Detection of Malpractice in E-exams by Head Pose and Gaze Estimation

https://doi.org/10.3991/ijet.v16i08.15995

Chirag S Indi, KCS Varun Pritham,
Vasundhara Acharya, Krishna Prakasha (✉)
Manipal Institute of Technology (MIT)
Manipal Academy of Higher Education (MAHE), Manipal, India
kkp.prakash@manipal.edu

Abstract—Examination malpractice is deliberate wrongdoing contrary to official examination rules designed to place a candidate at an unfair advantage or disadvantage. The proposed system depicts a new use of technology to identify malpractice in e-exams, which is essential due to online education growth. The current solutions for such a problem either require complete manual labor or have various vulnerabilities exploited by an examinee. The proposed application encompasses an end-to-end system that assists an examiner/evaluator in deciding whether a student passes an online exam without any probable attempts of malpractice or cheating in e-exams with the help of visual aids. The system works by categorizing the student's VFOA (visual focus of attention) data by capturing the head pose estimates and eye gaze estimates using state-of-the-art machine learning (ML) techniques. The system only requires the student (test-taker) to have a functioning internet connection and a webcam to transmit the feed. The examiner is alerted when the student wavers in his VFOA from the screen greater than X, a predefined threshold of times. If this threshold X is crossed, the application will save the person's data when his VFOA is off the screen and send it to the examiner to be manually checked and marked whether the student's action was attempted malpractice or just a momentary lapse in concentration. The system uses a hybrid classifier approach where two different classifiers are used. One when gaze values are being read successfully. On failing this due to various reasons like transmission quality or glare from his spectacles, the model falls back to the default classifier, which only reads the head pose values to classify the attention metric. It is later used to map the student's VFOA to check the likelihood of malpractice. The model has achieved an accuracy of 96.04 percent in classifying the attention metric.

Keywords—Online proctoring system, Visual focus of attention, Head post estimation, Gaze estimation, Machine learning, Malpractice detection, Hybrid classifier, Automated proctoring model

1 Introduction

Online education has come into the picture and revolutionized the education market, especially after introducing platforms like Coursera, Edx, Udacity, where institutions like MIT, Stanford University provide courses with world-class content accessible by anyone. The effect of COVID-19 on education has caused many schools and universities to switch their medium of instruction from in-person lectures to the online mode to adhere to public safety regulations. Due to the pandemic, the number of courses available online and the number of users accessing this content has exponentially grown, which is depicted in Figure 1

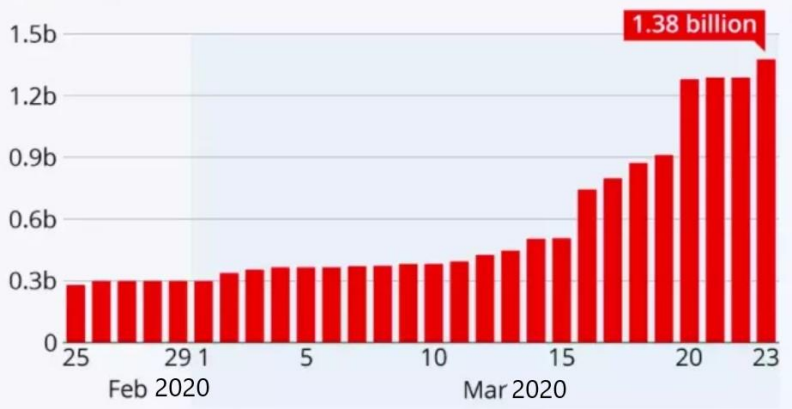

Fig. 1. Number of learners impacted by national school closures worldwide due to COVID [4]

This above-stated number in Figure 1 has been growing since March. It has impacted other similar fields like pre-employment assessments and corporate training certifications, allowing students to take their exams from home instead of a test center with in-person proctoring. Assessments that are carried out usually have limited supervision, making it extremely difficult to regulate and control cheating [5]. Due to the pandemic, Educational Testing Service (ETS), the non-profit educational organization which offers standardized tests including GRE and GMAT, has announced that these tests would continue to stay available to students with the option to take it from home. It is planning to continue this even after the current global scenario has changed [3].

In the current online education market, the current market leaders like Coursera, edX, and Canvas rely on the Code of Honor pledged by the test-taker to maintain integrity. Other websites like HackerRank that involve e-exams try to reduce malpractice by forcing the student's browser to full-screen mode and isolating his/her access to other applications on the system. This cheat prevention system can be easily bypassed by using a secondary device, as many online exams have no restrictions regarding the physical location where the student takes the test. The problem of ghostwriting, where a third party would take the test on the student's behalf, is also an increasing problem

in this industry. These cheating forms have reduced the intrinsic value of these certifications offered by prestigious institutions across the globe.

Several research works have been done in this field to detect malpractice. The existing methods focus on capturing faces from surveillance videos and detecting suspicious activities like peeping and object exchange. The advanced models are capable of ensuring the focus level of candidates. It checks for suspicious activities in video and background voice activity. Candidates are authenticated by utilizing a face recognition algorithm to prevent any impersonation. Some of the models can detect eye movement as even the most subtle movement of eyes suggests malpractice. These systems offer many advantages. They eliminate the schedule and location constraints and are scalable.

The proposed model introduces an end-to-end system that assists an examiner/evaluator in deciding whether a student passes an online exam without any probable attempts of malpractice or cheating in e-exams with the help of visual aids. It operates by categorizing the student's VFOA (visual focus of attention) data by capturing the head pose estimates and eye gaze estimates using state-of-the-art machine learning techniques. The main advantage of this system is the minimal requirement of resources and hence is cost-effective. It expects the student (test-taker) to have a functioning internet connection and a webcam to transmit the feed. The examiner is alerted when the student wavers in his VFOA from the screen greater than X, a predefined threshold of times. If this threshold X is crossed, the application will save the person's data when his VFOA is off the screen and send it to the examiner to be manually checked and marked whether the student's action was attempted malpractice or just a momentary lapse in concentration. Hence, it can help reduce human oversight in online proctoring and increase efficiency. With more and more exams being proctored with Artificial intelligence (AI) and machine learning, AI and ML systems will continue to learn in the near future. They will be able to judge the seriousness of their findings.

2 Literature Survey

Methods for malpractice detection have been proposed in various forms like source-code plagiarism detection [6], or to detect a common exploited strategy called CAMEO (Copying Answers using Multiple Existences Online), and various methods are being researched to combat such practices [7–9]. There have been methods that suggest and implement a part of our pipeline that include facial recognition to detect ghostwriting [10]. Like our proposed system, intelligent applications and algorithms were proposed in [11, 12] that worked well but were not flexible. In [11], the authors developed an intelligent inference system that took both audio and video as input. The dataset used in [11] contained three individuals with 13 recordings, which simulated 16 malpractice attempts. Features were then extracted and fed into an inference system, which was later used to detect malicious activities. It was built to detect changes in the yaw angle, along with audio and active window capture.

The system in [12] also followed a very similar approach mentioned in [11]. The dataset created had 12 different videos of 10 minutes with 10 malpractices each. The

model in [12] worked based on a rule-based heuristic system. The model calculated the yaw angle using cylindrical and ellipsoidal face models. It had the added advantage of detecting basic hand gestures compared to the system in [11] and measured system usage to detect anyone tampered with the system.

3 Problem Statement

The system proposed in [11] uses yaw angle variations, audio presence, and active window capture to classify malpractice. Such a system can fail to produce accurate and consistent results, as the only contributing factor need not be just the yaw angles. The examinee can cheat by looking at a different plane of view without varying his head pose estimates. The use of gaze estimates makes the proposed system more robust and foolproof to such methods of malpractice. The use of audio presence in [11, 12] and detecting malpractice by the use mean value of ambiance can be skewed due to network disruptions or minor changes or shifts in microphone placement, which will lead to an inaccurate result.

The proposed system is robust and does not use time-varying bursty factors like audio that could be influenced by network or connectivity issues. It uses state-of-art machine learning and computer vision algorithms to extract data such as head pose estimates or gaze estimates and is dynamic and robust.

4 Materials and Methods

4.1 Data collection

The authors created the dataset used to train this application, as there is no publicly available dataset of students being recorded from the screen's perspective. Ten videos were collected from volunteers, each averaging 4 to 5 minutes, each second being sampled ten times, thus about 300,000 frames before pre-processing. Each of these frames was passed through Haar feature-based cascade classifiers [13]. All the frames which had recognizable faces and eyes were kept, whereas the rest were discarded. These frames were individually and manually classified as 1, if the student's VFOA is on the screen, else 0, if the student is looking away. Each image is resized to 1024x768 (WidthxHeight) before being fed into the pre-trained FSA-Net [15] model to get the estimated head pose angles, and followed by finding the gaze coordinates if eye detection is successful. Figure 2 depicts the head pose angles.

Fig. 2. Head pose angles depicted by the blue (yaw), red (pitch), and green (roll) lines.

4.2 Proposed model

The proposed application extracts and uses individual frames from a live video stream. Each extracted frame is fed into a classifier, which classifies whether a face is detected or not. This classifier works on the Haar Cascade algorithm is a machine learning object detection algorithm used to identify objects in an image or video and based on the concept of features proposed by Paul Viola and Michael Jones [13]. This is necessary as, during the head's natural movements, the extracted frames might be blurry beyond detection, which may cause erroneous results when fed into the model. The frames in which no face is detected are discarded. The valid ones are now passed to the pre-trained FSA-Net model, which outputs the hose pose angles, yaw, pitch, and roll. The frame is then passed to another Haar Cascade classifier, which is used to detect whether the eyes are visible and open. If it is, the gaze estimates are extracted, and the calculation of a few other derived features from these estimates. All these features, along with the head pose angles, are fed as input features to Classifier 1, which works on the XGBoost Algorithm [14], a decision tree-based ensemble machine learning method that used gradient boosting. If the eyes are not detected, the frame is passed to Classifier 2, which also works on the XGBoost algorithm but only taking in the head pose estimates as the features. This dual classifier approach helps counteract the insufficient capture resolution or low transmission quality of the live video stream. The flow of the model is depicted in Figure 3.

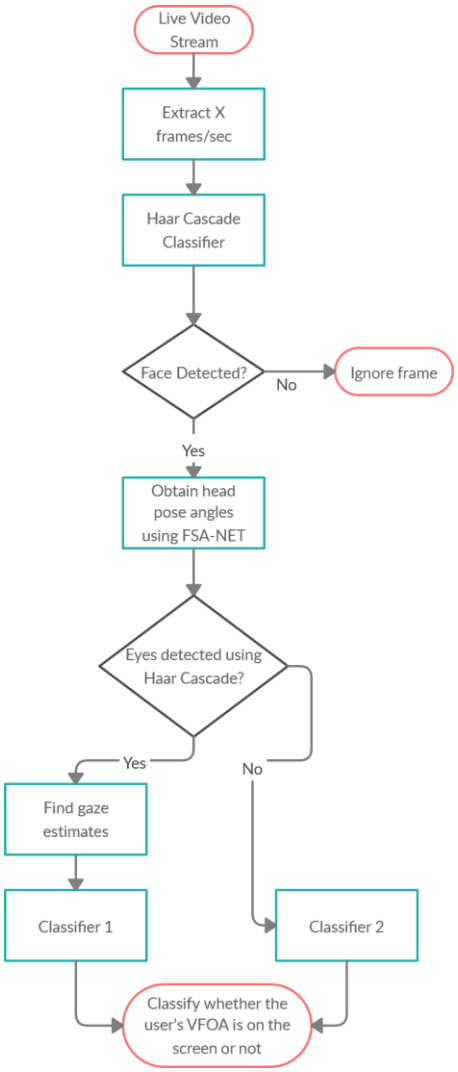

Fig. 3. The flow diagram represents the flow of the proposed model

4.3 Head pose detection

As discussed earlier, the proposed system uses head pose detection, which outputs a 3-D vector containing yaw, pitch, and roll angles. The proposed method uses the current state of the art FSA-Net [15], a deep learning algorithm that can generate the angles from a single image, based on regression and feature aggregation, to extract head pose estimates.

Given a set of training images $\{x_n | \text{ n } = 1,..,\text{N}\}$ and the head pose 3-D vector y_n. It tries to find a function F that can map $\tilde{y} = F(x)$ by minimizing the mean absolute error (MAE).

$$J(X) = \frac{1}{N}\sum_{n=1}^{N} \|\tilde{y}_n - y_n\|_1 \quad (1)$$

Where $\tilde{y}_n$ is the pose angles predicted for the image x_n.

FSA-net uses the architecture depicted in Figure 4. The neural network uses the SSR-Net (Soft Stagewise Regression) architecture [16], which works based on a hierarchical classification approach. The network at each stage performs an intermediate classification by using the class probability distribution and uses the stage-wise regression, shown in Equation (2) to predict the vector $\tilde{y}$.

$$\tilde{y} = \sum_{k=1}^{K} \vec{p}^{(k)}.\ \vec{\mu}^{(k)} \quad (2)$$

K = number of stages

$\vec{p}^{(k)}$ = probability distribution of the angles at the k^{th} stage

$\vec{\mu}^{(k)}$ = is the vector representing the age groups at the k^{th} stage

$\vec{\eta}^{(k)}$ = shift vector to adjust the center of the distribution

Δ_k = the width of the probability distribution

Before performing the above operation, FSA-Net performs feature generation by passing the image through two streams comprising convolution layers. It combines the feature maps obtained from each stage by element-wise multiplication, followed by 1x1 convolution and average pooling. After receiving K feature maps, it performs aggregation without losing the spatial information within the feature map.

To achieve the spatial grouping, it first generates an attention map A_k using a scoring function, which can be seen in Figure 4.

An ensemble of three models using three different scoring functions is used to make the results more robust. They are:

1. 1 × 1 convolution layer as a learnable scoring function, which learns from the training data to weigh features.
2. Variance, which allows the selection of features based on variance.
3. Uniform, which treats all features equally.

The three options are said to provide complementary information by exploring learnable, non-learnable, and constant alternatives.

After the attention maps are created, they are passed through a mapping module where $n'\ c - d$ representative features are generated. Later these features are passed through a capsule network for feature aggregation where the final set of features are obtained to generate representative features for regression, *V*, containing $K\ c' - d$ features. The vector V_k is used to generate the stage outputs $\{\vec{p}^{(k)}, \vec{\eta}^{(k)},\ \Delta_k\}$ for the k^{th} stage through a fully connected layer. These outputs are then substituted into the SSR function for obtaining the pose estimation.

All these special considerations and tuned architecture help the model produce excellent results when compared to its predecessors, even in the case of Occlusion, Extreme lighting conditions, Face rotation and Extreme head pose angle.

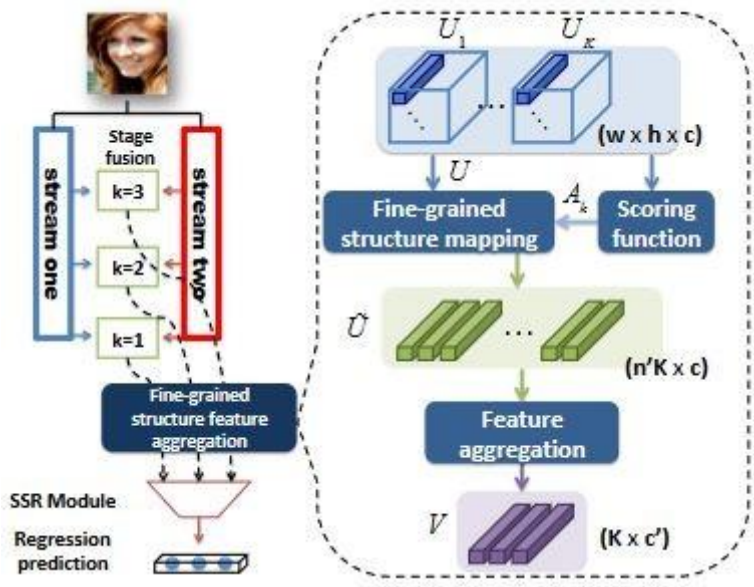

Fig. 4. FSA-Net Architecture

4.4 Gaze estimation

With the help of Dlib [17], a landmark's facial detector with pre-trained models, the proposed system follows a different approach to formalize a numerical value for the gaze. Using shape_predictor_68_face_landmarks.dat, which estimates the location of 68 coordinates (x, y) that map the facial points on a person's face. The model makes use of the coordinates labeled 37 to 46 (shown in Figure 5).

Thresholds values: Hard coding the threshold value would give inaccurate results; thus, an automatic calibration algorithm is used to find the right threshold value for the user/webcam. According to some approximations and statistics, the iris' size is around 48% of the eye's surface when a person's line of sight is horizontal and directed towards the camera. Thresholds values to binarize images can differ significantly from person to person, but iris sizes are very stable. The threshold's automatic calibration is found using the first 20 frames, provided as input to the algorithm. The frames are binarized with different thresholds values with the multiples of 5, from 5 to 100, after which the iris size is calculated for each frame. For each frame, the value that gives the closest iris size to 48% is saved. The final threshold value is the average of the best 20 values. The steps to achieve Pupil Pose is given in Algorithm 1.

Algorithm 1. Pupil Pose Estimation Algorithm

```
Procedure: Pupil_Pose_Estimation
Input: Input Image
Output: Pupil Position
1. Blur the image to a slight degree to remove any
noise using the    BilateralFilter method in OpenCV [19]
library
2. Erode the resulting image to remove backlights using
the erode method in OpenCV [19] library.
3. Convert the image to binary to have only black and
white pixels (no grayscale). A threshold value is to be
passed to the algorithm to separate white and black pix-
els. This value varies not only from person to person but
also with the image quality from the webcam. The accuracy
of gaze estimation depends highly on the threshold value.
4. Contours are detected, and the centroid is calcu-
lated, which is estimated to be the pupil's position.
5. End Procedure
```

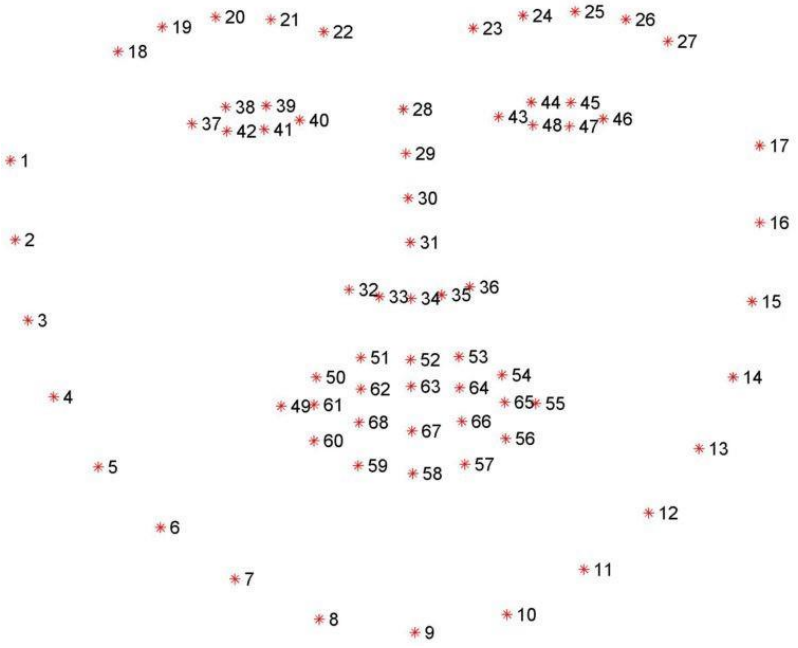

Fig. 5. Shows the 68 landmarks extracted using the pre-trained Dlib model [17, 18]

5 Results

The system proposed is robust and helps with online proctoring, reducing manual intervention, and automating the process with impressive accuracy. The total accuracy achieved in the entire model classification is 96.04%, a weighted average of the accuracy of the individual classifiers. The individual accuracy scores for Classifier 1 and Classifier 2 are 96.59% and 91.64%, respectively. The total is accuracy is lower than the accuracy of Classifier 1 because some portion of the data relies on Classifier 2 to produce a label as gaze estimates cannot be read accurately for all input images. Each video's accuracy score is shown in Table 1. The confusion matrix consisting of 30% of the total test data is shown in Table 2. Figure 6 visualizes the output of a few frames used in the training of the model.

When the model correctly predicts the positive or 1 class, the positive class is a TP (true positive outcome). A TN (true negative) outcome occurs when the model correctly predicts the negative class as a negative class.

Similarly, when the model predicts the negative or 0 class as positive, it is an FP (false positive outcome). An FN (false negative) outcome occurs when the model predicts positive or class 1 as the negative class. The performance metrics computed for the proposed application are tabulated in Table 3. The comparison of the state of the art techniques with the proposed work is tabulated in Table 4.

Table 1. Accuracy scores for individual videos

Accuracy for Classsifier1	Accuracy for Classsifier2	Total Accuracy (Classifier 1 + Classifier 2)	No. of frames with a detected face
98.19%	88.66%	95.96%	1438
96.94%	93.21%	95.97%	2356
98.01%	95.39%	97.60%	1548
99.43%	98.66%	99.25%	1240
99.64%	98.59%	99.61%	942
98.94%	97.55%	98.19%	679
100.00%	100.00%	100.00%	231
98.25%	90.07%	97.27%	1944
99.56%	97.81%	99.38%	3331
96.69%	91.72%	96.09%	5026

Table 2. Confusion Matrix for Classifier 1 + Classifier 2 (30% of test data)

Matrix	Predicted 0	Predicted 1
Actual 0	501	173
Actual 1	50	4913

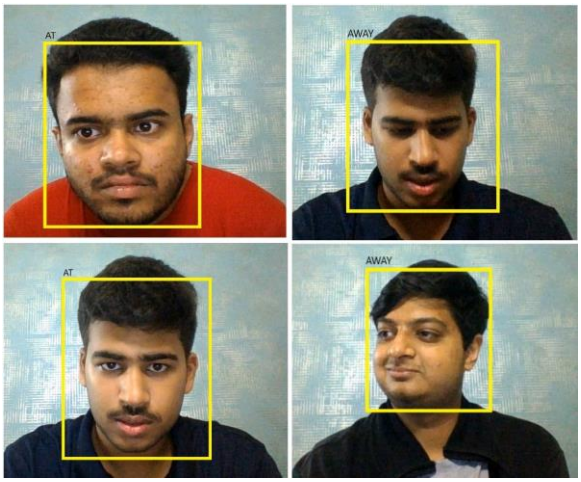

Fig. 6. Visualizing the output of the classifier

Table 3. Performance Metrics

Metrics	Formula	Values
Sensitivity (Recall) or True Positive Rate (TPR)	$\frac{tp}{tp + fn}$	0.989
Specificity (SPC) or True Negative Rate (TNR)	$\frac{tn}{tn + fp}$	0.743
Precision	$\frac{tp}{tp + fp}$	0.965
Accuracy	$\frac{tp + tn}{tp + tn + fp + fn}$	0.96
F1 Score	$\frac{2tp}{2tp + fp + fn}$	0.977

Table 4. Comparison of the proposed model with state of the art techniques

Method	Result	Reference
Detection using yaw angle variations, audio presence and active window capture	Accuracy = 80%	[11]
Detection using heuristic based inference system with face detection, tracking, and active window capture	False Positive Rate = 0.08 True Negative Rate = 0.13	[12]
Proposed System	Accuracy = 96.04%	-

6 Conclusion and Future Work

The model proposed can act as a strong baseline for institutions that want to build an online proctoring system that works with minimal intervention and 96.04% accuracy. As the model has minimal requirements, it would be inexpensive to implement. With the hybrid classifier approach's help, the proposed system would tackle both ghostwriting problems and malpractice attempts using a secondary device.

The model can be further enhanced by the addition of audio and checking audio queues to identify possible malpractice. With audio, it would be possible to isolate voices sources and determine whether the examinee is cheating using auditory aids. It would also be possible to use a modular component in the pipeline that detects whether the student is wearing a headphone.

Another metric that would provide more insight would be the distance of the subject from the screen. It can be precisely measured if the testing condition has more than one lens/camera, enabling depth estimation calculation. Adding both these metrics would result in a more accurate model.

7 Data Availability

The data that support the findings of this study are not made public yet. However, interested readers can obtain data by emailing the corresponding author(kkp.prakash@manipal.edu). The dataset will be sent upon a reasonable request.

8 Ethical Approval

All procedures performed in the study involving human participants were in accordance with the 1964 Helsinki Declaration.

9 Conflict of Interest

The authors declare that they have no conflict of interest.

10 References

[1] McGill, A Brief History of MOOCs (2018). URL. https://www.mcgill.ca/maut/current-issues/moocs/history

[2] A. A. Carr-Chellman, Global Perspectives on E-Learning: Rhetoric and Reality, in: Global Perspectives on E-Learning: Rhetoric and Reality, SAGE Publications, 2004. https://doi.org/10.4135/9781452204390.n16

[3] ETS GRE (November 2020). URL https://www.ets.org/s/cv/gre/institutions/update/

[4] UNESCO (March 2020). URL https://en.unesco.org/news/137-billion-students-now-home-covid-19-school-closures-expand-ministers-scale-multimedia/

[5] Arkorful, V. and Abaidoo, N. (2014) The Role of e-Learning, the Advantages and Disadvantages of Its Adoption in Higher Education. International Journal of Education and Research, 2, 397-410.

[6] Katta, J.Y.B., 2018. Machine Learning for Source-code Plagiarism Detection (Doctoral dissertation, International Institute of Information Technology Hyderabad).

[7] J. A. Ruiperez-Valiente, P. J. Munoz-Merino, G. Alexandron, D. E. Pritchard, Using Machine Learning to Detect 'Multiple-Account' Cheating and Analyze the Influence of Student

and Problem Features, IEEE Transactions on Learning Technologies 12 (1) (2019) 112–122. URL. https://doi.org/10.1109/tlt.2017.2784420

[8] C. G. Northcutt, A. D. Ho, I. L. Chuang, Detecting and Preventing "Multiple-Account" Cheating in Massive Open Online Courses, CoRR abs/1508.05699 (2015). https://doi.org/10.1016/j.compedu.2016.04.008

[9] G. Alexandron, S. Lee, Z. Chen, D. E. Pritchard, Detecting Cheaters in MOOCs Using Item Response Theory and Learning Analytics., in: UMAP (Extended Proceedings), 2016.

[10] O. Ojo, Y. Nureni, Deterring malpractice in a Networked Computer Based Examination Using Biometric Control Attendance Register, International Journal of Advanced Networking Applications 10 (05 2019). https://doi.org/10.35444/ijana.2019.10062

[11] S. Prathish, K. Bijlani, An intelligent system for online exam monitoring, in: 2016 International Conference on Information Science (ICIS), 2016, pp. 138–143. https://doi.org/10.1109/infosci.2016.7845315

[12] R. S. V. Raj, S. A. Narayanan, K. Bijlani, Heuristic-Based Automatic Online Proctoring System, in: 2015 IEEE 15th International Conference on Advanced Learning Technologies, 2015, pp. 458–459. https://doi.org/10.1109/icalt.2015.127

[13] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Vol. 1, 2001. https://doi.org/10.1109/cvpr.2001.990517

[14] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, CoRR abs/1603.02754 (2016).

[15] T. Yang, Y. Chen, Y. Lin, Y. Chuang, FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation from a Single Image, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1087–1096. https://doi.org/10.1109/cvpr.2019.00118

[16] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, Y.-Y. Chuang, SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation, 2018, pp. 1078–1084. https://doi.org/10.24963/ijcai.2018/150

[17] D. E. King, Dlib-ml: A Machine Learning Toolkit, Journal of Machine Learning Research 10 (2009) 1755–1758.

[18] A. George, Image based Eye Gaze Tracking and its Applications, CoRR abs/1907.04325 (2019).

[19] G. Bradski, The OpenCV Library, Dr. Dobb's Journal of Software Tools (2000).

11 Authors

Chirag S Indi., Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), Manipal, India Email: si.chirag@gmail.com

KCS Varun Pritham is currently working as an Analyst at Deloitte USI, Hyderabad. His research interests lie in Natural Language Processing, Computer Vision and Deep Learning Email: kcs.varun@gmail.com

Vasundhara Acharya is currently located in the USA to pursue her Ph.D. Her research interests lie in Artificial Intelligence and Machine Learning. Email: vasundhara.acharya93@gmail.com

Krishna Prakasha K is currently serving as an Assistant Professor-Senior Scale in the Department of Information and Communication Technology, Manipal Institute of Technology (MIT), Manipal Academy of Higher Education (MAHE), Manipal, India. His research interests lie in Network Security, Algorithms and Technology Acceptance Models, Real Time Systems, Wireless Sensor Networks, Machine learning, Computer Networks and Protocols. Email: kkp.prakash@manipal.edu

Article submitted 2020-06-03. Resubmitted 2020-11-19. Final acceptance 2020-11-20. Final version published as submitted by the authors.