Massive Open Online Proctor: Protecting the Credibility of MOOCs Certificates

Xuanchong Li Kai-min Chang Yueran Yuan Alexander Hauptmann
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, USA
xcli@cs.cmu.edu kkchang@cs.cmu.edu yueranyuan@gmail.com alex@cs.cmu.edu

ABSTRACT
Massive Open Online Courses (MOOCs) enable everyone to receive high-quality education. However, current MOOC creators cannot provide an effective, economical, and scalable method to detect cheating on tests, which would be required for any certification. In this paper, we propose a Massive Open Online Proctoring (MOOP) framework, which combines both automatic and collaborative approaches to detect cheating behaviors in online tests. The MOOP framework consists of three major components: Automatic Cheating Detector (ACD), Peer Cheating Detector (PCD), and Final Review Committee (FRC). ACD uses webcam video or other sensors to monitor students and automatically flag suspected cheating behavior. Ambiguous cases are then sent to the PCD, where students peer-review flagged webcam video to confirm suspicious cheating behaviors. Finally, the list of suspicious cheating behaviors is sent to the FRC to make the final punishing decision. Our experiment show that ACD and PCD can detect usage of a cheat sheet with good accuracy and can reduce the overall human resources required to monitor MOOCs for cheating.

Author Keywords
Education; MOOC; Cheating Detection; Crowdsourcing; Machine Learning; Human Computer Interaction

ACM Classification Keywords
K.3.1 Computers and Education: Computer Uses in Education; H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

General Terms
Human Factors; Design; Experimentation

INTRODUCTION
Massive Open Online Courses (MOOCs) enable people from all over the world to receive high-quality education.

CSCW '15, March 14 - 18 2015, Vancouver, BC, Canada

http://dx.doi.org/10.1145/2675133.2675245

However, current MOOC creators cannot provide an effective, economical, and scalable method to detect cheating on tests, which would be required for any certification. Although MOOC providers have not officially disclosed detailed statistical data of dishonest behaviors in MOOC, lots of cheating behaviors have been reported by both students and instructors[19][7]. Researchers have clearly shown that unproctored online testing is vulnerable to cheating behavior[14][13][12].

We introduce the Massive Open Online Proctor (MOOP) framework to proctor online exams of MOOC which tackles the test proctoring scalability problem through sensor-detection and crowd-sourcing. MOOP consists of 3 components:

- *Automatic Cheating Detector (ACD)* - an automatic, machine learning approach that utilizes different sensor technologies to detect cheating behaviors.
- *Peer Cheating Detector (PCD)* - a collaborative, crowd-sourcing approach to review ambiguous cases detected by ACD.
- *Final Review Committee (FRC)* - an authoritative committee that reviews all suspicious cheating cases detected by ACD and PCD, and make the final punishing decision.

In the following sections, we first review existing cheating detection methods. Then, we describe each MOOP component in detail. To evaluate our system, we ran a simple experiment where students are encouraged to cheat with a cheat sheet. Finally, we discuss both advantages and limitations of our proposed MOOP and suggest future directions.

EXISTING CHEATING DETECTION METHODS
The traditional solution to preventing cheating is on-site proctoring. Many standardized exams, such as SAT, GRE, etc., have adopted this solution. Students are asked take a proctored test in a specific test center and at a specific time. Since the test is administered by test center staff, the test result is credible. Nonetheless, on-site proctoring is unsuitable for large scale MOOC testing due to expense and inconveience. The cost of renting the test venue and hiring proctors translates into a test fee that is passed on to the students. Furthermore, students who are far from established test centers will need to travel great distances to take tests.

Online proctoring (a.k.a. remote proctoring) uses human proctors to monitor online exams remotely[5]. For example, ProctorU[1] enables students to take the exam remotely by assigning a human proctor to monitor the exams through a webcam. Although online proctoring increases the availability of cheating detection, scalability and cost are still issues. Because it relies largely on human judges, cost can still be relatively high. The price for a one-hour exam is $10 to $20[2], which is still too expensive for students from poor areas.

The common limitation of both on-site proctoring and online proctoring is scalability since every student action must be monitored by paid human proctors. With millions of student using MOOCs[11], hiring human proctors to detect cheating in every MOOC exam is implausible for both students and MOOC providers.

MOOP is designed to address the scability problem. By combining automated cheating detection through sensors/machine learning and a crowd-sourced verification, the number of cheating cases that must be reviewed by the central committee of hired proctors can be drastically reduced. A similar mechanism is adopted in the game industry. In *League of Legends*, the Tribunal system receives cases of possible misconduct through automatic detectors (e.g., "away from keyboard") and player reports (e.g. player harrassment, negative attitude). Players are then invited to peer-review these mal-behaviors and vote in a court-like system, before the final punishment decision is decided by the game developer. The Tribunal system has been proven effective in regulating conduct in [3].

MASSIVE OPEN ONLINE PROCTOR

MOOP has three parts: ACD, PCD and FRC (see Figure 1).

Automatic Cheating Detector (ACD)

During the exam, students are monitored by sensors (e.g. webcam). The recorded data are then subject to automated machine learning models to flag whether students are likely to be cheating. Since ACD works completely automatically, the human resource cost is nearly zero. However, there may be an overhead cost to purchase or rent sensors if students do not already own them. In this paper, we explore two versions of ACD - a webcam-only version that seeks to reduce overhead cost and a multi-modal version which incorporates additional sensors (gaze tracker and EEG) to improve detection performance.

For the webcam-only version of ACD, we position one webcam in front of the student to capture his face, another webcam right-hand side of the student to capture his profile. Since the webcams are becoming ubiquitous and cheap, the cost of sensor will be low (or free) for many students. Students, even in poor areas, can easily set up the proctoring environment by using the low-price webcams or cameras on cellphones or laptops.

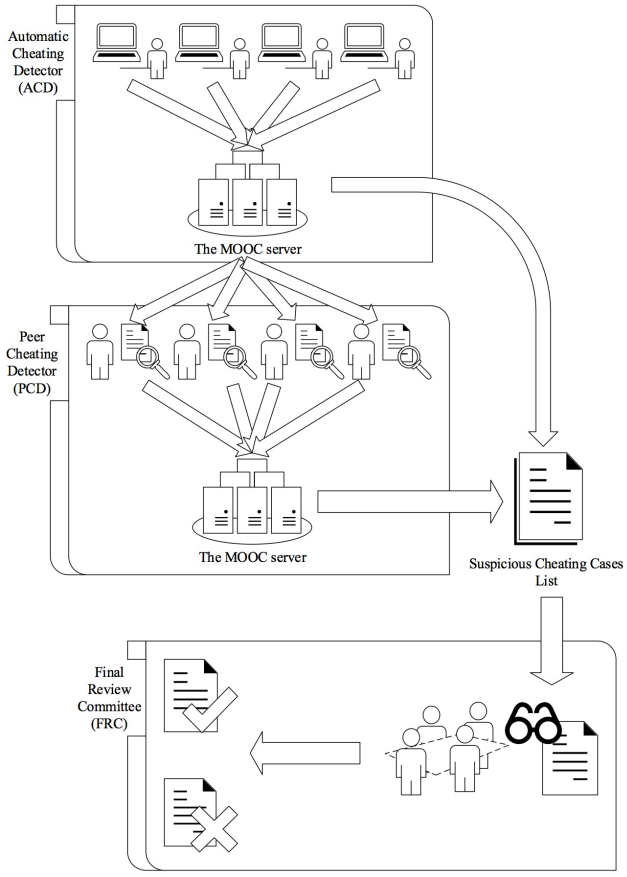

Figure 1: The MOOP framework

The multi-modal ACD is based on three different kinds of sensors: webcam, EEG sensor, and gaze tracker. Though EEG sensors and gaze trackers are more expensive and less available than webcams, they may be able to detect cheating in cases where webcams are insufficient. Cheating takes many forms[6] and different sensors offer complementary ability to detect different types of cheating. For example, A student who is using a mobile phone to ask a friend for answers might frequently take their gaze off the question on screen. This behavior may be captured by webcams but may be more easily captured with a gaze tracker. But a student who's been told the answers before the exam would not show the same scattered gaze, making it difficult if not impossible to detect his cheating behavior visually. In that case, EEG sensors might detect that his mental workload is too low for the problem he is supposedly working on.

In this paper, we test the ACD on one specific cheating behavior - usage of a cheat sheet. In practice, to cope with the many different ways of cheating, different ACDs should be trained to cover the wide variety of possible cheating behavior MOOCs may face. Figure 2 shows our experimental setup for evaluating ACD.

Figure 2: A user is taking an online test, proctored by multi-modal ACD. The EEG sensor mark is by blue circles, the gaze tracker is marked by green circles, and the two Kinect cameras are marked by red circles. In the right-hand picture, the user is cheating by referring to a cheat sheet.

- *Camera.* Two Microsoft Kinects are used as webcams. One Kinect is mounted above the computer monitor, monitoring the face. One Kinect is placed to the right hand side of the subject, monitoring the profile.

  Action recognition using the camera has been actively researched in past years[4]. Among the various method for action recognition, we adopt the trajectory-based approach[8][9], specifically the Dense Trajectory approach[18][17][16], which is the state-of-the-art method. We follow the typical steps of action recognition which can be roughly divided into four steps. a) extracting Dense Trajectory features (raw features) from each video or shot, b) quantizing and encoding raw features using a pre-trained k-means codebook to get the encoded feature vector for each video or shot, c) training a χ^2 kernel SVM using these encoded feature vectors and d) applying the model to test data to produce a prediction result.

- *Gaze tracker.* Gaze trackers use the relative position of the pupils and viewing area to compute the precise gaze point where someone is looking. Calibrating a gaze tracker typically requires the viewer to look at a sequence of known reference points on a computer monitor. The gaze tracker then reports subsequent gaze points in terms of computer display coordinates to translate into meaningful targets such as words in a text, items on a menu, or icons on a screen.

  Recently, portable eye-trackers have become available at a much lower cost (around $5000) than the high-end eye-trackers used in high-precision laboratory studies. In our experiment, we used the Mirametrix S2 eye-tracker because of its portability (35 cm width, 4 cm height, 3 cm depth, and 0.7 lb weight), and simple set-up (the eye-tracker simply sits below the computer screen).

  In our detector, only eyes position information from the gaze tracker is used. We trained binary K-Nearest Neighbor (KNN) classifiers to distinguish gaze data from cheating versus non-cheating. The classifier parameter K (number of nearest neighbor to consider) of KNN is set to be 8. As features for classifier we used features: average x, average y, variance of x, and variance of y, where x and y are coordinates of gazes on the screen.

- *EEG sensor.* The EEG (electroencephalogram) signal is a voltage signal that can be measured on the surface of the scalp, arising from large areas of coordinated neural activity. This neural activity varies as a function of development, mental state, and cognitive activity, and the EEG signal can measurably detect such variation.

  The recent availability of relatively inexpensive (less than $300), portable EEG monitoring devices suddenly makes it feasible to take this technology from the lab into the real world. The NeuroSky MindSet, for example, is an audio headset equipped with a single-channel EEG sensor which could be used to distinguish two fairly similar mental states (neutral and attentive) with 86% accuracy. In out experiment, the EEG sensor is placed on the forehead of the subject, measuring frontal brain activity.

  In our detector, we trained binary Gaussian Nave Bayes classifiers to distinguish EEG data from cheating vs. non-cheating. The dependent variable estimated by each classifier was the probability that a subject cheated. As features for classifier we used all available measures: NeuroSkys attention and meditation, the raw and filtered signals, and power spectrum across different frequency bands. We averaged each measure over the time interval of each stimulus, excluding times where the Mindset reported poor signal quality.

In the MOOP system, independent cheating detectors are trained for each sensor. Each sensor flags a *cheating alarm* for possible cheating behavior - the more number of alarms is flagged for a case, the more likely it is that the student has cheated. The cases which with zero or small number of cheating alarms will be dropped as non-cheating and won't go to PCD and FRC. The cases with a large number of cheating alarms are placed in the *suspicious cheating list* and reviewed by FRC directly. The cases with a moderate number of cheating alarms are placed in the *ambiguious cheating list*, which are first sent to PCD for crowdsourced peer review before it is dropped as non-cheating or sent to FRC for final review.

There are two types of detection errors from ACD - false positive (a flagged non-cheating case) and false negative (a non-flagged cheating case). On one hand, a large ratio of false positive increases the workload (and subsequently human resource cost) for PCD and FRC. On

the other hand, a large ratio of false negative means that lots of cheating cases are missed. The threshold for what to send to FRC, what to send to PCD, and what to drop can be tuned depending on the resources of the MOOC provider, the size of the class, and the quality of ACD sensors used. We discuss the thresholds we use in our experiment and the limitations of higher or lower thresholds in later sections.

Peer Cheating Detector (PCD)

PCD crowdsources the preliminary verification of cheating alarms raised by the ACD by asking students to review video of their peers. After an exam, randomly selected students are shown various video clips which have been flagged by ACD. Students are asked to score each clip with a likelihood of cheating based on the criteria shown in Table 1. In addition, some "ground truth" video clips will be used to rate students' ability to peer-review. Students who were incorrect on the ground truth clips either due to malignance, apathy, or inability will be not be considered when tallying final PCD scores. All cases with an average PCD score above an adjustable threshold (3 in our experiment) would be considered as *suspicious cheating cases* and would be sent to the FRC for the final decision.

Notice that the instruction to "flag unnessary" behaviors prompts the peers to ground their ratings on evidence of cheating. These instructions are intended to give peers a "innocent-until-proven-guilty" mentality, in which they prioritize the reduction the false alarm rates over the false negative rate. By reducing the false alarm rates, the PCD which would further reduce the workload of FRC.

Score	Criterion
1	Strong non-cheating: based on the videos, you cannot find any unnecessary action by the examinee. You have no reason to believe the examinee is cheating
2	Weak non-cheating: based on the videos, although you can find some unnecessary action of the examinee, you do not believe those actions can contribute to cheating.
3	Border line: based on the videos, you saw some unnecessary action of the examinee and you are not sure whether that was related to cheating behavior.
4	Weak cheating: based on the videos, although you cannot find any clear visual proof of cheating, the examinee is taking some unnecessary actions which you believe are related to cheating behavior.
5	Strong cheating: based on the videos, you find clear visual proof of cheating.

Table 1: Score criteria for PCD

Final Review Committee(FRC)

FRC is responsible for making the final punishing decisions for suspicious cheating cases detected by ACD and PCD. It is a committee formed by staff from MOOCs provider, faculty members, and teaching assistants. Suspicious cheat cases need to be reviewed by an FRC member. We expect FRC to be comparable to online proctoring in accuracy while taking far less time and human resources than online proctoring because the ACD and PCD steps can drastically reduce the number of cases that FRC has to see.

To achieve this, we must be wary of two metrics: First, the precision of the suspicious cheating cases list generated by ACD and PCD should be high, so that FRC does not waste time judging lots of non-cheating cases; this reduction in wasted time is what gives FRC its advantage over online proctoring. Second, the recall rate of ACD and PCD should also be high so that virtually all controversial cheating cases are reviewed by the FRC; a high enough recall rate will make FRC comparable in accuracy to online proctoring.

EXPERIMENT

To validate our assumptions about ACD and PCD accuracy, we ran a controlled experiment where we induced cheating with a cheat sheet.

Experimental Setup

Exam Question Generation

We utilize an computerized test with 20 multiple choice questions. The test is implemented on the Reading Tutor platform[10], which logs and syncronizes the user's camera, gaze, EEG data with the exam questions it presents. Among the 20 questions, 10 of them are tricky questions requiring esoteric knowledge which we believe most participants will lack. The remaining 10 questions are relatively simple questions about computer science. Since most of our participants are members of the computer science department, they should not need to refer to the cheat sheet for most of the questions. Here are some examples of tricky and simple questions:

> James Bond was searching for what kind of airplane in 'Thunderball'?
> a) 707 b) Piper Cub c) Vulkan d) B-52

The answer is c) Vulkan.

> The complexity of the merge sort algorithm is
> a) $O(n)$ b) $O(\log n)$ c) $O(n^2)$ d) $O(n \log n)$

The answer is d) $O(n \log n)$.

Defining The Cheating Behavior

There are at least 53 defined forms of cheating[6] and new forms may be invented in a MOOC setting. It is therefore infeasible to include all possible cheating behavior in one experiment. In this experiment, we focus

on one of the more straightforward cheating methods: using a cheat sheet. However, MOOP is designed to be a general framework to detect cheating so it can theoretically be adapted to detect other cheating methods by training new ACDs.

We provide the participants with a cheat sheet - an A4-size paper on which about 50 questions are printed along with the answers. The questions during the test are shown in random order, so the ordering of questions on the cheat sheet will be different from the test. The cheat sheet contains 30 extra questions which do not appear on the test, this is intended to simulate the searching behavior students may engage in if they were really using a cheat sheet.

Experiment Process

Our subjects are students and researchers with computer-science backgrounds. We instructed each participant to take the test to the best of their ability. The participants are only permitted to cheat with the provided cheat sheet. During the exam, the participants' gaze data, video, and EEG signals are recorded by our sensors. After the exam, the participant fills out a survey indicating which question or questions he or she cheated on. We take the survey results as ground truth.

All participants completed all 20 questions of the test. Data from 9 participants were collected successfully and analyzed to make for 180 problems in total. As Table 2 shows, students cheated on 84 out of 180 questions and answered 135 out of 180 questions correctly.

Subject	Cheating	Correct
1	11	17
2	9	20
3	8	7
4	13	10
5	8	19
6	9	16
7	2	20
8	12	10
9	12	16
All	84 (46.7%)	135 (75.0%)

Table 2: Data set collected

RESULTS AND ANALYSES

Metrics

We performed leave-one-out between-subject evaluation. For each subject, we take all other subjects' data as the training set and apply the learned machine learning model on the left-out subject's data. The major metrics used in our experiment are the recall, precision, and accuracy. Recall is the ratio of the number of true cheating cases in detected cases to the total cheating cases number in the dataset. A high recall rate means most of cheating cases in the data set are detected. Precision is the ratio of the number of true cheating cases to the number of detected cases. A high precision means most of cases, which are detected, as cheating cases are true. Accuracy is the ratio of the number of correct classification to the number of all cases. A high accuracy means most of classification results are correct.

Performance

Webcam-only ACD

Here we analyze the MOOP with the ACD purely based on webcams. Figure 3 shows that the two webcams provide similar performance. The best recall, precision and accuracy are 76.2%, 83.3% and 86.6% respectively.

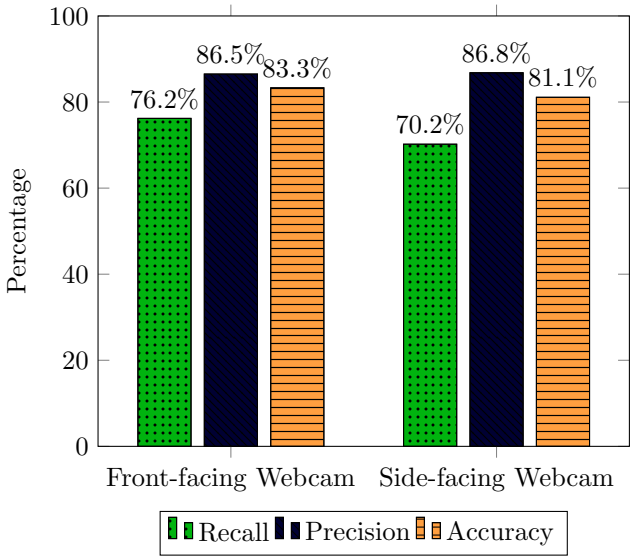

Figure 3: Recall and Precision of Webcams

We consider the number of *cheating alarms* as the indicator of whether the cases should be sent to FRC, PCD, or labled as non-cheating case directly. Figure 4 shows the relationship between the number of *cheating alarms* and the cheating cases ratio we used in our experiment. Naturally, the cases with more cheating alarms are more likely to be cheating cases. Using 1 as the threshold (i.e., cases with 1 or more cheating alarms are flagged as cheating), the recall, precision, and accuracy of the webcam-only ACD are 78.6%, 84.6% and 83.3%, respectively.

PCD

Our initial PCD consists of 2 members, who watched recorded video of the test-takers and rated whether the test-taker was cheating based on the criteria in Table 1. The front-facing video and the side-facing video are provided. Note that in order to access the PCD, we asked them to rate all 180 cases, as opposed to only those that are flagged by the ACD. As seen in Figure 5, the PCD gives clear separation between cheating and non-cheating

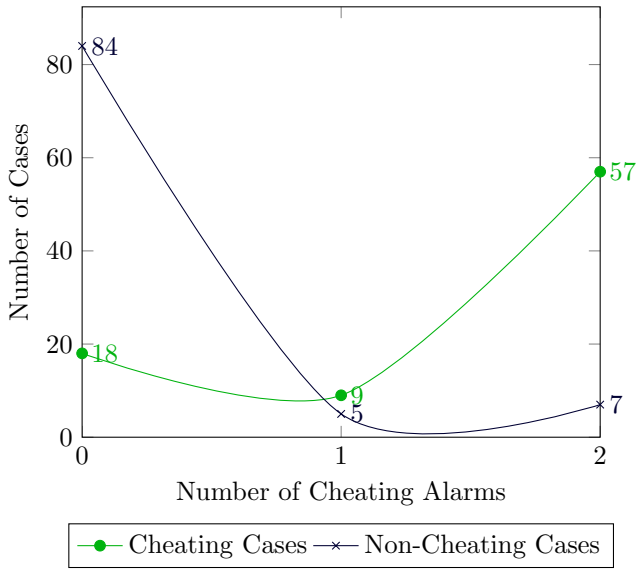

Figure 4: Number of cases with different number of cheating alarms from webcams ACD

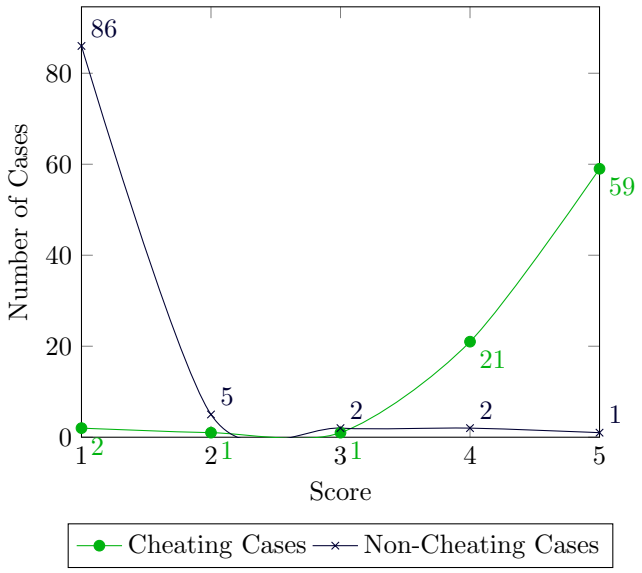

Figure 5: Human score distribution

cases. Using a score of 3 as the threshold (i.e., cases with 3 or more cheating alarms are flagged as cheating), the recall, precision, and accuracy of the PCD is 96.4%, 94.1% and 95.6%, respectively. Note that PCD is more accurate than ACD, but comes at the cost of human resource.

Combination of ACD and PCD

Due to the limited time and patience of crowdsourced labor, we only send cases to PCD which were determined to be ambiguous by ACD.

In ACD, each webcam may raise cheating alarm. According to the number of cheating alarms, we can divide the dataset into 3 subsets - cases with zero cheating alarm (likely non-cheating), one cheating alarm (ambiguous on cheating), and two cheating alarms (likely cheating). There are 102 zero-alarm cases, 14 one-alarm cases and 64 two-alarm cases. Note that the ambiguous case is the smallest subset.

Figure 6 shows a comparison of accuracy of ACD and PCD. PCD performance is good across all subsets but ACD performance varies dramatically, in particular in the one alarm ambiguous cases. By sending the ambiguous cases to PCD for verification, we can take advantage of the high accuracy in PCD while not overwhelming the students with many cases to label.

Performance of the Whole System

To summarize, in ACD, the cases that received 0 alarms are dropped by the ACD, whereas cases that received 1 alarm are verified by PCD before being dropped or sent to FRC. In PCD, cases that received a human rating of 2 or less are dropped by the PCD, whereas cases that received 3 or more are sent to FRC for the final punishing decision. Since FRC is comparable to online proctoring, we do not evaluate its performance and consider it a perfect cheating detector. Considering the whole MOOP system as a cheating detector, the recall, precision and accuracy of the whole system is 78.6%, 100% and 90.0%.

As for the human resource cost, considering there are 84 cheating cases in the dataset, traditional proctoring or online proctoring would need to manually check all of 180 cases to detect the cheating cases. In MOOP, the number of cases that needs to be manually checked in PCD and FRC are 14 and 78, respectively. Moreover, the workload of FRC can be further reduced to 14 by directly labeling the 64 two-alarm cases in ACD as the cheating cases and not sending them to FRC. In this case, the recall, precision and accuracy of the whole system is 78.6%, 90.4% and 86.1% respectively, while FRC only need to review 16.7% (14/84) of total cases. In theory, if there are 1% cheating cases as opposed to the 50% cheating cases in our experimental dataset, only 1.67‰ of total cases will be reviewed by FRC.

Multi-modal ACD and Multi-modal MOOP

We see that MOOP performs well in terms of accuracy and precision but its performance is relatively low in recall. We hypothesize that the low recall performance can be improved by including other sensor modalities. In this section, we show the experiment results of a multi-modal ACD system that utilize webcams, EEG sensor and gaze tracker.

Using the same evaluation method, we get the performance of each sensor. As can be seen in Figure 7, webcam is our best performing sensor. We believe this is at least partly due to the immaturity of EEG and gaze tracking technologies and their susceptibility to noise. For instance, EEG sensors may produce noisy data if the user performs lots of head movements. The gaze

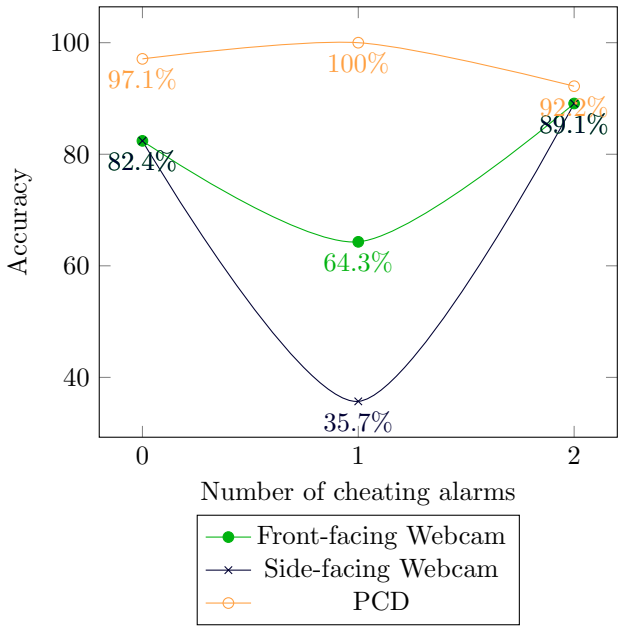

Figure 6: Comparison of PCD and ACD

tracker cannot track the users well if they are wearing glasses. We also note that automatic video processing is a more mature field than EEG and gaze processing and we used more sophisticated algorithms in processing webcam data than in EEG and gaze data.

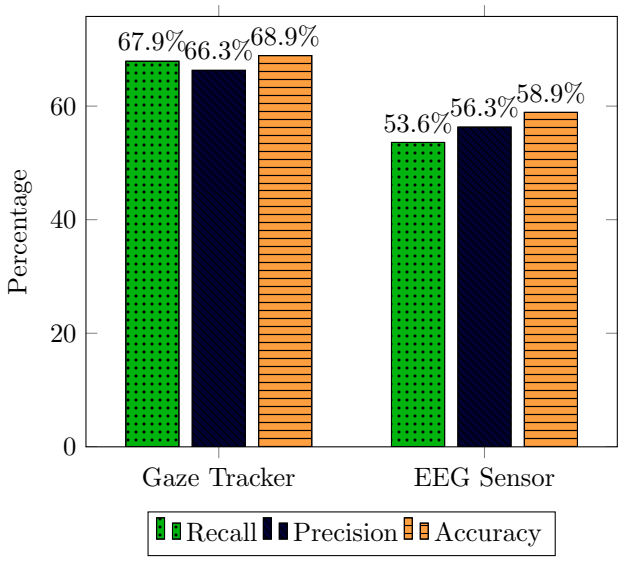

Figure 7: Performance of Gaze Tracker and EEG Sensor

Combined multi-modal ACD output can be split into 5 subsets according to the number of cheating alarms (i.e., each subsets with 0 to 4 cheating alarms). Figure 8 presents the distribution of cheating and non-cheating cases on those 5 subsets. Using a score of 2 as the threshold (i.e., cases with 2 or more cheating alarms are flagged

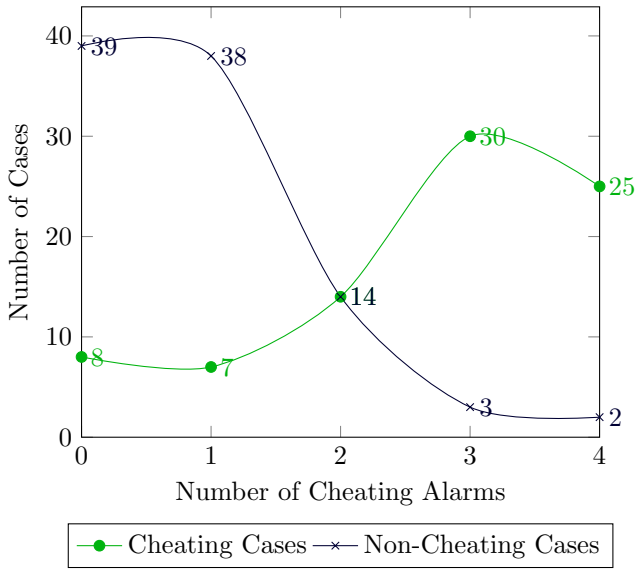

Figure 8: Number of cases with different number of cheating alarms from multi-modal ACD

as cheating), the recall, precision, and accuracy of multi-modal ACD are 82.1%, 60.0% and 81.1%, respectively. Although the recall is improved over the webcam ACD, the precision and accuracy decrease. This may be due to the added noise of gaze tracker and EEG sensor.

Combining the multi-modal ACD with PCD and FRC gives the Multi-modal MOOP. From the multi-modal ACD, we take the cases with 1 or 2 cheating alarms into the PCD. Those cases with 3 or 4 cheating alarms are directly judeged as cheating and those cases without cheating alarms are directly judged as non-cheating. Then the recall, precision and accuracy of the mutli-modal MOOP are 90.5%, 93.8%, and 92.7% respectively. It improves the performance of webcam-only MOOP, with the cost of more expensive detectors. Table 3 presents a summary of performance.

Cheating Classifier	Recall	Precision	Accuracy
Webcam-only ACD	76.6%	84.6%	83.3%
Multi-modal ACD	82.1%	60.0%	81.1%
PCD	96.4%	94.1%	95.6%
Webcam-only MOOP	78.6%	90.4%	86.1%
Multi-modal MOOP	90.5%	93.8%	92.7%

Table 3: Summary of Cheating Classifier

DISCUSSION

Automated proctoring in MOOCs is a relatively new problem and the MOOP system we present in this paper is an initial attempt with much room for improvement.

Challenges of ACD

Diversity of Cheating Behaviors

In our experiment, we trained the ACD on just one kind of cheating behavior. There are many other ways to cheat[6] and many more may be invented in the future. Thus, we consider the ACD to be an evolving component of the MOOP system and that combining complementary data streams from multiple sensors is critical to its success. In addition to the camera, gaze-tracker and EEG sensors that are already deployed, other sensors like galvanic skin response, keystroke patterns, microphone, etc. may also provide useful information.

Cost of sensors

Though additional sensors hold promise for better performance, we must beware that increasing the number and sophistication of sensors increases the cost of the ACD system.

For now, it is not realistic for MOOC students to buy an EEG sensor and gaze tracker for exams. Initially, MOOC providers can buy sensors in batch and rent them to students. As sensors become more ubiquitous in the future, it may become more feasible for students to provide their own. Webcams would not have been considered ubiquitous 10 years ago; we remain optimistic that sensor technologies like portable EEG and gaze tracking will one day become more available. In the meantime, work must be done to improve webcam-only recognition rates for various cheating behaviors.

Diversity of Test Subjects

In our experiment, we trained the ACD on a small participant pool. Future work remains to test if the ACD is generalizable across the diverse learning communities in MOOCs. More importantly, cultural and interpersonal differences in behavior, movement pattern, and brain activity may affect the ACD. For example, culutural differences in expressing attention may trigger the camera or gaze alarms; students with ADHD may trigger EEG alarms. It is not appropriate to subject non-majority cultural or non-neurotypical students to additional scrunity, especially when such status make them more prone to unfair discrimination. The ACD will have to be tuned for different participant pools.

Challenges of PCD

Quality Control

PCD is a typical crowdsourcing system. In the initial system, we suggested relying on moral incentives for students' participation in the crowd-sourcing effort. There have been a lot of previous research about how to improve the work quality of crowdsoucing tasks by using designed incentive protocols.[20] presents a reputation-based incentive proctols to optimize the work quality.[15] compares different kind of incentives, such as financial incentive, social incentive, etc., that will infere with the work quality. They are all applicable for PCD.

Another issue is avoiding the bad peer proctor who makes random decisions. Our current proposed solution is adding some known cases as the validate set into each PCD task. If the student gives wrong detection for the validate set, the weight of the student's detection is lowered. Another solution can be assigning the weight to students' detection based on their previous record. For example, a student caught cheating before should not be assigned to PCD.

Privacy

Since MOOP records the video of students in the privacy of their own home, it may record sensitive information and in some cases, that information may be sent to PCD. The first line of defense is to ask students to try to keep their video clear of private information. Students who do not allow their peers to watch the recorded video may be charged the additional fee for using the FRC instead of PCD. But the fee should still be lower than traditional online proctoring since a lot of non-cheating cases are filtered out by the ACD.

CONCLUSION

In this paper, we propose a possible solution to the MOOC proctoring problem. MOOP combines machine learning with crowdsourcing to deal with the scalability issue of MOOC proctoring.

We design and conduct an experiment to validate the MOOP based on webcams. The full MOOP system achieves 78.6% in recall, 90.4% in precision and 86.1% in accuracy. MOOP uses sensor hardware that is easy to get (e.g. webcam) and the human resource cost is greatly reduced compared to online proctoring.

We also test a multi-modal ACD which combines webcams, gaze tracker and EEG sensor. The combined system improved recall rate but lowered precision and accuracy and increased the sensor cost.

We demonstrate that MOOP has potential to be a more scalable solution to MOOC proctoring. We acknowledge that the question of MOOC proctoring is a relatively new research problem with many open questions and we believe MOOP is a promising first step in answering these questions.

ACKNOWLEDGEMENT

We thank Joseph Valeri and Anders Weinstein for helping set up the experiment. We are grateful to the anonymous reviewers for their thoughtful suggestions. We also thank the volunteers in the experiment for their participation and feedback.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-12511827 and Grant No. IIS-1124240. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. ProctorU. http://www.proctoru.com/.
2. ProctorU: Overview and Technology Requirements. http://www.ao.uiuc.edu/support/source/student_services/proctoru_tech.html.
3. THE TRIBUNAL. http://na.leagueoflegends.com/tribunal/.
4. Aggarwal, J., and Ryoo, M. S. Human activity analysis: A review. *ACM Computing Surveys (CSUR) 43*, 3 (2011), 16.
5. Case, R., and Cabalka, P. Remote proctoring: Results of a pilot program at western governors university. *Proceedings of the 25th Annual Conference on Distance Teaching and Learning 10* (2009), 2010.
6. Dick, M., Sheard, J., Bareiss, C., Carter, J., Joyce, D., Harding, T., and Laxer, C. Addressing student cheating: Definitions and solutions. In *Working Group Reports from ITiCSE on Innovation and Technology in Computer Science Education*, ITiCSE-WGR '02, ACM (New York, NY, USA, 2002), 172–184.
7. Eisenberg, A. Keeping an Eye on Online Test-Takers. The New York Times, 2013.
8. Jiang, Y.-G., Dai, Q., Xue, X., Liu, W., and Ngo, C.-W. Trajectory-based modeling of human actions with motion reference points. In *Computer Vision–ECCV 2012*. Springer, 2012, 425–438.
9. Matikainen, P., Hebert, M., and Sukthankar, R. Trajectons: Action recognition through the motion analysis of tracked features. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, IEEE (2009), 514–521.
10. Mostow, J., Chang, K.-M., and Nelson, J. Toward exploiting eeg input in a reading tutor. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, AIED'11, Springer-Verlag (Berlin, Heidelberg, 2011), 230–237.
11. Pappano, L. The Year of the MOOC. The New York Times, 2012.
12. Prince, D. J., Fulton, R. A., and Garsombke, T. W. Comparisons of proctored versus non-proctored testing strategies in graduate distance education curriculum. *Journal of College Teaching & Learning 6*, 7 (2009).
13. Richardson, R., and North, M. Strengthening the trust in online courses: A common sense approach. *J. Comput. Sci. Coll. 28*, 5 (May 2013), 266–272.
14. Rogers, C. F. Faculty perceptions about e-cheating during online testing. *J. Comput. Sci. Coll. 22*, 2 (Dec. 2006), 206–212.
15. Shaw, A. D., Horton, J. J., and Chen, D. L. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, ACM (New York, NY, USA, 2011), 275–284.
16. Wang, H., Klaser, A., Schmid, C., and Liu, C.-L. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE (2011), 3169–3176.
17. Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision 103*, 1 (2013), 60–79.
18. Wang, H., Schmid, C., et al. Action recognition with improved trajectories. In *International Conference on Computer Vision* (2013).
19. Young, J. R. Dozens of Plagiarism Incidents Are Reported in Coursera's Free Online Courses. The Chronicle of Higher Education, 2012.
20. Zhang, Y., and van der Schaar, M. Reputation-based incentive protocols in crowdsourcing applications. In *INFOCOM, 2012 Proceedings IEEE* (March 2012), 2140–2148.