Robotics Project: Structure From Motion: Accurate Optical Flow vs Active Vision

Feriel Amira

Supervisor: Aravind Battaje

***Abstract*— Structure from Motion (SfM) is a critical problem in computer vision that involves recovering the 3D structure of a scene from a set of 2D images captured from different angles. In this study, we compare the performance of two different approaches for SfM: deep learning-based optical flow and fixation-based active vision. We implement both methods and evaluate their efficacy in the same environment. For the optical flow approach, we tested four models: GMA, RAFT, PWC-NET, and IRR-PWC. We selected the model that provided the most accurate flow estimation as the primary representative for the rest of the experiments. Our results show that deep learning-based optical flow outperformed fixation-based active vision for estimating the 3D structure of distant targets, while fixation was more effective for closer targets.**

I. Introduction

Estimating the depth of objects in a scene is a fundamental problem in computer vision with practical applications such as robotics, autonomous driving, and 3D reconstruction. One traditional approach to estimate depth is through Structure from Motion (SfM), which uses the motion of a camera to reconstruct the 3D structure of a scene. As the camera moves, the objects that are closer appear to move faster than those that are farther away. This phenomenon, known as motion parallax, creates a relationship between the camera velocities and the apparent image velocities that can be exploited to estimate the depth of objects in the scene.

Optical flow methods are used to estimate the motion of pixels in the image across two neighboring frames of a video. However, traditional optical flow methods used in naive SfM often struggle with the aperture problem, resulting in distance overestimates due to ambiguity in local neighborhoods. To overcome this limitation, two pathways are of interest: we can either combine deep learning with optical flow algorithms, benefiting from its capability to automatically learn complex features and representations of data, or we can use active SfM, where we orchestrate the camera movements to fixate on one object at a time and simplify perception. This technique has been shown to provide superior depth recovery compared to naive SfM, as regularities in visual motion representation can help overcome the aperture problem [1]

This report compares state-of-the-art optical flow methods to estimate the apparent image velocity in the naive SfM approach and analyzes their performance in comparison to active SfM in terms of accuracy of distance estimation, runtime, and specifications. Our goal is to identify the strengths and limitations of both approaches and provide insights into which method may be more appropriate for different types of scenes or applications. Furthermore, we will discuss potential future developments for real-time depth estimation using optical flow and active vision.

II. Related work

This work mainly relates to fixation-based active vision, structure from motion and optic flow. In the following, we relate to these topics.

A. Structure from motion

Structure from Motion (SfM) is a technique used to recover the 3D structure of a scene from a series of 2D images captured by a camera. The process typically involves the following steps[2]:

- Feature extraction: Identify salient features (e.g., corners or edges) in each image Feature matching: Find corresponding features across the image sequence in order the estimation of the camera motion.
- Camera motion estimation: Estimate the motion of the camera between each pair of consecutive images.
- Triangulation: Compute the 3D coordinates of the matched feature points using the camera parameters and the corresponding image points.
- Bundle adjustment: Refine the camera parameters and 3D points by minimizing the reprojection error.

Structure from Motion with Optical Flow involves computing the 3D structure of a scene using the motion field between consecutive frames of a video sequence. [3] On the other hand, Structure from Motion from Fixation-Based Active Vision uses an active vision system to actively control the motion of a camera to achieve fixation on a point of interest, it does not require estimating the optic flow and relies fully on the proprioception of the robot. [1]

B. Optical Flow

Optical flow is a technique used in computer vision that estimates the apparent motion of image objects between two consecutive frames, caused by either the movement of the objects or the camera. The method relies on two assumptions: the pixel intensities of objects remain constant between consecutive frames, and adjacent pixels have similar motion.

One way to understand optical flow is to consider a pixel $I(x,y,t)$ in the first frame, which moves by a certain distance (dx,dy) in the next frame taken after a time dt. According to the first assumption, the intensity of these pixels remains the same:

$$I(x,y,t) = I(x+dx,y+dy,t+dt) \tag{1}$$

However, taking a Taylor series approximation of the right-hand side of the equation results in an underconstrained optical flow equation, leading to an ambiguity in determining the true motion of the pixels, known as the aperture problem.

To address this issue, traditional methods formulate an optimization function and rely on additional constraints to reduce the number of unknowns. The objective function is then optimized using numerical optimization techniques such as gradient descent or energy minimization. However, accounting for the complexity of the scene while maintaining an accurate prediction often leads to complex functions that are computationally expensive for real-time application. [4]

Deep learning-based optical flow methods, such as PWC-Net, can sidestep the need to formulate an optimization problem by training a convolutional neural network to predict flow directly. These methods have shown better generalization, improved accuracy, and faster processing times compared to traditional methods. However, deep learning methods are dependent on the quality and quantity of the labeled training dataset. They also come with a significant risk of overfitting due to the large number of parameters contained in the neural networks and may not always be adaptable across different datasets without retraining

This paper focuses on four state-of-the-art algorithms for deep learning based optical flow estimation : RAFT, GMA, PWC-Net, and IRR PWC-Net.

- The RAFT model employs per-pixel feature extraction and generates multi-scale correlation volumes for all possible pixel pairs. It also utilizes a recurrent unit to update the flow field by performing lookups on these correlation volumes. Despite its success, RAFT presents limitations such as difficulty at handling large displacements and ambiguity and the risk of overfitting due to the large number of parameters. [5]
- GMA overcomes these limitations by using self-attention mechanisms in motion aggregation and RNNs in motion feature extraction to model long-range dependencies and temporal dynamics of the video. However, it is effective only in situations where the flow of the attended areas is relatively consistent. [6]
- PWC-Net uses a deep CNN to extract features from input images, which are then processed at multiple scales in a pyramid structure to capture motion information at different levels of abstraction. The features are used to warp the second image to match the first image, producing a cost volume to calculate the optical flow. The final optical flow is estimated from the cost volume using a convolutional neural network. [7]
- IRR PWC-Net is a variation of PWC-Net that includes additional modules to handle occlusions and refine flow estimates. The refinement process uses iterative residual refinement blocks to compare the current flow estimate with the previous flow estimate and refine it based on the residual difference. This allows IRR-PWC to have less parameters than its previously mentioned counterparts and makes it therefor more adapted to real-time use and less prone to overfitting [8]

Please refer to their respective original papers for further details. Each of these algorithms has its own strengths and limitations, and the choice of algorithm depends on the specific application requirements.

C. Fixation-Based Active Vision

In their paper [1], Battaje and Brock propose a new approach to depth estimation that relies on the robot proprioception. In their paper, they show that gaze fixation on a target creates a coupling between the camera translational and rotational velocities: the camera must produce a rotation that counteracts the effect of translation to maintain an object of interest in the camera's field of view. The authors take advantage of the robustness achieved with fixation to generate accurate depth estimates of the point of interest.

In this paper, we will provide a comprehensive evaluation of these two methods in terms of accuracy, robustness, and computational efficiency, and build on [1]

III. Methodology

A. Depth Estimation From Camera Motion

Let $[X,Y,Z]$ be any 3D point in the camera frame. We are looking for the coordinate Z, in other words, the distance of the point to the camera. As described in [1] and [9], for a moving pinhole camera, the velocity of projected point $\begin{bmatrix}\dot{x} & \dot{y}\end{bmatrix}$ and its 2D position $\begin{bmatrix}x & y\end{bmatrix}$ on the image plane can be related to the camera velocity $\begin{bmatrix}v_x & v_y & v_z & w_x & w_y & w_z\end{bmatrix}$ in the world frame with the help of time derivative of projection equations:

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \begin{bmatrix}-\frac{f}{Z} & 0 & \frac{x}{Z} & \frac{xy}{f} & -\frac{f^2+x^2}{f} & y\\ 0 & -\frac{f}{Z} & \frac{y}{Z} & \frac{f^2+y^2}{Z} & -\frac{xy}{f} & -x\end{bmatrix}\begin{bmatrix}v_x\\ v_y\\ v_z\\ w_x\\ w_y\\ w_z\end{bmatrix} \tag{2}$$

where f is the focal length of the camera.

If the camera is translating laterally its rotational velocity is negligeable, 2 becomes:

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \frac{1}{Z}\begin{bmatrix}-f & 0 & x\\ 0 & -f & y\end{bmatrix}\begin{bmatrix}v_x\\ v_y\\ v_z\end{bmatrix} \tag{3}$$

The depth can then be directly recovered from the equation

If the camera is fixating on a point, it should maintain it at the center of its FOV at all times. We can then assume that: $\begin{bmatrix}\dot{x} & \dot{y}\end{bmatrix} \approx \begin{bmatrix}0 & 0\end{bmatrix}$ and $\begin{bmatrix}x & y\end{bmatrix} \approx \begin{bmatrix}0 & 0\end{bmatrix}$. 2 becomes:

$$Z = -\frac{v_x}{w_y} \qquad\qquad Z = \frac{v_y}{w_x} \tag{4}$$

Therefor, using fixation, the depth can be recovered using only robot propriocetion thanks to the coupling between rotation and translation.

B. Workflow of Depth Estimation

This section provides an overview of the workflow used to estimate the distance of the object of interest from the camera. Given a pair of images, our implementation follows the following steps:

1) Detects the 2D coordinates of the edges object on the right image. In this work, we consider the center of the object to portray 2D position of the object as a whole.
2) Performs optical flow estimation between the image pair using one of the pre-trained model discussed in II-B
3) Retrieve the camera properties and robot proprioception
4) Accumulates the optical flow of random pixels within the detected object, and uses the interquartile range method to remove outliers [10]. Then, it computes the mean optical flow of the object using the cleaned data.
5) Estimates the distance between the camera and the object using the equations 3 and 4 depending on the camera movement.

In case of fixation, steps 2 and 3 can be skipped as depth estimation solely relies on the robot proprioception to estimate the depth.

IV. Experimental Setup

The setup consists of a Franka Emika Robot equipped with a RealSense D435 camera rigidly attached to the end-effector. The robot's movements were controlled with the servo laws presented in [1] and executed at a rate of 1000 Hz then recorded as rosbag trials .

Our distance estimation is performed at a rate of approximately 30 Hz on the recorded trials. Each trial involved moving the camera in front of a target object while maintaining a constant distance to it. The target object was a printed Aruco marker with two possible orientations: straight or tilted. We intentionally tilted the marker to add complexity to the scene and to evaluate the methods' robustness to motion ambiguities. Additionally, we printed the marker at different sizes and placed it at varying distances from the camera. The target distance ranged over two orders of magnitude, from 1 m to 5 m. At each target distance, the robot end-effector either used fixation (cyclic translation around the target) or translation-only motion. To evaluate the accuracy of our distance estimation, we report the mean and standard deviation of the error between ground truth and our estimated distance, accumulated over approximately 240 samples at each target distance. For optical flow estimation and Aruco marker detection, we used the MMflow [1] and the MMdetection[2] libraries.

To reproduce the results, please refer to the implementation on Github. All experiments are run on the Python 3 Google Compute Engine backend (GPU) offered by Google Colab[3].

[1]https://mmflow.readthedocs.io/
[2]https://mmdetection.readthedocs.io/
[3]https://colab.research.google.com/

V. Experimental Results

A. IRR-PWC offers the best trade-off

The results presented in Figure 1 demonstrate that the IRR-PWC algorithm outperforms PWC-NET, GMA, and RAFT in terms of accuracy. Notably, the optical flow estimations produced by GMA and RAFT tend to overestimate the distance significantly. While this could partially be due to differences in training datasets, as shown in Figure 2, it is more likely that overfitting is the primary contributing factor to this phenomenon. Overfitting arises when a model is trained on complex examples that contain patterns irrelevant to simpler examples, and the high number of parameters in optical flow algorithms like RAFT, GMA, and PWC-NET could exacerbate this problem.

Additionally, Table 2 reveals that IRR-PWC offers a favorable trade-off between accuracy and runtime, indicating that it is an efficient algorithm for real-time optical flow estimation. Therefore, we will use IRR-PWC for the remainder of the experiments.

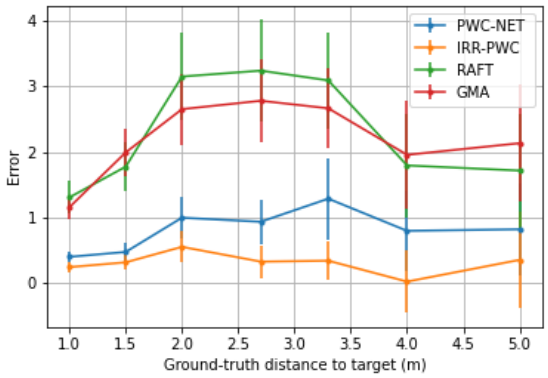

Fig. 1: Error and standard deviation of distance estimation of an Aruco marker from the camera at varying distances and with different deep learning based optic flow algorithms

	Dataset	Error	Confidence	Runtime
IRR-PWC	FC+FT+ STF	0.13	1.127	41.16
PWC-NET	FC+FT+ STF	0.62	1.49	14.1
GMA	FC+FT+ST	2.03	2.21	51.2
RAFT	FC+FT+ST	1.68	2.4	46.7
Fixation	-	-0.225	0.71	5.1

Fig. 2: The average of the accumulated errors, standard deviations and runtimes over the different trials of different deep learning based optic flow algorithms pre-trained on different datasets. FC: FlyingChairs, FT:FlyingThings3D, ST: Sintel, STF: Sintel Final. The error and confidence are measured in meter

B. Optical flow and motion ambiguity

The estimation of optic flow is intuitively simpler for a straight object compared to a tilted object. This is attributed to the greater symmetry of a straight object, which facilitates the visual system's processing. In the present study, we introduce complexity by tilting the object to assess the model's ability to estimate depth. Our experimental results, as illustrated in Figure 5, demonstrate that the model performs

Fig. 3: Optical flow estimated by IRR-PWC. On the right, optical flow of a straight marker. On the left: optical flow of a tilted marker

better in estimating depth for a tilted marker, in contrast to the aforementioned expectation. To further substantiate our findings, we have included snapchots of the estimated flow fields in 3 that depicts the optic flow for an oblique marker and a straight one. It appears that the target feature, namely the Aruco marker is more distinguishable when it is tilted. This could be due to the fact in general, a tilted marker produces more image motion than straight one, making it appear closer even when moving at the same speed from the same distance. As a result, the deep learning-based optic flow algorithm may assign a higher magnitude to the image motion of the titled object, which results in a smaller distance between the target and the camera.

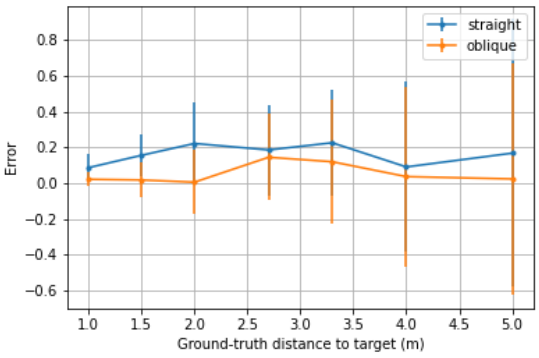

Fig. 4: Error and standard deviation of distance estimation of an a straight vs tilted aruco marker at varying distances and with IRR-PWC

C. Structure from motion from optical flow vs fixation

The graph reveals that fixation outperforms naive Sfm with optic flow for shorter distances, while deep learning-based optic flow outperforms fixation for a more distant target. This can be attributed to the fact that, with fixation, the accuracy of distance recovery is contingent upon the amount of camera rotation, which is reduced for objects located farther away. Thus, the model becomes sensitive noisy measurements small camera rotations. On the other hand, the pyramidal structure of IRR-PWC allows it to handle objects of different sizes and granularity. Based on these observations, it can be inferred that optic flow is a better approach for estimating the 3D structure of targets that are located at significant distances from the camera, while fixation is a more effective method for estimating the distance to targets that are situated closer to the camera. Furthermore, fixation has a faster runtime making it more suitable for real-time applications.

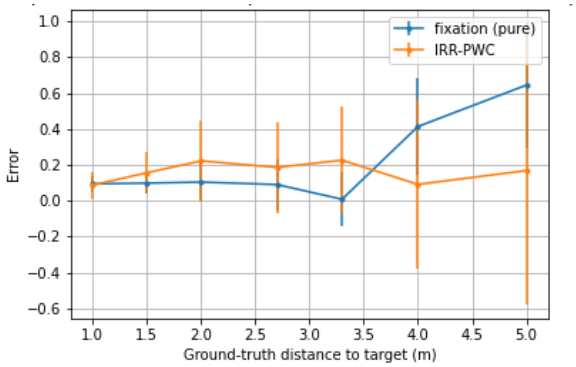

Fig. 5: Error and standard deviation of distance estimation of an a aruco marker at varying distances and with fixation vs optic flow. For visual clarity, the sign of error for distance estimates under fixation was flipped

VI. Conclusion

Overall, our study demonstrated that IRR-PWC is a promising algorithm for estimating 3D distances, providing a superior trade-off between accuracy and runtime compared to other state-of-the-art methods. Our experiments also showed that the model can effectively estimate depth for complex shapes, such as tilted markers. Notably, deep learning-based optic flow outperformed fixation for distant targets, while fixation was more effective for closer targets. These findings suggest that optic flow is a more suitable approach for estimating the 3D structure of distant targets, where fixation may not be feasible due to the limited amount of camera rotation.

In contrast, fixation is a faster method and can provide accurate depth estimates for closer targets. It's important to note that our study did not address challenging scenarios such as occlusions and non-rigid motion, which may require the use of more sophisticated algorithms. Investigating these aspects in future research would provide a more comprehensive assessment of the capabilities of deep learning-based optical flow. Moreover, integrating different types of camera movements could further enhance the accuracy and efficiency of depth estimation. For instance, a hybrid approach that combines fixation and optical flow could be used, where the camera fixates on nearby targets and switches to translation movement for distant targets. However, developing a mechanism to automatically determine when to switch between fixation and optical flow would be a challenging task.

REFERENCES

[1] A. Battaje and O. Brock, "One object at a time: Accurate and robust structure from motion for robots," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, oct 2022. [Online]. Available: https://doi.org/10.11092Firos47612.2022.9981953

[2] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion." *Acta Numerica*, vol. 26, p. 305–364, 2017.

[3] W. A. Simpson, "Optic flow and depth perception," *Spatial Vision*, vol. 7, no. 1, pp. 35–75, 1993.

[4] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Comput. Surv.*, vol. 27, no. 3, p. 433–466, sep 1995. [Online]. Available: https://doi.org/10.1145/212094.212141

[5] Z. Teed and J. Deng, "RAFT: recurrent all-pairs field transforms for optical flow," *CoRR*, vol. abs/2003.12039, 2020. [Online]. Available: https://arxiv.org/abs/2003.12039

[6] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. I. Hartley, "Learning to estimate hidden motions with global motion aggregation," *CoRR*, vol. abs/2104.02409, 2021. [Online]. Available: https://arxiv.org/abs/2104.02409

[7] D. Sun, X. Yang, M. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," *CoRR*, vol. abs/1709.02371, 2017. [Online]. Available: http://arxiv.org/abs/1709.02371

[8] J. Hur and S. Roth, "Iterative residual refinement for joint optical flow and occlusion estimation," *CoRR*, vol. abs/1904.05290, 2019. [Online]. Available: http://arxiv.org/abs/1904.05290

[9] F. Chaumette and S. Hutchinson, "Visual servo control. i. basic approaches," *IEEE Robotics Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[10] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in *Information and Decision Sciences*, S. C. Satapathy, J. M. R. Tavares, V. Bhateja, and J. R. Mohanty, Eds. Singapore: Springer Singapore, 2018, pp. 511–518.