

Remote Assessment and Proctoring using Intelligent Devices - SENSE (RAPID-SENSE)

1st Wyvern Khiang Teng Wei
ICT, SIT, Singapore
2200577@sit.singaporetech.edu.sg

2nd Muhamed Fauzi Bin Abbas
ICT, SIT, Singapore
fauzi.abbas@singaporetech.edu.sg

3rd Goh Wei Han
ICT, SIT, Singapore
wei.han.goh@singaporetech.edu.sg

Abstract—RAPID-SENSE is an extended proctoring solution designed to preserve the integrity of online assessments, addressing the challenges that have arisen with the rapid shift to remote learning, particularly during the COVID-19 pandemic. RAPID-SENSE enhances traditional digital monitoring to detect suspicious behaviour during examinations by integrating audio-visual sensors. This system leverages AI algorithms to analyze camera and microphone data in real-time, identifying activities such as unauthorized communication or unusual movements, thereby ensuring comprehensive oversight. Built on a lightweight, IoT-based architecture, RAPID-SENSE is scalable for large-scale deployments and focuses on maintaining resource efficiency and robust proctoring capabilities. This paper details the system's technical architecture, including visual gaze tracking, facial recognition, and acoustic monitoring. It demonstrates its efficacy in providing secure and reliable online assessment proctoring on resource-constraint embedded systems.

Index Terms—AIoT, remote proctoring

I. INTRODUCTION

The COVID-19 pandemic accelerated the need for effective proctoring technologies to maintain academic integrity in remote assessments. Solutions like live monitoring, video recording, and automated systems address this demand but face significant challenges, especially in large-scale or real-time scenarios.

Live monitoring relies on human invigilators observing examinees via webcams, offering a familiar oversight experience but lacking scalability due to high resource demands and the risk of inconsistent decision-making. Cloud-based proctoring, while more scalable, imposes heavy bandwidth requirements, creating inequities for examinees with limited internet access and driving up costs for large-scale implementation. These limitations highlight the need for innovative, efficient, and scalable proctoring solutions.

To address the growing challenges of preserving assessment integrity in online assessments, which were intensified by the COVID-19 pandemic, our previous work introduced RAPID, an IoT-based proctoring solution [13]. RAPID was designed to monitor digital anomalies and detect cheating during remote assessments without requiring software installation on the examinee's device. While RAPID successfully showcased the feasibility of leveraging IoT and cybersecurity techniques for proctoring, this paper extends that work by introducing

RAPID-SENSE, which integrates audio-visual sensors to capture real-world activities, enhancing proctoring accuracy. This extended approach addresses previous data collection and analysis constraints and adapts to diverse operating environments.

RAPID-SENSE utilizes AI-driven algorithms to process camera and microphone feeds, detecting suspicious behaviour such as unauthorized communication, unusual movements, or external individuals in the testing environment. Furthermore, it tracks activity on connected peripherals like keyboards and unauthorised processes, providing insight into potential cheating attempts. Unusual data transfer patterns or unauthorised access to external resources can be quickly flagged for further investigation.

This paper details the technical architecture of RAPID-SENSE, illustrating its core functionalities and demonstrating its capacity to enhance online assessment security. By building on the original RAPID framework, RAPID-SENSE offers a robust solution to the challenges of large-scale online proctoring, paving the way for more secure and reliable remote examinations.

II. LITERATURE REVIEW

This literature review examines existing online proctoring methods—live monitoring, video recording, and automated systems—alongside algorithms suited for resource-constrained embedded systems, focusing on their application in image and sound analytics for proctoring. It establishes the foundation for developing AIoT-based solutions to enhance online assessment security.

Video recording offers an alternative to live monitoring, addressing scalability issues by enabling post-exam review. However, reviewing footage is time-intensive and may miss subtle cheating behaviours [15]. Privacy concerns, especially under GDPR regulations, further complicate its use [16]. Gaze detection aids in identifying suspicious behaviour through eye and head movement analysis [1]. Infrared cameras provide high accuracy but are costly [2], while webcams offer a cheaper alternative by estimating gaze direction from iris position [3], though they may affect examinee comfort [4].

Face verification confirms identity, with tools like the DeepFace library integrating models such as FaceNet, OpenFace, and DeepID, offering a unified framework for facial recognition tasks [10]–[12]. While FaceNet and OpenFace achieve high accuracy, their computational demands make

them unsuitable for constrained environments, unlike DeepID that balances accuracy and efficiency [9].

Acoustic monitoring complements visual methods by detecting unauthorized communication. Using techniques like FFT, speaker detection distinguishes vocal frequencies efficiently, though accuracy declines with overlapping frequency ranges [7]. Advanced methods like MFCC improve speaker mapping and identification [5] but are computationally demanding and less effective in real-time scenarios without prior speaker data. FFT remains a practical choice for embedded systems requiring real-time processing [8].

III. SYSTEM DESIGN AND ARCHITECTURE

This section outlines the architecture of RAPID-SENSE, an enhancement of the RAPID proctoring system. RAPID-SENSE enables real-time detection of unauthorized communication and gaze tracking during online assessments by integrating audio-visual sensors and intelligently scheduling machine learning models. Designed for lightweight IoT devices, it ensures resource efficiency and tamper resistance. The architecture's key components, subsystem interactions, and data flow are detailed below.

A. Overview of System Architecture

Figure 1 illustrates how the proposed solution works by monitoring the examinee's PC via RAPID [13] to monitor all digital activities through scripts obtained from the Invigilator's portal, which are executed directly in the PC's memory. These scripts collect relevant data on user interactions, including application usage and potential anomalies, which are then transmitted to RAPID for real-time analysis. The analyzed reports and supporting data are sent back to a central database via the PC, where the Invigilator can review the results and take appropriate actions when necessary. This process ensures comprehensive oversight of the examinee's digital environment during the examination.

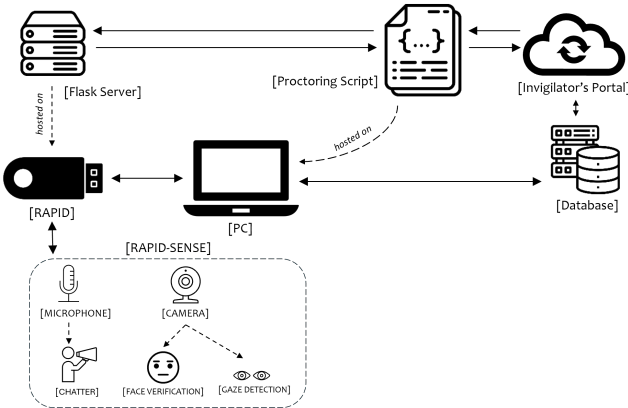


Fig. 1. Architecture of RAPID-SENSE within RAPID

While the original RAPID system primarily monitors digital activities during proctoring, RAPID-SENSE enhances these capabilities by integrating real-world monitoring

through acoustic and visual data collection via a unified device. This advanced system introduces three critical features: gaze tracking, facial recognition, and conversation detection, providing a more holistic view of the examinee's environment. By addressing both the digital and physical aspects, RAPID-SENSE ensures a more thorough approach to maintaining academic integrity during remote assessments.

1) *Visual Integrity: Gaze Monitoring:* To monitor gaze, a camera mounted atop the screen captures the examinee's face for real-time tracking. The system employs the MediaPipe library for face detection, generating a facial mesh that highlights key features. Mesh points near the eyes track gaze, while those on the nose, chin, and mouth monitor head rotation.

Gaze detection relies on the pinhole camera model, which maps the 3D world onto a 2D image. Using the camera feed, the model extracts a 2D facial mesh and maps 3D coordinates onto it, with the nose serving as the origin in 3D space. This mapping process allows us to estimate the head pose later in the calculations [14].

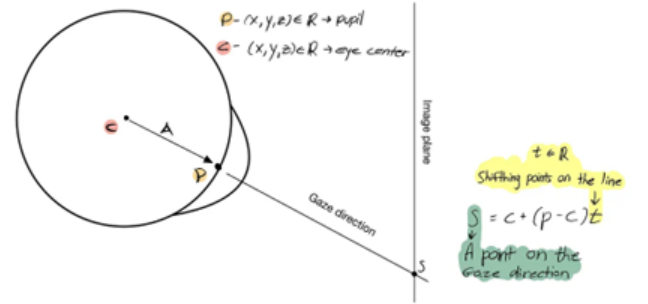


Fig. 2. Calculation of gaze direction.

The OpenCV library's solvePnP function computes rotation and translation vectors to map a 3D face model onto a 2D plane, enabling head pose estimation. MediaPipe provides 2D pupil coordinates, which OpenCV's estimateAffine3D maps into 3D space. The gaze direction vector is determined by subtracting the 3D coordinates of the eyeball center from those of the pupils, scaling the vector, and combining it with the eyeball center. Finally, the gaze direction is projected back onto the 2D plane using projectPoints, incorporating head rotation data derived from the head pose estimation.

Gaze Detection via angle: The system determines the examinee's screen size using the gaze direction vector. During calibration, the examinee first focuses on the screen's center to record the "correct" gaze. They then look at the four corners, and the system calculates angles between each corner gaze vector and the correct gaze using dot products. These angles help estimate the screen size and set the maximum permissible deviation for detecting off-screen gazes, triggering a colour-coded flag (red, yellow, or green) based on the degree of deviation. Inconsistent corner angles prompt recalibration.

Gaze Detection via coordinates: Alternatively, instead of relying on angles, projected coordinates onto a 2D plane

can detect whether the examinee is looking away from the screen. During calibration, blue circles will appear at the four corners of the screen, creating a rectangle on the 2D plane that represents the screen from the camera's perspective. The system will flag it as a reminder whenever the examinee's gaze coordinates fall outside this rectangle. A colour-coded system (red, yellow, or green) will indicate the distance between the gaze coordinates and the rectangle, reminding the examinee to focus on the screen.

2) *Visual Integrity: Face Verification:* The DeepFace library is a lightweight facial analysis framework that contains several face recognition models. The implementation of FaceNet, OpenFace, and DeepID was prioritized due to their suitability for resource-constrained environments and their balance of accuracy and efficiency. A fourth model, dlib, was omitted as the chosen platform did not support it [6].

Obtaining Images: Captured images are transmitted to the RAPID device for real-time monitoring. A dedicated script stores the images in shared memory within the device's RAM to support concurrent access by machine learning models like gaze tracking and face recognition. This enables multiple models to process the camera data simultaneously without resource conflicts, ensuring efficient and smooth operation.

3) *Acoustic Oversight: Chatter:* RAPID-SENSE uses FFT to analyze microphone data, converting audio signals into the frequency domain. By focusing on the 75–300 Hz range of human speech, it detects significant frequency shifts (over 100 Hz) as potential multi-speaker activity, indicating unauthorized collaboration. Upon detection, the Vosk library transcribes speech for invigilator review, optimizing processing by activating speech-to-text conversion only during significant speech activity. The Vosk library is chosen for its offline functionality and reliable transcription accuracy.

4) *Intelligent Task Management:* Microcontroller cores are dedicated to specific tasks: visual processing (camera reading, image capturing, gaze detection), acoustic processing, communication protocol management, and load balancing for efficient resource distribution.

Image handling and verification processes have varying execution times, with camera reading taking 80 ms, image storage 200 ms, gaze detection 160 ms, and face verification 2500 ms. A round-robin scheduling approach ensures that camera reading, image capturing, and gaze detection complete their cycles in each iteration, while the less frequent, computationally intensive face verification process runs periodically. Sleep timers in each process loop further optimize resource allocation by enabling smooth task execution. The calculated sleep intervals for each process are summarized in Table I.

This scheduling framework optimises the computational resources on the monitoring device, ensuring that the system can operate efficiently even in resource-constrained environments. By limiting the activation of high-demand models to situations

TABLE I
CALCULATED PROCESS SLEEP TIMES

Process	Time (milliseconds)
Obtain Image from Camera (R)	$100+100+100+100+60=460$
Store Image in microSD (C)	$60+80=140$
Gaze Detection (G)	$80+100=180$
Face Verification (F)	$100+60+80+100+100=440$

where they are genuinely required, RAPID-SENSE ensures the scalability of the proctoring system without necessitating high-performance hardware. This approach allows for effective large-scale deployment across multiple devices while maintaining the integrity and accuracy of the monitoring process.

IV. RESULTS AND ANALYSIS

This section analyses the various components of RAPID-SENSE to detect potential academic misconduct during online assessments. The results presented here provide insights into the efficacy of these analytics in enhancing the security and integrity of online assessments.

A. Hardware Setup

The RAPID system, built on the Raspberry Pi Zero 2 W for its processing power and affordability, integrates a Pi Camera v2.1 via the CSI for high-speed video capture and a ReSpeaker 2-Mics Pi Hat via the HAT interface for audio processing with noise filtering and sound localization. This configuration frees the USB port for PC communication and uses off-the-shelf components for easy deployment. The camera achieves 16 fps, enabling accurate gaze direction estimation. Performance tests in diverse scenarios demonstrated the system's real-world effectiveness.

1) *Visual Integrity: Gaze Detection:* The system's gaze detection was tested under various scenarios, including when the examinee stays within screen boundaries, looks just outside or entirely away, multiple faces are present, or the examinee leaves the frame. Outcomes are flagged using a colour-coded system (green, yellow, red) based on predefined thresholds, with additional detections for multiple or no faces. While the system performed as expected in most cases, it failed in scenarios where the examinee moved back after calibration, as significant gaze angle changes led to incorrect outcomes.

The gaze detection system performs as expected in most scenarios, accurately identifying on-screen focus, off-screen gazes, multiple faces, and absences from the camera frame. However, it fails in cases where the examinee moves back from the camera after calibration, resulting in incorrect outcomes due to significant changes in gaze angle.

2) *Visual Integrity: Face Verification:* Table II compares the performance of three face verification models. All models accurately identified whether the same person was involved, but processing times varied significantly. Facenet was the slowest (193.21s for correct and 154.68s for incorrect verification), OpenFace was faster (67.41s and 62.78s), and DeepID was the fastest (21.05s and 38.63s). While all models are effective, their efficiency differs greatly.

TABLE II
PERFORMANCE OF FACE VERIFICATION MODELS

Model	Verified	Same Person	Avg Time (sec)
Facenet	True	Yes	193.21
Facenet	False	No	154.68
OpenFace	True	Yes	67.41
OpenFace	False	No	62.78
DeepID	True	Yes	21.05
DeepID	False	No	38.63

Consequently, the DeepID model was selected for facial verification, providing an optimal balance between speed and accuracy to meet the system's performance requirements.

TABLE III
SPEAKER DETECTION PERFORMANCE IN VARIOUS SCENARIOS

Case	Description	Flagged	Recorded
1	One person speaking during proctoring process	No	No
2	Multiple people speaking during proctoring process	Yes	Yes
3	Multiple people whispering when pi is operating	No	No
4	One person speaks and another person speaks from distance	Yes	Yes
5	One person speaks and another person whispers from distance	No	No

3) *Acoustic Monitoring: Chatter*: To assess the effectiveness of the speaker detection, different scenarios were designed for the algorithm to proctor, as illustrated in Table III.

Compared to its strong performance in processing regular speech, the script's accuracy may decrease when dealing with the complexities of multiple people whispering, such as Case 3 and 5. However, it consistently delivers reliable results in scenarios with clearly audible speech

The Vosk library processes speech-to-text conversion with varying accuracy and efficiency. For example, a short sentence is processed in 41 milliseconds, while a longer sentence takes 132 milliseconds, approximately three times the duration of the original speech. Although some words are inaccurately transcribed or omitted, the system is sufficient for post-analysis to determine if more than two people are engaged in a discussion.

V. CONCLUSION

This paper introduced RAPID-SENSE, an extension of the RAPID project, integrating audio-visual sensors with AIoT to enhance online proctoring security. By combining data from connected devices and employing intelligent analytics, RAPID-SENSE effectively detects suspicious behaviours, environmental anomalies, gaze direction, unauthorized speech, and verifies identity, addressing key aspects of academic misconduct.

The findings highlight the feasibility of deploying multiple machine learning models on resource-constrained devices for secure online assessments. However, improvements in gaze detection under movement and varying lighting, as well as

enhanced audio monitoring for low-frequency sounds and speaker mapping, are needed to further optimize the system.

RAPID-SENSE demonstrates significant potential as a reliable proctoring tool, but continued testing and refinement are necessary to ensure robust performance in diverse conditions. As demand for secure online assessments grows, IoT-based solutions like RAPID-SENSE offer a promising approach to upholding academic integrity in the digital era.

REFERENCES

- [1] Burton L, Albert W, Flynn M. A Comparison of the Performance of Webcam vs. Infrared Eye Tracking Technology. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2014;58(1):1437–41.
- [2] Ebisawa Y, Fukumoto K. Head-Free, Remote Eye-Gaze Detection System Based on Pupil-Corneal Reflection Method With Easy Calibration Using Two Stereo-Calibrated Video Cameras. IEEE Trans Biomed Eng. 2013;60(10):2952–60.
- [3] Cuong NH, Hoang HT. Eye-gaze detection with a single WebCAM based on geometry features extraction. In: 2010 11th International Conference on Control Automation Robotics & Vision. 2010;2507–12.
- [4] Atoum Y, Chen L, Liu AX, Hsu SDH, Liu X. Automated Online Exam Proctoring. IEEE Trans Multimedia. 2017;19(7):1609–24.
- [5] Nwe TL, Foo SW, De Silva LC. Detection of stress and emotion in speech using traditional and FFT based log energy features. In: Fourth International Conference on Information, Communications and Signal Processing, and the Fourth Pacific Rim Conference on Multimedia. 2003;1619–23.
- [6] Venkatesan R, Shirly S, Selvarathi M, Jebaseeli TJ. Human Emotion Detection Using DeepFace and Artificial Intelligence. Eng Proc. 2023;59:37.
- [7] Islam MR, Azam MS, Ahmed S. Speaker identification system using PCA eigenface. In: Proceedings of the 12th International Conference on Computers and Information Technology. 2009;261–6.
- [8] Wang Y, Nishizaki H. A Lightweight End-to-End Speech Recognition System on Embedded Devices. IEICE Transactions on Information and Systems. 2023;106(7):1230–1239.
- [9] Sun Y, Liang D, Wang X, Tang X. DeepID3: Face Recognition with Very Deep Neural Networks. arXiv preprint arXiv:1502.00873. 2015 Feb 4.
- [10] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE conference on computer vision and pattern recognition 1 2014 Jun 23 (pp. 1891-1898)
- [11] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2015 Jun 7 (pp. 815-823).
- [12] Amos B, Ludwiczuk B, Satyanarayanan M. Openface: A general-purpose face recognition library with mobile applications. In: Proceedings of the IEEE international conference on computer vision workshops 2016 (pp. 1-8).
- [13] Ho, Jubilian Hong Yi, et al. "IoT-Enhanced Remote Proctoring: A New Paradigm for Remote Assessment Integrity." 2023 IEEE 35th International Conference on Software Engineering Education and Training (CSEE&T), IEEE. 2023.
- [14] Hartley, R., Zisserman, A. Multiple View Geometry in Computer Vision (2nd ed.). Cambridge University Press, 2003
- [15] Respondus. (2019). LockDown Browser and Respondus Monitor Quick Start Guide (Instructor) [User guide]. Retrieved from <https://web.respondus.com/wp-content/uploads/2019/08/RLDB-Quick-Start-Guide-Brightspace-Instructor.pdf>
- [16] Smuha, N. The impact of the GDPR on video surveillance: Towards a risk-based approach. *Computer Law & Security Review*, *36*(3), 105582. 2020.