MOTOR FOCUS: EGO-MOTION PREDICTION WITH ALL-PIXEL MATCHING

Hao Wang †
School of Computing
Clemson University
Clemson, SC, USA
hao9@g.clemson.edu

Jiayou Qin †
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ, USA
jqin6@stevens.edu

Xiwen Chen
School of Computing
Clemson University
Clemson, SC, USA
xiwenc@g.clemson.edu

Ashish Bastola
School of Computing
Clemson University
Clemson, SC, USA
abastol@g.clemson.edu

John Suchanek
School of Computing
Clemson University
Clemson, SC, USA
jsuchan@g.clemson.edu

Zihao Gong
School of Cultural and Social Studies
Tokai University
Tokyo, Japan
0CPD1206@mail.u-tokai.ac.jp

Abolfazl Razi *
School of Computing
Clemson University
Clemson, SC, USA
arazi@clemson.edu

ABSTRACT

Motion analysis plays a critical role in various applications, from virtual reality (VR) and augmented reality (AR) to assistive visual navigation. Traditional self-driving technologies, while advanced, typically do not translate directly to pedestrian applications due to their reliance on extensive sensor arrays and non-feasible computational frameworks. This highlights a significant gap in applying these solutions to human users since human navigation introduces unique 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 range of humans.

In this project, we introduce an image-only method that applies motion analysis using optical flow with ego-motion compensation to predict **Motor Focus**—where and how humans or machines focus their movement intentions. Meanwhile, this paper addresses the camera shaking issue in handheld and body-mounted devices which can severely degrade performance and accuracy, by applying a Gaussian aggregation to stabilize the predicted motor focus area and enhance the prediction accuracy of movement direction. This also provides a robust, real-time solution that adapts to the user's immediate environment. Furthermore, in the experiments part, we show the qualitative analysis of motor focus estimation between the conventional dense optical flow-based method and the proposed method. In quantitative tests, we show the performance of the proposed method on a collected small dataset that is specialized for motor focus estimation tasks.

Keywords Image Process · Motion Analysis · Assistive Visual Navigation · Augmented Reality · Vision Enhancement

† Equal contribution

1 Introduction

The rapid development of mobile computing has significantly enhanced real-life applications through advanced visual technologies. Technologies such as object detection, augmented reality (AR), and visual navigation have benefited immensely from the integration of AI into mobile devices. Among these, visual navigation and related enhancements have become particularly prominent, as they enhance user experiences by blending digital elements with the physical world [1–4]. Consequently, as reliance on these visual technologies increases, ensuring user safety has emerged as a critical concern. Specifically, enhanced environment sensing can significantly improve safety by helping users detect nearby dangers and avoid potential emergencies, making these technologies indispensable in our increasingly digital world [2, 5–8].

Motion analysis plays a critical role in environmental sensing by utilizing both spatial and temporal information to create a comprehensive visual understanding of dynamic surroundings. By analyzing how objects and features move over time within a given space, motion analysis is employed in a suite of technologies to enhance autonomous systems and interactive applications [9, 10]. For instance, visual odometry [11, 12] and ego-motion estimation [13–15], rely on analyzing consecutive images captured by the camera and extracting features or keypoints that can be tracked across frames, determining the motion of a camera or a vehicle relative to its environment. Video saliency identifies the most visually conspicuous or attention-grabbing regions within a video sequence, which aims to predict where human observers are likely to focus their attention [16, 17]. Further, optical flow is often used for video stabilization [18], object detection [2, 19–22], image segmentation [23], and depth prediction [24, 25]. These technologies collectively augment this framework, allowing the system to effectively parse and interact with its surroundings by identifying and categorizing environmental elements, thereby ensuring safe and effective navigation. Furthermore, in the domains of healthcare and sports, motion analysis helps in monitoring human activities and movements, offering valuable insights for training, rehabilitation, and overall well-being assessments [26]. By effectively analyzing spatial and temporal data, motion analysis not only enhances environmental awareness but also significantly improves the interaction capabilities of various technology applications in real-world settings.

Figure 1: Concept of Assistive Visual Navigation

While motion analysis methods have been extensively developed for autonomous driving, their direct implant into mobile devices encounters significant challenges. Firstly, although mobile computing capabilities have rapidly advanced, the sensors integrated into mobile devices are typically less sophisticated and more constrained compared to those used in autonomous vehicles. For instance, autonomous vehicles are equipped with a range of high-end sensors such as LIDAR, radar, and multiple cameras that provide detailed, 360-degree environmental feedback [27]. In contrast, mobile devices generally rely on more basic components such as monocular cameras and inertial sensors, which offer less comprehensive data. Additionally, autonomous systems are designed to operate continuously and consume considerable power, which is feasible in a vehicle with a substantial power supply. However, for mobile devices, preserving battery life is critical, necessitating energy-efficient solutions [28]. The extensive processing power and energy requirements of autonomous driving technologies are thus incompatible with the limited capabilities and energy constraints of typical consumer mobile devices. Consequently, there is a pressing need for cost-effective, energy-efficient motion analysis solutions that benefit the hardware and energy limitations of mobile devices, ensuring that users can enjoy enhanced sensing capabilities without compromising device performance or battery life. 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 standard in autonomous systems [29].

Recognizing these limitations, there is a pressing need to develop a navigation aid from the perspective of pedestrians. While numerous studies have focused on visual attention, movement analysis is also vital, as the visual focus and body movement can often diverge significantly [30]. 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 human-based research more focuses on visual attention and head direction prediction [17, 31]. Motor focus can be useful across a wide array of applications including augmented reality (AR), visual navigation, photography, video stabilization, action cameras (such as 360-degree cameras), drones, and robotics [32]. These fields all benefit from understanding and predicting user movement and orientation, which can enhance interaction quality and system responsiveness. For instance, in AR and visual navigation, accurately predicting a user's movement can improve the alignment and relevance of augmented content, making the experience more intuitive and immersive. In the context of photography and video, understanding motor focus can lead to more effective stabilization techniques and dynamic framing methods. Similarly, in robotics and drone operations, anticipating human motion can enhance collaborative and interactive capabilities. This paper aims to explore motor focus extensively to fill this research void, proposing new methodologies and applications that leverage this underexplored aspect of human-machine interaction.

To bridge this gap, we present **Motor Focus** prediction, the study of how users physically move and orient themselves in space. Specifically, this study analyzes and predicts human movement patterns, particularly in scenarios captured from a first-person perspective using wearable or point-of-view cameras.

Specifically, this paper makes several contributions as follows:

- Motor Focus Prediction: We introduce a method to predict the moving direction of users by analyzing ego-motion. Utilizing this prediction, we project an attention area onto the captured image.
- Fast Ego-Motion Compensation: Our model processes video frames to display the optical flow. Specifically, we apply SVD to divide the transform matrix to boost the compensation for ego-motion.
- Specialized Dataset: To expand and evaluate this study, we have collected a small, specialized dataset aimed at fostering motor focus. By using this dataset to showcase our implemented methods, we demonstrate the practical application of motion information and attention mapping in improving navigational aids.

2 Methodology

This study employs a comprehensive video processing framework designed to enhance visual navigation by analyzing and visualizing motion dynamics in video sequences. The methodology encompasses several key phases, each crucial for extracting meaningful motion data from video frames to aid in navigation tasks.

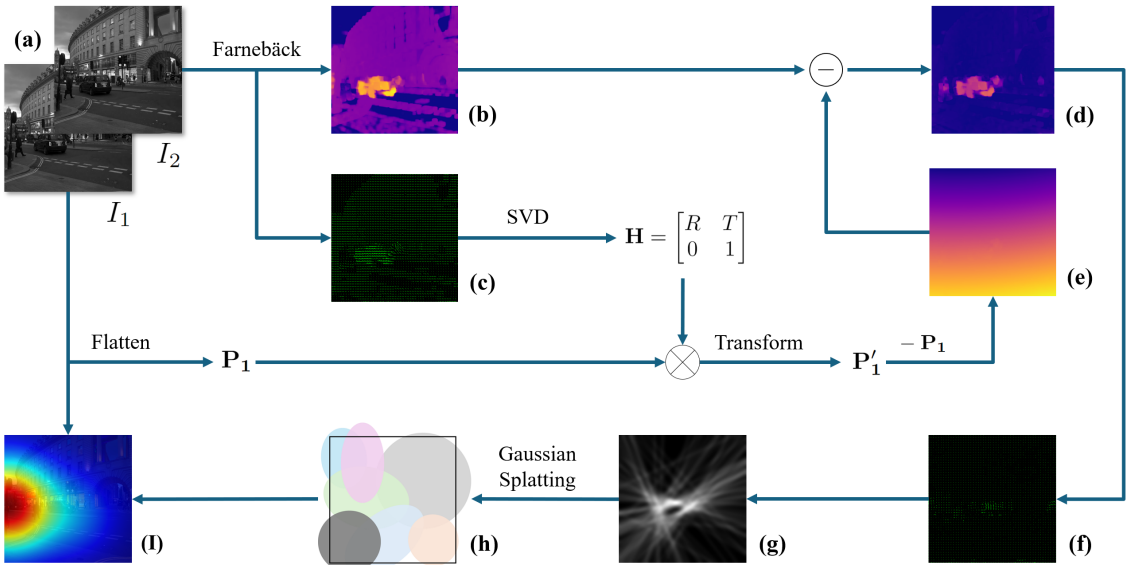

Figure 2: Framework, (a) is a two consecutive frame pair, (b) is the original optical flow, (c) is the original optical flow field, (d) is the compensated optical flow, (e) is the camera noise ϵ, (f) is the compensated optical flow field, (g) is the probability map of motor focus for I_2, (h) is the aggregated gaussian distribution of consecutive frames, and (i) is the attention map for motor focus of frame I_2.

The proposed project is fully open-sourced and available at: GitHub

2.1 Pre-Processing

The video capture process is initiated with continuous frame reading from a video stream. Initial frames undergo preprocessing, where they are resized and converted to grayscale to facilitate further analysis. The Scale-Invariant Feature Transform (SIFT) algorithm is implemented early to detect and compute key points and descriptors in the initial frame, establishing a baseline for subsequent frames. The optical flow, specifically 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.

Specifically, optical flow vectors (dx, dy) are calculated using the brightness constancy assumption between two consecutive frames I_1 and I_2:

$$I_1(x, y) \approx I_2(x + dx, y + dy)$$

The flow field $\mathbf{f}$ is computed using classical Farneback's [33] method:

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

where $\mathbf{f}(x, y) = (v_x, v_y)$ represents the displacement vector at position (x, y) in image I_1 that aligns with I_2.

The optical flow field $\mathbf{f}$ between two consecutive images I_1 and I_2 provides a dense vector field where each vector points from a pixel in I_1 to its corresponding location in I_2, assuming brightness constancy and spatial coherence. This field inherently includes motion from both moving objects and the camera's own motion (ego-motion), which we use the noise term ϵ to represent in the rest of the paper.

2.2 Transformation Matrix Estimation using SVD

Classical image homography estimation such as RANSEC requires a set of matched keypoints $\mathbf{P}_1$ and $\mathbf{P}_2$ through feature detectors (e.g., SIFT, ORB) to extract the transform matrix that describes the image transform relationships between images I_1 and I_2.

In our proposed framework, we sample pixels directly from image I_1. Specifically, $\mathbf{P}_1$ is 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 $\mathbf{p}_i^1 = (x, y)$ in $\mathbf{P}_1$, we find its corresponding point in I_2 using the displacement (dx, dy) obtained from the optical flow field $\mathbf{f}$ from the previous step:

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

This process ensures that $\mathbf{P}_2$ 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 [34]. Specifically, the cross-covariance matrix M is computed:

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

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

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

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

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

2.3 Ego-Motion Compensation

The transformation matrix $\mathbf{H}$ 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:

$$\mathbf{P_1'} = \mathbf{H} \times \mathbf{P_1}$$

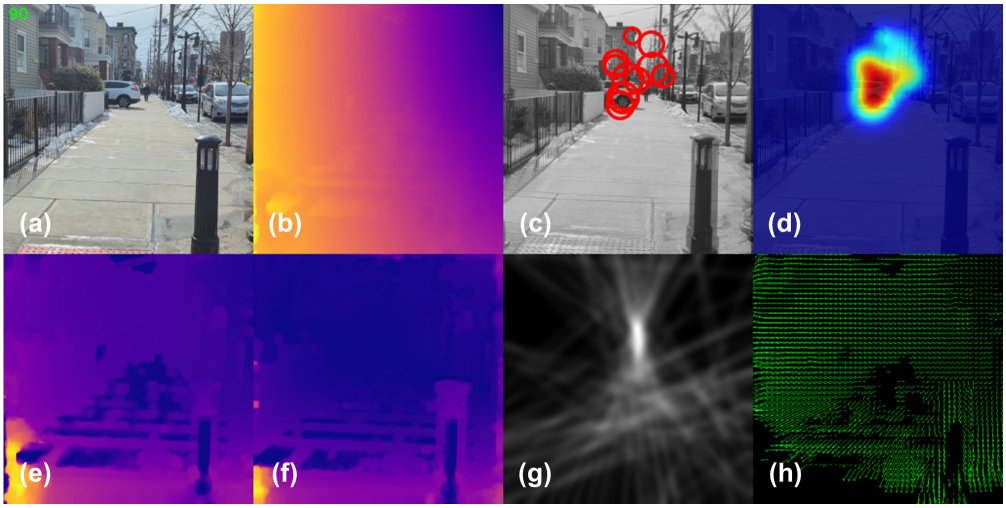

Figure 3: Motor focus visualization, (a) is the raw RGB image, (b) is the camera noise ϵ, (c) is the attention area of 10 consecutive frames, (d) is the attention map aggregated by the gaussian distributions of 10 recent frames, (e) is the original optical flow, (f) is the compensated optical flow, (g) is the probability map of focus center, and (h) is the compensated optical flow field.

The noise term ϵ of this study is defined by the difference between the predicted positions of the pixels due to ego-motion and the predicted positions of the pixels without ego-motion, which represents the displacement caused exclusively by the camera's motion, ignoring any independent object movements:

$$\epsilon = \mathbf{P'_1} - \mathbf{P_1}$$

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 own motion. This is expressed mathematically as:

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

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

2.4 Prediction of Ego-Movement Direction

Given the initial set of vectors $\mathbf{v}_i = \mathbf{p}_1^i + \mathbf{f}_i'$, assuming that each $\mathbf{v}_i$ 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:

$$\min_{\mathbf{c}} \sum_i \|\mathbf{c} - \text{proj}_{\mathbf{v}_i}(\mathbf{c})\|^2$$

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 $\mathbf{v}_i$, we formulate an optimization problem to minimize these deviations for each cluster. Specifically, for each candidate focus $\mathbf{c}_k$:

$$\mathbf{c}_k = \sum_{i=1}^{N} \frac{\left\|\mathbf{c}_k - \text{proj}_{\mathbf{v}_i}(\mathbf{c}_k)\right\|^2}{N}, i \in \text{cluster}_k$$

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

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

$$\text{Score}(\mathbf{c}_k) = \sum_{i=1}^{N} i, i \in \text{cluster}_k$$

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

To stabilize the attention area across frames and reduce the effect of transient shaking or focus shifts, the Gaussian splatting 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, $\mathbf{u}_i$, representing the activity level or movement intensity in the frame. The formulation can be described as follows:

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

$$\sigma_i \propto \frac{1}{u_i} \tag{2}$$

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

The final Gaussian splatting mask effectively smoothes and stabilizes the attention area, as the aggregation helps in filtering out the ambiguous focus due to camera movement, thus providing a more consistent and reliable attention signal over time.

3 Experiments

In this section, we evaluate the proposed method both qualitatively and quantitatively.

3.1 Data Collection

To evaluate the performance of the proposed method, we collected a small dataset that is specialized for visual navigation. The collected dataset consists of two parts: in-person collection, and online collection.

The in-person collection is a subset of Urban Cruising Dataset which is used for visual navigation testing [35]. Specifically, in this subset, data collectors move by walking, biking, and electric scooters and capture videos in first-person. The dataset consists of a total of 50 clips of human users' movements and has been used for assistive visual navigation tasks. Table 1 shows the detailed dataset description. The collected data is then annotated with the state-of-the-art open-world object detection model (Yolo-World), with classification labels that are designed for visually impaired navigation.

Location	Scene	Movement	Weather	Clips	Total length	Unique Classes	Total detected objects
Urban	Sidewalk	Scooter	Cloudy	8	10 mins	31	16944
Suburban	Bikeline	Scooter	Cloudy	5	6 mins	26	8394
Urban	Park	Scooter	Cloudy	6	5 mins	23	15310
City	Road	Biking	Sunny	5	5 mins	21	5464
City	Sidewalk	Biking	Sunny	7	6 mins	27	9569
City	Park	Biking	Cloudy	5	5 mins	19	4781
Town	Park	Walking	Cloudy	6	4 mins	18	5156
Town	Sidewalk	Walking	Sunny	8	7 mins	14	8274
City	Coast	Walking	Sunny	2	5 mins	37	29280
Suburban	Theme Park	Walking	Rain	3	6 mins	34	24180

Table 1: Action type and length of Urban Cruising Dataset.

Furthermore, to test the motor focus estimation performance, we annotated these clips so that their movement directions are estimated by 3 researchers, each video clip is observed frame by frame, and a pixel location (x, y) of moving direction is annotated for each frame. The final annotation is averaged among 3 different annotations. As shown in Figure 4, each video clip is annotated with 3 different labels, and the ground truth is averaged by these 3 different annotations. Figure 4 shows the sample of the Urban Cruising Dataset. Specifically, Figure 4 (a) is a biking scene, (b) and (c) is a scooter riding scene, and (d) is a walking scene.

3.2 Ego-Motion Compensation

To evaluate the compensation performance of the proposed method, we compare the original optical flow and the compensated optical flow in a series of scenes to showcase its difference from the classical optical flow and test its robustness to ego-motion and sensitivity to objects that are individually moving.

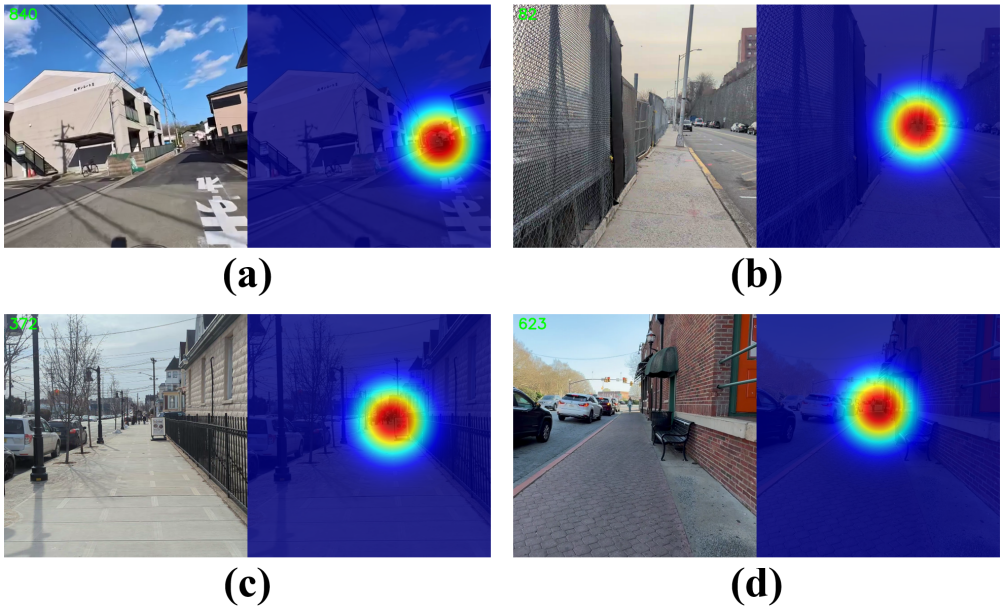

Figure 4: The samples of the Urban Cruising dataset. For each image, the left is the original RGB image, and the right is the visualized ground truth moving direction.

As shown in Figure 5, the compensated optical flow can filter the ego-motion caused by the camera shift and self-moving, which distinguishes the objects that are moving relatively. Meanwhile, the moving speed and direction of nearby objects can also be estimated from the compensated optical flow map. Furthermore, the camera noise ϵ shows the camera's moving direction, which can be used to estimate the motor direction, as proved in Section 2.

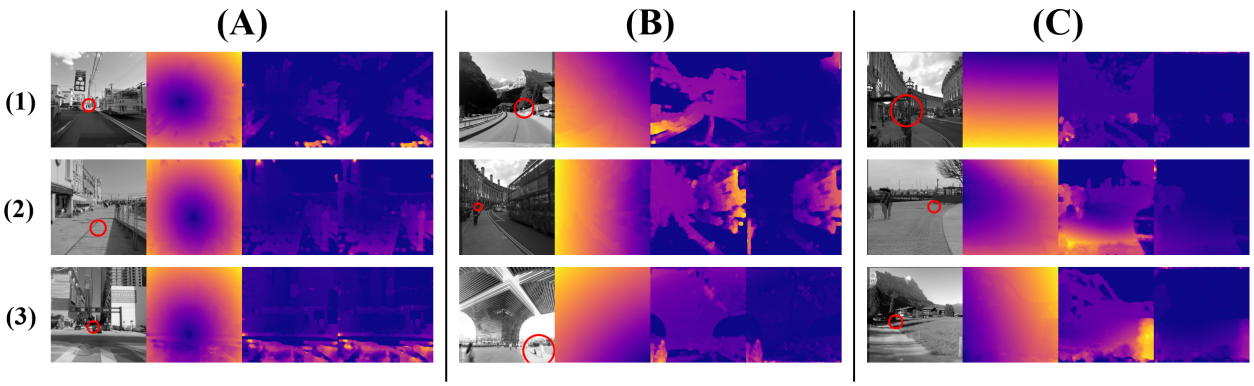

Figure 5: Visualization of ego-motion compensation, each image consists of four cells, from left to right: grayscale image with predicted moving direction, the magnitude of camera noise ϵ (ego-motion), raw optical flow, and optical flow with ego-motion compensation.

More importantly, Figure 5 indicates that the visual focus can be very different than the motor focus. For instance, in Figure 5 group B, while the compensated optical flow and camera noise indicated the camera moving potential, the actual motor movement is different. Specifically, in B1 and B3, the camera is moving toward the left corner, while the user is moving toward the right side. However, in B2, both the camera and the user are moving toward the left. On the other hand, in Figure 5 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. Specifically, in C1, the user is standing statically, but the camera is slowly pitching up.

Interestingly, when both the camera and the user move straightforwardly, the raw optical flow and the compensated optical flow are theoretically identical, and the low point of the camera noise map tends to overlap with the motor focus area, as shown in Figure 5 group A.

3.3 Ego Motion Prediction Performance

We applied the proposed method to predict the center of moving direction and we usin MAE and MSE to compare the predicted location $(\hat{x}, \hat{y})$ with the annotated ground truth(x, y). Figure 6 shows the results of the proposed method compared with different approaches. As expected, the proposed method achieves minimum errors compared with other methods.

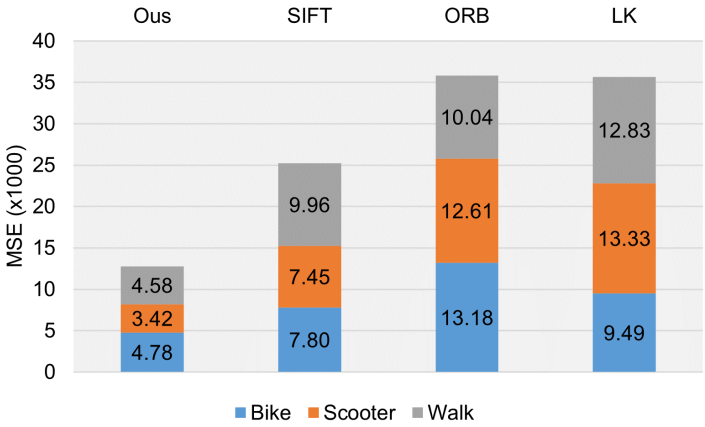

Figure 6: MSE benchmark on the collected dataset.

Furthermore, we evaluate the computation cost of the implemented method with the classical motion analysis methods. Table 2 shows the computation cost comparison results. Notably, our All-Pixel Matching method achieves extraordinary results compared to other methods. Specifically, our proposed All-Pixel Matching requires no searching for keypoints as we consider each pixel as a keypoints and take the optical flow map as the matching map. Our matching time stands out from others by taking advantage of Singular Vector Decomposition, the linear operation processes a total of $262,144$ elements without additional computation cost.

	Total Keypoints	Keypoints Searching Time (ms)	Matching Time (ms)	Total Time (ms)
LK	101	3.74	4.86	8.60
ORB	134	4.42	5.34	9.76
SIFT	501	31.03	35.49	66.52
Ours	**262,144**	0	0.91	**0.91**

Table 2: Performance comparison

The results from our experiments indicate that our method successfully achieves real-time processing capabilities without sacrificing accuracy or robustness. The stabilization algorithm effectively mitigates the impact of camera shake, allowing for clear and consistent detection of nearby objects and their motions. This capability is particularly important in densely populated urban environments where accurate and timely information about surrounding dynamics is critical for safe navigation.

4 Conclusion

In conclusion, our study presents a novel image-only method for motion analysis, specifically designed for predicting motor attention in both humans and machines. By utilizing optical flow with ego-motion compensation, our approach effectively tackles the challenges posed by handheld and body-mounted devices, such as camera shakes, enhancing movement prediction accuracy and stability. The introduction of Gaussian aggregation further improves the robustness and real-time applicability of our system, making it suitable for a wide array of applications from consumer electronics to personal navigation. Our experimental results, both qualitative and quantitative, validate the superiority of our method over traditional dense optical flow-based techniques, especially in accurately estimating motor attention with a specialized small dataset. This work not only advances the understanding of motor attention but also offers a scalable, adaptable solution that redefines its application in technology-driven environments.

References

[1] Askat Kuzdeuov, Shakhizat Nurgaliyev, and Hüseyin Atakan Varol. Chatgpt for visually impaired and blind. Authorea Preprints, 2023.

[2] Ashish Bastola, Md Atik Enam, Ananta Bastola, Aaron Gluck, and Julian Brinkley. Multi-functional glasses for the blind and visually impaired: Design and development. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 67, pages 995–1001. SAGE Publications Sage CA: Los Angeles, CA, 2023.

[3] Ashish Bastola, Aaron Gluck, and Julian Brinkley. Feedback mechanism for blind and visually impaired: a review. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 67, pages 1748–1754. SAGE Publications Sage CA: Los Angeles, CA, 2023.

[4] Ashish Bastola, Julian Brinkley, Hao Wang, and Abolfazl Razi. Driving towards inclusion: Revisiting in-vehicle interaction in autonomous vehicles. arXiv preprint arXiv:2401.14571, 2024.

[5] Jonathan Donner. After access: Inclusion, development, and a more mobile Internet. MIT press, 2015.

[6] Fatma Al-Muqbali, Noura Al-Tourshi, Khuloud Al-Kiyumi, and Faizal 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), pages 1–4. IEEE, 2020.

[7] Muiz Ahmed Khan, Pias Paul, Mahmudur Rashid, Mainul Hossain, and Md Atiqur Rahman Ahad. An ai-based visual aid with integrated reading assistant for the completely blind. IEEE Transactions on Human-Machine Systems, 50(6):507–517, 2020.

[8] Bing Li, Juan Pablo Munoz, Xuejian Rong, Qingtian Chen, Jizhong Xiao, Yingli Tian, Aries Arditi, and Mohammed Yousuf. Vision-based mobile indoor assistive navigation aid for blind people. IEEE transactions on mobile computing, 18(3):702–714, 2018.

[9] Uwe Franke, Clemens Rabe, Hernán Badino, and Stefan Gehrig. 6d-vision: Fusion of stereo and motion for robust environment perception. In Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, August 31-September 2, 2005. Proceedings 27, pages 216–223. Springer, 2005.

[10] Seungwon Lee, Nahyun Kim, Kyungwon Jeong, Kyungju Park, and Joonki Paik. Moving object detection using unstable camera for video surveillance systems. Optik, 126(20):2436–2441, 2015.

[11] Davide Scaramuzza and Friedrich Fraundorfer. Visual odometry [tutorial]. IEEE robotics & automation magazine, 18(4):80–92, 2011.

[12] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In 2017 IEEE international conference on robotics and automation (ICRA), pages 2043–2050. IEEE, 2017.

[13] Christopher D Monaco. Ego-Motion Estimation from Doppler and Spatial Data in Sonar, Radar, or Camera Images. The Pennsylvania State University, 2019.

[14] Jun Zhang, Mina Henein, Robert Mahony, and Viorela Ila. Robust ego and object 6-dof motion estimation and tracking. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5017–5023. IEEE, 2020.

[15] Jiexiong Tang, Rares Ambrus, Vitor Guizilini, Sudeep Pillai, Hanme Kim, Patric Jensfelt, and Adrien Gaidon. Self-supervised 3d keypoint learning for ego-motion estimation. In Conference on Robot Learning, pages 2085–2103. PMLR, 2021.

[16] Yi Tang, Wenbin Zou, Zhi Jin, and Xia Li. Multi-scale spatiotemporal conv-lstm network for video saliency detection. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pages 362–369, 2018.

[17] Alexander Makrigiorgos, Ali Shafti, Alex Harston, Julien Gerard, and A Aldo Faisal. Human visual attention prediction boosts learning & performance of autonomous driving agents. arXiv preprint arXiv:1909.05003, 2019.

[18] Anli Lim, Bharath Ramesh, Yue Yang, Cheng Xiang, Zhi Gao, and Feng Lin. Real-time optical flow-based video stabilization for unmanned aerial vehicles. Journal of Real-Time Image Processing, 16:1975–1985, 2019.

[19] Mouna Afif, Riadh Ayachi, Yahia Said, Edwige Pissaloux, and Mohamed Atri. An evaluation of retinanet on indoor object detection for blind and visually impaired persons assistance navigation. Neural Processing Letters, 51:2265–2279, 2020.

[20] Hernisa Kacorri, Kris M Kitani, Jeffrey P Bigham, and Chieko Asakawa. People with visual impairment training personal object recognizers: Feasibility and challenges. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 5839–5849, 2017.

[21] Abinash Bhandari, PWC Prasad, Abeer Alsadoon, and Angelika Maag. Object detection and recognition: using deep learning to assist the visually impaired. Disability and Rehabilitation: Assistive Technology, 16(3):280–288, 2021.

[22] Fahad Ashiq, Muhammad Asif, Maaz Bin Ahmad, Sadia Zafar, Khalid Masood, Toqeer Mahmood, Muhammad Tariq Mahmood, and Ik Hyun Lee. Cnn-based object recognition and tracking system to assist visually impaired people. IEEE access, 10:14819–14834, 2022.

[23] Jianying Yuan, Tao Jiang, Xi He, Sidong Wu, Jiajia Liu, and Dequan Guo. Dynamic obstacle detection method based on u–v disparity and residual optical flow for autonomous driving. Scientific Reports, 13(1):7630, 2023.

[24] Baigan Zhao, Yingping Huang, Wenyan Ci, and Xing Hu. Unsupervised learning of monocular depth and ego-motion with optical flow features and multiple constraints. Sensors, 22(4):1383, 2022.

[25] Vitor Guizilini, Kuan-Hui Lee, Rareş Ambruş, and Adrien Gaidon. Learning optical flow, depth, and scene flow without real-world labels. IEEE Robotics and Automation Letters, 7(2):3491–3498, 2022.

[26] Dominic FL Southgate, Joe AI Prinold, and Robert A Weinert-Aplin. Motion analysis in sport. In Sports innovation, technology and research, pages 3–30. 2016.

[27] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3):3292–3310, 2022.

[28] Hao Wang, Xiwen Chen, Natan Vital, Edward Duffy, and Abolfazl Razi. Energy optimization for hvac systems in multi-vav open offices: A deep reinforcement learning approach. Applied Energy, 356:122354, 2024.

[29] Vladyslav Usenko, Jakob Engel, Jörg Stückler, and Daniel Cremers. Reconstructing street-scenes in real-time from a driving car. In 2015 International Conference on 3D Vision, pages 607–614. IEEE, 2015.

[30] Yuki Tamaru, Yasunori Ozaki, Yuki Okafuji, Junya Nakanishi, Yuichiro Yoshikawa, and Jun 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), pages 1064–1068. IEEE, 2022.

[31] Miao Liu, Siyu Tang, Yin Li, and James 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, pages 704–721. Springer, 2020.

[32] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In Proceedings of the 26th ACM international conference on Multimedia, pages 1190–1198, 2018.

[33] Gunnar 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, pages 363–370. Springer, 2003.

[34] Hao Wang, Xiwen Chen, Abolfazl Razi, and Rahul Amin. Fast key points detection and matching for tree-structured images. In 2022 International Conference on Computational Science and Computational Intelligence (CSCI), pages 1381–1387. IEEE, 2022.

[35] Hao Wang, Jiayou Qin, Ashish Bastola, Xiwen Chen, John Suchanek, Zihao Gong, and Abolfazl Razi. Visiongpt: Llm-assisted real-time anomaly detection for safe visual navigation. arXiv preprint arXiv:2403.12415, 2024.