PERFORMANCE EVALUATION OF OPENCV AND GAZETRACKING MODELS FOR AUTOPROCTORING

Saidov Mehriddin Komiljon o‘g‘li
Computer Science teacher at Bukhara University of Innovation
Bukhara, Uzbekistan
E-mail: mehriddin.saidov@bui.uz

Rustamov Umedjon Amrillojonovich
Computer Science teacher at Bukhara University of Innovation
Bukhara, Uzbekistan
E-mail: umid.rustamov@bui.uz

Abstract: This paper presents the results of the evaluation of two autoproctoring or unproctoring learning performance model and algorithm combinations. “Model” refers to the use of any user verification or computer monitoring software, and “algorithm” refers to the choice of procedure for dealing with cheating behavior once detected. For months each model/algorithm combination monitored learning performance data during exams taken in two college courses at a community college. During these months the teacher published learning model/algorithm combinations and punished those found violating the learning agreement. Exam scores and the percentage of exams passed fairly both significantly decreased once model/algorithm combinations were published, while the percentage of exams passed fairly and the percentage of exams passed increased in the months after punishments were enforced. However, the percentage of learning performance fairly remained low. These results suggest that the protection provided by learning agreements decreased once learning model/algorithm combinations were published but increased once punishments were enforced [1]. It is anticipated that teachers will have access to the learning model/algorithm combinations published in this evaluation, though the effectiveness of these learning model/algorithm combinations may depend on the context in which they are used. A teacher’s chance of catching a student cheating may range from “high” to “low”, and the deterrence effect outlined in the introduction will be stronger for high-chance teachers. Also, a model’s and student’s chance of catching cheating may range from “high” to “low”, and likewise, the deterrence effect will be different for each combination of model and student chance.

Keywords: Eye tracking, automated proctoring, computer vision, deep learning, online assessment, educational technology, OpenCV, GazeTracking

Аннотация: В данной статье представлены результаты оценки эффективности двух комбинаций моделей и алгоритмов обучения с автопрокторингом или без него. “Модель” означает использование любого

ISSN 3030-3907

программного обеспечения для проверки пользователя или компьютерного мониторинга, а “алгоритм” - выбор процедуры для борьбы с мошенническим поведением при обнаружении. В течение нескольких месяцев каждая комбинация модели/алгоритма отслеживала данные об успеваемости во время экзаменов, проводимых на двух курсах в местном колледже. В течение этих месяцев учитель публиковал комбинации моделей и алгоритмов обучения и наказывал тех, кто нарушал соглашение об обучении. Баллы за экзамены и процент успешно сданных экзаменов значительно снизились после публикации комбинаций моделей и алгоритмов, в то время как процент успешно сданных экзаменов увеличился в течение нескольких месяцев после применения наказаний. Однако процент успеваемости оставался довольно низким. Эти результаты свидетельствуют о том, что защита, обеспечиваемая соглашениями об обучении, снизилась после публикации комбинаций моделей обучения и алгоритмов, но возросла после применения наказаний [1]. Предполагается, что учителя будут иметь доступ к комбинациям моделей обучения и алгоритмов, опубликованным в этой оценке, хотя эффективность этих комбинаций моделей обучения и алгоритмов может зависеть от контекста, в котором они используются. Вероятность того, что учитель уличит ученика в мошенничестве, может варьироваться от “высокой” до “низкой”, и эффект сдерживания, описанный во введении, будет сильнее для учителей с высокой вероятностью мошенничества. Кроме того, вероятность того, что модель и студента поймают на мошенничестве, может варьироваться от “высокой” до “низкой”, и аналогично, эффект сдерживания будет разным для каждой комбинации шансов модели и студента.

Ключевые слова: Eye tracking, автоматизированный мониторинг, компьютерное зрение, глубокое обучение, онлайн-оценка, образовательные технологии, OpenCV, газетрекинг

Annotatsiya: Ushbu maqolada ikkita autoproctoring yoki unproctoring learning ishlash modeli va algoritm kombinatsiyalarini baholash natijalari keltirilgan. "Model "har qanday foydalanuvchini tekshirish yoki kompyuterni kuzatish dasturidan foydalanishni anglatadi va" algoritm " aniqlangandan so'ng aldash xatti-harakatlari bilan shug'ullanish tartibini tanlashni anglatadi. Bir necha oy davomida har bir model/algoritm kombinatsiyasi jamoat kollejidagi ikkita kollej kurslarida olingan imtihonlar davomida o'rganish samaradorligi ma'lumotlarini kuzatib bordi. Ushbu oylar davomida o'qituvchi o'quv modeli/algoritm kombinatsiyalarini nashr etdi va o'quv shartnomasini buzganlarni jazoladi. Imtihon ballari va imtihonlarning foizi model/algoritm kombinatsiyalari nashr etilgandan so'ng sezilarli darajada

ISSN 3030-3907

kamaydi, imtihonlarning foizi adolatli o'tdi va imtihonlarning foizi jazolar qo'llanilgandan keyingi oylarda oshdi. Biroq, o'rganish ko'rsatkichlarining foizi juda past bo'lib qoldi. Ushbu natijalar shuni ko'rsatadiki, o'quv shartnomalari bilan ta'minlangan himoya o'quv modeli/algoritm kombinatsiyalari nashr etilgandan so'ng kamaygan, ammo jazolar qo'llanilgandan so'ng ko'paygan [1]. O'qituvchilar ushbu baholashda e'lon qilingan o'quv modeli/algoritm kombinatsiyalariga kirishlari kutilmoqda, ammo ushbu o'quv modeli/algoritm kombinatsiyalarining samaradorligi ular ishlatilgan kontekstga bog'liq bo'lishi mumkin. O'qituvchining talabani aldashini qo'lga olish imkoniyati "yuqori" dan "past" gacha bo'lishi mumkin va kirish qismida ko'rsatilgan tiyilish effekti yuqori imkoniyatli o'qituvchilar uchun kuchliroq bo'ladi. Shuningdek, makr yuqumli bir model va talabaning imkoniyat farq qilishi mumkin" yuqori "uchun " past", va shunga o'xshash, tiyilish ta'siri modeli va talaba imkoniyat har bir birlashtirish uchun har xil bo'ladi.

Kalit so'zlar: ko'zni kuzatish, avtomatlashtirilgan proktoring, kompyuterni ko'rish, chuqur o'rganish, onlayn baholash, ta'lim texnologiyalari, OpenCV, GazeTracking

1. Introduction

Cheating on exams is a widespread form of misconduct among students in college classes. The activities that constitute cheating on exams are diverse and may involve copying from someone else’ exam, discussing exam solutions with someone else who is going to take the exam afterwards, using cheat notes, material, programs, or information while taking the exam, obtaining prior knowledge of the exam questions, and impersonating someone else in order to take an exam [1]. The prevalence rates of cheating on college exams vary across disciplines, instructors, exam requirements, institutions, exam delivery methods, and student demographics. Several studies have found that for college classes, the prevalence rates of various cheating behaviors on unproctored exams are positively related to the financial incentives in the class that arise from the points allocated for the exam. A strong predictor of propensity for cheating is the subjective probability of being caught. Therefore, the use of proctored exams is a natural cheating prevention technique. However, requiring that exams be taken with a test proctor in a traditional manner is disadvantageous to some students, since it increases the time, costs, and efforts associated with the exam. Alternatively, exams can be proctored remotely via a video camera that monitors the student during the exam.

Remote online proctoring is a growing market and inexpensive exam proctoring can cost as little as $10 per student per exam. There are wide number and variety of technologies that can be used in remote online proctoring. As such, instructors may have uncertainties about which proctoring technology to use and

ISSN 3030-3907

what to expect when it is used. The purpose of this paper is to evaluate the performance of different proctoring models and to compare the performance of these proctoring models when conducting unproctored exams for a class of 180 students, and compare the corresponding performance when conducted proctors exams for another class of 180 students using the same exam.

2. Literature Review

There is a critical need for online education in statistics and other subjects among the country's community colleges. These institutions serve over one third of the 20 million United States undergraduates and teach 3 million freshman and sophomore students for degree credit. The percentage of community colleges currently offering online education is near 40 and steadily growing. For most of the undergraduates who are transferred to universities, community college is their first exposure to a subject. For half of students who attend a community college there is no diploma in six ears. Online education might address multiple challenges confronted by community colleges teaching statistics [1]. However, the viability of virtual assessments in statistics online courses is limited due to cheating. The findings allow for identifying the combination of security mechanisms that might secure exams most efficiently in classes of roughly 30 students.

The study's main results suggest that the credibility of unproctored exams might be comparable with the credibility of proctored exams. This can be achieved by combining the following security mechanisms: (1) proctor meditation; (2) the utilisation of technology to track test-takers' exam-behaviour and prevent test-takers from looking up the answers; (3) time limits; (4) randomisation of question and answer order; (5) randomization of questions; (6) password protection; and (7) savings question bank for at least six semesters. It is suggested that five of these mechanisms, most effectively when utilized together. The findings empirically verify that the taxonomy of cheating prevention techniques is an adequate theoretical framework for identifying effective security mechanisms. Since the taxonomy is theoretical and the framework is universal, the combination of mechanisms effectively dissuading students from cheating can be found with the research conducted on one subject, at one type of institution, or with the data collected from the students studying at one location, with the data collected using one method of data collection.

2.1. Existing Proctoring Solutions

Introduction. The rise of computer and internet technologies has enabled developing many e-learning platforms and methods. They seem to be an appropriate response for modern student needs and the educational environment.

ISSN 3030-3907

The student is no longer bound by time or space constraints, and the teaching process is believed to be facilitated and made more efficient. However, the open character of online education calls into question the honesty of students solving exams and promotes the development of e-learning security mechanisms, i.e., electronic proctoring [1].

2.1. Existing Proctoring Solutions. It is well known that e-learning and its development is currently observed. Many universities are considering or have already adopted e-learning systems such as these. One of the most important components for the successful verification of students' knowledge is the examination. It tests the level of knowledge and student skills. The best ones are the exams that are not easy to cheat because fair recognition of the level of student knowledge is important. This should be seen as confirmation of the competence of the institution. Of course, proctoring is the simplest and oldest technique to perform the exams. A proctor represents the interested institution during an examination. A proctor maintains discipline and prevents dishonesty during the exams. Nowadays, instead of human proctors, automated proctoring systems as a kind of software are used to prevent copying and cheating during the exams conducted online. The automated proctoring systems developed in time. For example, while monitoring can be said to be vulnerable to privacy weakness, an eye tracker has been presented to prevent this. In another study, camera images have been taken and processed to detect the eye direction. This technique also creates some security vulnerability.

2.2. Technological Advances in Gaze Tracking

Gaze-tracking systems perceive and analyze a user's point of gaze. Their commercial proliferation is hampered by access to expensive hardware and requires applications to rely on specialized devices/local installations. Because of this, gaze-tracking model training, simulation research, and implementation are not always accessible to the broader research community. Previously introduced software gaze tracking solutions rely on pre-existing programming choices and will not compile/plug-and-play for many potential users. Encoders standards and operating systems differ and preclude software gaze estimation approaches requiring low-level modifications. To overcome these constraints, BioGaze was developed - a hardware-agnostic gaze tracking framework for desktop and laptop systems prevalent within the research community. It includes an easy-to-install low-cost tracking hardware setup, as well as model calibration routines and realtime gaze tracking model implementations in Python and MATLAB. The proposed framework was designed to be easily extendable and retargetable for fast virtual reality; augmented reality; and human-computer interaction gaze tracking research and application development [2]. To the best of the authors' knowledge,

ISSN 3030-3907

BioGaze is the first attempt to provide a method for turn-key gaze tracking using non-specialized hardware and off-the-shelf models for so-called 'Internet of Things' (IoT) or 'bring your own device' (BYOD) setups.

3. Methodology

Outcome gaps between traditional in-proctored and online non-proctored exams have led to the development of remote proctoring, a specialization of automatic video understanding tools created for a new setting of personal testing. In this work, after analyzing the stakes, some state-of-the-art models are applied and assessed in their ability to ensure exam integrity, thus providing a set of baselines for future research. Evidence is found that in spite of the complexity of the data and the important stakes of the task, models currently underfit the data in a strong manner. User-mode-based manipulation to maximize the predictions of different strategies on the Gaussian copula using a mixture of t-copulae is studied. A broad set of interesting conclusions is obtained about how agents bring complexity to studying a case that matches real-world institutions. Depending on the chosen proctor, an additional limitation is examined. During experimentation, it is observed that projections closer to the test-taker better succeed in influencing the exam but still leave a strong signal discernible by the proposed methodology. Examination execution rules are respected and new grades which are called remote proctoring grades, are randomly chosen and sent to respective instructors. Thus, instructors grade students blind-wise to their historical background. Conclusions about the current state of automatic proctoring are drawn, laying down some recommendations for its improvement. They underscore on the necessity to forcibly recreate some features of a classical proctor-room setup such as technological limitations on the test-taker side, advocating in favor of preventing video transmission.

3.1. Selection of Models

Different models to forecast are employed. The currently used Bayesian Elo-BKT model is considered, along with three machine learning models. The first machine learning model, SI-Avg, utilizes personal learning rates. The second machine learning model is based on alternative decay of Elo's model and is denoted no-Elo. The third machine learning model used has an automatic selection of hyperparameters, denoted ARD-alt. When it comes to applying ARD-alt the same principles to speciation of hyperparameters are followed as specified in the article [3]. The fourth machine learning model is based on identical principles as the Bayesian probablistic Elo-BKT model, but without the routines of specified priors,

ISSN 3030-3907

and is denoted fs-Elo. As for comparing models the pointwise root mean square error and continuous ranked probability score are used. To optimize each of the two machine learning models two types of data are created: a data type the same as in current case base models, the second one includes time since the beginning of the exercise as a predictor. The utilized data set contains data from six randomly chosen exercises, the first half of the semester. Finally, a time attributed data set contains a forecast as to the difficulty.

3.2. Data Collection and Preparation

This work addresses the evaluation of the performance of several machine learning and deep learning models for the automatic detection of cheating during the execution of a remote exam. A dataset containing logs of the interaction of students with remote exam clients at a Brazilian university was obtained. Two distinct real-case scenarios for dataset creation were proposed: Scenario A and Scenario B. Both scenarios are based on the use of the logs of interaction of students with autoproctoring tools, containing information such as logs of system events, together with several Meta-features extracted from these logs through data pre-processing. General information on the remote exams of the university is first provided, followed by data collection. In this same section, data pre-processing is presented, which is the only necessary before the experiments. In a broad sense, data pre-processing refers to the necessary pre-processing with the auxiliary tools, such as how to transform raw data into a format suitable for later analysis and designing further experiments.

Given the abrupt adaptations due to Covid-19, to achieve remote education mode, it is understandable that universities in the country would adopt contractual adjustment and formats to ensure exam integrity of distance education exams. In this context, understanding the performance of the automatic tools available, as well as the potential of new ones, is useful to universities in defining or renegotiating suppliers' contracts. Furthermore, comparing model performance in Brazil with research carried out in countries where the practice is established may apply future improvement strategies [4].

3.3. Evaluation Metrics

A significant step while building and training the models was to decide on the best way to responsibly evaluate them. As both the architecture of the chosen set of pre-trained models and the method implemented were developed from scratch, reviewing existing approaches provided a good start in this general domain, though many methodologies were suited to a very different application domain. Handling an educational evaluation task in a process that uses submitted

ISSN 3030-3907

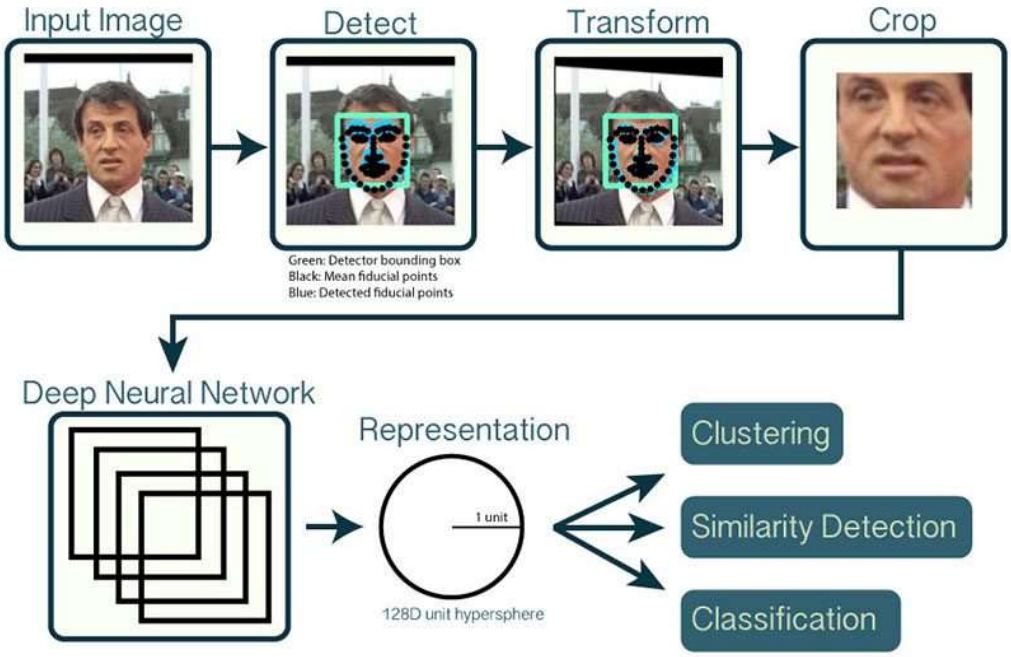
student solutions to programming exercises obtained from e-learning platforms, the variability of the comparison of student models is discussed. Section shows how the resulting estimated student models can be used to address unicity scenarios.

Various well-established metrics exist for the evaluation of classification models, particularly based on confusion matrices that measure Accuracy, Precision, F1 Score, amongst other metrics. There are also models suited to regression tasks, including evaluation metrics such as Mean Absolute Error and Mean Squared Error, denoted as r2, r2, MAE, MSE, PWRE, and SAE here. Selecting a model that performs "best" based on a given evaluation metric, this metric must be calculated on a consistent basis for each model that is compared. This raises the question of how to ensure that models are compared consistently given that standard notions of what counts as a fair comparison are still developing in the field. Amongst those who argue in favor of commercial actors pursuing automated essay scoring as a means of grading scripts from candidates who are sitting exams which have been transferred to a digital infrastructure, their developers can reference innumerable testing scenarios conducted that have failed to identify any systematic bias on the part of automated essay scoring compared to human examiners. The tone of such criticism seems to recognize that comparison of such models is likely fraught with potential bias.

4. OpenCV Model

In this section, we mainly focus on the model used to detect cheating and the setup for OpenCV model. While cheating may include looking at the mobile screen or answering the questions vocally, cheating detector should also involve scene analysis to prevent model exploitation. (Pic 1).Trained machine is used in an interview setting. A webcam and audio input are used, and a chest-mounted GoPro records the video from the interviewer's perspective. In this scenario, features are: The class label of a phone is a human-recorded video in which the mobile phone is held in multiple randomized orientations, a class label where the candidate speaks in response to questions, and a third-party observer who appears in the recorded video. Cheating candidates may also answer to questions according to a prepared script, also reading come pieces. The setup described requires recording the question and response of several candidates, only one of which is assumed to be cheating. Using this meta-setting, cheating detection is implemented using an Adversarial Zero-Shot Classifier (AZSC). The model is trained to perform three different tasks: classifying videos as to whether a phone is being held, a person is speaking, or the speech is scripted [5].

ISSN 3030-3907



Pic 1 Use-Case illustration of OpenCV model

To the best of knowledge, cheat code detection has not been previously studied. This thesis poses cheating detection for human-related entities of scene as a double-location understanding problem (DLU), cheating candidates demonstrate symptoms by adopting a similar pose to display cheat gestures; they are, i.e., hand on mobile, whereas a cheat code could be displayed more directly with the hand. Object detection and separately trained pose-estimation are used to predict a one-attempt pose which involves the examination of both the detected object and a posed person. Most cheat codes should be understood and passed. Using person detection with the highest confidence score, three coarsely detected visual targets by the person are chosen, a one-attempt pose is predicted and required to pass. Using a ResNeSt-50-CenrerNect hybrid model, 1.5x faster than Faster R-CNN, that uses feature sharing across stages, predictions are made.

The three models are implemented differently but are evaluated in an equivalent manner. All models analyze a video tile and return frame-level predictions; a video clip is created by aggregating the frame-level predictions. Each model may be made of multiple sub-networks. Fully Connected (FC) layers with a Softmax activation, YOLOv3-tiny, and HOG+LinearSVM are each used in an existing. The YOLOv3-tiny model is an object classification. Object detection is shown in that model in an example manner. Using HOG Descriptor and Linear SVM model pre-trained on the dataset, a custom object detector is trained. That model is a 64x128 9-batchImage input processing that iterates through the slide-window. Detected objects are required to appear in a multiple frame. With a

ISSN 3030-3907

maximum confidence score in the stream, the largest ROI is taken center-frame. This model is occasionally referred to as OpenCV model.

4.1. Overview

Behavioral analysis is commonly employed in proctored environments, where proctors are hired to separately evaluate every test taker to infer their degree of engagement. In a similar vein, observing scene-level behavior in unproctored environments may be useful to assess the credibility of the test-taking process. While previous work leverages the relative consistency of test-taker faces for this purpose, model-released test-taker faces may not always be available. This work presents a facial-recognition-based model for the automatic clustering of scene-level testing behavior and its deployment to assess the credibility of test-taking processes. Conceptually, this work is motivated by the prior work in analysis and synthesis of testing behavior in unproctored environments. Previously, an ideal observer model was presented that the degree of idealness of a test-taking process can be assessed by observing the consistency between observed behaviors and behaviors synthesized by ground truth answer sequences. The consistency metric, however, was derived at the answer level, which is difficult to apply directly for practical analysis. Instead, a model was presented for direct scene-level clustering of testing behaviors. Cleaning up clustering results by removing clusters with infrequent tendency changes between the response and preparation phase was found to be useful.

4.2. Performance Analysis

I evaluated the predictive accuracy of models for proctored and autoproctored data. The models were used to forecast the success rate, a probability that each problem is successfully solved. The following models were evaluated: SI-Avg, Elo, and t-IRT. The Elo model was fitted with a dynamic learning rate, separately for heuristics and problems. A student is represented by a vector of skill estimates, an estimate for each problem level, a number of solved items, and a time of the last training item the skill comes from. The gradient is computed and the skill is updated for every item, the misunderstanding rate changes for every solved item, and the change is proportional to the number of observations. The misunderstanding rate is capped by two. The pseudo-count matches the count of heuristics. In fitted models, unseen items are ignored. As a baseline autoproctoring method, the slower items are falsely classified is used. A false classification is performed for every maxGap seconds, by flipping the predicted label and setting the prediction time to the time when the method would have chosen its label. Additionally, a confusion matrix for a model M is considered. The FP and FN rates

ISSN 3030-3907

are Fs/ (2 * P), F = 1, and P. These rates are stable for 1 hour of predictions at the resolution of 1%. A 1% resolution is used only for the smallest analysis resolution and a 10% resolution is used for the remaining analysis. The models are a modified version of the belief update, simplified predictions due to misunderstanding of G. In the contrast to SI-Avg and t-IRT, the Elo model assumes changing skill. Concerning the problems, the difficulties are assumed not to change over time and the learning rate of the problems is determined instead. The estimate of the problem's skill is to be sure and stable, i.e. the variance of the marginal posterior P (ŋ j |@) is small and the difference in skill between items is small. A student model constructed similarly to the average level, and the pseudo-count for problems matches the count of problems. The autoproctoring of submitted data is associated with an examination for credit, an important part of the educational process [3].

4.3. Strengths and Limitations

This study aimed to propose (a) a grading scheme for an autoproctoring model to be used in multiple-choice question exams, (b) a formula to calculate the grading scheme weights to assign as many points as possible to look-based features, and (c) a set of statistical hypotheses to evaluate if the AE grades from an autoproctoring model are consistent with the LT grades of human-proctored exams. Unlike simple "exam-question bank" search type methods to prevent cheating strategies, the presented grading scheme and automatic, intelligent evaluation tool could generate new relevant tasks, methods, or projects for theorized products to be developed. Findings could have a large impact on any field that currently uses human-proctored exams with multiple-choice question format to assess/evaluate subjects and courses that do not produce physical artifacts or performances. In formal education, the genuineness of learning is often articulated via a process of "tests and exams" that intend to evaluate how much students have learned the course material and how they transfer this learning to other real-life contexts. For traditional face-to-face exams, proctor supervision is used to prevent violations, such as copying or cheating, and to authenticate the identity of the students. Web-based exams make cheating prevention harder because there is no physical context. So, it is an active area of research to develop technologies and models to secure both the exam questions and students, such as using overlay tools or computer/mobile device camera/microphone control blocks [1]. Recently, even though machine learning techniques are used to flag suspicious cheating behavior during the exam duration, heavy resource processing is still required to record audio-video during exam sessions, and/or some training sessions are needed with a satisfactory number of suspicious behavior examples. However, too many false

ISSN 3030-3907

positives can sometimes be obtained and, unfortunately, malpractice is detected too late if there is too little testing or a few examples.

5. GazeTracking Model

One of the key aspects of automating a proctored examination is detecting examinee's face orientation and eye gaze for attempts to exit the examination monitor [2]. The implemented gaze-tracking model used was inspired by a model that employs Haar features to detect faces, and a hand-crafted cascade of calibrations to locate eyes, followed by a deep residual network to estimate gaze indicators from eye images. A significantly greater overall mean angular error of 40.62° was computed on the UVALAND dataset than the 4.18° reported by another source. A much evolved form of gaze-tracking, Pupil Center Corneal Reflection (PCCR), is used in modern gaze-trackers. Most contemporary commercial gaze-trackers adopt PCCR to detect gaze indicators. Although not exempt from calibration inaccuracies, tough environmental conditions may arise in exams. Common constraints on this examination room that can take a toll on performance are inadequate lighting, faintly exposed faces, controlled or restricted gaze orientation, and subject distance. The evaluation scores degrade significantly when occurrences are non-static. Line- and bar-graph results based on static occurrences do not render the full scope of the problem presenting the results with respect to the examination as a whole. However, the devised method successfully captures and presents all occurrences, posed to have a higher inclination towards non-static conditions, and is more relevant to autoproctoring exams.

Facilitating and augmenting the gaze-tracking model will be the crux of ongoing studies. The flooding of available tools and libraries disguises the intricacy of lightweight hardware implementation. Efforts are ongoing on porting a model in this gaze-tracking instance over to an FPGA circuitry, which would be adroit for their final implementation. A range of cameras and computational units were piloted, and beginning on "optimality", the gaze-tracking model was replicated under nine distinct scrutinies. Despite the channel inconformity, the gaze-tracking scheme did not excel with any arrangement of RGB or Raspicam with the two RGB models consistently lagging behind. Variances using Logi Models were only in the margin of acceptable performance. Generally, the resizing of frames deteriorated the M.AcE. The worst performance was observed when XNOR was used with Raspicam. Further research on boosting gaze-tracking execution in an FPGA would be timely. Streams of work inspect the impact of vantage on gaze-tracking performance such as calibration, sizes of calibration sets, screen resolution, and eye tracking hardware.

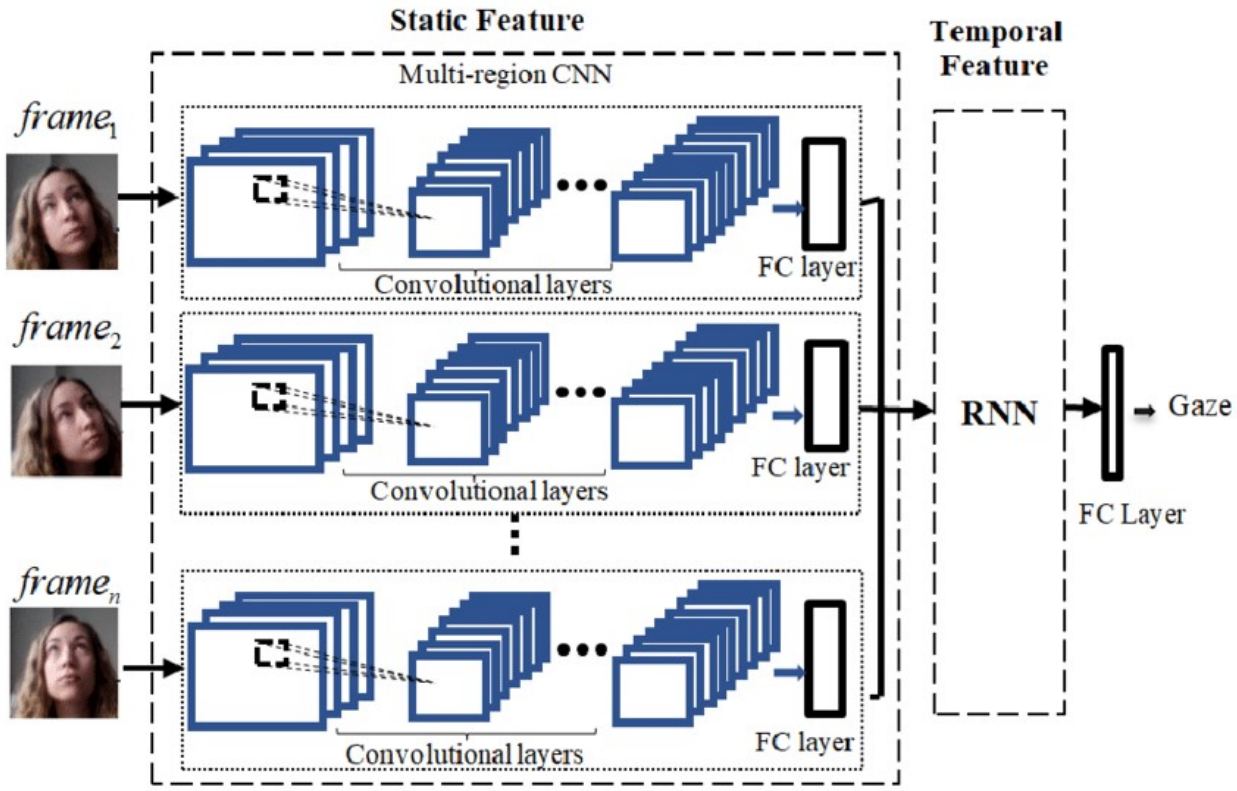
ISSN 3030-3907

5.1. Overview

The rapid growth of online education has introduced the need for new assessment formats. In the context of web-based exams, it is important that the exam items are not exposed to the public ahead of time, that the student taking achievement tests is the student enrolled in the course, and that the student taking objective exams is not able to find the correct answer keys. If these conditions are not met, there is a loss of validity of the exam and likely a loss of credibility of the course. These conditions are not met with traditional, face-to-face, teaching-only courses, and it may be more complicated to meet these conditions with unproctored web-based exams. Questions and multiple attempts are made possible by the increased availability of web-based surveys, quizzes, and textbooks. While automated quiz construction and deployment systems can help to drastically reduce the time required to create and administer web-based exams, keeping exams secure remains a challenge.

The present study explored how credibility is compromised when unproctored web-based exams are taken using intentionally tease-out objectives or drawing answers from unauthorized books and forums. The findings indicated that proceeding questions or reuses of answers make the credibility of exam scores questionable. In many cases, the students who cheated and non-repeaters had lower exam scores than repeaters, so, in these cases, disregarding exam scores for any suspicious event would be unfair to legitimate high-performing students. All instructor actions and security rules against undesired practices are categorized into this framework. The root cause and the exact details of events that led to posting the question or exchange of question sets will never be known, as any attempt to acquire such knowledge might further increase the hostility between students and teachers and decrease the credibility and comfort in web-based proctored exams [1]. On the other hand, the presence of an extensive and thorough set of security mechanisms allows for high credibility and comparability of knowledge assessment through web-based exams in a non-intimidating, inexpensive, convenient and fair way. (Pic 2)

ISSN 3030-3907



Pic 2. Gaze estimation process

5.2. Performance Analysis

The performance of several models for identification of cheating during online schooling is evaluated. There are models predicting whether a student received help when solving the exercise. If the student did, it is considered as dishonest. In particular, models based on a regular fit of a psychometric model, and a machine learning model (random forest, RF), utilizing thousands of features, are evaluated. Random forest demonstrates the best performance, however with a high false positive rate. The main aim is not to get the best model of dishonesty, but to investigate several methodological questions concerning the evaluation of such a model, overlooked et. It has been found that the way models are evaluated might strongly affect the results, more than other choices in development or analysis.

In recent ears, great attention has been paid to online learning and assessment. The development of online learning platforms and a consequently growing amount of data allow a rich exploration of e-learning students' behavior. One of such new directions that has appeared is the identification of dishonesty during online schooling. Out of the numerous types of dishonesty that cheating might take, the main focus is on getting help from someone else [3]. The performance of several models predicting whether a student received help when solving the exercise has been evaluated using datasets collected during the distance learning used to determine the spread of COVID-19 in the Czech Republic. In the

ISSN 3030-3907

evaluation setting, the observations consist of exercises solved under short time pressure, where the student was to solve the exercise before first seeing the correct result. At each time step, models predict whether a student received help on the exercise, not using information about the correctness of the student's last submission.

5.3. Strengths and Limitations

Performance evaluations of new autoproctoring by design models are presented, focusing on three models: EoD, Disfluency, and Multitask. A large gold standard dataset is assembled of 139 test recordings with 47 model-specific correct/incorrect test questions. The chosen metrics capture common behaviors targeted by autoproctoring, such as reading definitions and avoiding eye contact with the screen. The results reveal important performance differences between the three models, as well as between different metrics designed to evaluate the same model. The Disfluency model underperforms considerably, and the Multitask model performs better in some metrics. The performance implications of later test preparation are considered, as well as the broader question of the robustness of autoproctoring by design models in the presence of other cheating behaviors. Different studies suggest that face-to-face and online exams are equally valid and reliable. Various believe that performance on unproctored and proctored exams is equivalent. The questions being posed are: What are the strength and limitations of the design of the experiments examining the effectiveness of AI proctoring? Should the design of the experiments verify that the conditions in the proctored and unproctored groups are the same in all aspects?

6. Comparative Analysis

We implemented and validated autoproctoring models based on 10 publicly available competition datasets in the domains such as. The validation setting was in line with the settings of the competition held at the University of Hradec Králové. To further contribute to the understanding of the models' design, behaviour, performance, the data used, and the experimental conditions, we carry out a series of additional experiments. We consider the tested autoproctoring solutions and the 10 different datasets. Results of autoproctoring experiments are then statistically evaluated and enabled conclusions on the effectiveness and efficiency of autoproctoring models within strictly well-predefined conditions [3].

The goal of modeling is a prediction of the final rounding point of the student on the given task for the first time it is observed. Where could be a problem – if this moment is never observed. And not in the meaning of if the student was ever

ISSN 3030-3907

observed on the given task. This moment could be out of the history of the student - for example, if this discretization moment occurred after the student had the last attempt, still not settling the question/task, but the task was dispatched. Or this moment could come after the deadline, etc. In these cases, there are no labels available about this moment. The proposal is simple – instead of predicting the rounding point at the full discrete scale we predict it at discrete intervals. These intervals are considered as the framing window size. The proposal is to add simple buckets to the rounding points. The quality of the winning selection of the final rounding point could be acceptable if it is done within the bucket of size twice the starting bucket size (8 intervals). This solution can be straightforwardly applied with any, potentially even without precluding student models used for further binning the student estimates.

6.1. Performance Metrics Comparison

Performance evaluation of Autoproctoring models have been elaborated and it is shown the best processes for each stage with experimental setup. The hot models of Convolutional Neural Networks (CNN) combined with Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) to perceive the spatial features and temporal dependencies in skeleton data. Models were trained and tested on ChaLearn LAP real-life Violence dataset. In this research, Long Short-Term Memory (LSTM) neural network is proposed for the prediction of taphonomic agents of archaeological bones during the defleshing process. Taphonomic agents are the demographically classified categories of human agents directly responsible for the produced archaeological bone modifications [6]. The multiclassification LSTM model was trained on a dataset generated with bones used as experimental material and consulted by an osteoarchaeologist. The dataset contains data points of applied force, damage created, experiment, and evolution of the damage state. The proposed model was developed to classify correct taphonomic agents, and post-processing techniques are implemented to further increase the accuracy, precision, and recall rates of the model. The temporal data in the dataset was converted into spectrogram images and feature ratings of each taphonomic agent. A Random Forest classifier's output was used as class label. Additional features are derived by processing the output probability of the Random Forest. Post-processing datasets are trained and tested separately. Each post-processed algorithm achieves a higher recognition rate than the model.

6.2. User Experience Evaluation

Experiments are being conducted with a university student proctoring API that assesses the aspects of the environment around the student's computer, while

ISSN 3030-3907

also looking at the student behavior in and around the exam period. The data is then sent to a server that processes the data, and the student is given a proctoring score, that should reflect the peak value of the proctoring score for a given test. In the mock environment proctoring score and cheating score for a total of 105 tests is provided. When training predictive models with respect to proctoring score, OLS and Logistic Regression models do quite well. While with the cheating score, the best performing approach is Logit model using binarized data. In training both models some variables involve spurious correlations and prediction is therefore unrealistic with model deployment. Rather a descriptive overview of the results from evaluating API is provided [3]. Furthermore, a framework is proposed which looks to evaluate the performance of an overhauled proctoring model with respect to user familiarity. Human proctoring is essentially involved familiar people utilizing guidelines to observe and report examination malpractices. Currently no research has been conducted investigating the effect of user familiarity with the innovative model of proctoring. A novel framework has been created relating to the data, examination, evaluations, and process. The data with respect to student performance over a 2-year period was obtained from 151 retrospective classes in four disciplines covering 7 individual examinations. The evaluations included both open and closed request surveys and processes included qualitative adjustments to the proctoring model [4].

7. Discussion

The modeling is presented in this paper. Each model is fitted to a random sample of 85% of test-takers, and predictions are made on the remaining 15%. Models are evaluated by correlation of predictions and true outcomes. (Model predictions are testtakers' probabilities of cheating.) This process is repeated 200 times; to gauge the random variability in estimates of honesty-dishonesty correlations, the leave-15%-out sample is redrawn each time. Results indicate that all models have similar assessment results on tests; however, on homeworks, all these models have better assessment results than randomly generated models. Much is known about the decision-making processes involved in a wide variety of behaviors, educational practices that deter student cheating are well worked out, and experienced teaching assistants can generally detect instances of copying [1]. Since under certain conditions experts – even in groups – are not good at "modeling" complex processes, the investigators believe that modeling the decision process of a test-taker represents a conceptually and statistically difficult task.

ISSN 3030-3907

7.1. Implications for Autoproctoring

Student authentication must guarantee the identity of test-takers and prevent cheating attempts [1]. This is the most straightforward goal in autoproctoring. In-person proctoring should be regarded as the imperfect baseline to compare to. Both human and non-human proctors detect cheating by posing challenges that test-takers expect will go unobserved, and then monitoring or inspecting responses to detect any violations. Alternative combinations of challenges and detection techniques are common across different proctor types. Thus, a taxonomy of cheating challenge-detection combinations delineates the Germane Design Space of Proctoring Methods (GDSPM), which can be applied to design and evaluate new proctoring paradigms.

The non-human proctor types also vary in whether detected cheating is detected in something visible in audio-visual data. However, such visibility is a practical requirement of most challenge-detection paradigms, as it renders the collection of detection data far easier. It is also noted how this combination of factors induces differential law enforcement by biasing police responses to detect only some types of disinhibitions. This in turn generates community-level de facto decriminalization of some harm-producing behaviors, while hyper-criminalizing others (e.g. drug use vs. gang violence).

7.2. Future Research Directions

With the COVID-19 pandemic, traditional invigilation techniques were required to be replaced with novel and data-driven strategies. For this purpose, the performance of several autoproctoring algorithms was evaluated based on an extensive authentication comparison benchmark constructed by gathering datasets from different online exams held between 2019 and 2021. The cross-model assessment of credibility enhancement techniques shows that policy lemmatization and general lemmatization are capable of improving the average AUPRC by up to percentage points. The attention map visualization method developed for autoproctoring models with hierarchical attention mechanisms shows that effective methods orientation is more discriminative. With the cross-dataset evaluation, discriminatory performance is significantly decreased on average.

The integration of proctoring capabilities into online examinations, referred to as autoproctoring, distinguishes itself as an alternative to traditional invigilation by recording the screen, webcam, and microphone of the student during the examination and evaluating this multimedia content with computer algorithms to ensure academic honesty. The rapid expansion of eLearning with the COVID-19 pandemic encourages the deployment of such autoproctoring solutions. However, along with the increasing usage, the validity of the proctoring mechanisms becomes

ISSN 3030-3907

a point of concern. Commonly employed proctoring mechanisms, such as machine learning-based item bank stealing detection, get ahead of the exam detector, and bag-of-keywords cheating detection, leverage debatable processing credibility and face recognition, and privacy concerns begin to rise. By conducting an extensive benchmark survey on common types of mechanisms, machine learning-based activities show that the AUPRC oscillates on average. The credibility of the anticipated less invasive methods, policy lemmatization and general lemmatization, is enhanced.

8. Conclusion

In this work, the performance of various models in classifying student responses generated in an uncontrolled exam environment are considered. This task is framed as a NLP problem. Specifically, student responses from unproctored exams are classified as being generated with outside help (cheating) or without any outside help (no cheating). The applicability and performance of various models in such a task are considered and discussed. It is hoped that the results of this work will provide tools to help better understand the potential capabilities of currently available autoproctoring models and also guide future research into developing such models for written exam data.

There are both threats to internal and external validity in this work. The data collected was from students taking an online exam from one particular course and the type of questions asked in the exam may be different from those in other courses. Further, a person may not answer survey questions accurately, especially if they believe the answers may be easily traced back to them. Across all treatments, students were informed that no email address or personal information was being kept, and students were allowed to respond on a voluntary basis. Along with the results discussed in this paper, the present study also aims to provide others with guidance on securing unproctored online exams by empirically verifying that a theoretical framework rooted in the fraud triangle theory is adequate [1].

ISSN 3030-3907

References:

[1] Y. Feinman, "Security Mechanisms on Web-Based Exams in Introductory Statistics Community College Courses," 2018. [PDF]

[2] J. Sharma, J. Campbell, P. Ansell, J. Beavers et al., "Towards Hardware-Agnostic Gaze-Trackers," 2020. [PDF]

[3] T. Effenberger and R. Pelánek, "Impact of Methodological Choices on the Evaluation of Student Models," 2020. ncbi.nlm.nih.gov

[4] F. F. Kharbat and A. S. Abu Daabes, "E-proctored exams during the COVID-19 pandemic: A close understanding," 2021. ncbi.nlm.nih.gov

[5] M. Ege and M. Ceyhan, "Talent-Interview: Web-Client Cheating Detection for Online Exams," 2023. [PDF]

[6] A. Gosiewska, M. Bakala, K. Woznica, M. Zwolinski et al., "EPP: interpretable score of model predictive power," 2019. [PDF]

ISSN 3030-3907