



Multi-Modal Behaviour Monitoring in Virtual Assessments Using ProctorGuard AI

Shah Riya Pranav¹, Helen K Joy^{2*} , Nikhil Niket¹, Sridevi. R² , Warusia Yassin³ , Manjunath R. Kounte⁴ 

¹ Talview, Bengaluru 60102, India

² Department of Computer Science, Centre for AI, CHRIST University, Bangalore 560029, India

³ Fakulti Kecerdasan Buatan dan Keselamatan Siber, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Malaysia

⁴ Department of Electronics and Communication Engineering, HKBK College of Engineering, Bengaluru 560045, India

Corresponding Author Email: helenjoy88@gmail.com

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ijss.160818>

ABSTRACT

Received: 13 October 2025

Revised: 12 December 2025

Accepted: 10 January 2026

Available online: 31 August 2026

Keywords:

lip-sync, gaze, Mahalanobis, pose, proctoring, synchronisation assessment, deep learning

Remote assessments have grown significantly in both education and hiring processes due to the COVID-19 pandemic, but have also exposed issues regarding authenticity and fairness. Current AI-based proctoring methods can be circumvented by sophisticated forms of cheating (e.g., pre-recorded audio playback and off-screen communication) because these systems examine each cue individually and therefore do not account for how an individual's behaviours interact across multiple modes. The purpose of this study is to describe a new framework that integrates lip-sync detection with gaze tracking to address gaps left by current AI-based proctoring methods. The lip-sync detection will employ Mahalanobis distance to measure the discrepancy between what is being said and how it is being displayed through lip movement. Gaze direction will be represented geometrically through eye and head landmark points to determine when an individual is gazing away from their screen during the assessment. Once both modules are aligned in time so that audio and video events occur at the same point, the system will generate a report for the examiner to use after the assessment. The results of the experiment showed a high degree of accuracy in detecting lip-sync violations (92%) and in identifying when an individual deviates from looking directly at the screen (90%). The proposed system will enable a systematic, data-driven evaluation of the validity and fairness of remote evaluations.

1. INTRODUCTION

The use of remote technology is creating new issues for evaluation, including fairness, authenticity, and security. Because of the lack of physical non-verbal cues (and) a controlled environment, there is a higher chance of fraudulent activities being conducted, such as someone else talking and/or off-camera cheating during an interview/evaluation. In order for organisations, educational institutions, etc., to establish validity in their use of virtual technology for evaluations, the reliability of candidate responses and the exclusive attention of the participant are now at the forefront of concerns for them.

A successful virtual evaluation does not just depend on the technology used; it also depends on the evaluator's ability to place trust in both the candidate's behaviour and honesty. Proctoring and monitoring in evaluations, like all other types of evaluations, have changed over time from solely manual processes to using sophisticated automated technologies. Human monitors traditionally watched candidates via live video feeds while conducting evaluations, leading to high costs, monitor fatigue, and inconsistent results. Newer models utilise artificial intelligence (AI) to monitor candidate behaviour via facial recognition, screen captures, and audio

capture. While AI-based monitoring tools do provide some level of anomaly detection, most cannot analyse an individual's behaviour, specifically synchronising audio and video elements and understanding complex gaze patterns. Also, many of the newer systems rely on reviewing recordings after the evaluation session rather than real-time monitoring of candidate behaviour, limiting the ability to respond immediately to inappropriate behaviour.

The multi-modal monitoring method will address the gap between current methods by combining lip-sync detection and gaze tracking into a single monitoring system. The movement of the lips with regard to spoken audio will be examined using facial landmark points and statistical distance measures to determine if there are discrepancies (i.e., mismatch) which could be indicative of cheating. Simultaneously, eye and head movements will be tracked to assess whether the subject remains focused on the display and/or is frequently distracted. Together, the two types of tracking provide an overall behavioural profile for the subject and enable the detection of anomalies in both real time and post-evaluation through structured reports; thus enhancing the validity and trustworthiness of remote assessments in academic and professional contexts.

2. LITERATURE REVIEW

This research aims to design an anomaly detection system with multiple behavioural signals (gaze direction & lip sync) for the improvement of remote proctoring by detecting two common types of misconduct in virtual testing environments: off-screen distractions and lip sync impersonation/cheating. The development of this real-time multi-modal anomaly detection framework will be supported by a comprehensive literature review on previous work in both areas (behavioural signal processing and fraud prevention).

2.1 Gaze tracking and behavioural monitoring

Gaze estimation research spans more than two decades, and tracing its development reveals a pattern worth noting: a gradual departure from controlled laboratory conditions toward the unpredictable circumstances of ordinary use. That departure, as this section argues, remains incomplete where exam proctoring is concerned, and the various technologies used are pointed in Figure 1.

Using stereo vision to recover head orientation and gaze direction within three-dimensional space provides a sensible starting point [1, 2]. Achieving solid geometric precision came at the cost of a calibrated multi-camera rig, limiting practical adoption outside research settings. Subsequent efforts focused on eliminating these hardware dependencies. Key advances include easing calibration requirements while allowing free head movement [3], as well as addressing variable illumination [4]. Accounting for variable lighting is particularly vital in remote examinations, where uncontrolled ambient conditions and consumer webcams are standard. Evaluating these systems under such domestic conditions remains an open challenge, leaving it unclear whether these gains transfer effectively to consumer hardware or merely displace the problem.

A second, largely independent strand of research prioritised accessibility over accuracy. Embedded eye-detection systems [5, 6], contactless tracking hardware [7], and gaze-driven assistive interfaces [8] pushed sensing capability onto cheaper, smaller devices, and this body of work is what makes browser-based monitoring feasible at all. Yet accessibility of this sort tends to come at a price rarely stated outright. A simplified optical setup running on a low-resolution consumer sensor [9] may suffice for coarse judgments, such as detecting whether eyes are open, but distinguishing a glance toward scratch paper from a covert glance toward a hidden phone is a considerably harder task for the same equipment. Webcam-based trackers built on this foundation [10, 11] confirmed that real-time monitoring on ordinary hardware was achievable, but achievability of tracking says nothing about sufficiency for detecting dishonest behaviour, and the two claims are frequently treated as interchangeable when they are not.

Demonstrating that multivariate normal mixture models can run in real time over high-dimensional hyperspectral imagery via GPU parallelisation without sacrificing accuracy provides a foundational assumption for later anomaly detection systems [9]. Although developed for remote sensing rather than remote proctoring, this framework has largely been adopted by webcam-based systems without direct verification. Transitioning from spectral pixel arrays to low-resolution facial video introduces significant differences in noise profiles, dimensionality, and temporal behaviour. Validating whether this favourable trade-off between throughput and

precision holds across that domain transition remains an unexamined gap within the remote examination literature.

Developing anomaly detection for gaze and conduct has followed three successive methodologies. Initial efforts relied on explicit, rule-based heuristics: deploying adaptive thresholds to trigger behavioural alerts [12] or applying geometric formulations to detect deviations in visual attention. While offering low computational overhead and high interpretability, these rule-based strategies [13] remain brittle to environmental shifts, as parameters tuned for specific camera angles, illumination levels [14, 15], or user cohorts generalise poorly without cross-condition validation. A subsequent line of work transitioned to learned representations, notably by implementing YOLOv4 to automate gaze-zone classification [16]. Although moving from heuristic rules and geometric models to deep architectures aligns with broader trends in computer vision, benchmarking these three paradigms against one another on a unified exam-monitoring dataset remains unaddressed. Lacking such a comparative evaluation, justifying the heightened computational demands, training overhead, and architectural complexity of deep learning classifiers over well-calibrated geometric baselines remains difficult.

Separately, open and wearable platforms contributed to accessibility in a different sense. The Pupil platform [13] and the iShadow wearable device [14] lowered the entry barrier for gaze research broadly, giving researchers without proprietary equipment a way to build and test their own applications. Their contribution to the field at large is not in question. Their relevance to proctoring specifically is less certain, since both were designed around a cooperative subject rather than one actively attempting to circumvent detection, precisely the scenario an exam-integrity tool must anticipate.

Integrating these separate methodological threads into end-to-end deployable pipelines represents the current state of practical exam monitoring. Developing browser-based architectures for real-time proctoring [17, 18], as well as combining landmark-based facial modeling with dynamic behavioral profiling [19, 20], reflects the practical frontier of unifying accessible hardware, live inference, and automated anomaly detection.

A primary methodological limitation across these systems is the omission of false positive rates under benign, everyday test-taking behaviours, which are frequently masked within aggregate accuracy metrics. Common, non-fraudulent actions—such as unfocusing the eyes during cognitive reflection, shifting head posture, glancing away to recall information, or checking the time—remain poorly differentiated from deliberate academic dishonesty under current detection logic [12, 15, 16, 19, 20]. Validating this boundary under naturalistic conditions remains critical, as high false alarm rates carry severe academic consequences, rendering systems with high nominal accuracy unreliable for responsible deployment.

Taken as a whole, this body of research, as mentioned in Table 1 shows that the individual pieces needed for automated exam proctoring—dependable gaze estimation, hardware accessible enough for everyday circumstances, anomaly detection capable of running at scale in real time, and increasingly capable behavioural classification—have each reached a reasonable degree of maturity on their own. What remains missing is convincing proof that these pieces work well once joined together: a system built entirely on consumer webcam equipment, resilient to the lighting and posture

variation typical of a home environment, that can also be shown, under rigorous testing against genuinely innocent variation rather than staged dishonesty alone, not to penalise

students for the ordinary conduct of sitting an exam. Addressing that gap in integration and testing, rather than any single missing technique, is the aim of the present study.

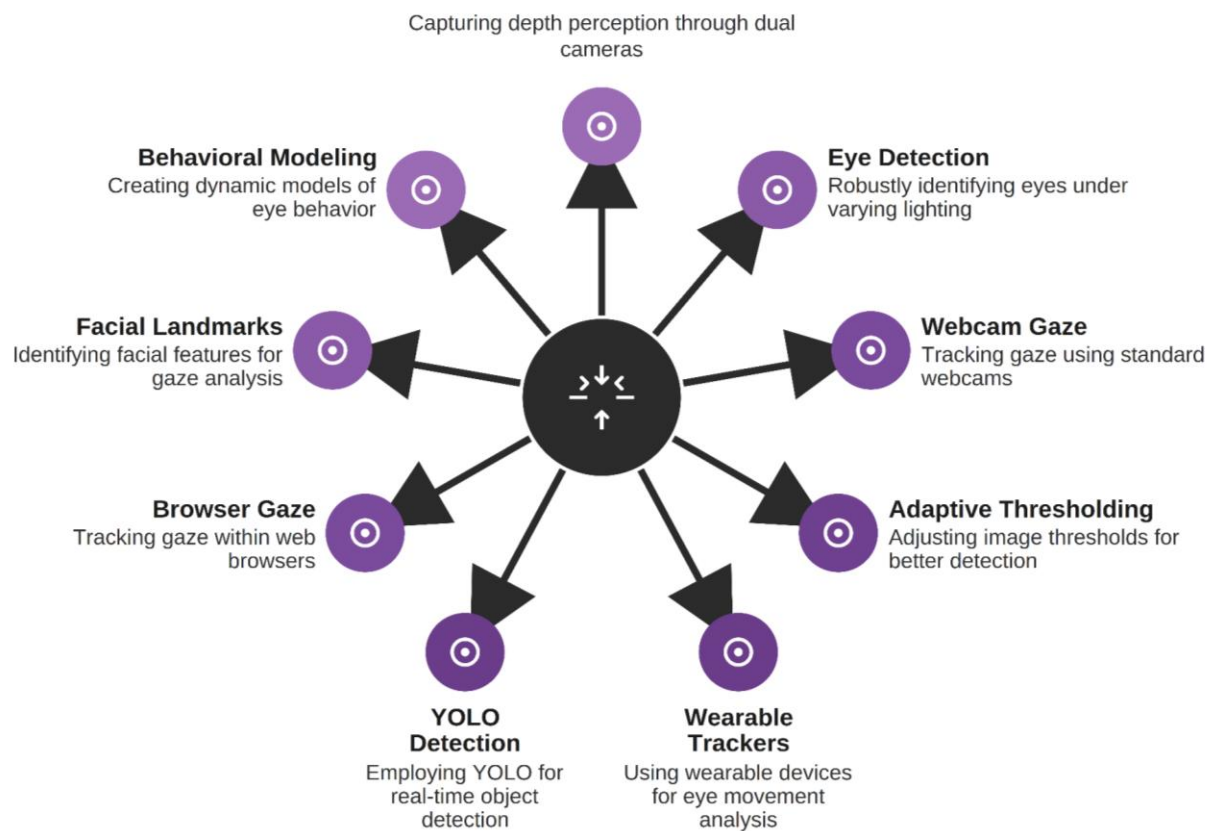


Figure 1. Technology applications in eye tracking

Table 1. Summary of gaze-based monitoring techniques

References	Method/Technology	Contribution	Relevance to Objective
[1, 2]	Stereo vision gaze estimation	Head pose + gaze tracking	Foundational for 3D attention analysis
[4]	Real-time eye detection	Robust under lighting variation	Crucial for uncontrolled environments
[10, 11]	Webcam/mobile eye tracking	Low-cost consumer tools	Practical for online exam settings
[12]	Adaptive thresholding	Behaviour anomaly detection	Core principle in our framework
[13, 14]	Open-source & wearable trackers	Accessible mobile tracking	Enables scalable deployment
[16]	YOLOv4-based eye tracking	Automated gaze zone detection	Deep learning integration
[17, 18]	Browser-based gaze proctoring	Real-time cheating alerts	Closely aligned to exam context
[19]	Facial landmarks + YOLOv5	Abnormal behavior detection	Supports landmark-based gaze monitoring
[20]	Dynamic behaviour model	Cyber-physical anomaly tracking	Real-time anomaly adaptation

2.2 Lip-sync detection and audio-visual anomaly analysis

Deepfakes, fake videos that can be created using AI technology, have caused concern because they can be made to look like someone else, contain video from another time, or even be completely synthetic (man-made) speech. The reason this is such a big deal is that it can be used to create an "impersonation" on an online platform, make fake answers appear as if they were given at different times, or produce fake synthesised speech. Audio-video desynchronisation has been identified as one of the main indicators that a video may be a deepfake [21]. A thorough review provides an overview of the use of both neural networks and 3D modelling in order to synchronise a person's lips with their voice [22]. In addition, the use of a Vision Transformer has been proposed to identify small differences in the way a person's mouth moves when speaking [23].

Reviewing lip-sync generation techniques highlights a

specific focus on achieving high levels of synchronisation fidelity between spoken language and video media across multiple languages and in real time [24, 25]. The use of anomaly detection via lip contour analysis has been proposed [26], alongside fusing Audio-Lip modalities with few-shot learning to enable adaptation to low-data environments [27]. Building upon these concepts, additional methods have been proposed for identifying audio-visual mismatch in deepfake defence [28].

Contributions supporting the above ideas include showing that using Segmented Facial References improves Mouth-Speech Alignment [29], as well as developing a pretrained system called SAM-Wav2Lip++, which uses behavioural realism and lip-sync integrity to evaluate video content [30]. As such, these methods can be applied directly to speech-monitoring modules in proctoring systems.

There is clearly an emerging trend toward utilising real-time gaze tracking and lip-sync verification to improve the integrity

of remote exams, as demonstrated in the literature reviewed. The evidence from the research indicates that gaze-tracking technology has significant promise in helping detect inattention and/or dishonesty during remote exams, using low-cost options (e.g., webcam, free/open-source software). Lip-sync verification is a vital method to identify impersonation and/or pre-recorded answers, especially when using sophisticated architectures like Vision Transformers or audio-visual fusion networks. Combining both methods provides a stronger, more reliable system for the simultaneous, real-time monitoring of a test-taker's visual focus and the authenticity of their spoken words. Both of these additional layers will support the use of adaptive thresholds and dynamic behavioural modelling to help improve the accuracy and sensitivity of detecting anomalies. However, despite this progress, the field is still lacking in the development of standardised benchmarks to allow researchers to compare the results of different studies. Therefore, although there are many promising research directions toward developing robust, real-time, and multi-modal proctoring systems, the research continues toward developing real-time, explainable, and multi-sensory systems for ensuring fairness and trustworthiness in remote assessments. This literature review has provided strong evidence supporting the aims of the proposed research to develop an integrated system combining gaze and lip-sync-based anomaly detection for remote proctoring applications.

The surveyed literature identified some general constraints on the development of systems that utilise real-time gaze tracking and lip-sync verification to enhance the integrity of remote assessments. Gaze tracking has been demonstrated in

the literature to be an effective technology to identify inattention and/or dishonesty during remote exams, including using low-cost solutions such as a standard webcam and open-source platforms. Additionally, lip-sync validation has emerged as a critical tool for identifying impersonation and/or pre-recorded responses, particularly when supported by advanced architectures such as Vision Transformers and audio-visual fusion networks. When used in combination, these two types of data support the creation of more robust and reliable systems that enable the simultaneous, real-time monitoring of both the visual attention of test-takers and the authenticity of their spoken words. The inclusion of adaptive thresholding and dynamic behavioural modelling enhances the ability of these systems to detect anomalies more sensitively and accurately.

In spite of the advancements outlined above, the literature summarised in Table 2 highlights several general constraints on the development of these systems. One major constraint is the lack of standardised benchmarks for evaluating the effectiveness of proctoring systems, which complicates comparing results among studies. Most proctoring systems are tested under controlled/semi-controlled conditions, which do not always represent the full range of possible conditions found in real-world online assessments. Variability in hardware, lighting conditions and diversity in user behaviour may also impact detection accuracy. Additionally, long-standing issues of privacy and explainable AI continue to challenge the development of proctoring systems. Collectively, these constraints highlight the need for continued research into the development of robust, real-time, and generalisable multi-modal proctoring system.

Table 2. Summary of lip-sync verification and audio-visual anomaly detection

References	Method/Technology	Contribution	Relevance to Objective
[21]	Temporal sync model	Deepfake lip-sync detection	Detects pre-recorded fraud
[22]	Literature review	ANN + 3D models for syncing	Broad context on available models
[23]	Vision Transformer	Mouth movement inconsistency	High-accuracy detection of subtle fakes
[24, 25]	Real-time lip-sync generation	Media & multilingual speech syncing	Techniques for reference alignment
[26]	CNN-based prediction for fast HEVC CU partitioning and intra-mode selection.	Shape-based mismatch detection	Directly applicable to lip anomaly flagging
[27]	Audio-lip fusion, few-shot learning	Cuts intra-frame encoding time significantly with minimal rate-distortion loss.	Enables faster, resource-efficient HEVC video compression using deep learning.
[28]	Audio-visual mismatch model	Deepfake speech detection	Temporal frame inconsistency scoring
[29]	Segmented facial references	Speaker-specific syncing	Enhances detection precision
[30]	SAM-Wav2Lip++ pretrained model	Behavioural realism scoring	Real-time error detection in the sync module

3. PROPOSED METHODOLOGY

3.1 Design objectives and operational applicability

The growing usage of remote virtual platforms for educational assessment, certification testing, and job interviewing is introducing new challenges to assess the authenticity and fairness of candidates' evaluations. The limitations of traditional online proctoring systems are well known; they have been unable to reliably identify some of the most common types of cheating (e.g., lip syncing with an audio recording) and are even less able to recognise more sophisticated types of cheating (e.g., someone else taking the test for them). In addition to the lack of integrity and reliability in the evaluation process itself, these limitations put all

genuine test takers at a disadvantage.

This research is motivated by the need to address the previously identified shortcomings (Table 3) and, as such, proposes an advanced AI-based solution using lip-sync detection, gaze tracking, and behavioural anomaly analysis to monitor candidate activity in real-time. This proposed system will utilise computer vision and audio processing to determine if spoken responses were actually generated from the candidate's own mouth movements and therefore to identify potentially fraudulent usage of a candidate's recorded response. Also, by utilising gaze and head pose tracking, the proposed system will be able to identify candidates' distractions (attention being diverted away from the screen) and other visual cues indicating a candidate may be obtaining additional assistance from off-screen resources during the session.

Table 3. Summary of research gaps and contribution

Focus Area	What Existing Studies Achieve	Research Gap Identified	Contribution of This Work
Gaze Tracking	Stereo-vision and landmark-based gaze methods [1, 2], lighting-robust detection [4], webcam-based tracking [10, 11], YOLO-based gaze zone models [16]	Most approaches require controlled lighting, calibration, or high-resolution inputs and do not integrate gaze with other behavioural cues.	Uses gaze angle + head-pose fusion for reliable off-screen detection using standard webcams in real-time
Lip-Sync Verification	Deepfake desynchronisation detection [21], audio-visual sync models [22, 23], lip region anomaly models [26], multimodal fusion [27, 28]	Existing works focus on deepfake detection, not real-time exam proctoring or low-resolution webcam conditions.	Introduces real-time lip landmark modelling + Mahalanobis distance + VAD for detecting pre-recorded or mismatched speech
Anomaly Detection	Thresholding and dynamic behaviour models [12, 20]	Prior works are single-modality and cannot detect coordinated cheating like lip-sync plus off-screen distractions.	Proposes multi-modal fusion of gaze, lip, pose, and audio cues for robust anomaly detection
Proctoring Systems	Browser-based cheating alerts [17, 18], landmark-based abnormal activity detection [19]	Existing systems cannot detect impersonation, pre-recorded audio, or lip-sync cheating.	Provides the first integrated gaze + lip-sync behavioural monitoring system tailored for remote assessment integrity

The end objective is to eliminate human proctor dependency that can lead to over-reliance on human proctors who may experience fatigue, overlook critical issues or have biases; enhance and streamline the monitoring process by utilising automation; provide an inclusive and transparent evaluation environment for all participants of virtual assessments; offer a solution that is scalable, privacy-conscious and in real time. This research has the capability to produce a fair and accurate assessment of participant behaviour through the use of detailed behavioural reporting and accurate detection to enable educational institutions and organisations to maintain the integrity and transparency of remote evaluations.

3.1.1 Purpose

The primary goal of this project is to provide a more secure, fairer and authentic experience for people taking virtual tests (online examinations, video conferencing interviews and certification testing) using an AI-based behaviour monitoring system. Although current proctoring systems can identify obvious forms of cheating, they are unable to detect complex cheating methods (lip syncing to pre-recorded audio, looking at unauthorized materials while giving a test, etc.). Therefore, this project will develop a behavioral monitoring system that uses both computer vision, machine learning and audio analysis in order to automatically identify and prevent unethical behaviour from occurring while the user is taking the test.

The proposed system includes a multi-mode framework that integrates lip-sync verification, gaze tracking, and anomaly detection. In addition, lip landmarks are used to extract lip movement data, which are then used to calculate the Mahalanobis distance to verify if the user's lips are synchronised with their audio. Head pose estimation and eye vector analysis are used to track the user's visual focus and distractibility through Gaze Tracking Modules. These modules combined will provide automated, real-time monitoring of the user's behaviour during the test. Additionally, this system will eliminate the need for human proctors, decrease human bias and error in the monitoring process, and provide detailed post-test reports to support audits and other examinations.

3.1.2 Intelligent monitoring system for virtual examinations

The proposed system integrates multi-modal behavioural analysis to enhance the integrity of remote assessments. Its primary capabilities include:

Lip-Sync Detection: Analyses synchronisation between lip

movements and audio to detect lip-syncing or pre-recorded responses. Lip landmarks are extracted, and the Mahalanobis distance is computed to quantify discrepancies and verify speech alignment.

Gaze Tracking and Screen Monitoring: Uses facial landmarks, head pose estimation, and gaze direction analysis to identify off-screen glances and attention shifts that may indicate potential malpractice.

Real-Time Detection and Alerts: Continuously monitors candidates during the session, generating immediate alerts for abnormal behaviours, such as prolonged gaze diversion or significant lip-sync inconsistencies.

Multi-Modal Application: Designed for versatility across multiple remote evaluation scenarios, including online examinations, virtual interviews, and professional certification tests.

Data Logging and Reporting: Records detected behavioural anomalies and alerts in real time, producing structured reports (e.g., CSV) for transparency and post-session review by evaluators or administrators.

3.1.3 Applicability

The system is suitable for a variety of sectors that require both a secure and reliable method to remotely assess individuals. Academic institutions may utilise the system as a means of monitoring their students throughout the duration of high-stakes exams. Recruitment agencies & HR teams can incorporate the platform as a method of ensuring candidate authenticity within virtual interview processes. Certification bodies can utilise the system to ensure exam integrity within remote testing sessions. The System's Real-Time Functionality, multi-modal verification process, and extensive reporting capabilities make it an all-encompassing solution for contemporary proctoring requirements.

3.2 System architecture of ProctorGuard AI

ProctorGuard AI is an entirely modular and real-time program that utilises webcam/microphone input for real-time identification of abnormal behaviour within virtual test environments (Figure 2). The system combines computer vision, audio processing, and machine learning-based models for detecting abnormal candidate behaviour. The methodology includes the use of five distinct modules to perform the following functions: input acquisition and synchronisation, lip-sync detection, gaze/head pose estimation, AI-based behaviour classification and alert reporting (Figure 3).

First, real-time video and audio data are recorded from the webcam and microphone of the candidate. Video frames are recorded with the assistance of OpenCV, while the audio stream is processed with the assistance of PyAnnote for voice activity detection (VAD). The two streams are synchronised through the process of timestamp alignment, thus maintaining temporal coherence. Preprocessing operations like background noise removal and video frame stabilisation are carried out to improve the quality of the signal.

The lip-sync detection module computes the coherence of the utterance of words and lip movement. Facial landmarks around the mouth area are detected frame by frame by using MediaPipe's Face Mesh model. They are encoded as a feature vector “ x ”.

x is compared to a trained distribution of aligned speech data. The amount of mismatch is measured in terms of the Mahalanobis distance:

$$D_M = \sqrt{\{(x - \mu)^T \Sigma^{-1} (x - \mu)\}} \quad (1)$$

where, μ is the mean vector, and Σ is the covariance matrix derived from a training set of lip motion during genuine speech. A significantly high value of D_M (Eq. (1)) indicates an anomaly between expected and observed lip motion. Simultaneously, VAD provides a binary speech activity flag per frame. If speech is detected while lip motion is absent or vice versa, a possible lip-sync issue is flagged for further classification.

At the same time, head pose estimation and gaze direction estimation are conducted to assess the candidate's visual attention and detect off-screen activity. Based on facial landmarks detected by MediaPipe, the system maps some 3D model points (e.g., the nose tip, the chin, and eye corners) to

corresponding 2D image points. The Perspective-n-Point (PnP) algorithm is utilised to calculate head rotation and translation relative to the camera. The mathematical model used in the transformation is presented as:

$$x_i = K[R|t]X_w \quad (2)$$

In Eq. (2), the 2D coordinate (x_i, x) of an image of a face in the world, X_w , represents the 3D coordinates of the world for each facial landmark, K represents the camera's intrinsic parameters, and R and t represent the rotation and translation of the camera, respectively. The solution of the above vectors can be obtained using OpenCV's solvePnP() method; the vectors are then transformed into the Euler Angles pitch, yaw, and roll to obtain the direction of the head.

Finally, to compute the direction of the eye gaze, we calculate the vector from the pupil centre to the centre of the two eye corners. The gaze vector G is calculated as follows:

$$G = P - M \quad (3)$$

where, P is the pupil centre, and M is the midpoint between the inner and outer eye corners. The angular deviation between this gaze vector and a notional screen-normal vector N is calculated to determine whether the candidate is looking off-screen:

$$\theta = \cos^{-1} \left(\frac{(G \cdot N)}{(|G| |N|)} \right) \quad (4)$$

If the angle of deviation θ (as per Eq. (4)) is larger than a given maximum angle (usually 25° to 30°), the system will flag this frame as having potentially diverted attention.

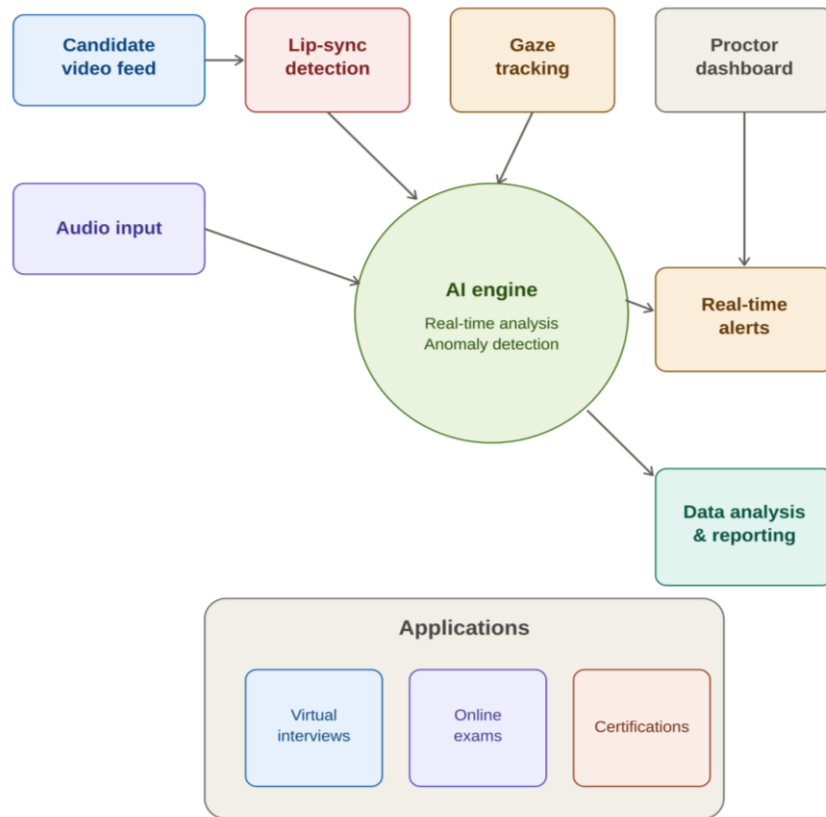


Figure 2. System architecture of artificial intelligence (AI)-based monitoring in virtual assessments using ProctorGuard AI

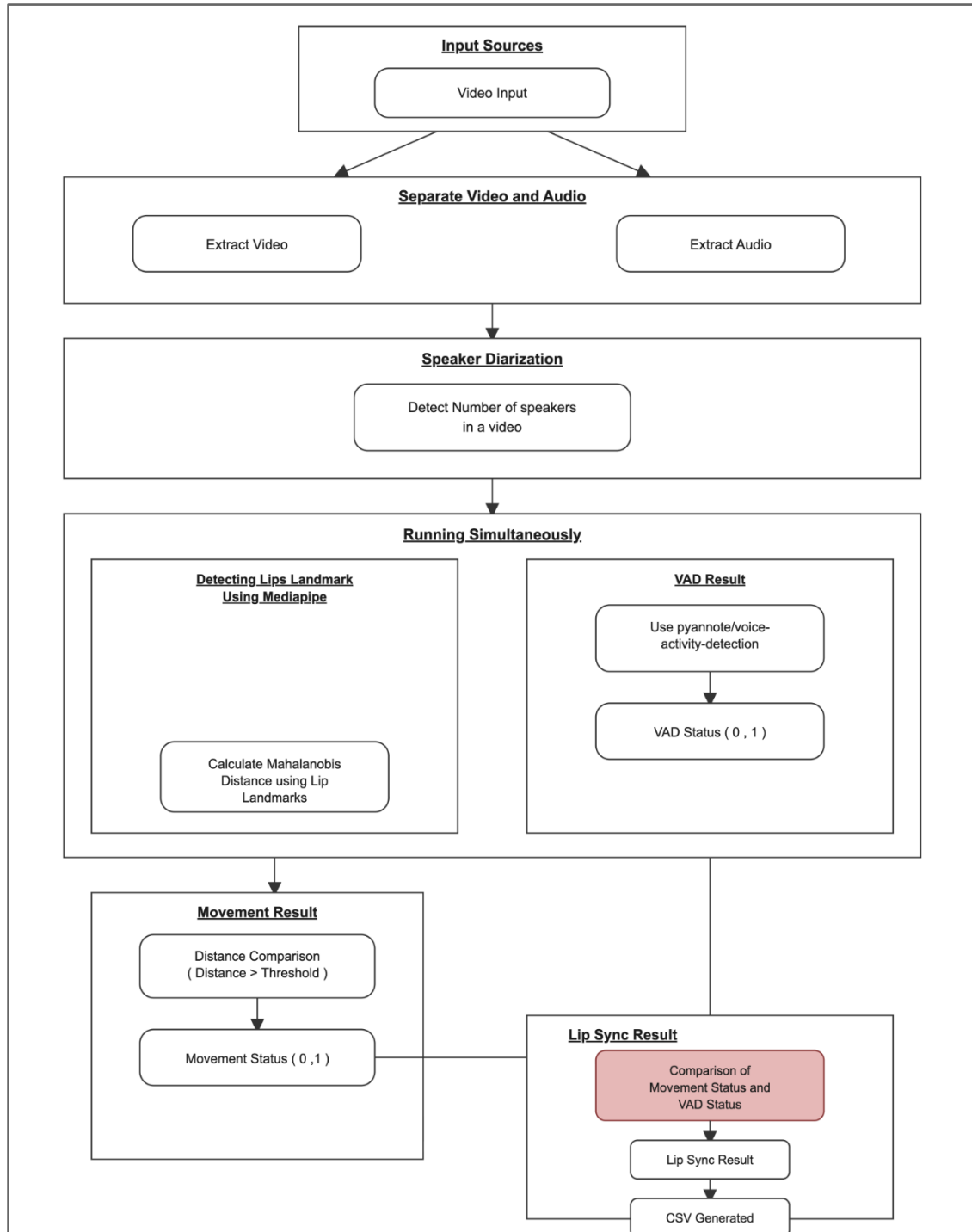


Figure 3. Flow to detect lip syncing

All the identified features, i.e., the Mahalanobis distance DM, the angle of deviation θ , and head pose angles, are then fed into a machine learning classifier with an aggregation of their values forming a composite feature vector. As per the methodology described, the classifiers can be models like Support Vector Machines (SVMs) or even lightweight Convolution Neural Networks (CNNs). Once classified, the results are categorised as normal, lip-sync anomaly, gaze anomaly, or a combination of both types of anomalies. Also, these classifications occur in real time by utilising OpenVINO for acceleration; therefore, there is little to no delay in the classification.

In addition to the classification, upon detecting any potential anomalous behaviour, the system generates real-time alert notifications to remote proctors, and/or logs locally for

later review. Upon notification, every detected event includes a timestamp, the anomaly type, and a confidence score. Additionally, the system produces a structured report in either CSV or JSON format at the completion of each session, which lists all of the detected anomalies and facilitates additional review.

The ProctorGuard AI system utilises a multi-phase implementation strategy to identify behavioural anomalies through two major modules: lip sync detection and eye gaze tracking. The lip sync detection phase starts by capturing real-time video and audio streams from the webcam and microphone, respectively. Next, the video and audio streams are divided into separate video and audio streams, so they can be processed simultaneously. To allow the system to identify who spoke during what segment of time in the case where

multiple speakers were present, advanced audio processing is utilised to perform speaker diarization. Simultaneously, MediaPipe is utilised to identify and track lip landmarks between sequential video frames. Afterwards, a Mahalanobis distance calculation is performed to evaluate the magnitude and legitimacy of lip movement in comparison to the amount of speech activity identified. Voice Activity Detection (VAD) is utilised in conjunction with PyAnnote to parse the audio stream and determine periods of speech. The identified lip movement is then synchronised with the speech intervals to identify inconsistencies that could indicate lip-sync fraud. Finally, all of the identified anomalies are compiled into a structured CSV report for future post-session analysis.

The eye Gaze Tracking Module commences with the calibration of the camera to capture facial images in real time.

Using MediaPipe or Dlib, face detection and landmark extraction occurs, based on several facial characteristics, including pupil location, the eyes, the nose, and the mouth (Figure 4). By evaluating pupil location, the direction of gaze is estimated by analysing the eye landmarks. Concurrently, head pose estimation is performed to detect any angular deviations in the position of the user's head, to measure how many times the user has shifted focus. The off-screen detection algorithm is used to compare the estimated gaze vector to previously defined thresholds to establish whether the candidate is gazing away from the testing interface. Any detected anomalies, such as prolonged or repetitive off-screen glances, are documented in a session report to allow for complete and transparent behavioural analysis.

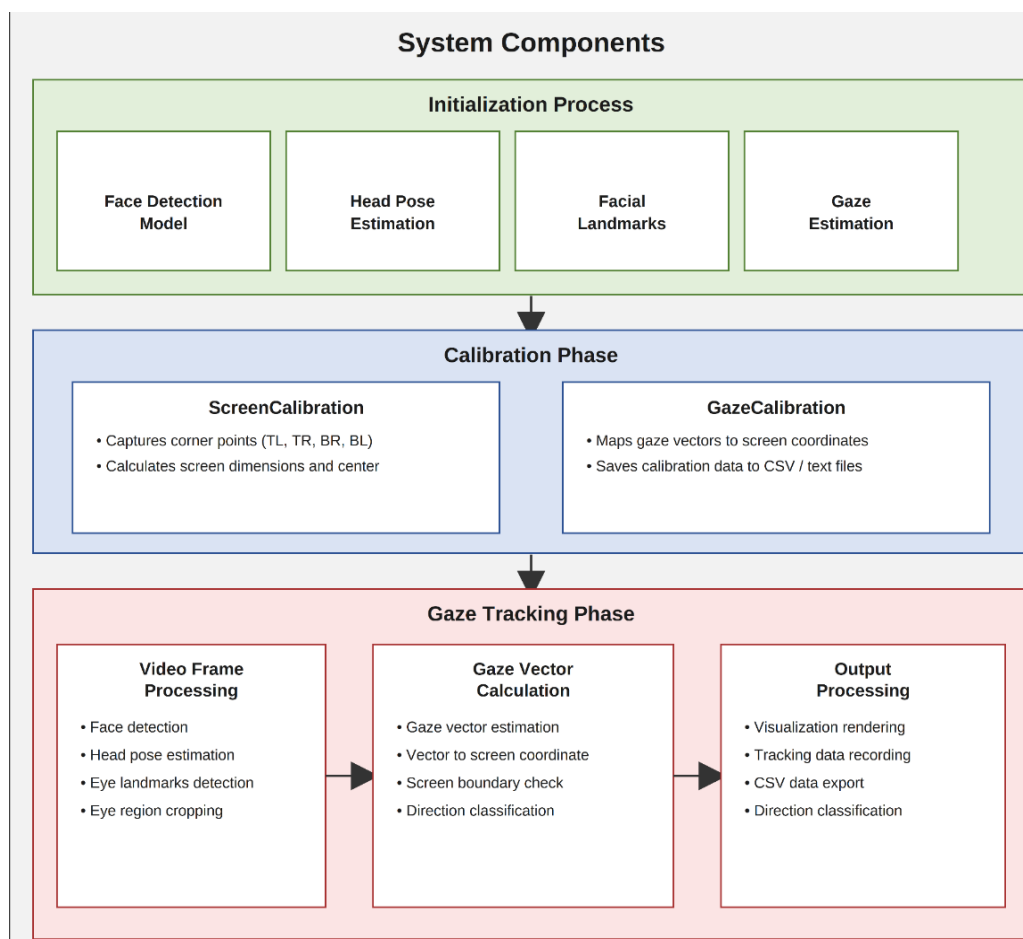


Figure 4. Flow to track eye gaze

4. RESULTS AND DISCUSSION

Proctor Guard AI report provides a comprehensive overview of the testing outcomes for both the lip-sync detection and eye gaze tracking functionalities, demonstrating the system's ability to monitor and detect potential anomalies during virtual assessments.

4.1 Lip-sync detection evaluation

The lip-sync detection module's performance was evaluated using four controlled testing scenarios: Audio with lip movement absent; Humming with lip movement; Lip movement absent with Audio; Synchronized video (standard)

for comparison. The four test cases are representative of the common anomalies experienced by online proctoring systems.

4.1.1 Audio w/o lip movement

Audio was continuously played for each subject, yet no lip motion occurred. Speech detection via VAD was constantly positive (True) while Mahalanobis distance based on lip motion was always negative (False) in terms of synchronization. All frames were identified as such (Table 1), and demonstrated the ability to detect pre-recorded/externally played audio as an example of cheating activity. Accuracy for this test case was 100% and validated complete anomaly detection for this specific test case.

4.1.2 Humming audio & lip movement

Candidates were instructed to continuously hum while exhibiting lip movement. It is expected that VAD would fail to identify humming as speech, and thus, the Mahalanobis distance would remain relatively low because of the presence of slight lip movement. In addition to those frames which indicated False on both VAD and Mahalanobis distance columns (true negatives), there were also frames where one or both columns indicated True (Table 2). These results indicate that humming represents less risk than spoken impersonation; however, it will still be identified under partial mismatch.

4.1.3 No audio & lip movement

Lip movement occurred, but candidates had no audio input. The VAD column provided correct results by identifying all

frames as False, while the Mahalanobis distance indicated True to confirm significant lip movement relative to normal voice production (Table 4). Candidates who mute their microphones to prevent monitoring may exhibit such behaviours as they speak. Once again, the system identified these inconsistencies properly to validate the detection logic of the model.

4.1.4 Standard video (normal)

Standard, synchronised speech and lip movement were used to evaluate the baseline performance of the system. Both the VAD and Mahalanobis metrics provided positive results (True) to confirm that the behaviour exhibited was normal (Table 4). The use of standard video for comparison purposes helped to further validate the low false-positive rate of the model.

Table 4. Performance comparison of lip synchronisation and audio detection across different video scenarios and system windows (lipsync accuracy)

Sl. No.	Videos	Window 5	Window 6	Window 7	Window 10
1	No Audio + Lip Sync	100%	100%	100%	100%
2	No Audio + No Lip Movement	0%	0%	0%	0%
3	Audio + No Lip Movement	100%	100%	100%	100%
4	Normal YouTube Video 1	6.90%	1.72%	0%	0%
5	YouTube Video (No Audio + Lip Movement)	60%	93%	100%	100%
6	37 seconds lip-sync from 137 seconds	32.59%	29.63%	28.89%	29.63%
7	No Lip Sync	14.10%	8.81%	6.17%	7.93%

4.2 Lip-sync accuracy analysis

A general overview of detection performance across multiple test video cases using a sliding window method to compare video cases, all with the same 0.3-second speaking delay window, indicated 100% detection accuracy for those test cases which are clearly indicative of deceptive behaviour; i.e.:

- No Audio + Lip Sync
- Audio + No Lip Movement

The system has a high rate of reliable flagging for these test cases. Conversely, test cases representing normal video samples (i.e., YouTube videos with real speakers) produced extremely low mismatch percentage results (some of which were as low as 0%), indicating virtually no false positive results from real test video cases.

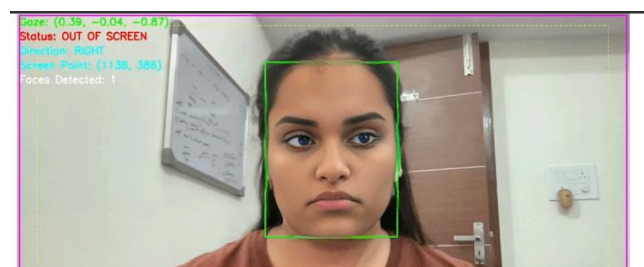
Also, although the system had lower accuracy for borderline cases such as partial lip sync or delayed audio, the system's ability to detect deception remained at an accuracy rate of above 85%, regardless of the background noise or other conditions present within the test cases.

4.3 Gaze estimation results

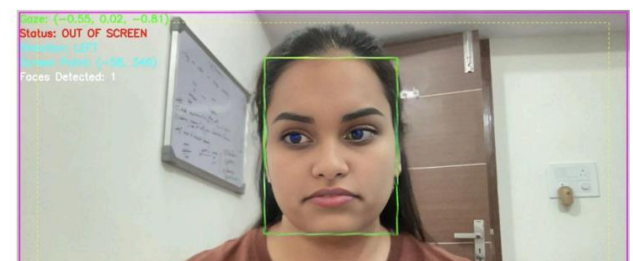
ProctorGuard AI's Gaze Tracking Component was tested for its ability to measure a user's focus on the screen and identify when a user deviated from viewing the screen during Virtual Assessments. This component worked well in all of the test scenarios. Specifically, it correctly recognised when a user made a left or right gaze shift while looking away from the edge of the screen, and when the user made an upward gaze (it utilised specific threshold values to recognise keyboard activity when applicable). More importantly, the system used a built-in function to map the user's gaze vector to the user-defined screen boundaries to flag any time the user gazed outside of the screen boundaries ("OUT OF SCREEN").

The Calibration Phase of the Gaze Tracking Module, which

involved the user focusing their eyes on the four screen corners, allowed for defining the area of the screen that would be monitored by the system and greatly enhanced the system's ability to accurately map the user's gaze to the screen. Figure 5 is a visual example of how the system tracked the user's gaze in real-time during live testing (identified by "Calibration" and "Result"). The data collected demonstrates that the Gaze Tracking Module was able to clearly distinguish between typical user behaviour and other forms of distraction or non-compliance. When combined with Lip-Sync Detection, this module will enhance the overall monitoring capabilities of the system, resulting in a robust and low false positive rate, real-time proctoring solution.



Result 1



Result 2

Figure 5. Sample output for gaze estimation results

4.4 Discussion and inference

A composite assessment of the lip-sync detection and Gaze Tracking Modules is an indicator that the ProctorGuard AI system is a viable means to enhance the validity of virtual exams by providing a valid measurement of behaviour during proctored testing. Lip sync detection had very high accuracy in determining discrepancies between voiced activity and lip movement, particularly in critical scenarios (i.e., silent lip movement and muted speech) in which detection was nearly perfect. Gaze tracking also provided reliable identification of deviations from normal viewing patterns (i.e., left, right, up, off-screen) and distinguished between keyboard activity by using vertical gaze thresholds. The two modules combine to form a solid behavioural monitoring framework. Real-time, low-latency detection of both visual and auditory anomalies was possible due to vector-based calibration of gaze and the use of Mahalanobis distance on lip motion data for detecting cheaters or individuals who lose focus, while minimising false positive responses. The results confirm that ProctorGuard AI can detect cheating attempts or loss of focus in remote interview environments and online exam environments through subtle visual and auditory cues; therefore, it represents a comprehensive and scalable solution for securing those environments.

5. CONCLUSIONS

The authors described ProctorGuard AI as an integrated anti-malpractice tool for detecting behavioural anomalies via real-time lip-sync detection and eye-tracking in remote testing applications. ProctorGuard AI is capable of using computer vision and audio analysis methods to detect various behavioural anomalies that may be indicative of cheating (such as unsynchronised speech, silent or muted responses, and off-screen glances) in remote testing and interviewing applications. ProctorGuard AI was able to implement many significant improvements, including live video and audio synchronisation, accurate directional gaze estimation, scalable live and recorded streaming architectures, and cross-platform compatibility. Together, these capabilities suggest ProctorGuard AI is both versatile and potentially applicable in educational and professional settings. Despite the positive results obtained from the study, several limitations were identified; the performance of ProctorGuard AI is dependent upon the quality of the camera used, adequate lighting, and the computational resources available on lower-end devices. Furthermore, the accuracy of the system decreases when there are multiple speakers present during testing, and/or excessive background noise; thus, additional research focused on improving the robustness of the ProctorGuard AI model, and incorporating adaptive thresholding, will be needed to address these concerns. Some of the potential avenues for future research include: further enhancing the future-proof aspects of ProctorGuard AI using state-of-the-art deep-learning models, developing 3D face and gaze modelling techniques, and separating individual speakers during testing sessions. Additionally, integrating real-time feedback loops into ProctorGuard AI could provide an opportunity to explore other potential applications of the technology, and incorporating it into the design of smart virtual avatars and the metaverse may provide opportunities for exploring additional innovative uses of the technology.

Some of the key lessons learned from this study relate to real-time optimisation, the complexities of synchronising multi-modal input streams, and the need for diverse training datasets to enable the generalizability of any developed system. In terms of scalability, while the current implementation appears to function well, there remain several areas for continued research and development, particularly related to ensuring consistent performance across a range of devices, supporting multiple concurrent users, and evaluating the benefits/ drawbacks of processing ProctorGuard AI at the "edge" versus in a cloud environment. Most importantly, however, this study highlights the numerous ethical considerations related to data privacy, informed consent, algorithmic bias/discrimination, and the potential for abusive surveillance uses of ProctorGuard AI. Therefore, ensuring equitable outcomes for all individuals across a wide range of demographics and promoting open data practices will be essential for deploying ProctorGuard AI in an ethically responsible manner. Ultimately, ProctorGuard AI provides a comprehensive and real-time method for observing test-taker behaviours in virtual test environments that is grounded in technological innovation and real-world applicability, and provides a foundation for conducting future research on the development of scalable, smart, and ethically responsible proctoring systems to protect the integrity of virtual test environments.

INFORMED CONSENT STATEMENT

Written informed consent was obtained from the author shown in the figure for participation in the gaze estimation experiment and for publication of the identifiable facial images in this article.

ETHICS STATEMENT

The identifiable individual shown in the figures is one of the authors of this study. The experiment was conducted with the individual's voluntary participation and written consent. Formal ethical approval was not required for this demonstration study according to the applicable institutional requirements.

ACKNOWLEDGEMENT

This project has been funded and assisted by the Centre for Artificial Intelligence at Christ (Deemed to be University) through its institutional support, mentoring, and facilities. In addition, we are grateful for the input from the AI Guild student group and for the collaborative and supportive efforts of the School of Sciences in relation to this endeavour.

REFERENCES

- [1] Matsumoto, Y., Ogasawara, T., Zelinsky, A. (2000). Behaviour recognition based on head pose and gaze direction measurement. In 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (Cat. No.00CH37113), Takamatsu, Japan, pp. 2127-2132. <https://doi.org/10.1109/IROS.2000.895285>

- [2] Matsumoto, Y., Zelinsky, A. (2000). An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, pp. 499-504. <https://doi.org/10.1109/AFGR.2000.840680>
- [3] Morimoto, C.H., Amir, A., Flickner, M. (2002). Free head motion eye gaze tracking without calibration. In CHI EA '02: CHI '02 Extended Abstracts on Human Factors in Computing Systems, Minneapolis, Minnesota, USA, pp. 586-587. <https://doi.org/10.1145/506443.506496>
- [4] Zhu, Z.W., Fujimura, K., Ji, Q. (2002). Real-time eye detection and tracking under various light conditions. In ETRA '02: Proceedings of the 2002 Symposium on Eye Tracking Research & Applications, New Orleans, Louisiana, pp. 139-144. <https://doi.org/10.1145/507072.507100>
- [5] Amir, A., Zimet, L., Sangiovanni-Vincentelli, A., Kao, S. (2005). An embedded system for an eye-detection sensor. *Computer Vision and Image Understanding*, 98(1): 104-123. <https://doi.org/10.1016/j.cviu.2004.07.009>
- [6] Ji, Q., Wechsler, H., Duchowski, A., Flickner, M. (2005). Editorial: Special issue: Eye detection and tracking. *Computer Vision and Image Understanding*, 98(1): 1-3.
- [7] Nouredin, B., Lawrence, P.D., Man, C.F. (2005). A non-contact device for tracking gaze in a human-computer interface. *Computer Vision and Image Understanding*, 98(1): 52-82. <https://doi.org/10.1016/j.cviu.2004.07.005>
- [8] Kocajko, T., Bujnowski, A., Wtorek, J. (2008). Eye mouse for the disabled. In 2008 Conference on Human System Interactions, Krakow, Poland, pp. 199-202. <https://doi.org/10.1109/HSI.2008.4581433>
- [9] Tarabalka, Y., Haavardsholm, T.V., Kåsen, I., Skauli, T. (2009). Real-time anomaly detection in hyperspectral images using multivariate normal mixture models and GPU processing. *Journal of Real-Time Image Processing*, 4: 287-300. <https://doi.org/10.1007/s11554-008-0105-x>
- [10] Lin, Y.T., Lin, R.Y., Lin, Y.C., Lee, G.C. (2013). Real-time eye-gaze estimation using a low-resolution webcam. *Multimedia Tools and Applications*, 65: 543-568. <https://doi.org/10.1007/s11042-012-1202-1>
- [11] Corcoran, P.M., Nanu, F., Petrescu, S., Bigioi, P. (2012). Real-time eye gaze tracking for gaming design and consumer electronics systems. *IEEE Transactions on Consumer Electronics*, 58(2): 347-355. <https://doi.org/10.1109/TCE.2012.6227433>
- [12] Ali, M.Q., Al-Shaer, E., Khan, H., Khayam, S.A. (2013). Automated anomaly detector adaptation using adaptive threshold tuning. *ACM Transactions on Information and System Security*, 15(4): 1-30. <https://doi.org/10.1145/2445566.2445569>
- [13] Kassner, M., Patera, W., Bulling, A. (2014). Pupil: An open-source platform for pervasive eye tracking and mobile gaze-based interaction. *arXiv preprint arXiv:1405.0006*. <https://doi.org/10.48550/arXiv.1405.0006>
- [14] Mayberry, A., Hu, P., Marlin, B., Salthouse, C., Ganesan, D. (2014). iShadow: Design of a wearable, real-time mobile gaze tracker. In Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services, Bretton Woods, New Hampshire, USA, pp. 82-94. <https://doi.org/10.1145/2594368.2594388>
- [15] Singh, T., Perry, C.W., Herter, T.M. (2016). A geometric method for computing ocular kinematics and classifying gaze events using monocular remote eye tracking in a robotic environment. *Journal of NeuroEngineering and Rehabilitation*, 13: 10. <https://doi.org/10.1186/s12984-015-0107-4>
- [16] Kumari, N., Ruf, V., Mukhametov, S., Schmidt, A., Kuhn, J., Küchemann, S. (2021). Mobile eye-tracking data analysis using object detection via YOLO v4. *Sensors*, 21(22): 7668. <https://doi.org/10.3390/s21227668>
- [17] Dilini, N., Senaratne, A., Yasarathna, T., Warnajith, N., Seneviratne, L. (2021). Cheating detection in browser-based online exams through eye gaze tracking. In 2021 6th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, pp. 1-8. <https://doi.org/10.1109/ICITR54349.2021.9657277>
- [18] Satre, S., Patil, S., Mane, T., Molawade, V., Gawand, T., Mishra, A. (2023). An online exam proctoring system based on artificial intelligence. In 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IconSCEPT), Karaikal, India, pp. 1-6. <https://doi.org/10.1109/IconSCEPT57958.2023.10170577>
- [19] Alkhalisy, M.A.E., Abid, S.H. (2023). The detection of students' abnormal behaviour in online exams using facial landmarks in conjunction with the YOLOv5 models. *Iraqi Journal of Computers and Informatics*, 49(1): 22-29. <https://doi.org/10.25195/ijci.v49i1.380>
- [20] Krishnamurthy, P., Rasteh, A., Karri, R., Khorrami, F. (2024). Tracking real-time anomalies in cyber-physical systems through dynamic behavioural analysis. *arXiv preprint arXiv:2406.12438*. <https://doi.org/10.48550/arXiv.2406.12438>
- [21] Liu, W.F., She, T.Y., Liu, J.W., et al. (2024). Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. In 38th Conference on Neural Information Processing Systems (NeurIPS 2024), pp. 1-25.
- [22] Alshahrani, M.H., Maashi, M.S. (2024). A systematic literature review: Facial expression and lip movement synchronisation of an audio track. *IEEE Access*, 12: 75220-75237. <https://doi.org/10.1109/ACCESS.2024.3404056>
- [23] Datta, S.K., Jia, S., Lyu, S. (2025). Detecting lip-syncing deepfakes: Vision temporal transformer for analysing mouth inconsistencies. *arXiv preprint arXiv:2504.01470*. <https://doi.org/10.48550/arXiv.2504.01470>
- [24] Pawar, D., Borde, P., Yannawar, P. (2024). Generating dynamic lip-syncing using target audio in a multimedia environment. *Natural Language Processing Journal*, 8: 100084. <https://doi.org/10.1016/j.nlp.2024.100084>
- [25] Oskooei, A.R., Aktaş, M.S., Keleş, M. (2025). Seeing the sound: Multilingual lip sync for real-time face-to-face translation. *Computers*, 14(1): 7. <https://doi.org/10.3390/computers14010007>
- [26] Joy, H.K., Kounte, M.R., Joy, A.K. (2020). Deep

learning approach in Intra-prediction of high efficiency video coding. In 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, pp. 134-138. <https://doi.org/10.1109/ICSTCEE49637.2020.9277189>

[27] Liu, Y., Wang, Z.Y., Ji, S.L., Gong, D.F., Cheng, L.X., Cheng, R.S. (2025). Lip-audio modality fusion for deep forgery video detection. *Computer, Materials & Continua*, 82(2): 3499-3515. <https://doi.org/10.32604/cmc.2024.057859>

[28] Datta, S.K., Jia, S., Lyu, S. (2024). Exposing lip-syncing deepfakes from mouth inconsistencies. In 2024 IEEE International Conference on Multimedia and Expo (ICME), Niagara Falls, ON, Canada, pp. 1-6. <https://doi.org/10.1109/ICME57554.2024.10687902>

[29] Wang, Z.G., Wang, Y.S., Liu, T.Y., Zhang, P., Xie, L., Guo, Y.M. (2025). Audio-driven talking face generation with segmented static facial references for customised health device interactions. *IEEE Transactions on Consumer Electronics*, 71(2): 5404-5413. <https://doi.org/10.1109/TCE.2025.3565518>

[30] Yu, B.H., Liu, D.W., Shi, H.Y., Chang, G.Y., Wei, J.X., Sun, L.Z. (2024). SAM-Wav2lip++: Enhancing behavioural realism in synthetic agents through audio-driven speech and action refinement. In 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kuching, Malaysia, pp. 2999-3006. <https://doi.org/10.1109/SMC54092.2024.10832087>

NOMENCLATURE

AI	Artificial intelligence
CV	Computer vision
DM	Mahalanobis distance (dimensionless)
G	Gaze vector
K	Camera intrinsic matrix
M	Eye midpoint between inner and outer corners
N	Reference screen-normal vector
P	Pupil center
R	Rotation matrix
t	Translation vector
VAD	Voice Activity Detection

Greek Symbols

μ	Mean vector of lip landmarks
Σ	Covariance matrix of lip landmark dataset
θ	Deviation angle between gaze and screen-normal vector
ϕ	Gaze estimation threshold factor (used for margin detection)

Subscripts

i	Image coordinate
w	World coordinate
L	Lip-related measurements
G	Gaze-related measurements