Dynamic head pose estimation in varied conditions using Dlib and MediaPipe

Rusnani Yahya[1,2], Rozita Jailani[1], Nur Khalidah Zakaria[1], Fazah Akhtar Hanapiah[3]
[1]Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Selangor, Malaysia
[2]Center for Medical Electronic Technology, Politeknik Sultan Salahuddin Abdul Aziz Shah, Shah Alam, Malaysia
[3]Faculty of Medicine, Universiti Teknologi MARA, Sungai Buloh, Malaysia

Article Info

ABSTRACT

This paper presents the formulation and validation of a dynamic head pose estimation (HPE) algorithm, addressing challenges related to diverse conditions, complex poses, and partial obstructions. The study aims to create a robust algorithm that maintains high accuracy in real-time applications across varying conditions. The algorithm was implemented and assessed using Dlib and MediaPipe models. The study involved 30 participants in face and head without obstacles, face with obstacles and head with obstacles conditions. The results demonstrated impressive performance in both controlled and spontaneous head movement categories. The algorithm achieved an average accuracy of 93% for head pose estimation and 88% in detecting visual attention under spontaneous head movement categories. A correlation coefficient of 0.866 indicates a strong positive linear association between performance and attention accuracy, indicating that performance improvements are intricately linked to proportional increases in attention accuracy. However, this does not necessarily imply causation. The findings provide valuable insights into the effectiveness of the proposed algorithms in assessing visual attention and demonstrate their potential applications in healthcare monitoring, educational intervention, and driver monitoring systems. The significance of these results lies in the ability to advance human-computer interaction, enhance healthcare diagnostics, and offer innovative solutions across various domains.

Corresponding Author:

Nur Khalidah Zakaria
Faculty of Electrical Engineering, Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia
Email: nkhalidah@uitm.edu.my

1. INTRODUCTION

Real-time head pose estimation (HPE) is a crucial component in computer vision systems designed to interpret human attention, including applications such as driver monitoring [1]–[4], adaptive e-learning interfaces [5], [6], and human-robot interaction [7], [8]. These systems can infer engagement and attention levels by tracking head orientation, allowing for dynamic and responsive interactions. However, achieving high accuracy in varied conditions, such as changes in lighting, occlusions such as wearing glasses or hats, and spontaneous head movements, remains a technical challenge.

Recent studies have proposed different methods to address these challenges such as using deep learning with data augmentation to handle diverse expressions, orientations and lighting environments [9], [10], applying attention mechanisms to focus on key facial regions despite occlusions [11], [12], employing three-dimensional (3D) morphable models to improve accuracy under partial visibility [13], [14], leveraging

robust facial landmark detection and tracking for occlusion resilience [15], [16], incorporating temporal models for handling spontaneous movements [17], and combining multi-modal data such as red-green-blue (RGB) and depth for enhanced robustness [18], [19]. Some algorithms employ advanced machine learning techniques to accurately determine an orientation of user head by analyzing facial features from video input [20], [21]. The capability is vital for enhancing interactive systems, such as educational technologies and sophisticated monitoring systems [22], where understanding gaze direction and head orientation can significantly improve system responsiveness and personalization [23]. While the studies demonstrated promising results, many relied on controlled environments or required specialized hardware, such as depth cameras or multiple sensors [24]. This limited their practicality in real-world scenarios and enabled more refined and accurate user interaction capabilities. Despite substantial advancements in the field, critical research gaps remain. Notably, few study have explored head pose estimation relying solely on RGB video under diverse and unconstrained real-world conditions [25]. Additionally, the robustness of existing methods against partial occlusions and spontaneous head movements has not been sufficiently validated [26]. Moreover, direct integration of head pose estimation with attention detection using low cost hardware remain limited [27], indicating a need for further investigation.

This study addresses these gaps by proposing a dynamic HPE algorithm that integrates Dlib and MediaPipe models, evaluated under real-time conditions using a standard RGB webcam. The proposed method demonstrates effectiveness across three conditions: i) face and head without obstacles, ii) face with obstacles such as the presence of glasses, and iii) head partially obstructed by wearing a hat, under both controlled and spontaneous head movements conditions. The key contributions of this study are as follows: a development of a robust, low-cost head pose estimation method utilizing facial landmarks and Euler angles; an experimental design that incorporates both controlled and spontaneous head movements under varying obstacle conditions; a detailed analysis of visual attention through head pose estimation accuracy; and a correlation analysis between performance and attention detection accuracy.

The rest of the paper is structured as follows: section 2 outlines the methodology, including algorithm development and experimental setup. Section 3 presents and discusses results regarding detection accuracy, visual attention, and comparisons with related works. Section 4 concludes with implications and recommendations for future research.

2. METHOD

This section follows a structured approach as illustrated in Figure 1, to systematically describe the research process in a reproducible and logical sequence. Four steps namely algorithm development, hardware setup, data collection, and data analysis, are described in detail, accompanied by proper justifications and standard procedures. The algorithm development and hardware setup subsections provide a comprehensive explanation of all tools and technologies utilized in this research. The experimental setups were designed to reflect real-world conditions thus, addressing the identified gaps for meaningful findings.
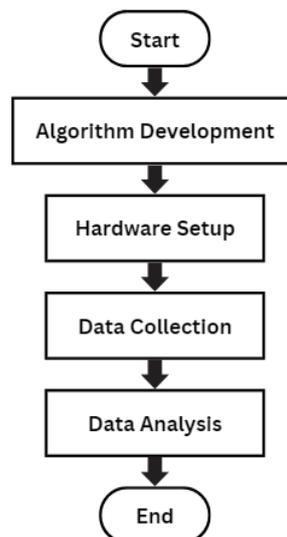

Figure 1. A detailed step-by-step description of the research procedure

2.1. Algorithm development

OpenCV is a free and open-source library for computer vision and machine learning, offering interfaces for multiple programming languages, including C/C++, Python, Java, and MATLAB. It is compatible with Windows, Linux, Android, and Mac OS platforms. With more than 2,500 algorithms for computer vision and machine learning, the library includes functionalities for tasks like face detection and recognition [28]–[31]. Dlib is a community-driven C++ library that offers a variety of tools for machine learning, computer vision, and image processing. It includes functionalities for developing complex software, such as facial landmark detection, object detection and deep learning-based tasks. Dlib is widely utilized across various fields, including robotics, mobile devices, and high-performance computing environments [32], [33]. Essential HPE system development libraries include OpenCV for image processing, Dlib for facial landmark detection, NumPy for mathematical computations, and MediaPipe for real-time machine learning across multiple modalities. The time library, CSV module, DateTime, and OS also aid data logging, time handling, and system interactions.

Figure 2 shows the four key stages of the HPE, which involve using Dlib and MediaPipe models. The process starts with Dlib's face detector and predictor, which are applied to identify facial landmarks. Next, MediaPipe is utilized to calculate the 3D orientation of the head [34]. The third stage involves solving the perspective-n-point (PnP) problem to obtain rotation and translation vectors [35]. Finally, these vectors are transformed into Euler angles, determining the direction of the head pose.

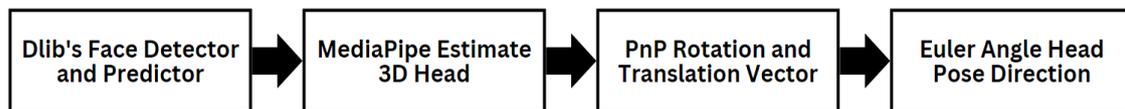| Dlib's Face Detector and Predictor	→	MediaPipe Estimate 3D Head	→	PnP Rotation and Translation Vector	→	Euler Angle Head Pose Direction

Figure 2. Four stages of the head pose estimation (HPE)

Face detector and landmark predictor are trained models distributed by Dlib that detect distinct facial landmarks. In real-time, the face mesh model will complete the face mapping with an estimated 478 3-dimensional landmarks. The PnP problem is a classic problem in the field of visual computing. The camera pose is determined using a set of 3D points from the world coordinate system and their corresponding two-dimensional (2D) projections on the image plane. In HPE, the 3D points represent facial landmarks with predefined positions, while the 2D points refer to their identified positions in the image. The rotation vector is a compact representation of the rotation. It uses Rodrigues' rotation formula to represent a 3D rotation by a single vector rather than a matrix. The direction of this vector indicates the rotation axis, while its magnitude signifies the rotation angle, measured in radians. The translation vector describes the object's position, in this case, the head, relative to the camera. It indicates how far and in which direction the object is from the camera in 3D space. The rotation matrix is a 3×3 matrix that comprehensively represents rotation. It can be derived from the rotation vector using Rodrigues' formula and is used to transform coordinates from one space to another. The rotation matrix R can be decomposed into three Euler angles: pitch (rotation around the x-axis), yaw (rotation around the y-axis), and roll (rotation around the z-axis), which provide a more intuitive representation of orientation.

2.2. Hardware setup

The experiment utilized a high-definition 1080p 2-megapixel webcam, produced by AUKEY with 1/2.9" CMOS image sensor and integrated stereo microphones. The frame rate was 30 frames per second (fps). The system specifications for running the program included: Windows 11 (64-bit), a 13th Gen Intel(R) Core (TM) i7-13700HX CPU, and 16.00 GB of memory.

Figure 3 illustrates the experimental setup, where a high-resolution webcam is employed to detect and record. The webcam is integrated into the system framework to ensure enhanced HPE. Throughout the experiment, the computer is connected to the webcam to monitor and track the progress. Notice that the built-in computer webcam is not used due to its lower resolution. A chair is positioned in front of the camera to facilitate subject participation. Subjects are instructed to sit comfortably in the chair before beginning the study. Maintaining a distance of 0.5 meters or less between the chair and the camera is crucial to ensure precise face detection. This distance parameter is carefully selected to optimize system performance. The researcher oversees the study's progress using the external webcam.
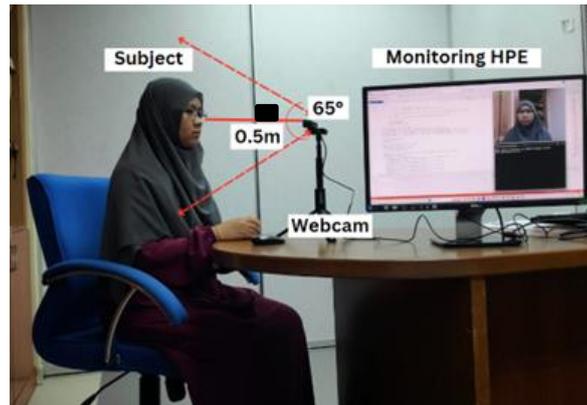
Figure 3. Experiment setup for performing head pose estimation (HPE)

2.3. Data collection

Thirty students from Polytechnic Sultan Salahuddin Abdul Aziz Shah (PSSAAS), Shah Alam, aged 19 and 27, participated in this study. Institutional approval was obtained, and written consent (REC4/2020/BM Pind. 2(2020)) was taken from all subjects. Two categories of data collection were conducted for each subject. First, the controlled category where the subject performed controlled head movements according to the instructions by the researcher. In the second phase, the subject carried out spontaneous head movements. Both categories were evaluated under three distinct conditions: face and head without obstacles, face with obstacles, and head with obstacles. Each condition involved a real-time head pose detection session lasting approximately 300 seconds, during which three types of outputs were simultaneously generated: video recording, extracted images, and corresponding CSV files containing head pose data.

In the controlled category, the subjects were asked to sit comfortably on the provided chair. The researcher instructed the subjects to perform head movements once the system program was initiated. During this category, five specific head movements were recorded, each lasting between 20 and 30 seconds, resulting in the collection of more than 60 data points. Subjects were instructed to orient their head forward, upward, downward, and to the right and left, with horizontal rotations exceeding 75 degrees and vertical tilts exceeding 45 degrees. Subjects followed a structured sequence: beginning by facing forward (0°), then slowly tilted their heads upward to a full tilt of 45°. Afterwards, they tilted their heads downward to a full tilt of -45°. The next step required them to turn their heads to the right until reaching a tilt of -75°, and finally, they turned their heads to the left to a maximum tilt of 75°. In the second category, subjects could move their heads freely in front of the camera for 120 seconds without any specified directional constraints.

The flowchart in Figure 4 illustrates detecting and estimating head pose using an RGB camera, Dlib's face detector and MediaPipe's face mesh. The process starts with the camera detecting a face. Once a face is detected, Dlib is employed to locate the facial features. MediaPipe face mesh detects landmarks, contributing to determining the orientation of the head. The algorithm then determines whether the head pose is in the specified position. If the head pose is detected, the system evaluates the direction in which the head is oriented. Depending on the direction, the appropriate text is displayed: "Looking Up" for head upward, "Looking Down" for head downward, "Forward" for head forward, "Looking Right" for head-turning right, and "Looking Left" for head-turning left. If the face is not in any true position, the system displays the text "Face not detected". After determining the head orientation, the system saves the output image, video and CSV data, which are likely to contain details of the detected head poses. The process then concludes and is ready to start over with new detections.

2.4. Data analysis

The effectiveness of the system is assessed based solely on the accuracy percentage. It is important to note that the predicted data output by the system may not always be entirely accurate, particularly when the head turn remains within a narrow range of degrees. Nonetheless, the system categorizes these movements as up, down, forward, right, or left head direction. To further evaluate the performance of the system, the results were classified based on four standard metrics: true positive (TP), false positive (FP), true negatives (TN) and false negatives (FN). A manual verification process was conducted by comparing the predicted head direction by the system against the actual head direction recorded in the video. Subsequently, the performance accuracy for each subject was calculated as the percentage of true detections relative to the

total number of detections, including both correct and incorrect predictions, across all conditions. The definitions of each metric are summarized in Table 1. These metrics provide a detailed view of the detection system capabilities and help identify specific errors contributing to overall accuracy and system robustness.
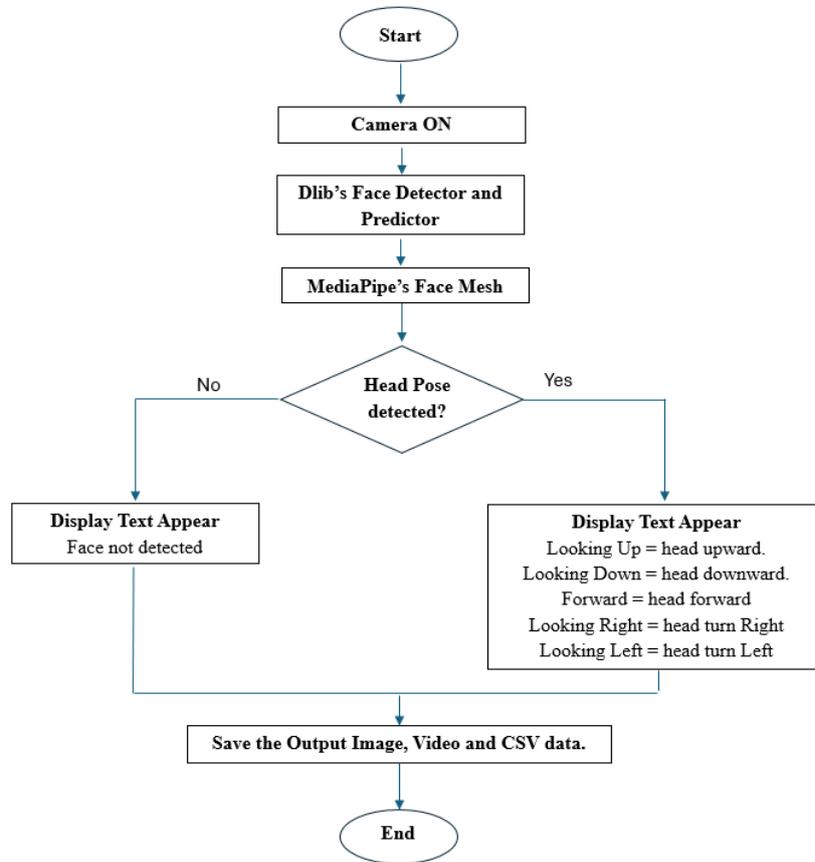
Figure 4. Flowchart of head pose estimation data collection

Table 1. Definitions of TP, FP, TN and FN for head pose estimation evaluation

Category	Definition
True Positive (TP)	The system correctly identifies the head pose, matching the actual direction.
False Positive (FP)	The system incorrectly predicts a head pose that did not occur.
True Negative (TN)	The system correctly recognizes that no change in head pose occurred when there was no significant movement.
False Negative (FN)	The system fails to detect an actual head pose, classifying it incorrectly or missing it altogether.

The performance of the system was evaluated based on the accuracy metric. Accuracy is the proportion of correct predictions, including true positives (TP) and true negatives (TN), relative to the total number of predictions. It is calculated using (1):

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

This metric reflects the overall correctness of the head pose estimation system across all conditions and participants. On the other hand, accurate detections are where the prediction by the system corresponds with the exact pose, thus enhancing the accuracy of the system. The performance accuracy related to visual attention will also be calculated and visualized, providing insights into the visual attention levels under different conditions. Visual attention was quantified by calculating the proportion of forward head pose detections $(N_{f,i})$ relative to the total number of detections $(N_{t,i})$ recorded by the system. Visual attention was computed using (2):

$$Visual\ Attention = (N_{f,i} \ / \ N_{t,i}) \times 100 \tag{2}$$

These metrics were used to quantify the individual level of visual attention during experimental sessions. This method was chosen due to its low computational cost and real-time efficiency, making it ideal for practical applications. Unlike deep-learning-based approaches that require large datasets and high processing power, our algorithm leverages facial landmarks and Euler angles for accurate and robust head pose estimation using a simple webcam setup.

3. RESULTS AND DISCUSSION

This study aimed to develop a real-time head pose estimation system using Dlib and MediaPipe that remains effective under various head movement conditions and facial obstructions. In this study, thirty students participated. Two categories of data collection were conducted for each subject: controlled head movement and spontaneous head movement. Next, both categories were evaluated under three conditions: face and head without obstacles, face with obstacles, and head with obstacles. Further, this section provides an in-depth analysis of the study data to achieve that objective. To evaluate the validity of the algorithm, the analysis focuses on the accuracy rate, which is categorized into controlled and spontaneous head movements. Additionally, the effectiveness of the algorithm in detecting visual attention will also be analyzed.

3.1. Controlled head movement

In this experiment, the subjects were directed to move their heads to achieve five specific head orientations gradually. The results shown in Figure 5 illustrate that the average performance was highest, reaching 99%, in conditions where both the face and head were visible without obstacles. The accuracy decreased slightly to 98% in both face and head with obstacle conditions, indicating that obstacles marginally impacted the overall performance but maintained high accuracy levels across all conditions. The slight decrease in accuracy happens because obstacles can block light, causing shadows or uneven lighting on the face, which makes it harder for the system to recognize head positions correctly. Even so, the system still shows a prominent level of efficiency.
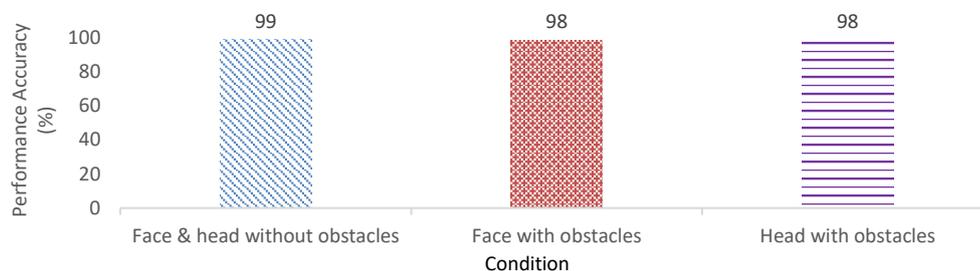

Figure 5. Average of performance accuracy in the controlled head movement category

3.2. Spontaneous head movement

This section will examine the results of an experiment in which subjects move their head freely and unpredictably without any specific guidance. The experiment was repeated three times under different conditions: face and head without obstacles, face with obstacles, and head with obstacles.

Figure 6 shows the average performance accuracy in percentages of the system under different conditions: face and head without obstacles, face with obstacles, and head with obstacles. The highest accuracy of 90% is observed when there are no obstacles, indicating optimal performance with clear visibility of the face and head. The accuracy decreases slightly to 87% when obstacles are present on the face, such as glasses, and further to the head with obstacles, such as a cap. Despite these reductions, the system maintains high accuracy, demonstrating robustness in various real-world scenarios. The slight drop in accuracy with face obstacles suggests that these might hinder landmark detection more than head obstacles. Sudden head movements or poor lighting conditions also affected landmark detection. These limitations highlight the need for further refinement, such as incorporating multi-view cameras or training with a more diverse dataset. Overall, 88% of the accuracy for spontaneous head movement categories indicated that the system is highly reliable. Still, future improvements could focus on increasing accuracy in the presence of obstacles to further enhance performance.

Figure 7 presents performance accuracy in percentages for head pose estimation across thirty subjects, S1 to S30, in three conditions. The face and head without obstacles category consistently show higher performance across most samples than the other two conditions, with many scores nearing or exceeding 90%. This demonstrates that the system reliably detects both face and head with no obstructions, exhibiting stable and consistent results. The scores in this category are tightly grouped, indicating predictable and dependable performance in explicit, obstacle-free situations.

The face with obstacles category shows more variation in scores, ranging from as low as 80 in some cases, such as S4, S5, and S17, to higher values in others like S1, S20, and S23. Performance drops significantly for some subjects like S10, S17, and S27, indicating that obstacles affect face detection more than head detection. This suggests that face detection is susceptible to visual obstructions, leading to lower accuracy than in obstacle-free situations. However, 7 of the 30 subjects such as S1, S2, S16, S20, S23, S26, and S28 achieved higher accuracy in the face with obstacles condition, likely because of distinct facial features, such as prominent cheekbones or jawlines, which made it easier for the system to recognize their faces despite the obstacles.

For the head with obstacle conditions, detection accuracy fluctuates similarly to face detection but tends slightly lower overall. However, head detection outperforms face detection in samples like S7 and S26, because obstacles significantly impact face visibility more than head orientation. This implies that head detection is more resilient to obstructions than face detection, maintaining better accuracy in challenging conditions.

In another observation, up to 98% of the highest accuracies are observed in the face & head without obstacles, particularly for subjects like S1, S3, S4, and S8. This suggests that the absence of obstacles significantly improves head pose estimation accuracy. However, some subjects, such as S7, S13, and S25, display more balanced accuracy across all conditions, indicating less impact from obstacles. Conversely, subjects like S3, S6, and S11 experience lower accuracy in the head with obstacle conditions, where accuracy dips to around 80%, although they still perform well in other conditions.
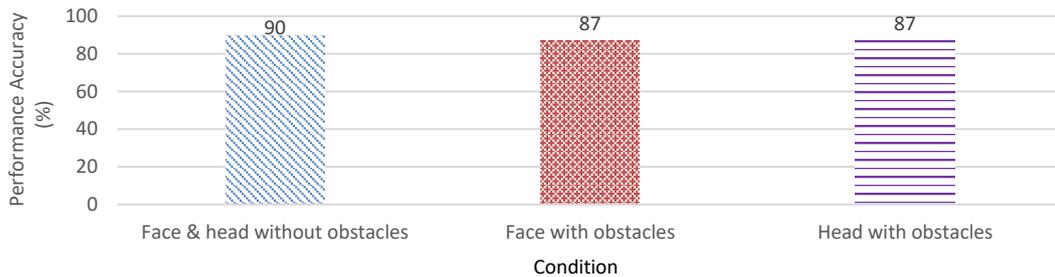

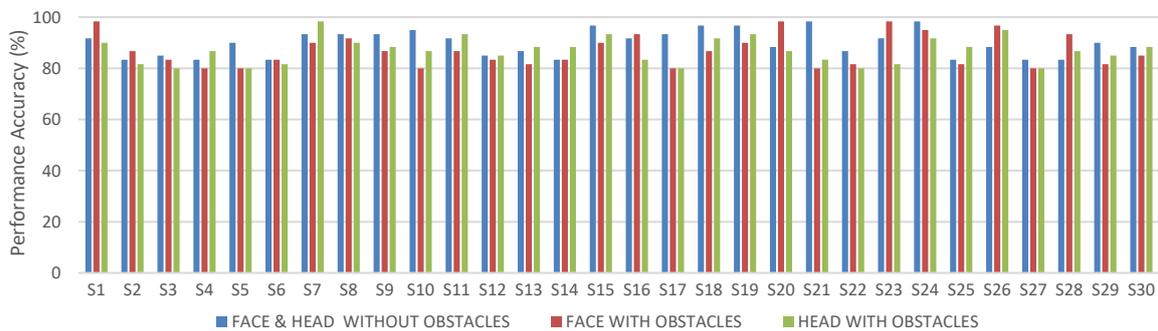Figure 6. Average of performance accuracy for a spontaneous head movement category

Figure 7. Performance accuracy for thirty subjects in the spontaneous head movement category

3.3. Visual attention

The effectiveness of the algorithm in detecting visual attention is evaluated using the dataset from the spontaneous head movement setting under three different conditions. Its accuracy is measured by analyzing prediction outputs when participants directly face the camera, indicating visual attention. The algorithm's results are cross-validated against video recordings to ensure accuracy.

Figure 8 compares the attention levels in percentage for thirty subjects under three conditions. It shows significant variations across subjects and conditions. Several subjects, such as S7, S8, S9, and S24, achieve perfect or near-perfect attention levels 100% across all conditions, indicating a robust attention estimation system for these individuals. However, attention levels drop noticeably for others, such as S14, S29, and S30, especially in the face with obstacles, where values dip as low as 70%. The presence of obstacles, particularly affecting the head, reduces the system's ability to estimate attention accurately, such as for S15, S24, and S29, where the value reduces by around 26%. In comparison, attention estimation remains high in most subjects with face and head without obstacles. The obstacles create a more challenging environment, resulting in lower performance. This suggests the system could be sensitive to visual obstructions, with accuracy dropping in more occluded conditions, especially when the head is partially blocked. Improving performance under these challenging conditions could enhance the system's overall reliability.

Figure 9 depicts visual attention accuracy for three conditions: face and head without obstacles is 90%, face with obstacles is 88%, and head with obstacles is 86%. Based on spontaneous head movement, these results indicate that obstacles slightly affect attention accuracy, but the variations are minimal. Overall, 88% of the accuracy of visual attention for spontaneous head movement categories indicated that the system is highly reliable. The high accuracy percentages across all conditions suggest that the system effectively detects visual attention, even when subjects face obstructions, demonstrating its robustness and reliability.
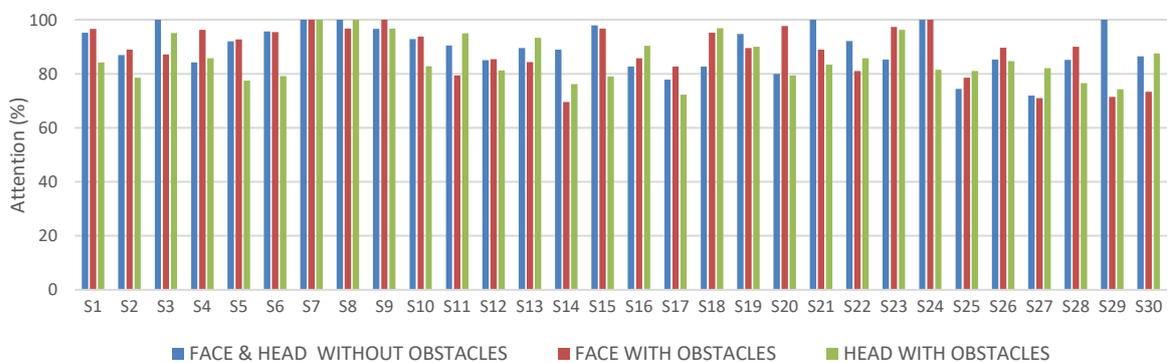

Figure 8. Visual attention accuracy while the subject's face and head without obstacles, face with obstacles and head with obstacles
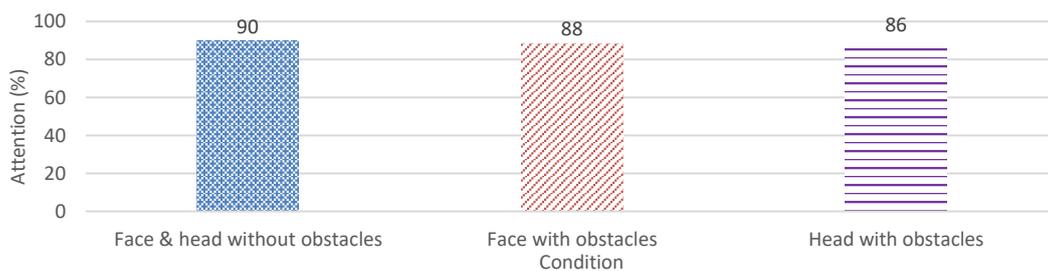

Figure 9. Visual attention accuracy for three varying conditions

3.4. Correlation analysis

Figure 10 shows the performance and attention accuracy percentages under three conditions: face and head without obstacles, face with obstacles, and head with obstacles. Both values remain consistently high across all conditions, with performance accuracy at 90% in the first condition and slightly lower (87%) for the other two. Attention closely follows performance accuracy, matching it at 90% in the first condition, slightly surpassing it in the second (88% vs. 87%), and lower in the third (86% vs. 87%). The correlation between performance and attention accuracy, calculated at 0.866, indicates a strong positive relationship, meaning changes in one variable are mirrored by changes in the other. This consistency suggests the system is resilient to obstacles, maintaining high performance and attention levels even when the face or head is

partially obscured. While head obstacles have a slightly more substantial effect on performance and attention accuracy than facial obstacles, the system remains robust overall.

The proposed algorithm demonstrates improved real-time head pose estimation accuracy using only RGB input compared to prior research. For instance, Owen *et al.* [32] achieved 86.85% accuracy in HPE in the sleepiness detection system, while our method achieved over 93% accuracy. Hammadi *et al.* [24] reported high performance using multimodal data but required pose tracking equipment. In contrast, our method uses only a simple webcam setup, making it more accessible. Although previous studies demonstrated good accuracy using OpenCV-based methods, the systems tended to become unstable during fast or unexpected head movements [5]. In this research, integrating MediPipe with its stable facial landmark tracking capability helped maintain accuracy, especially in spontaneous head movement category. These results align with our initial expectation that combining Dlib and MediaPipe would improve real-time head pose estimation accuracy, even under natural head movements and minimal hardware conditions.

Moreover, our strong correlation coefficient (0.866) between head pose and attention estimation supports similar findings by Liu *et al.* [36] and Afroze *et al.* [7], who noted the importance of facial orientation in human-computer interaction. These findings are valuable for developing accessible and cost-effective attention tracking systems, which can benefit educators, therapists, and researchers working with children, especially those with special needs.

Combining high accuracy, low hardware requirements, and robustness in obstacle-rich conditions can help develop real-time attention tracking systems for classrooms, therapy sessions, or interactive learning environments. Future research should explore integrating deep learning for occlusion robustness, multi-camera setups for enhanced tracking, and eye-tracking technology to refine attention analysis further. Combining head pose with facial expression analysis can also be extended to detect emotional response or engagement levels. However, the system showed some sensitivity to rapid lighting changes and complex occlusions, indicating the need for further refinement in more diverse environments.
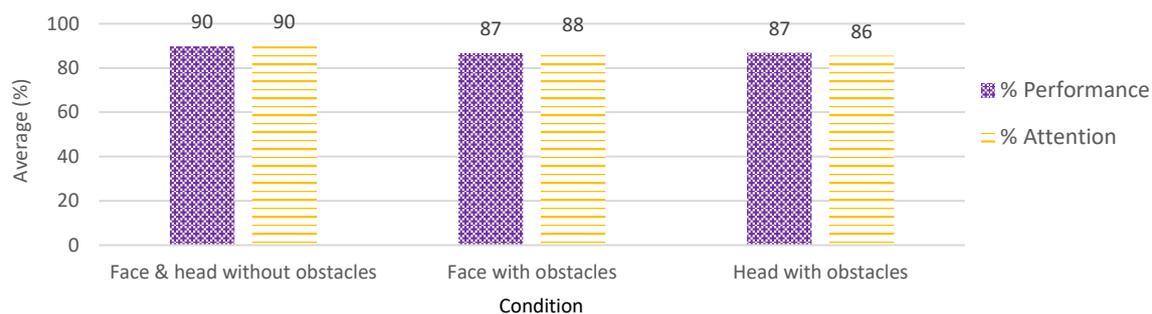

Figure 10. Comparison of performance and attention accuracy across different conditions

4. CONCLUSION

This study successfully achieved real-time head pose estimation using Dlib and MediaPipe, demonstrating reliable facial landmark detection across a broad range of head orientations. Compared to the standard OpenCV algorithm, the proposed method showed superior performance in terms of robustness and adaptability. Two experiment setups were used to evaluate the algorithm: one with controlled head movement and another with spontaneous head movements, both tested under three conditions (face and head without obstacles, face with obstacles and head with obstacles). The algorithm achieved over 93% accuracy in head pose estimation and at least 88% in detecting visual attention, even with facial obstacles. These findings suggest that the proposed HPE algorithm is practical and adaptable, with minimal performance degradation under real-world conditions, such as occlusions and natural movements. This has significant implications for the research field, particularly in applications related to human-computer interaction, assistive technologies, and educational or therapeutic settings involving visual attention monitoring. For the community, particularly in robot-assisted therapy or classroom engagement monitoring contexts, this research opens the door to more accessible, noninvasive tools for understanding attention and interaction patterns. Moving forward, multi-view estimation and advanced feature extraction enhancements could improve performance, making the technology even more resilient and applicable in diverse situations and uncontrolled environments. Ultimately, this study contributes to a growing body of work to bridge the gap between computer vision and real-world human behavior analysis, offering practical solutions with the potential for wide-reaching social impact.

AUTHOR CONTRIBUTIONS STATEMENT

To ensure transparency and clarity in research contributions, the roles of each author involved in this study are specified based on the Contributor Roles Taxonomy (CRediT).

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rusnani Yahya	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Rozita Jailani		✓				✓	✓	✓	✓	✓	✓	✓		
Nur Khalidah Zakaria	✓		✓	✓			✓			✓		✓	✓	✓
Fazah Akhtar Hanapiah										✓		✓		

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

Research related to human use has been conducted in accordance with all relevant national regulations and institutional policies, as outlined in the principles of the Declaration of Helsinki. It has been approved by the Research Ethics Committee (REC), Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia, with Approval number REC/06/2024 (PG/FB/21).

DATA AVAILABILITY

The data supporting this study's findings are available on request from the corresponding author. However, due to certain restrictions, the data, which contains information that could compromise the privacy of research participants, is not publicly available.

REFERENCES

[1] Y. Albadawi, A. AlRedhaei, and M. Takruri, "Real-time machine learning-based driver drowsiness detection using visual features," *Journal of Imaging*, vol. 9, no. 5, 2023, doi: 10.3390/jimaging9050091.

[2] R. K. Shukla, A. K. Tiwari, and A. K. Jha, "An efficient approach of face detection and prediction of drowsiness using SVM," *Mathematical Problems in Engineering*, vol. 2023, no. 1, pp. 1–12, 2023, doi: 10.1155/2023/2168361.

[3] W. Rahmaniar, Q. M. U. Haq, and T.-L. Lin, "Wide range head pose estimation using a single RGB camera for intelligent surveillance," *IEEE Sensors Journal*, vol. 22, no. 11, pp. 11112–11121, Jun. 2022, doi: 10.1109/JSEN.2022.3168863.

[4] Anitta D and A. Fathima A, "Human head pose estimation based on HF method," *Microprocessors and Microsystems*, vol. 82,

2021, doi: 10.1016/j.micpro.2020.103802.

[5]   I. Ahmad, F. AlQurashi, E. Abozinadah, and R. Mehmood, "A novel deep learning-based online proctoring system using face recognition, eye blinking, and object detection techniques," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 10, pp. 847–854, 2021, doi: 10.14569/IJACSA.2021.0121094.

[6]   R. Daza, L. F. Gomez, J. Fierrez, A. Morales, R. Tolosana, and J. Ortega-Garcia, "DeepFace-attention: multimodal face biometrics for attention estimation with application to e-learning," *IEEE Access*, vol. 12, pp. 111343–111359, 2024, doi: 10.1109/ACCESS.2024.3437291.

[7]   S. Afroze, M. R. Hossain, and M. M. Hoque, "DeepFocus: A visual focus of attention detection framework using deep learning in multi-object scenarios," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 10109–10124, 2022, doi: 10.1016/j.jksuci.2022.10.009.

[8]   H. D. Vankayalapati, S. Kuchibhotla, M. S. K. Chadalavada, S. K. Dargar, K. R. Anne, and K. Kyandoghere, "A novel zernike moment-based real-time head pose and gaze estimation framework for accuracy-sensitive applications," *Sensors*, vol. 22, no. 21, 2022, doi: 10.3390/s22218449.

[9]   Z.-W. Hong and Y.-C. Lin, "Improving facial landmark detection accuracy and efficiency with knowledge distillation," *arxiv.org/abs/2404.06029*, 2024.

[10]   F. Becattini, C. Bisogni, V. Loia, C. Pero, and F. Hao, "Head pose estimation patterns as deepfake detectors," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 11, 2024, doi: 10.1145/3612928.

[11]   S. Jha, N. Al-Dhahir, and C. Busso, "Driver visual attention estimation using head pose and eye appearance information," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 4, pp. 216–231, 2023, doi: 10.1109/OJITS.2023.3258184.

[12]   C. Giannetti, "Advancing robot-assisted autism therapy: a novel algorithm for enhancing joint attention interventions," *arXiv:2406.10392*. 2024.

[13]   H. Kim, S. Lee, and M. Sohn, "3D facial landmarks detection and head pose estimation using multi-task learning and vision transformer," *Journal of Industrial Information Technology and Application*, vol. 7, no. 1, pp. 666–670, 2023.

[14]   R. Algabri, H. Shin, and S. Lee, "Real-time 6DoF full-range markerless head pose estimation," *Expert Systems with Applications*, vol. 239, p. 122293, 2024, doi: 10.1016/j.eswa.2023.122293.

[15]   S. Malek and S. Rossi, "Head pose estimation using facial-landmarks classification for children rehabilitation games," *Pattern Recognition Letters*, vol. 152, pp. 406–412, 2021, doi: 10.1016/j.patrec.2021.11.002.

[16]   T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "Toward robust and unconstrained full range of rotation head pose estimation," *IEEE Transactions on Image Processing*, vol. 33, no. 3, pp. 2377–2387, 2024, doi: 10.1109/TIP.2024.3378180.

[17]   T. Hu, "Temporal head pose estimation from point cloud in naturalistic driving conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8063–8076, 2022, doi: 10.1109/TITS.2021.3075350.

[18]   C. Thai, "Real-time masked face classification and head pose estimation for RGB facial image via knowledge distillation," *Information Sciences*, vol. 616, pp. 330–347, 2022, doi: 10.1016/j.ins.2022.10.074.

[19]   A. Hosamani and M. Phirke, "Real-time head pose estimation based on face geometry," in *ACM International Conference Proceeding Series*, 2020, vol. 2, pp. 38–42, doi: 10.1145/3381271.3381296.

[20]   S. Huang *et al.*, "A new head pose tracking method based on stereo visual SLAM," *Journal of Visual Communication and Image Representation*, vol. 82, p. 103402, 2022, doi: 10.1016/j.jvcir.2021.103402.

[21]   A. Saeed, A. Al-Hamadi, and A. Ghoneim, "Head pose estimation on top of Haar-like face detection: A study using the Kinect sensor," *Sensors (Switzerland)*, vol. 15, no. 9, pp. 20945–20966, 2015, doi: 10.3390/s150920945.

[22]   A. Patti *et al.*, "Training attention skills in individuals with neurodevelopmental disorders using virtual reality and eye-tracking technology," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14697 LNCS, 2024, pp. 368–381.

[23]   N. Aunsri, "Novel eye-based features for head pose-free gaze estimation with web camera: New model and low-cost device," *Ain Shams Engineering Journal*, vol. 13, no. 5, 2022, doi: 10.1016/j.asej.2022.101731.

[24]   Y. Hammadi, "Evaluation of various state of the art head pose estimation algorithms for clinical scenarios," *Sensors*, vol. 22, no. 18, 2022, doi: 10.3390/s22186850.

[25]   Z. Wang *et al.*, "Learning to detect head movement in unconstrained remote gaze estimation in the wild," in *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, 2020, pp. 3432–3441, doi: 10.1109/WACV45572.2020.9093476.

[26]   Y. Liu, Z. Gu, S. Gao, D. Wang, Y. Zeng, and J. Cheng, "Mos: a low latency and lightweight framework for face detection, landmark localization, and head pose estimation," in *32nd British Machine Vision Conference, BMVC 2021*, 2021, pp. 1–14, doi: 10.5244/c.35.162.

[27]   B. Li and P. Liu, "Online learning state evaluation method based on face detection and head pose estimation," *Sensors*, vol. 24, no. 5, 2024, doi: 10.3390/s24051365.

[28]   M. Khan, S. Chakraborty, R. Astya, and S. Khepra, "Face detection and recognition using OpenCV," in *Proceedings - 2019 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2019*, 2019, pp. 116–119, doi: 10.1109/ICCCIS48478.2019.8974493.

[29]   A. Kumari Sirivarshitha, K. Sravani, K. S. Priya, and V. Bhavani, "An approach for face detection and face recognition using OpenCV and face recognition libraries in python," in *2023 9th International Conference on Advanced Computing and Communication Systems, ICACCS 2023*, 2023, pp. 1274–1278, doi: 10.1109/ICACCS57279.2023.10113066.

[30]   R. R. P. H. Sejati and R. Mardhiyyah, "Facial landmark based face detection using OpenCV and Dlib," *[in Bahasa] Jurnal Teknologi Informasi*, vol. 5, no. 2, pp. 144–148, 2021, doi: 10.36294/jurti.v5i2.2220.

[31]   R. T. H. Hasan and A. B. Sallow, "Face detection and recognition using OpenCV," *Journal of Soft Computing and Data Mining*, vol. 2, no. 2, pp. 86–97, 2021, doi: 10.30880/jscdm.2021.02.02.008.

[32]   V. Owen and N. Surantha, "Computer vision-based drowsiness detection using handcrafted feature extraction for edge computing devices," *Applied Sciences (Switzerland)*, vol. 15, no. 2, pp. 1–18, 2025, doi: 10.3390/app15020638.

[33]   L. L. R. Reinoso, F. L. G. López, J. C. Gutiérrez, G. Bressan, and W. V. Ruggiero, "Real-time head pose estimation with SVM model for frontal face classification," in *IADIS Conferences*, 2020, vol. 2020, no. Its, pp. 63–67, doi: 10.33965/its_ste2020_202001c008.

[34]   B. Thaman, T. Cao, and N. Caporusso, "Face mask detection using MediaPipe facemesh," in *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology, MIPRO 2022 - Proceedings*, 2022, pp. 378–382, doi: 10.23919/MIPRO55190.2022.9803531.

[35]   M. Ye, "Driver fatigue detection based on residual channel attention network and head pose estimation," *Applied Sciences (Switzerland)*, vol. 11, no. 19, 2021, doi: 10.3390/app11199195.

[36]   H. Liu, "Precise head pose estimation on HPD5A database for attention recognition based on convolutional neural network in human-computer interaction," *Infrared Physics and Technology*, vol. 116, 2021, doi: 10.1016/j.infrared.2021.103740.

BIOGRAPHIES OF AUTHORS

Rusnani Yahya 🆔 🎓 SC ⬡ received her B.Eng. in electrical engineering (medical electronics) from Universiti Tun Hussein Onn (UTHM), Malaysia, and M.Sc. in mechanical engineering (biomechanics) from Universiti Teknologi MARA, Malaysia. She is pursuing her Ph.D. in electrical engineering at the Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia. She can be contacted at: 2022692778@student.uitm.edu.my.

Rozita Jailani 🆔 🎓 SC ⬡ received her Ph.D. in automatic control and system engineering from Sheffield University, UK. She is an associate professor at the Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia, and the Chief Executive Officer of UiTM Technoventure Sdn. Bhd. Her research interests lie in biomedical engineering, focusing on assistive technology for children with special needs and patients with brain impairments, healthcare robotics, advanced image and signal processing, and artificial intelligence systems across various applications. She can be contacted at: rozita@ieee.org / rozitaj@uitm.edu.my.

Nur Khalidah Zakaria 🆔 🎓 SC ⬡ received her Ph.D. in electrical engineering from Universiti Teknologi MARA. She is currently a lecturer in the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Malaysia. Her research interests include artificial intelligence, the internet of things, and advanced signal and image processing techniques. She can be contacted at: nkhalidah@uitm.edu.my.

Fazah Akhtar Hanapiah 🆔 🎓 SC ⬡ received her Master of Rehabilitation Medicine from Universiti Malaya, Malaysia, and B.A. of Medicine, B.A. of Surgery, and B.A. of Obstetrics from the Royal College of Surgeons in Ireland. She is a professor at the Faculty of Medicine, Universiti Teknologi MARA, Malaysia. Her research interests include medical and health science, clinical medicine, and rehabilitation. She can be contacted at: fazah@uitm.edu.my.