the evaluated methods perform well, but there are also a significant number of outliers. Hence, proper outlier detection and removal will be necessary when using the optical flow in any navigation solution.

The proposed method to evaluate the different dense optical methods using sparse optical flow as ground truth seems to work. However, there are situations where it could be argued that the feature points are incorrectly associated, resulting in strange ground truth. More work is needed to eliminate these points. Furthermore, using sparse ground truth provides less information, as it is impossible to say anything about the optical flow in between the points in the sparse optical flow points, but at the same time it adds realism to the evaluation. A combination of both artificial dense ground truth and real sparse ground truth is to be preferred.

IV. Conclusions

Optical flow can be a vital component when designing GNSS independent navigation solutions. In this paper, different state-of-the-art learning based dense optical flow methods have been evaluated, and the results have been compared to those obtained with sparse feature-based optical flow methods. In the process, a method utilizing sparse optical flow to obtain ground truth for the comparison, has been proposed. Overall, the dense methods provide results similar to the sparse method, and no specific method stands out as significantly better than any other. The quality of the dense optical flow would seem to make the estimates suitable for supporting measurement in GNSS independent navigation solutions but will require careful outlier handling.



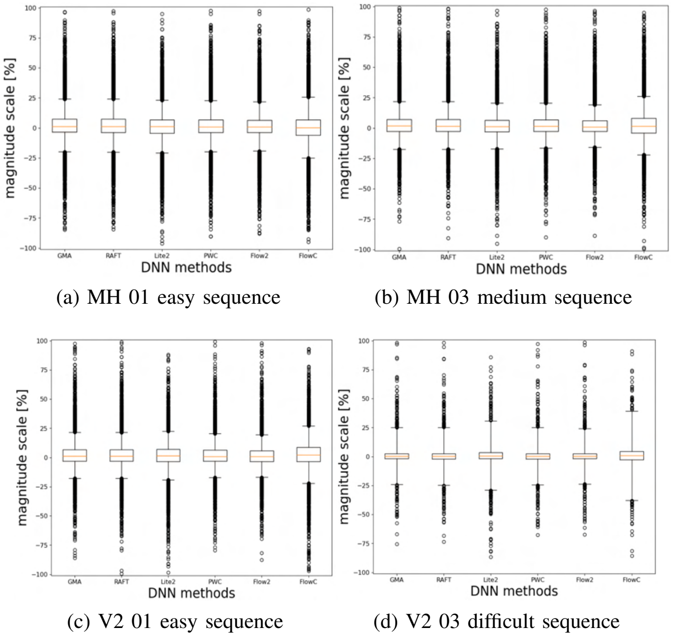
(a) MH 01 easy sequence

(b) MH 03 medium sequence

(c) V2 01 easy sequence

(d) V2 03 difficult sequence

Fig. 4: Angle comparison of all image frames between feature-based and learning-based methods. The default value of whisker is 1.5 corresponds to Tukey's original definition of boxplots.



(a) MH 01 easy sequence

(b) MH 03 medium sequence

(c) V2 01 easy sequence

(d) V2 03 difficult sequence

Fig. 5: Magnitude comparison of all image frames between feature-based and learning-based methods. The default value of whisker is 1.5 corresponds to Tukey's original definition of boxplots.

Future work includes exploiting the computed dense optical flow to aid inertial navigation systems in visual-inertial odometry type of system.

References

[1] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–110, 2006.

[2] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, 2004, pp. I–I.

[3] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular slam with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[5] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[6] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679.

[7] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2758–2766.

[8] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2462–2470.

[9] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

[10] T.-W. Hui, X. Tang, and C. C. Loy, "LiteFlowNet: A lightweight convolutional neural network for optical flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8981–8989.

[11] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proceedings of the 16th European Conference on Computer Vision–ECCV*, Glasgow, UK, Aug. 2020, pp. 402–419.

[12] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9772–9781.

[13] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[14] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, pp. 1–31, 2011.

[15] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[16] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, 2012, pp. 611–625.

[17] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS 2017 Workshop on Autodiff*, 2017.

[19] T.-W. Hui, X. Tang, and C. C. Loy, "A lightweight optical flow CNN—revisiting data fidelity and regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2555–2569, 2020.