Motor Focus: Fast Ego-Motion Prediction for Assistive Visual Navigation

Hao Wang[†,1], Jiayou Qin[†,2], Xiwen Chen[1], Ashish Bastola[1], John Suchanek[1], Zihao Gong[3], and Abolfazl Razi[*,1]

[1]*School of Computing, Clemson University*
[2]*Department of Electrical and Computer Engineering, Stevens Institute of Technology*
[3]*School of Cultural and Social Studies, Tokai University*

***Abstract*—Assistive visual navigation systems for visually impaired individuals have become increasingly popular thanks to the rise of mobile computing. Most of these devices work by translating visual information into voice commands. In complex scenarios where multiple objects are present, it is imperative to prioritize object detection and provide immediate notifications for key entities in specific directions. This brings the need for identifying the observer's motion direction (ego-motion) by merely processing visual information, which is the key contribution of this paper. Specifically, we introduce Motor Focus, a lightweight image-based framework that predicts the ego-motion —the humans' (and humanoid machines') movement intentions based on their visual feeds, while filtering out camera motion without any camera calibration. To this end, we implement an optical flow-based pixel-wise temporal analysis method to compensate for the camera motion with a Gaussian aggregation to smooth out the movement prediction area. Subsequently, to evaluate the performance, we collect a dataset including 50 clips of pedestrian scenes in 5 different scenarios. We tested this framework with classical feature detectors such as SIFT and ORB to show the comparison. Our framework demonstrates its superiority in speed ($> 40\mathrm{FPS}$), accuracy ($\mathrm{MAE} = 60\mathrm{pixels}$), and robustness ($\mathrm{SNR} = 23\mathrm{dB}$), confirming its potential to enhance the usability of vision-based assistive navigation tools in complex environments. The code is publicly available at https://arazi2.github.io/aisends.github.io/project/VisionGPT.**

***Index Terms*—Assistive Visual Navigation, Motion Analysis, Vision Enhancement, Image Processing**

I. Introduction

The rapid development of mobile computing has significantly enhanced real-life applications. Technologies including object detection, augmented reality (AR), and assistive visual navigation, have benefited immensely from integrating AI into mobile devices. As reliance on these digital experiences increases, ensuring user safety through visual assistive technologies has become a critical priority. [1]–[3].

Motion analysis plays a critical role in assistive visual navigation by utilizing both spatial and temporal information to create a comprehensive visual understanding of dynamic surroundings. While existing motion analysis methods have been extensively developed for autonomous driving, their direct use in mobile devices faces significant challenges, including the unpredictable nature of human movement, limited processing capabilities of portable devices, and the need for directional responsiveness due to the limited perception capability of humans. Additionally, human users typically carry devices that are constrained by size, weight, and computational capabilities, limiting the use of intensive real-time processing methods such as Simultaneous Localization and Mapping (SLAM) that are commonplace in autonomous systems [4], [5].

Due to these limitations, there is a pressing need to develop a navigation aid customized for humans. While numerous studies have only focused on visual attention, movement analysis is also vital, as the visual focus and body movement can often diverge significantly [6]. Notably, works in visual-based movement prediction in point-of-view camera settings are heavily missing, as vehicle-based research mostly focuses on visual odometry functionality on ego-location recording, while pedestrian-oriented research focuses more on visual attention and head direction prediction [7], [8].

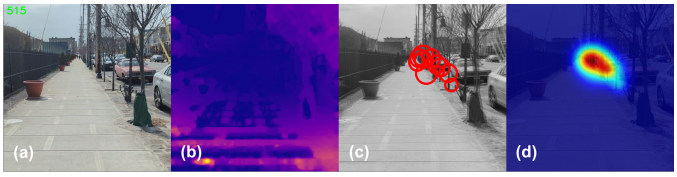

Fig. 1. Motor focus visualization, (a) is the raw RGB image, (b) is the compensated optical flow map, (c) shows the identified attention points of 10 consecutive frames, (d) is the attention map aggregated by the Gaussian distributions of attention points from (c).

To bridge this gap, we present **Motor Focus**, a novel framework for predicting how users physically move and orient themselves in space. Specifically, we introduce an optical flow-based pixel-wise temporal analysis that can predict the movement direction of users and simultaneously filter out the unintended and noise-like camera motion without any camera calibration. We also combined the Gaussian aggregation method to smooth out the projected movement attention area to address the camera shake issue in pedestrian applications. Then, to obtain the transform matrix from two consecutive frames, we apply Singular Vector Decomposition (SVD) instead of classical feature mapping to reduce the computation load. Finally, we validate the proposed framework using our self-collected visual navigation-oriented dataset.

† Equal contribution
* Corresponding author

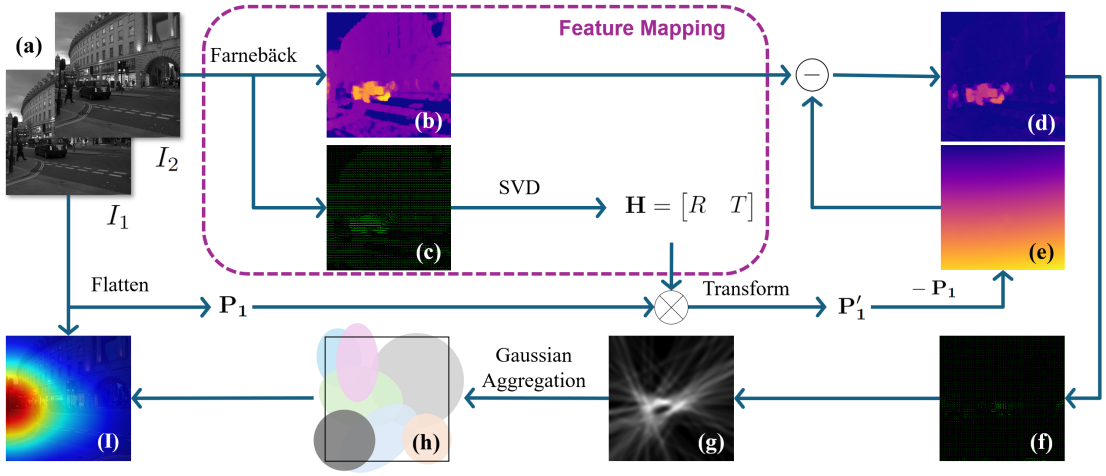

Fig. 2. The proposed framework, (a) is a two consecutive frame pair, (b) is the original optical flow map (magnitude), (c) is the original optical flow field (vector), (d) is the compensated optical flow map, (e) is the camera motion ϵ, (f) is the compensated optical flow field, (g) is the probability map of attention point for I_2, (h) is the aggregated gaussian distribution of attention points from (g), and (i) is the attention map for motor focus of frame I_2.

II. Methodology

A. Pre-Processing

The optical flow, i.e. the motion between matched points across frames, is calculated using Farneback's method to provide a dense flow field, highlighting the texture and movement patterns within the scene.

The optical flow field $\mathbf{f} \in \mathbb{R}^{H\times W\times 2}$ is computed from two consecutive frames $I_1 \in \mathbb{R}^{H\times W}$ and $I_2 \in \mathbb{R}^{H\times W}$ using classical Farneback's [9] method:

$$\mathbf{f} = \text{Farneback}(I_1, I_2) \tag{1}$$

where $\mathbf{f}_{x_i,y_i} = (dx_i, dy_i)$ represents the displacement vector at position (x_i, y_i) in image I_1 that aligns with I_2.

B. Ego-Motion Estimation using SVD

In our proposed framework, we use all pixels directly from the gray-scale image $I_1 \in \mathbb{R}^{H\times W}$. Specifically, $\mathbf{P}_1 \in \mathbb{R}^{(H\times W)\times 2}$ is directly mapped from I_1, denoted as $\mathbf{P}_1 = \{\mathbf{p}_1^1, \mathbf{p}_2^1, \ldots, \mathbf{p}_N^1\}$, where N is the number of keypoints and is equal to the total pixel number of I_1. For each keypoint $\mathbf{p}_i^1 = (x_i^1, y_i^1)$ in $\mathbf{P}_1$, we find its corresponding point in I_2 using the displacement (dx_i, dy_i) obtained from the optical flow field $\mathbf{f}$ from the previous step:

$$\mathbf{p}_i^2 = (x_i^1 + dx_i^1, y_i^1 + dy_i^1), i = 1, 2, \ldots, N \tag{2}$$

This process ensures that $\mathbf{P}_2 \in \mathbb{R}^{(H\times W)\times 2}$ contains keypoints that are spatially aligned with $\mathbf{P}_1$ in I_2.

Singular Value Decomposition (SVD) is then used to compute an optimal rigid transformation (rotation R and translation T) that best aligns these points [10]. Specifically, the cross-covariance matrix M is computed:

$$M = \mathbf{P}_2^T \mathbf{P}_1 = U\Sigma V^T \tag{3}$$

The rotation matrix R and the translation vector T are derived from the SVD components to form the transformation matrix $\mathbf{H}$:

$$R = UV^T, \quad T = \text{mean}(\mathbf{P}_2) - R \times \text{mean}(\mathbf{P}_1) \tag{4}$$

Where $\text{mean}(\mathbf{P}_2)$ and $\text{mean}(\mathbf{P}_2)$ are the centroid of each keypoints set. The transformation matrix $\mathbf{H}$ is then assembled as:

$$\mathbf{H} = [R \quad T] \tag{5}$$

C. Ego-Motion Compensation

The transformation matrix $\mathbf{H} \in \mathbb{R}^{2\times 3}$ is then applied to a grid of pixel coordinates $\mathbf{P_1}$ from I_1, representing the original positions. The grid modified by the flow $\mathbf{f}$ gives the new positions $\mathbf{P_1}$.

Applying $\mathbf{H}$ to $\mathbf{P_1}$ provides a prediction of where each grid point would be if only the camera's motion (ego-motion) affected it:

$$\mathbf{P'_1} = \mathbf{P_1} \times \mathbf{R} + \mathbf{T} \tag{6}$$

The residual term $\epsilon \in \mathbb{R}^{H\times W\times 2}$ of this study is defined by the difference between the predicted positions of the pixels due to the camera motion and the predicted positions of the pixels without camera motion, which represents the displacement caused exclusively by the camera's motion, ignoring any independent object movements:

$$\epsilon = reshape(\mathbf{P'_1} - \mathbf{P_1}) \tag{7}$$

The compensated optical flow then can be derived by correcting the raw optical flow $\mathbf{f}$ for the motion attributable to the observer's (camera's) own campaign. This correction ensures that $\mathbf{f}'$ predominantly reflects the motion of objects relative to the observer, rather than due to the observer's motion. This is expressed mathematically as:

$$\mathbf{f}' = \mathbf{f} - \epsilon \tag{8}$$

Here, $\mathbf{f}'$ represents the motion vectors corrected for camera motion, highlighting only those movements that are due to objects moving in the scene rather than the camera itself.

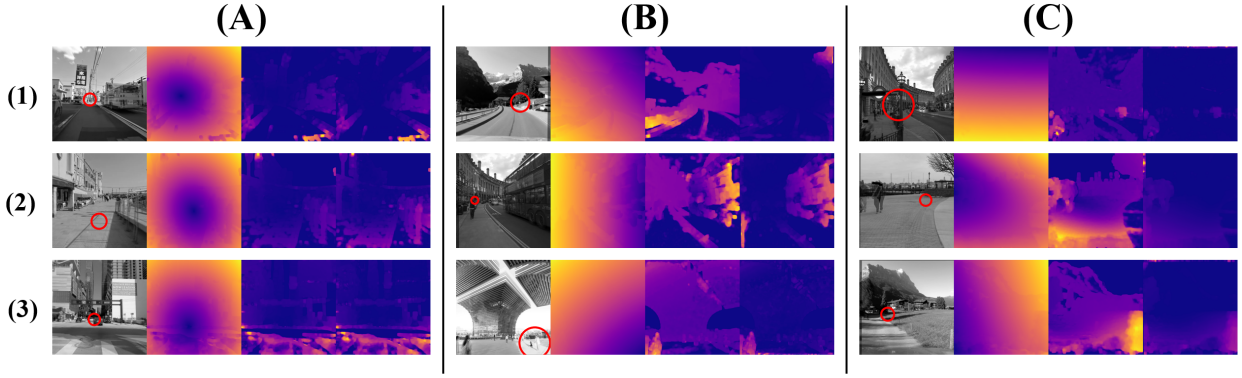

Fig. 3. Visualization of ego-motion compensation, each image consists of four cells, from left to right: grayscale image with predicted moving direction, the magnitude of camera motion ϵ (ego-motion), raw optical flow, and optical flow with ego-motion compensation.

D. Project Movement Direction

Given the initial set of vectors $\mathbf{v}_j = \mathbf{p}_1^j + \mathbf{f}'_j$, assuming that each $\mathbf{v}_j$ potentially indicates a trajectory toward a potential focus point, the goal is to find a point $\mathbf{c}$ that most of these vectors converge toward:

$$\min_{\mathbf{c}} \sum_j \|\mathbf{c} - \text{proj}_{\mathbf{v}_j}(\mathbf{c})\|^2 \tag{9}$$

However, for complex scenes, there might be multiple focuses, which shift the point $\mathbf{c}$ from the optimum location. To measure how well a point $\mathbf{c}_k$ (a candidate point from the k-th cluster) aligns with the vectors $\mathbf{v}_j$, we formulate an optimization problem to minimize these deviations for each cluster. Specifically, for each candidate focus $\mathbf{c}_k$:

$$\mathbf{c}_k = \sum_{j=1}^{M} \frac{\left\|\mathbf{c}_k - \text{proj}_{\mathbf{v}_j}(\mathbf{c}_k)\right\|^2}{M}, j \in \text{cluster}_k \tag{10}$$

Where $\text{proj}_{\mathbf{v}_j}(\mathbf{c}_k)$ is the projection of $\mathbf{c}_k$ onto the line defined by $\mathbf{v}_j$.

Then we calculate the score of each candidate focus $\mathbf{c}_k$:

$$\text{Score}(\mathbf{c}_k) = \sum_{j=1}^{M} j, j \in \text{cluster}_k \tag{11}$$

Select the $\mathbf{c}_k$ with the highest score, which indicates the maximum number of vectors $\mathbf{v}_j$ converging towards it.

E. Attention Area Smoothing

To stabilize the movement attention area across frames and reduce the effect of transient shaking or focus shifts, the Gaussian aggregation method is applied at each frame's focus point $\mathbf{c}_i$ to create a smooth attention mask over multiple frames. The spread of each Gaussian is determined by the inverse of the mean magnitude of the optical flow field, $\mathbf{U}_i$, representing the activity level or movement intensity in the frame. The formulation can be described as follows:

$$K_{\text{new}}(x,y) = \sum_{i=1}^{n} K_i(x,y|\mathbf{c}_i,\sigma_i), \quad \sigma_i \propto \frac{1}{U_i}, \tag{12}$$

$$U_i = \sqrt{u_1^2 + u_2^2 + ... + u_N^2} \tag{13}$$

Here, $K_i(x,y|\mathbf{c}_i,\sigma_i)$ is the Gaussian distribution from the i-th frame with its standard deviation σ_i, and n is the number of distributions (frames) considered.

The final Gaussian aggregation mask effectively smoothies and stabilizes the attention area by filtering out the ambiguous focus points due to the strong camera motion, as shown in Figure 1

III. Experiments

A. Camera-Motion Compensation

To evaluate the compensation performance of the proposed method, we compare the vanilla dense-optical flow and the proposed compensated optical flow in a series of scenes to showcase its capability to capture camera motion.

As shown in Figure 3, the compensated optical flow can filter the motion caused by camera shifting, distinguishing the relatively moving objects. Meanwhile, nearby objects' moving speed and direction can also be estimated from the compensated optical flow map.

More importantly, Figure 3 indicates that the visual focus can differ greatly from the motor focus. For instance, in Figure 3 group B, the compensated optical flow and camera motion indicated the camera moving potential is different from the actual body movement. In B1 and B3, the camera moves toward the left corner, while the user moves toward the right side. In B2, both the camera and the user move toward the left. In Figure 3 group C, the camera in both C2 and C3 are turning right, while the user in C2 is moving right and the user in C3 is moving left. In C1, the user stands statically, while the camera motion suggests that the camera is slowly pitching up. All results are proved in the actual video footprints.

Interestingly, when the camera moving direction and the user movement are aligned straightforwardly, the vanilla optical flow and the compensated optical flow become identical, and the camera motion map tends to overlap with the motor focus area, as shown in Figure 3 group A.

B. Ego-Motion Prediction

To test the performance of our framework, we collected a dataset that is specialized for visual navigation. In advance, each video clip is observed by three researchers frame by

frame, and a pixel location (x, y) of moving direction is annotated for each frame. Figure 4 (a) shows the sample of the collected dataset, where colored points represent the annotation from different researchers. The ground truth is calculated by the average of three different pixel locations.

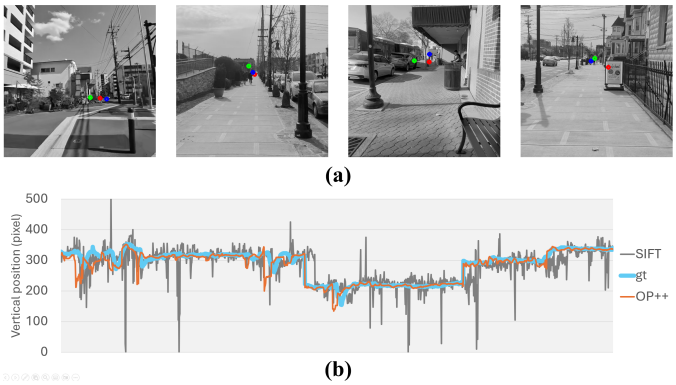

Fig. 4. (a) The samples of the collected dataset. (b) Plot of the focus points (vertical position $y_i \in \mathbf{c_i}$) on the selected scene.

We applied the proposed method to predict the center of moving direction and we used Mean Absolute Error (MAE), and Mean Squared Error (MSE) to compare the predicted location $(\hat{x}, \hat{y})$ with the annotated ground truth(x, y). Furthermore, we used the Signal-to-Noise Ratio (SNR) to compare the vertical motion $y_i \in \mathbf{c_i}$ along the time in a selected scene.

Our framework seamlessly works with other feature detectors, such as Scale-Invariant Feature Transform (**SIFT**) and Oriented FAST and Rotated BRIEF (**ORB**), and optical flow methods such as the Lucas-Kanade (**LK**) method, by simply replacing the Feature Mapping module of the proposed framework, as shown in Figure 2. We show the comparison between these methods in Table I, where $\mathbf{OP}$ represents the vanilla dense-optical flow method (Farneback), $\mathbf{OP}+$ represents the $\mathbf{OP}$ method with camera motion compensation, and $\mathbf{OP}++$ represents compensated $\mathbf{OP}$ with the Gaussian aggregation. All frames are resized to 512×512 for experiments.

Notably, the matching time ($< 1ms$) and FPS (> 40) of all three $\mathbf{OP}$ methods stand out from other classical methods as we directly use the optical flow field $\mathbf{f}$ for feature mapping. By taking advantage of SVD, the linear operation processes a total of $\mathbf{262,144}$ elements (all pixels) without additional computation cost. Meanwhile, the camera motion compensation helps to approach lower predicting error in $\mathbf{OP}+$ compared to vanilla optical flow ($\mathbf{OP}$). Furthermore, the proposed $\mathbf{OP}++$ method achieves even better results. Specifically, it obtained the lowest prediction error of motor focus in both MAE and MSE. More importantly, the Gaussian aggregation helps smooth the prediction especially the vertical location of the motor focus point $\mathbf{c_i}$. Figure 4 (b) confirms the accuracy (better alignment with ground truth) and stability (fewer fluctuations) of our method ($\mathbf{OP}++$) compared with the SIFT.

TABLE I
PERFORMANCE COMPARISON

Feature Detection Method	Matching Time (ms)	Total Time (ms)	FPS	MAE	MSE (x1000)	SNR (dB)
LK	4.86	27.60	36.24	103.89	11.88	16.47
ORB	5.34	26.76	37.36	112.35	11.95	16.44
SIFT	35.49	58.31	17.15	93.34	8.40	18.45
OP	0.91	19.38	51.59	108.21	11.37	14.27
OP+	0.91	22.49	44.47	90.17	8.82	19.45
OP++	**0.91**	23.45	42.64	**60.66**	**4.26**	**23.09**

IV. CONCLUSION

This study presents the **Motor Focus** — a novel image-based framework for motion analysis, specifically designed for predicting ego motion for pedestrians. By utilizing camera-motion compensation with Gaussian aggregation, our approach effectively tackles the camera shaking challenge, enhancing movement prediction accuracy and stability. Our experimental results, both qualitative and quantitative, validate the superiority of our method over classical feature detectors, especially in accuracy and computation cost. Our method can be utilized by blind assistive navigation tools to prioritize notifications based on the detected objects' alignment with ego-motion direction.

REFERENCES

[1] A. Kuzdeuov, S. Nurgaliyev, and H. A. Varol, "Chatgpt for visually impaired and blind," *Authorea Preprints*, 2023.

[2] F. Al-Muqbali, N. Al-Tourshi, K. Al-Kiyumi, and F. Hajmohideen, "Smart technologies for visually impaired: Assisting and conquering infirmity of blind people using ai technologies," in *2020 12th Annual Undergraduate Research Conference on Applied Computing (URC)*. IEEE, 2020, pp. 1–4.

[3] B. Li, J. P. Munoz, X. Rong, Q. Chen, J. Xiao, Y. Tian, A. Arditi, and M. Yousuf, "Vision-based mobile indoor assistive navigation aid for blind people," *IEEE transactions on mobile computing*, vol. 18, no. 3, pp. 702–714, 2018.

[4] V. Usenko, J. Engel, J. Stückler, and D. Cremers, "Reconstructing street-scenes in real-time from a driving car," in *2015 International Conference on 3D Vision*. IEEE, 2015, pp. 607–614.

[5] Y. Liao, J. Xie, and A. Geiger, "Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3292–3310, 2022.

[6] Y. Tamaru, Y. Ozaki, Y. Okafuji, J. Nakanishi, Y. Yoshikawa, and J. Baba, "3d head-position prediction in first-person view by considering head pose for human-robot eye contact," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 1064–1068.

[7] A. Makrigiorgos, A. Shafti, A. Harston, J. Gerard, and A. A. Faisal, "Human visual attention prediction boosts learning & performance of autonomous driving agents," *arXiv preprint arXiv:1909.05003*, 2019.

[8] M. Liu, S. Tang, Y. Li, and J. M. Rehg, "Forecasting human-object interaction: joint prediction of motor attention and actions in first person video," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 704–721.

[9] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 2003, pp. 363–370.

[10] H. Wang, X. Chen, A. Razi, and R. Amin, "Fast key points detection and matching for tree-structured images," in *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2022, pp. 1381–1387.