Robust Eye Blink Detection Using Dual Embedding Video Vision Transformer

Jeongmin Hong* Joseph Shin* Juhee Choi Minsam Ko†
Hanyang University ERICA Campus
{jeongminhong, joeshin3956, wpaak1004, minsam}@hanyang.ac.kr

Abstract

Eye blink detection serves as a crucial biomarker for evaluating both physical and mental states, garnering considerable attention in biometric and video-based studies. Among various methods, video-based eye blink detection has been particularly favored due to its non-invasive nature, enabling broader applications. However, capturing eye blinks from different camera angles poses significant challenges, primarily because the eye region is relatively small and eye blinks occur rapidly, necessitating a robust detection algorithm. To address these challenges, we introduce Dual Embedding Video Vision Transformer (DE-ViViT), a novel approach for eye blink detection that employs two different embedding strategies: (i) tubelet embedding and (ii) residual embedding. Each embedding can capture large and subtle changes within the eye movement sequence respectively. We rigorously evaluate our proposed method using HUST-LEBW, a publicly available dataset, as well as our newly collected multi-angle eye blink dataset (MAEB). The results indicate that the proposed model consistently outperforms existing methods across both datasets, with notably minor performance variations depending on the camera angles.

1. Introduction

In the realm of biometrics and human behavior analysis, blinking serves as a multifaceted indicator of both physical and mental states. Specifically, eye blink patterns have been leveraged for a variety of applications including, but not limited to, the detection of driver fatigue [1, 2], monitoring of stress and attention levels [3, 4], diagnosis of ocular conditions [5], and even as a non-verbal communication channel for individuals with disabilities [6, 7].

As the applications of eye blink detection continue to proliferate, numerous methodologies have been investigated for automating this process. One prevalent approach involves the utilization of bio-signals, such as electroencephalogram (EEG) waveforms or electrooculogram (EOG) sensors, to capture electrical changes in the vicinity of the eye [8–11]. Another non-contact strategy employs the analysis of the infrared (IR) spectrum reflected from the eyes [12]. However, these methodologies are often encumbered by limitations related to sensor placement and lack of adaptability in diverse contexts [13, 14].

Extensive research has been conducted on eye blink detection using RGB imagery, owing to its versatile and efficient applicability in diverse scenarios. Various computational models have been proposed, including Convolutional Neural Network (CNN) based binary classifiers [15, 16] and parallel network architectures that utilize images from both eyes as input data [17]. Additionally, some studies have incorporated time-series models like Long Short-Term Memory (LSTM) networks [18, 19] to capture the overall eyelid motion dynamics.

Nonetheless, a notable limitation of existing eye blink detection methodologies is their optimization primarily for scenarios where the subject is directly facing the camera. Prominent datasets employed in the field, such as mEBAL [15], Eyeblink8 [20], and ZJU [21], predominantly feature frontal face images. This creates an inherent constraint in generalizing these models to real-world situations where subjects may engage in natural movements and actions that deviate from a frontal orientation. While some research endeavors have sought to mitigate this limitation by incorporating various facial orientations and multi-camera setups [22, 23], these efforts still represent a minority in the extant literature. Moreover, the absence of label data concerning camera angles or environmental conditions poses further constraints on directly analyzing performance in specific scenarios. The present study aims to address these gaps by focusing on robust eye blink detection that accommodates a range of camera angles.

In this study, we introduce the Dual Embedding Video Vision Transformer (DE-ViViT) that estimates the probability of an eye blink occurring within a given eye-region frame sequence. Especially, DE-ViViT is designed to offer a robust framework for eye blink detection by concurrent

[*] Both authors contributed equally to this research.
[†] Corresponding author

utilization of two distinct embedding techniques—Tubelet Embedding and Residual Embedding. Given the minuscule observation area and rapid motion dynamics associated with eye blinks, Residual Embedding is implemented to capture subtle changes between adjacent frames effectively. This approach is especially potent in scenarios devoid of macroscopic body or facial movements. However, considering real-world variability, Tubelet Embedding is also employed, which directly takes the original video sequence as input.

We trained DE-ViViT using the HUST-LEBW Training set, an in-the-wild dataset in the field of eye blink detection. We evaluated the model's performance on two distinct datasets to assess its generalizability and robustness. The first test dataset was the HUST-LEBW Testset, which comprises eye blink instances collected from movie clips, encapsulating a variety of conditions. The second dataset is the multi-angle eye blink dataset (MAEB), a custom-compiled dataset specifically designed for this research. While MAEB consists of data collected under controlled laboratory conditions, it distinguishes itself by featuring simultaneous captures from nine different camera angles, thereby enabling the analysis of angle-specific performance variations. For a comprehensive performance assessment, we developed comparison models based on prior research and evaluated our approach in conjunction with results reported in existing literature.

The evaluation results indicate that our proposed method consistently achieved high performance across both test datasets. A detailed analysis of eye blink detection results on the MAEB dataset further revealed that our approach exhibited minimal performance variance across different camera angles compared to other methods. This research makes two significant contributions to the field:

- We introduce Dual Embedding ViViT, designed specifically for robust eye blink detection, which has proven effective across diverse testing conditions.

- We provide MAEB dataset, a valuable resource for assessing and researching the robustness of eye blink detection under varying camera angles.

2. Related Work

Eye blinking serves as a multifaceted indicator of various physiological and psychological states, contributing to multiple domains of human activity and healthcare. First, the rate of blinking has been employed as a reliable metric for evaluating cognitive states [24–26], as well as levels of fatigue and mental stress [1, 2, 20, 27]. The duration of blinks has additionally been shown to be indicative of concentration levels, particularly in drivers [28, 29] and students [3]. Second, eye blinks have been integrated into health monitoring schemes, ranging from smartphone-based visual acuity tests [30] to wearable devices that diagnose specific eye conditions [31]. Further, image-based blink detection methods have been developed to contribute to overall eye health [5, 32, 33]. Third, the communicative potential of eye blinking has been harnessed for interactions with individuals who are immobilized due to emergency situations [34, 35] or patients facing restrictions in movement and speech [36–38].

The detection of eye blinks has been addressed through a variety of methodologies, each with its own merits and limitations. While biometric signals [3, 5] and infrared (IR) data have been utilized for eye blink detection [12, 39], there is a growing interest in image-based approaches owing to their more straightforward applicability across diverse settings. Among the more conventional techniques for image-based eye blink detection is the eye aspect ratio (EAR), which quantifies the ratio between the vertical and horizontal distances of an eyelid as identified through eye landmark extraction [40, 41]. Subsequent analysis may involve either filtering the EAR value and comparing it against a pre-defined threshold [42, 43], or feeding the data into a Long Short-Term Memory (LSTM) network for blink detection [41]. Another notable approach leverages motion vectors to identify blinks [20, 44]. In this method, two images are processed to generate a motion field, representing the movement of the eyelid. The cosine similarity between motion vectors in disparate images is then computed to classify the eye's open or closed state. Nevertheless, it must be acknowledged that many image-based blink detection techniques have been primarily designed for controlled indoor settings, often assuming uniform lighting conditions and fixed viewing angles [45, 46]. Moreover, the efficacy of these approaches is frequently contingent upon the adaptability of their predefined thresholds [45].

To surmount the inherent limitations of traditional image-based techniques for eye blink detection, recent research has increasingly employed deep learning architectures. Convolutional Neural Networks (CNNs) have been particularly instrumental, often utilized in binary classifiers that operate on single-image inputs. These deep learning-based models have been tailored for an array of applications including student attention level assessment [15], neurological disorder diagnosis [47], and fatigue detection [16]. Notably, studies leveraging deep residual networks have exhibited enhanced performance when compared to their traditional CNN counterparts [48].

A subset of the research community has proposed methodologies that necessitate the inclusion of images of both eyes. Approaches such as dual-parallel CNN architectures [17, 49] and binary mask applications on both eye images [50] fall within this category. However, such techniques are susceptible to misclassification, particularly under conditions like low camera frame rates [51].

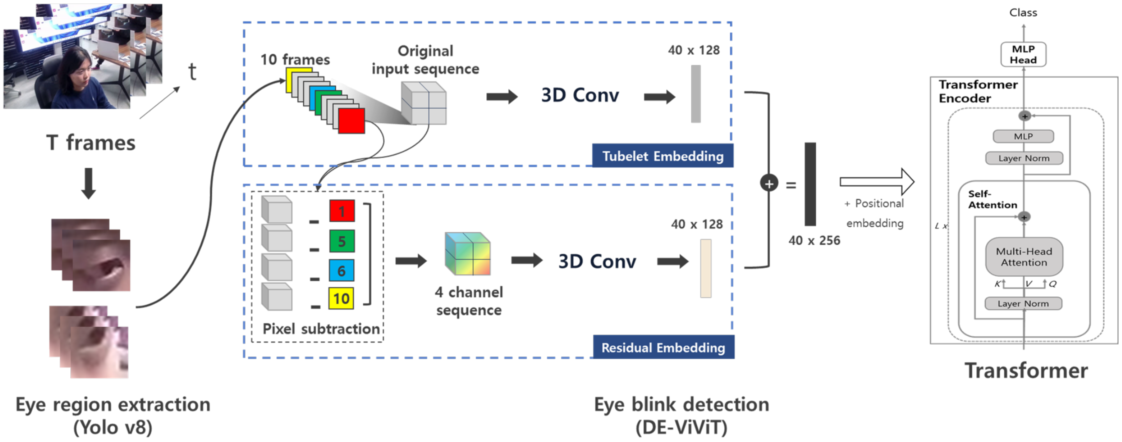


Figure 1. Workflow of Dual Embedding Video Vision Transformer.

Sequential models have also been introduced, exploiting image sequences constructed from multiple consecutive frames. These range from Long Short-Term Memory (LSTM) networks [45] to two-dimensional spectrograms [52] and Long-Term Recurrent CNNs (LRCNs) [18, 19]. While generally more accurate in blink detection compared to single-frame methods, they often fail to identify the exact timepoint of the blink event.

A majority of the existing approaches are grounded in the use of frontal-view images of the eye region. However, some pioneering works have begun to diversify this focus. For instance, studies by Bekhouche et al. [22] and Yang et al. [23] have delved into the impact of camera angles on blink detection accuracy. Others have ventured into testing these models in more robust and uncontrolled environments [22, 45, 46]. In the present study, we aim to extend this line of inquiry by focusing on robust eye blink detection that is applicable in both controlled and uncontrolled settings. Moreover, we introduce MAEB dataset, which distinguishes itself from existing datasets by incorporating diverse camera angles, thereby enabling the representation of the same scene from multiple viewpoints.

3. Proposed Method

3.1. Overview

Our method is based on the detection of blinks based on the spontaneous course of eye movements. One of the reasons is that blinking is a continuous behavior reflecting the eyelid's activity, rather than a simple binary response. To capture the robustness of blink information, it is important to detect blinking at a sequential level [52]. In this study, we developed a Dual Embedding Video Vision Transformer (DE-ViViT) for eye blink classification, leveraging the strong performance of ViT on computer vision tasks. As shown in Figure 1, our approach starts with video frames. When facial features are detected in the frame, both eye regions are cropped according to their coordinates. All 10 frames are passed to the dual embedding module, which includes tubelet embedding and residual embedding. After the input sequence is converted into embedding vectors, they can be processed by the transformer architecture. Finally, the multilayer perceptron (MLP) head predicts whether a blink will occur based on the encoder's output.

3.2. Eye Region Extraction

The first task for eye region extraction is face detection. We chose YOLOv8 [53] for object detection and image segmentation. After the face is detected in the input image, the bounding boxes and five coordinates for the eyes, nose, and mouth are returned. The locations of both eyes were used to crop the left and right eye regions, respectively. Finally, the eye patches are resized to 24 × 24 pixels and stored according to the timestamp.

3.3. Eye Blink Detection

We introduce DE-ViViT as a robust framework tailored to detect eye blinks in a broader spectrum of scenarios, extending beyond constrained conditions and frontal camera views. The architecture is fundamentally an augmentation of the Vision Transformer (ViT) [54], with a focus on capturing pairwise interactions between tokens derived from a single frame sequence. DE-ViViT is comprised of three main components: a dual embedding layer, a transformer encoder, and a classification head.

3.3.1 Dual Embedding

In the dual embedding phase, a video clip consisting of a 10-frame eye region is processed. Traditional embedding techniques are less effective for eye blink sequences due to the continuous and dynamic nature of eyelid movements. To address this, we implement a dual embedding approach that accommodates global context information and consists of two sub-embeddings: tubelet embedding and residual embedding.

Tubelet Embedding is the component which aims to capture macro changes in the frames via spatio-temporal 3D convolution operations. Unlike conventional ViT, our implementation incorporates overlapping stride operations, allowing the convolutional kernel to slide across neighboring pixels to extract significant features. The tubelet embedding network consists of three convolutional layers activated by rectified linear units (ReLU), with the exception of the final layer. These layers use 32/64/128 filters of $3 \times 3 \times 3$ dimensions and are interspersed with max-pooling layers that do not operate along the time axis, preserving temporal nuances. Through this, each convolutional layer facilitates interactions over time and spatial dependencies across frames, achieving efficient computation.

We introduce a novel concept of residual embedding, designed to capture micro-variations within the sequence. We observed substantial changes in eyelid positions at the start, middle, and end frames—deemed key frames—that correspond to pivotal moments in the blinking activity. Residual embedding is generated through a pixel-by-pixel subtraction between the original input sequence and each key frame, thus capturing residual information about the eye's behavior during a blink. The dual embedding phase culminates in two embedding vectors of size 128×40, which are concatenated and supplemented with positional embeddings before proceeding to the transformer encoder.

3.3.2 Transformer

The transformer encoder houses a Multi-Headed Self-Attention (MSA) mechanism [55] and an MLP block. Layer normalization [56] is applied preceding each block, and skip connections are incorporated post-block. We employ a two-headed self-attention architecture to allow the model to encapsulate information from diverse perspectives. Each attention head calculates the query, key, and value vectors from the embeddings to compute the attention score. The MLP block contains a single layer activated by Gaussian Error Linear Units (GELU) [57]. Finally, layer normalization and global average pooling are conducted to predict the class probability.

4. Experimental Details

4.1. Datasets

In this study, we utilized one benchmark dataset and one custom-collected dataset to train and evaluate our proposed model. The benchmark dataset, HUST-LEBW, was collected under in-the-wild conditions. Our custom-collected dataset was specifically designed to investigate the blink detection performance of our model under varying camera angles; this was achieved through controlled experiments conducted in-house. Table 1 provides the number of samples in each dataset used in this research.

Class	Dataset		
	HUST-LEBW		MAEB
	Training set	Test set	
Blink	740	392	720
Non-Blink	983	497	720
Total	1,723	889	1,440

Table 1. Number of Samples in Each Dataset.

4.1.1 HUST-LEBW

In the present study, the HUST-LEBW dataset serves as the primary resource for both training and evaluation of our proposed model. HUST-LEBW is a comprehensive dataset that compiles eye blink samples from 20 different commercial movies, encompassing a broad array of attributes such as the name of the movie, filming location, style, and premiere time. The dataset is comprised of 673 video samples, each extracted from unique movie scenes. Each video sample in HUST-LEBW is either made up of a sequence of 10 or 13 images. Given that our proposed model utilizes sequences of 10 frames, we adapted the dataset to fit this requirement. Specifically, for sequences composed of 13 images, we excluded the last three frames, thus standardizing the input to 10-frame sequences for all samples. This methodological decision not only ensures compatibility with our proposed model but also facilitates a more robust evaluation, accounting for various factors like changes in lighting, camera angles, and subject movement present in the diverse set of movie scenes within the HUST-LEBW dataset.

4.1.2 MAEB

In addition to the HUST-LEBW, this study also utilized another dataset for performance evaluation, MAEB, specifically assembled to analyze the robustness of our proposed model concerning varying camera angles.

To compile the MAEB dataset, we recruited 20 participants (6 females and 14 males), maintaining an even spectacle-wearing distribution. The average age of the participants was about 26 years. Each participant underwent three experimental sessions where they were asked to watch a 5-minute documentary video clip. A three-minute rest interval was provided between the sessions.

As illustrated in Figure 2, nine cameras simultaneously captured the facial expressions of the participants from different angles during the video viewing task. Camera 5 was positioned at a distance of approximately 80 cm directly in

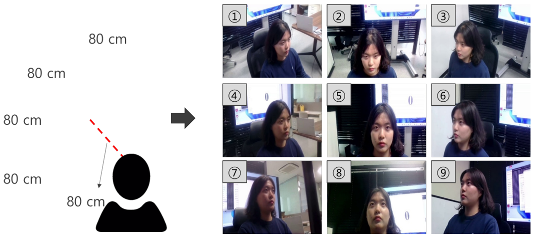


Figure 2. Experimental Setup for Data Collection.

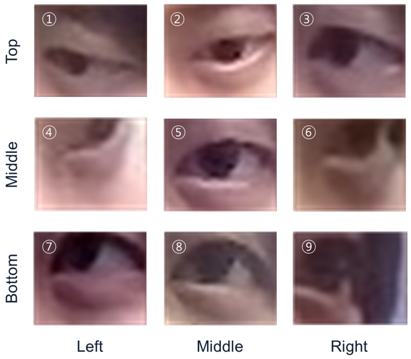


Figure 3. Image Samples in MAEB.

front of the participant. The captured videos were stored at a resolution of 640 x 480 pixels with a frame rate of 30 fps, resulting in a total of 540 video files (20 participants x 3 sessions x 9 cameras).

Initially, blink moments were automatically identified using Tobii Eye Tracker equipment. An eye-tracker signal interruption signified a blink event. For each sequence containing a blink, four frames preceding and five frames following the frame in which the blink was detected by the eye-tracker were extracted, forming a 10-frame sequence. For sequences without blinks, the remaining frames between each blink sequence were utilized to compile 10 consecutive frames. The extracted sequences underwent manual verification by three researchers to form the finalized dataset. We ensured an equal number of blink and non-blink sequences, using a total of 1,440 sequences to evaluate the model's performance. Figure 3 shows the sample images for each camera in the MAEB dataset.

4.2. Baselines

To comprehensively assess the performance of our proposed method, we conducted a comparative evaluation against existing deep learning models designed for image sequence inputs as follows:

- **3D CNN**: As the most straightforward baseline, we employed a 3D Convolutional Neural Network (CNN). This model was adapted from a binary 2D CNN classifier inspired by VGG16 [15], but modified to accommodate 10-frame image sequences instead of individual images.

- **CNN-LSTM**: The second baseline model integrated Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) units. In contrast to previous studies that evaluated the CNN-LSTM architecture under static conditions [18, 19], our implementation followed the approach advocated by Daza *et al.*, with considerations for the number of layers and the output shape. Also, considering the impact of learning rate on CNN-LSTM specifically, an additional CNN-LSTM model with its learning rate set to 0.001 was trained as well. We will refer to the CNN-LSTM model with the learning rate of 0.0001 as CNN-LSTM and the CNN-LSTM model with the learning rate of 0.001 as CNN-LSTM2.

- **Pyramidal Bottleneck Block Network (PBBN)**: Our final baseline was the PBBN, a specialized architecture utilizing pyramid bottleneck (PB) blocks. This model offers several configurations of PB blocks and branches. For the purposes of our study, we chose three configurations as baseline algorithms: one with 2 PB blocks and 2 branches (P2B2), another with 2 PB blocks and 3 branches (P2B3), and the last with 3 PB blocks and 3 branches (P3B3) [22].

Also, because our proposed model is fundamentally based on the Video Vision Transformer (ViViT) architecture. We evaluated its efficacy through an ablation study with the following variants:

- **ORG-ViViT [58]**: This variant employed the original ViViT model architecture, with the only modification being an adjusted image size to match the specifications of our dataset.

- **T-ViViT**: This version incorporated tublet embeddings into the original ViViT framework in line with our proposed method. However, residual embedding was not implemented in this variant.

Through this rigorous comparative evaluation, we aim to establish the advantages and robustness of our proposed model relative to existing methods and its own variations.

4.3. Training and Evaluation

In order to ensure a fair and consistent evaluation, all models were trained using the HUST-LEBW training dataset with random horizontal flip as data augmentation. The hyperparameters for each model were kept identical across all experimental conditions as follows: Image shape is 24×24×1, frame number per sequence is 10, batch size is 32, learning rate is 0.0001, epoch is 100. The exception

Dataset	Metric	3DCNN	CNNLSTM	CNNLSTM2	P2B2	P2B3	P3B3	ORG-ViViT	T-ViViT	DE-ViViT
HUST-LEBW (Testset)	Precision	.809	.691	**.856**	.796	.814	.775	.689	.861	.851
	Recall	.828	.760	.849	.839	.838	.810	.776	.817	**.858**
	F1-score	.818	.724	.852	.815	.825	.789	.730	.837	**.853**
MAEB	Precision	.949	.603	.803	.928	.939	.871	.739	.941	**.981**
	Recall	**.829**	.654	.705	.756	.797	.731	.663	.712	.821
	F1-score	.885	.627	.749	.828	.862	.793	.698	.811	**.894**

Table 2. Performance on HUST-LEBW (Testset) and MAEB.

Method	Eye	Recall	Precision	F1 score
Soukupova and Cech [59]	Left	.361	.647	.463
	Right	.302	.576	.396
Hu *et al*. [45]	Left	.541	.892	.674
	Right	.444	.767	.563
Blink detection+ [43]	Both	.590	.801	.679
InstBlink [46]	Both	.976	.566	.717

Table 3. Benchmark Scores on HUST-LEBW (Testset).

for the learning rate is the additional CNN-LSTM model (CNN-LSTM2) we trained with a learning rate of 0.001. The choice of the final model was determined based on the validation accuracy during the training process. Training was conducted on a computational setup featuring an Nvidia TITAN RTX GPU, complemented by an Intel i9-9900X CPU and 125GB of RAM. The best-performing model was saved for subsequent testing and evaluation, ensuring that our comparisons are both consistent and robust.

For performance evaluation, we primarily utilized the test dataset from HUST-LEBW. Metrics such as Precision, Recall, and F1 score were employed, with the blink event serving as the target class. These evaluation metrics are commonly used in existing eye blink detection studies, as corroborated by several prior works [15, 19, 22, 40, 45, 46]. Like some previous studies that reported performance separately for the left and right eyes when using the HUST-LEBW dataset, we have grouped the detection results into three categories: both, left eye, and right eye, to facilitate comprehensive comparison.

Furthermore, we extended our evaluation to include our in-house collected MAEB dataset. The MAEB dataset allows us to analyze the performance variation according to different camera angles. For each model, performance metrics were separately aggregated for all nine shooting angles. While it was evident that the highest performance would likely be achieved with the frontal shots from camera 5, the study aimed to analyze how much the performance would degrade at different angles. This multi-angle evaluation provides a robust measure of our model's adaptability and effectiveness across various recording conditions.

5. Results

5.1. Overall Accuracy

Table 2 presents the evaluation results on both the HUST-LEBW and MAEB datasets. For the HUST-LEBW dataset, our proposed method and the CNN-LSTM2 model exhibited the highest performance metrics. While a direct comparison may be challenging due to differences in the eye region extraction procedures (our study employs YOLOv8), both methods outperformed benchmark scores reported in prior works, as shown in Figure 3.

For the MAEB dataset, the proposed method led the performance charts, closely followed by the 3DCNN model. Intriguingly, the CNN-LSTM2 model, which excelled on the HUST-LEBW dataset, showed a relative performance drop on the MAEB dataset. This could indicate that the CNN-LSTM2 model may be overfitting to the HUST-LEBW dataset and is less robust to variations in camera angles.

In summary, our proposed method not only achieved impressive results on the well-known benchmark dataset for eye blink detection in wild scenarios, HUST-LEBW, but also maintained consistent performance across a newly acquired dataset featuring diverse camera angles. This underscores the generalizability and robustness of the proposed model in addressing eye blink detection problems across a variety of recording conditions.

5.2. Performance Variation Across Camera Angles

The unique design of the MAEB dataset, encompassing eye blink images captured from nine distinct camera angles, offers an unprecedented opportunity for an in-depth analysis of blink detection performance on a per-angle basis. Figure 4 delineates the angle-specific performance of the various models, including our proposed DE-ViViT and the 3DCNN model, both of which exhibited exceptional overall results. Notably, DE-ViViT showed a more consistent performance across the range of camera angles compared to other models.

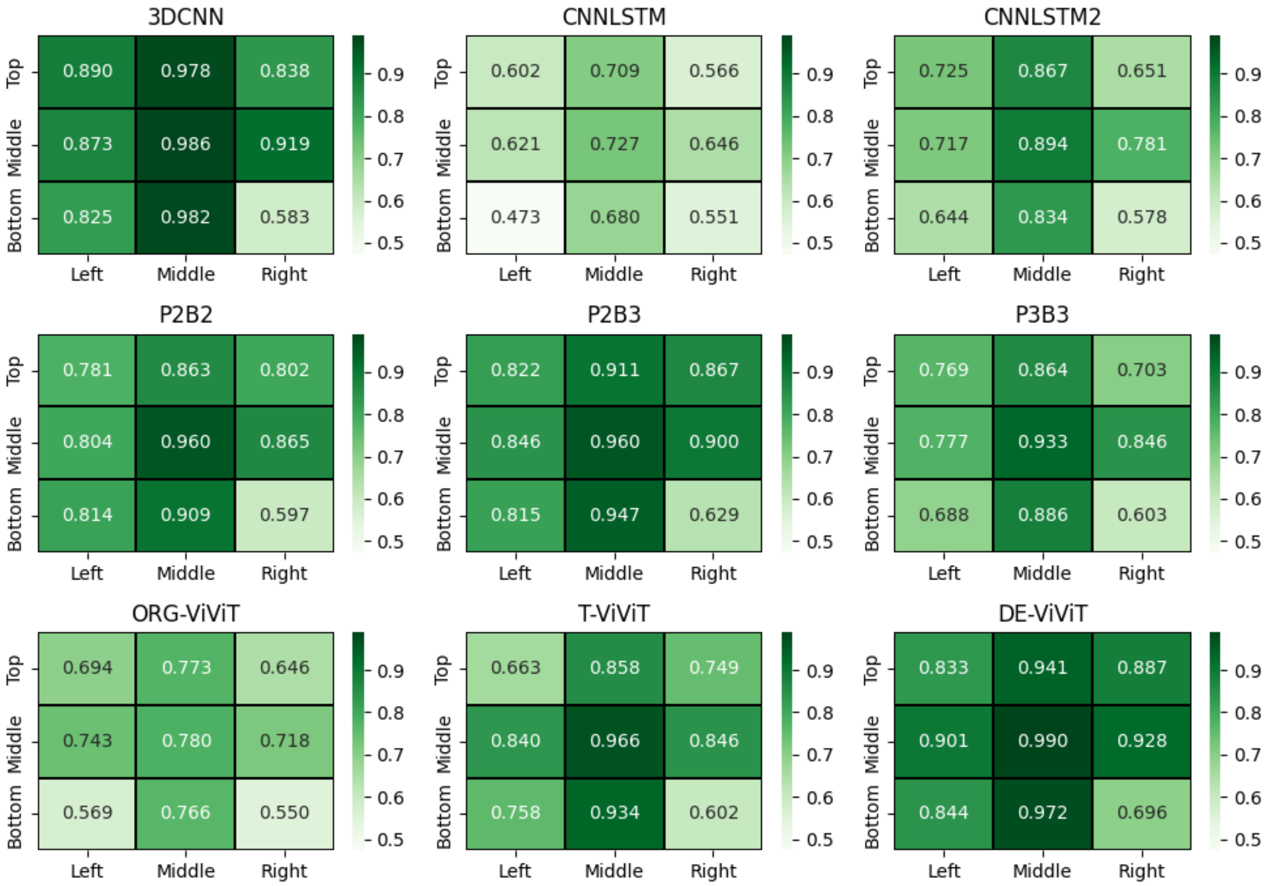


Figure 4. Camera-Specific Performance Variation on MAEB Dataset.

Despite the relatively uniform performance displayed by the proposed DE-ViViT model, a One-way Repeated Measures ANOVA test revealed statistically significant differences in detection outcomes across various camera angles ($F(8, 152) = 9.681$, $p < .001$, $eta^2 = 0.289$). The Post-hoc test results, demonstrated in Figure 5, further substantiate these variations through Tukey's HSD test. Detection performance was found to be significantly reduced when images were captured from cameras positioned at the 1st (top-left), 7th (bottom-left), and 9th (bottom-right) angles, in comparison to those from the frontal 5th camera.

These performance disparities align well with the broader trend observed across all models: the detection capabilities deteriorate when the camera is situated at corner angles, especially at the lower corners (namely, cameras 1, 3, 7, and 9). This reduction in performance is hypothesized to arise due to the decreasing visibility of the eye region as the camera angle deviates from the frontal view. Capturing the nuanced movements of the eyelids becomes increasingly challenging, particularly when the viewpoint is below the eye level.

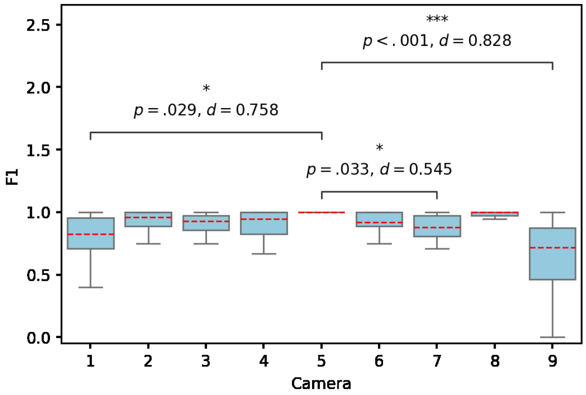
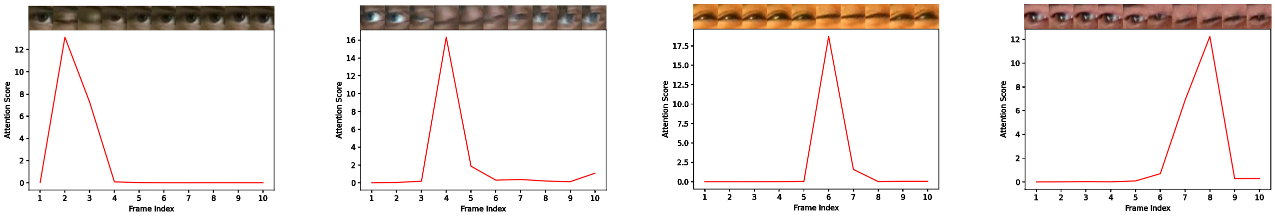


Figure 5. Results of the Post-hoc Test on DE-ViViT's F1 Scores Across Cameras.

In summary, while our proposed DE-ViViT model exhibits robust and consistent performance across a wide array of camera angles, the study identifies specific angles that constitute vulnerable points in blink detection. These vulnerabilities signify areas for future research and potential improvements in model design.



(a) Sample 1 (b) Sample 2 (c) Sample 3 (d) Sample 4

Figure 6. Visualization of Average Attention Scores.

6. Discussion

The results show that Based on our findings, the proposed model demonstrates robust performance in Eye Blink Detection. However, it is imperative to acknowledge that there are areas requiring further refinement and avenues for future research.

First, although DE-ViViT reduced the performance gap across various camera angles, a statistically significant drop in performance was still observed, particularly when the camera was placed at the corners. By leveraging the MAEB dataset, future research could focus on enhancing the robustness of eye blink detection algorithms to camera angle variations, thereby mitigating this drawback.

Second, the current study employs a two-stage pipeline: eye region detection followed by eyeblink identification. While such a pipeline is common in eyeblink detection research, its utility in real-time applications may be constrained due to the inherently fast nature of eyeblinks, which typically last between 0.1 to 0.4 seconds [60]. In our experiments, DE-ViViT took an average of 15.31 ms to process a sequence of 10 frames, suggesting its potential for real-time deployment. However, this time measure does not include the entire pipeline—from camera input to eye region detection to eyeblink detection—and thus, a more comprehensive real-time evaluation is needed.

Third, one promising avenue for future research is the utilization of attention scores for precise blink timing. Our model employs self-attention mechanisms. When the embedding vector is fed into the encoder, the query, key, and value for each head are calculated as follows:

$$[q, k, v] = zU_{qkv}, U_{qkv} \in R^{D \times 3D_h}, \tag{1}$$

where z is the embedding vector, U_{qkv} is weight matrix of query/key/value, and D_h represents the feature dimension per head. These are projections of the input onto the three other spaces. The attention score (A) of each frame is computed by Softmax probability distribution as follows:

$$A = softmax(\frac{qk^T}{\sqrt{D_h}}), A \in R^{N \times N}$$
$$SA(z) = Av, \tag{2}$$

We observed a correlation between such attention scores from various heads and the moments of eye blinking. Figure 6 showcases this relationship. Exploiting these attention scores could lead to more nuanced applications in eyeblink detection.

Lastly, exploring how eyeblink detection from various camera views can be applied in real-world scenarios presents an exciting research opportunity. Situations requiring user assistance [61] or driver drowsiness detection [62] on mobile devices are sensitive to the direction of the user's face. In deception detection, the discreet placement of cameras for data acquisition is often crucial [63]. In these contexts, the MAEB dataset, with its multi-angle eyeblink data, could prove invaluable.

7. Conclusion

We introduce a dual-embedding video vision transformer for robust eye blink detection across diverse environments. We further enrich the evaluation landscape with our MAEB dataset, capturing eye blinks from multiple camera angles. Empirical comparisons with other baseline methods affirm the robustness of our model. Our work also highlights areas for improvement, such as detection vulnerabilities at certain camera angles. Exploring the impact of other areas of robustness such as facial movement and facial expression may also provide a more comprehensive analysis. Finally, we suggest future research directions, including leveraging attention scores for more precise blink timing detection. Our contributions aim to advance both the academic and practical applications of eye blink detection technology.

8. Acknowledgments

This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023 (Project Name: Cultivation of Next-generation Convergence Culture Technology (CT) Experts through Education, Developments, and Demonstration Projects for the Real-time Adaptive Performance and Exhibition Technologies to the User Reactions in the Metaverse, Project Number: RS-2023-00224524, Contribution Rate: 100%).

References

[1] Elena Magán, M Paz Sesmero, Juan Manuel Alonso-Weber, and Araceli Sanchis. Driver drowsiness detection by applying deep learning techniques to sequences of images. *Applied Sciences*, 12(3):1145, 2022.

[2] Fang Bin, Xu Shuo, and Feng Xiaofeng. A fatigue driving detection method based on multi facial features fusion. In *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 225–229. IEEE, 2019.

[3] Andrea F Abate, Lucia Cascone, Michele Nappi, Fabio Narducci, and Ignazio Passero. Attention monitoring for synchronous distance learning. *Future Generation Computer Systems*, 125:774–784, 2021.

[4] Chandran Jyotsna and Joseph Amudha. Eye gaze as an indicator for stress level analysis in students. In *2018 International conference on advances in computing, communications and informatics (ICACCI)*, pages 1588–1593. IEEE, 2018.

[5] T Sree Sharmila, R Srinivasan, KK Nagarajan, and S Athithya. Eye blink detection using back ground subtraction and gradient-based corner detection for preventing cvs. *Procedia Computer Science*, 165:781–789, 2019.

[6] KS Ahmed. Wheelchair movement control via human eye blinks. *American Journal of Biomedical Engineering*, 1(1):55–58, 2011.

[7] Kristen Grauman, Margrit Betke, James Gips, and Gary R Bradski. Communication via eye blinks-detection and duration analysis in real time. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[8] Amanda Ferrari Iaquinta, Ana Carolina de Sousa Silva, Aldrumont Ferraz Júnior, Jessica Monique de Toledo, and Gustavo Voltani von Atzingen. Eeg multipurpose eye blink detector using convolutional neural network. *arXiv preprint arXiv:2107.14235*, 2021.

[9] Mohit Agarwal and Raghupathy Sivakumar. Blink: A fully automated unsupervised algorithm for eye-blink detection in eeg signals. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1113–1121. IEEE, 2019.

[10] Jürgen Schmidt, Rihab Laarousi, Wolfgang Stolzmann, and Katja Karrer-Gauß. Eye blink detection for different driver states in conditionally automated driving and manual driving using eog and a driver camera. *Behavior research methods*, 50:1088–1101, 2018.

[11] M Sanjeeva Reddy, B Narasimha, E Suresh, and K Subba Rao. Analysis of eog signals using wavelet transform for detecting eye blinks. In *2010 International Conference on Wireless Communications & Signal Processing (WCSP)*, pages 1–4. IEEE, 2010.

[12] Abdolhossein Fathi and Fardin Abdali-Mohammadi. Camera-based eye blinks pattern detection for intelligent mouse. *Signal, Image And Video Processing*, 9:1907–1916, 2015.

[13] Yash S Desai. Driver's alertness detection for based on eye blink duration via eog & eeg. *Int. J. Adv. Comput. Res*, 2(7):93–99, 2012.

[14] Arnab Ghosh, Tania Chatterjee, Sunny Samanta, Jayanta Aich, and Sandip Roy. Distracted driving: A novel approach towards accident prevention. *Adv. Comput. Sci. Technol*, 10(8):2693–2705, 2017.

[15] Roberto Daza, Aythami Morales, Julian Fierrez, and Ruben Tolosana. Mebal: A multimodal database for eye blink detection and attention level estimation. In *Companion publication of the 2020 international conference on Multimodal interaction*, pages 32–36, 2020.

[16] Israt Jahan, KM Aslam Uddin, Saydul Akbar Murad, M Saef Ullah Miah, Tanvir Zaman Khan, Mehedi Masud, Sultan Aljahdali, and Anupam Kumar Bairagi. 4d: a real-time driver drowsiness detector using deep learning. *Electronics*, 12(1):235, 2023.

[17] Roberto Daza, Daniel DeAlcala, Aythami Morales, Ruben Tolosana, Ruth Cobos, and Julian Fierrez. Alebk: Feasibility study of attention level estimation via blink detection applied to e-learning. *arXiv preprint arXiv:2112.09165*, 2021.

[18] Ronan Bennett and Shantanu H Joshi. A cnn and lstm network for eye-blink classification from mri scanner monitoring videos. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3463–3466. IEEE, 2021.

[19] Gonzalo de la Cruz, Madalena Lira, Oscar Luaces, and Beatriz Remeseiro. Eye-lrcn: A long-term recurrent convolutional network for eye blink completeness detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[20] Tomas Drutarovsky and Andrej Fogelton. Eye blink detection using variance of motion vectors. In *European conference on computer vision*, pages 436–448. Springer, 2014.

[21] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[22] Salah Eddine Bekhouche, I Kajo, Y Ruichek, and Fadi Dornaika. Spatiotemporal cnn with pyramid bottleneck blocks: Application to eye blinking detection. *Neural Networks*, 152:150–159, 2022.

[23] Cong Yang, Zhenyu Yang, Weiyu Li, and John See. Fatigueview: A multi-camera video dataset for vision-based drowsiness detection. *IEEE Transactions on Intelligent Transportation Systems*, 24(1):233–246, 2022.

[24] Andrea Casanova, Lucia Cascone, Aniello Castiglione, Michele Nappi, and Chiara Pero. Eye-movement and touch dynamics: a proposed approach for activity recognition of a web user. In *2019 15th international conference on signal-image technology & internet-based systems (SITIS)*, pages 719–724. IEEE, 2019.

[25] Alfonso Magliacano, Salvatore Fiorenza, Anna Estraneo, and Luigi Trojano. Eye blink rate increases as a function of cognitive load during an auditory oddball paradigm. *Neuroscience Letters*, 736:135293, 2020.

[26] Danilo Avola, Luigi Cinque, Gian Luca Foresti, and Daniele Pannone. Automatic deception detection in rgb videos using facial action units. In *Proceedings of the 13th International Conference on Distributed Smart Cameras*, pages 1–6, 2019.

[27] R Martins and JM Carvalho. Eye blinking as an indicator of fatigue and mental load—a systematic review. *Occupational safety and hygiene III*, 10:231–235, 2015.

[28] Zuojin Li, Shengbo Eben Li, Renjie Li, Bo Cheng, and Jinliang Shi. Online detection of driver fatigue using steering wheel angles for real driving conditions. *Sensors*, 17(3):495, 2017.

[29] Turker Tuncer, Sengul Dogan, Fatih Ertam, and Abdulhamit Subasi. A dynamic center and multi threshold point based stable feature extraction network for driver fatigue detection utilizing eeg signals. *Cognitive neurodynamics*, 15:223–237, 2021.

[30] Hillary Rono, Andrew Bastawrous, David Macleod, Emmanuel Wanjala, Stephen Gichuhi, and Matthew Burton. Peek community eye health-mhealth system to increase access and efficiency of eye health services in trans nzoia county, kenya: study protocol for a cluster randomised controlled trial. *Trials*, 20(1):1–12, 2019.

[31] Dillam Díaz, Nicholas Yee, Christine Daum, Eleni Stroulia, and Lili Liu. Activity classification in independent living environment with jins meme eyewear. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9. IEEE, 2018.

[32] Akihiro Kuwahara, Rin Hirakawa, Hideki Kawano, Kenichi Nakashi, and Yoshihisa Nakatoh. Eye fatigue prediction system using blink detection based on eye image. In *2021 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–3. IEEE, 2021.

[33] Akihiro Kuwahara, Rin Hirakawa, Hideki Kawano, Kenichi Nakashi, and Yoshihisa Nakatoh. Blink detection using image processing to predict eye fatigue. In *Human Interaction, Emerging Technologies and Future Applications III: Proceedings of the 3rd International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET 2020), August 27-29, 2020, Paris, France*, pages 362–368. Springer, 2021.

[34] Puneet Singh Lamba, Deepali Virmani, and Oscar Castillo. Multimodal human eye blink recognition method using feature level fusion for exigency detection. *Soft Computing*, 24(22):16829–16845, 2020.

[35] Puneet Singh Lamba and Deepali Virmani. Information retrieval from emotions and eye blinks with help of sensor nodes. *International Journal of Electrical and Computer Engineering*, 8(4):2433, 2018.

[36] Emanuele Cardillo, Gaia Sapienza, Changzhi Li, and Alina Caddemi. Head motion and eyes blinking detection: A mm-wave radar for assisting people with neurodegenerative disorders. In *2020 50th European Microwave Conference (EuMC)*, pages 925–928. IEEE, 2021.

[37] William C Francis, C Umayal, and G Kanimozhi. Brain-computer interfacing for wheelchair control by detecting voluntary eye blinks. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 9(2):521–537, 2021.

[38] Emmanuel Jadesola Adejoke and Ibiyemi Tunji Samuel. Development of eye-blink and face corpora for research in human computer interaction. *Development*, 6(5), 2015.

[39] Artem A Lenskiy and Jong-Soo Lee. Driver's eye blinking detection using novel color and texture segmentation algorithms. *International journal of control, automation and systems*, 10:317–327, 2012.

[40] Simone Dari, Nico Epple, and Valentin Protschky. Unsupervised blink detection and driver drowsiness metrics on naturalistic driving data. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2020.

[41] Reza Ghoddoosian, Marnim Galib, and Vassilis Athitsos. A realistic dataset and baseline temporal model for early drowsiness detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[42] Afraa Z. Attiah and Enas F. Khairullah. Eye-blink detection system for virtual keyboard. In *2021 National Computing Colleges Conference (NCCC)*, pages 1–6, 2021.

[43] Tran Thanh Phuong, Lam Thanh Hien, Do Nang Toan, and Ngo Duc Vinh. An eye blink detection technique in video surveillance based on eye aspect ratio. In *2022 24th International Conference on Advanced Communication Technology (ICACT)*, pages 534–538, 2022.

[44] Jaikrishna Mohanakrishnan, Satoshi Nakashima, Junichi Odagiri, and Shanshan Yu. A novel blink detection system for user monitoring. In *2013 1st IEEE Workshop on User-Centered Computer Vision (UCCV)*, pages 37–42. IEEE, 2013.

[45] Guilei Hu, Yang Xiao, Zhiguo Cao, Lubin Meng, Zhiwen Fang, Joey Tianyi Zhou, and Junsong Yuan. Towards real-time eyeblink detection in the wild: Dataset, theory and practices. *IEEE Transactions on Information Forensics and Security*, 15:2194–2208, 2019.

[46] Wenzheng Zeng, Yang Xiao, Sicheng Wei, Jinfang Gan, Xintao Zhang, Zhiguo Cao, Zhiwen Fang, and Joey Tianyi Zhou. Real-time multi-person eyeblink detection in the wild for untrimmed video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13854–13863, 2023.

[47] Sharik Ali Ansari, Koteswar Rao Jerripothula, Pragya Nagpal, and Ankush Mittal. Eye-focused detection of bell's palsy in videos. *arXiv preprint arXiv:2201.11479*, 2022.

[48] Ki Wan Kim, Hyung Gil Hong, Gi Pyo Nam, and Kang Ryoung Park. A study of deep cnn-based classification of open and closed eyes using a visible light camera sensor. *Sensors*, 17(7):1534, 2017.

[49] Kevin Cortacero, Tobias Fischer, and Yiannis Demiris. Rt-bene: A dataset and baselines for real-time blink estimation in natural environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[50] Ritabrata Sanyal and Kunal Chakrabarty. Two stream deep convolutional neural network for eye state recognition and blink detection. In *2019 3rd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, pages 1–8. IEEE, 2019.

[51] Essa R Anas, Pedro Henriquez, and Bogdan J Matuszewski. Online eye status detection in the wild with convolutional neural networks. In *International conference on computer vision theory and applications*, volume 7, pages 88–95. SciTePress, 2017.

[52] Youngjun Cho. Rethinking eye-blink: Assessing task difficulty through physiological representation of spontaneous blinking. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–12, 2021.

[53] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023.

[54] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[56] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[57] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[58] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[59] Jan Cech and Tereza Soukupova. Real-time eye blink detection using facial landmarks. *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, pages 1–8, 2016.

[60] Yuang Zhang, Xiangwei Zheng, Weizhi Xu, and Hong Liu. Rt-blink: A method toward real-time blink detection from single frontal eeg signal. *IEEE Sensors Journal*, 23(3):2794–2802, 2023.

[61] Md Talal Bin Noman and Md Atiqur Rahman Ahad. Mobile-based eye-blink detection performance analysis on android platform. *Frontiers in ICT*, 5:4, 2018.

[62] B Rajkumarsingh and D Totah. Drowsiness detection using android application and mobile vision face api. *R&D Journal*, 37:26–34, 2021.

[63] Brandon S Perelman. Detecting deception via eyeblink frequency modulation. *PeerJ*, 2:e260, 2014.