Journal of Engineering Science and Technology
Vol. 15, No. 6 (2020) 4032 - 4047
© School of Engineering, Taylor's University

CHEATING ACTIVITY DETECTION ON SECURE ONLINE MOBILE EXAM

ARIEF SETYANTO*, BAYU SETIAJI, MARDHIYA HAYATY, KRISNAWATI

Faculty of Computer Science, Universitas Amikom Yogyakarta
Ring Road Utara, Depok-Sleman, Yogyakarta, Indonesia
*Corresponding Author: arief_s@amikom.ac.id

Abstract

The online exam is one of the important tasks in online learning systems. Online exam proctoring, therefore, is important to ensure the credibility of the exam. The online exam used to ascertain students' knowledge on a given topic, irrespective of their locations. This research proposes an online exam system developed using an android application, with a standard camera and audio recorder installed to capture human activity during the test. Audio-video data were obtained from a total of 20 students, which recorded seven cheating and seven non-cheating activities. The method used in this research, is a system used captured students' faces using the front camera, while the audio recorded the sound. A mid-level representation of the audio-video was conducted before the classification task with data normalization performed into uniform units for each parameter between zeros to ten. Finally, the multilayer perceptron (MLP) carried out the classification of a midlevel signal into cheating and non-cheating activity. The result showed a classification accuracy, precision, recalls, and F1 score of 91.73%, 91.73%, 92.9 % and 91.68%, respectively.

Keywords: Action detection, Eye detection, Face detection, Mobile exam, Multilayer perceptron.

1. Introduction

Electronic learning is currently one of the game-changing applications in the education sector due to its flexibility and convenience. The commonly used traditional way of acquiring knowledge, in classrooms has gradually changed into synchronous/ asynchronous online learning system. This new technique has been enabled by online video meeting applications, facilitated by power point presentations, animation, and the use of more interactive programming tools. According to a survey conducted by [1], 14.3% of American students take online courses without a traditional degree, while 15 % take a combination of both. However, there was a significant improvement in the 2013 survey [2], with an increase from 411,000 to 7.1 million students, while the online education provider improved from 2.6% to 5% over the past years. Although the survey was carried out in the USA, it shows the general trend all over the world.

Examination plays an important role in ensuring students' have understood the learning material. It is either synchronous or asynchronous. Synchronous examination is defined as the process whereby the exam occurs on set schedules and timeframes through online channels. Conversely, the asynchronous online exam is carried out by an interactive questionnaire filled by the participants. It permits the occurrence of the process at any time and place. However, online real time proctoring is not possible in this mode. Therefore, an automatic online synchronous examination system is needed to ensure the test is properly conducted.

Several studies have reported that online exam has a higher possibility of malpractice [3-5]. According to King et al. [4] and King and Case [5], 74% of students stated that it is easy to carry out exam malpractice in online examination, while 29% stated that they participated in the act. According to Corrigan-Gibbs et al. [6], the provision of an honor code tends to provide an insignificant effect of reduction in online malpractice, while warning aids in its reduction.

Therefore, online proctoring is important due to the growth of massive open online courses (MOOC). Some of the online/offline proctoring products available in the market are Kryterion, Web assessor and ProctorU [7]. These applications usually put a surveillance camera on the user site, however, they still rely on humans watching over through the video streaming. Kaiiali et al. [8] proposed a mobile exam system to be embedded in the Moodle system to identify various vulnerabilities capable of violating exam security in m-learning environments and to design the appropriate services and counter measures to ensure exam security.

Bawarith et al. [9] utilised a front face camera to track the participant eye movement. The exploitation of eye movement leads is associated with users' focus, and when it leaves the testing sheet, it is assumed that malpractice has occurred. The multimedia analysis helps online proctoring automation with the Gaze tracker, webcam and EEG installed to assist in the process [10]. Atoum et al. [11] installed a webcam and camera on specially designed eyeglasses in order to record the user's face while capturing the participant's view. However, although the design functions properly, a specially installed camera is needed to reduce the practicality.

This, therefore, led to the development of the audio visual tool for activity/event recognition, social interaction, object tracking and sports analysis. This tool was also designed for social cohesion among group members extracted from video and classified by support vector machine [12]. It is specially designed to identify the good behaviour of job seekers with interview videos as proposed by Nguyen et al. [13]. The process of identifying unwanted behaviour in audio visual data has also been conducted by Zajdel et al. [14] and Lefter et al. [15]. In addition, sport activities such as attack, defence, and other numerous reports associated with matches has been proposed by Wang et al. [16], while goals, saves and kick-of activities are recognized using Bayesian network [17].

Various research has also been conducted to detect cheating behaviour on online exams, however, human supervision still plays a vital role in this process, to achieve detailed surveillance. Moreover, in some research such as in [10, 11], special devices such as additional camera, electroencephalogram (EEG) and gaze tracker need to be installed. This research aims to simplify the online exam proctoring task with standard devices only for easier implementation. This research investigates minimum set of data acquired by the standard camera and audio recorder of low-end smartphone to recognize pre-defined cheating activities using artificial neural network (ANN). The developed online test system was designed on a smart phone with a front camera, and voice recorder. These specifications are standard available currently available on most smart phones to provide multimodal data captured during the exam. The camera and the voice recorder are used for test surveillance.

2.Related Work

There are three attributes associated with online proctored exam model. Firstly, it is conducted without human controls [18-20]. Secondly, humans are involved either online through a webcam or physically in a certain location [21]. Thirdly participant are allowed to take the test without supervision, with the exam process recorded by the multimedia analysis system [11, 22].

Online examinations need to ascertain user identity, and there are many efforts used to ensure computer user identity such as keystroke behaviour or biometrics. Kumar and Rathi [23] used keystroke dynamics for online exam authentication, while in [24], iris recognition was used for biometrics authentication. Furthermore, face recognition was periodically used during the exam to ensure the authentication of the participant [25].

Besides participant identity, user action during an exam is an important clue for identifying cheating behaviour. There are many researches dealing with action recognition such as [26-28]. Head movement is one important clue for activity recognition [26]. Detection of human activity is conducted by analysing the socio-communicative and affective behavioural characteristics of interacting partners using their head motion. There are many techniques for head pose estimation as reported in [27-30]. Liu et al. [27] used the multilevel structured hybrid forest to detect the head and estimate the pose. Dornaika and Raducanu [28] proposed a 3D face pose estimation using 2D Eigen faces and DE algorithm. Qing et al. [29] reported the use of Mutual information (MI) exploited to deal with pose estimation in an uncontrollable environment. Another technique proposed by Drouard et al. [30] utilized a mixture of linear regression with partially-latent output.

Eye-tracking plays an important role in understanding human attention. It is used to point out a user's point of view in accordance with its direction. Currently, most eye movement technology are developed using near-infrared light, which lights up the human pupil. This technology is however expensive and lacks practicability due to the need for a special device. In many medical purposes, this technology is viable due to the accuracy requirement and facility availability.

However, this research was not designed to provide a special camera with near-infrared sources, therefore, the simple technique of the web camera was used to detect and track the pupil. Cheung and Peng [31] and Krafka et al. [32] proposed the installation of a webcam on the desktop to recognize the pupil, eye corner and movements. Krafka et al. [32] proposed a normal camera for eye movement recognition. Subsequently, Meng and Zhao proposed CNN based eye movement detection which was tested in a half-million frames and successfully recognized activities such as reading, browsing and watching video [33].

Sound has been identified as a strong clue for activity recognition [11] and cheating behaviour detection in the presence of a human voice. Therefore, speech extraction is performed in modelling the cheating action and has attracted as the attention of [34, 35].

This research made the following contribution:

- It developed a mobile exam video dataset consisting of cheating and non-cheating activities from twenty subjects, 280 actions and a total of 36.120 frames.
- It defined a midlevel data modeling for audio-video cheating dataset which consisted of head tracking, eye tracking and audio.
- Developed an artificial neural network to perform cheating activity detection based on the mid-level dataset.

The proposed method of the paper is explained in section 3, while section 4 described the results and discussed the achievement with conclusion drawn in the last section.

3. Multimodal Action Detection

3.1. Research framework

The research is divided into four main tasks which are, defining the cheating actions, developing a dataset for pre-defined actions, mid-level feature extraction, and action recognition task. Figure 1 depicts the entire research framework.

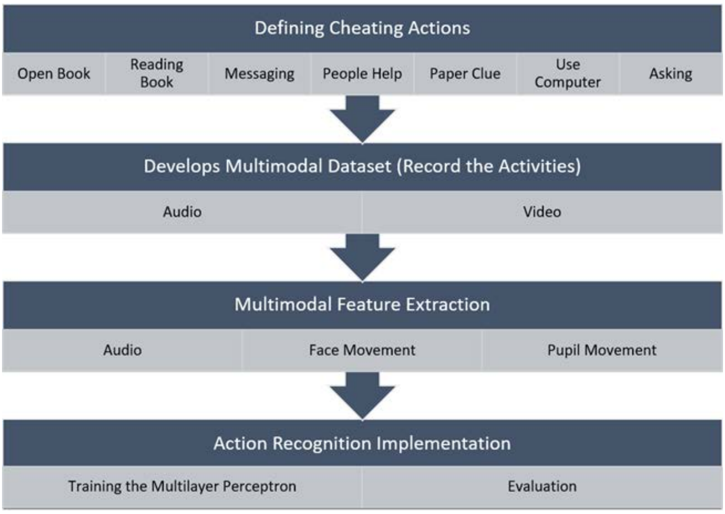
Defining Cheating Actions
Open Book
Reading Book
Messaging
People Help
Paper Clue
Use Computer
Asking
Develops Multimodal Dataset (Record the Activities)
Audio
Video
Multimodal Feature Extraction
Audio
Face Movement
Pupil Movement
Action Recognition Implementation
Training the Multilayer Perceptron
Evaluation

Fig. 1. Research framework.

Online exam cheating activities is defined into seven types. They consist of opening/reading a book, sending/receiving messages through a smart phone, accepting help from people, receiving clues through paper, using the computer to browse testing material, and asking for answers from someone. In order to build a dataset, these activities were recorded in audio and video through standard smart phones. An additional video of non-cheating participants was also recorded with a standardized recording duration of 21.5 seconds in 12 frame per second (fps). A feature extraction was carried out of all the positive and negative cheating videos. The visual and motion data were extracted into face and pupil movement, while the audio was sampled into 12 data per second. The three modalities were fused before entering the network in artificial neural network (ANN), thereby, leading to a total of 1290 data length for each recorded activity. Multi perceptron (MLP), training and evaluation were carried out by using the available data.

3.2. Multimodal input

Multimodal cheating behaviour in online exams involves audio and video data, which is pre-processed from the front camera of the smart phone. Before further processing, a standardization of the visual signal was carried out to reduce the volume of the data while maintaining the quality for feature extraction purposes. The sizes of the frames were under-sampled into 180 x 320 pixels and a reduction of the frame rate was made into 12 per second. The data was considered enough because the video is in slow motion and the distance between the object under surveillance was close. The audio was continually recorded during the data collection for further processing. Figure 2 shows the normal position of a smart phone, with a front camera and the participant’s face.

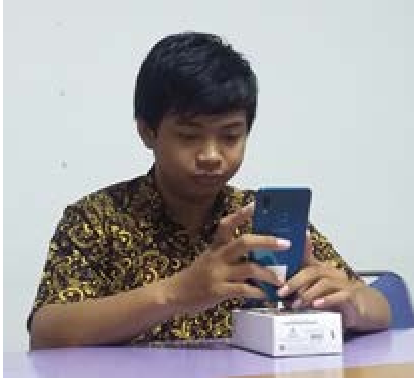

Fig. 2. Normal position of a mobile test examinee.

The video data was extracted into two time series features namely face and pupil movements. Two modalities were obtained in the data collection tasks. The video was explored into the movement using motion vector calculation and Lucas-Kanade methods. Before the tracking started, the face and eye recognition were carried out. This produced a three time series data which are face tracking, eye tracking and audio data. In order to enable the modality fusion for all the data, normalization was carried out in a range of -10 to 10. The periodicity of the data sampled was in 12 data point per second. During the data collection, the original duration was not precisely uniform, while pre-processing and was standardized into 21.5 seconds for each captured audio video. The face tracking produced two variables x and y, which was considered as 1 unit for every 25 pixels movement. Furthermore, the pupil tracking yielded two x and y variables in pixel. Sound on the other hand produced one variable and was sampled in exactly the same frequency at 12 data per second. During the feature extraction, a zero padding is implemented to ensure standard length of each feature for each audio video data. Figure 3 shows the multimodal data pre-processing before a multi-layer perceptron was applied.

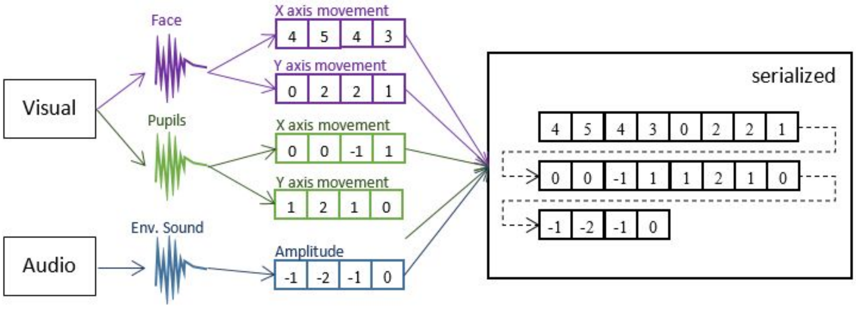

Fig. 3. Multi modal feature extraction and fusion.

3.2.1. Face and eye recognition

The region of interest in the data transformation to the mid-level data representation is the face. Therefore, a number of tasks were needed to analyse the face and the eyes. The first task was to resize the frames into 180 x 320 pixels per frame with 12 frames per second (fps). Secondly, the face location was identified, and its landmark calculated in order to identify its key points such as the nose tip and the eyes.

The face was extracted from the whole image in each frame, using the histogram of gradients HOG [36]. The image was divided into a 16x16 patch and the orientation of the gradient was calculated in 9 bins of unsigned orientation between 0 to 360 degrees [36]. The vector was then normalized to acquire L2-Norm, L1-Norm and L1-Sqrt in Eqs. (1), (2), and (3).

$$L2 - Norm : f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}} \quad (1)$$

$$1 - Norm : f = \frac{v}{\|v\|_1 + e} \quad (2)$$

$$L1 - Sqrt : f = \sqrt{\frac{v}{(\|v\|_1 + e)}} \quad (3)$$

where v is the non-normalized vector containing all histograms in a given $\|v\|_k$block, $\|v\|_k$ is the k-norm (k = 1,2,..) and e is a constant with a small value.

According to Dalal and Triggs [36], the histogram orientation with L2-Norm and L1-Sqrt provided a good performance. The result of the procedure was a localized face area.

After the face location was determined, its landmark was identified and implemented as proposed by Kazemi and Sullivan [37]. The face orientation in each frame was detected and the relative position of its front direction recorded.

The basic idea was to determine 68 points within the face. Kazemi and Sullivan [37] trained the detector and provided the right model to be used in detecting the task. A cascading ensemble regression tree (ERT) was used to estimate the position of the points in the face landmark as shown in the following formula with an S shape, where the points is a member of the S vector.

In the training session, *n* face image (*I*) and the shape *S* were provided. The training data consists of $(I_{i,}S_i)i \in \{1 \ldots n\}$ $(I_{i,}S_i)$ where $i \in \{1 \ldots n\}$, n is the number of training images. The cascaded regression trained using Eqs. (4), (5) and (6).

$$\pi_i \in \{1, \ldots., n\} \tag{4}$$

$$\hat{S}\,_i^{(0)} \in \{S_1, \ldots, S_n\} \mid S_{\pi_i} \tag{5}$$

$$\Delta S_i^{(0)} = S_{\pi_i} - \hat{S}_i^{(0)} \tag{6}$$

This is repeated for *I* = 1… *N* where *N* =*nR* where *R* is the initialization for each image *I*I_i. Based on this data, regression 0 (*r0*) yield $S_i^{(1)}$. The regression is cascaded from *t* = 0 to *T*. The formula for cascade regressing and estimating the shape is shown in Eqs. (7) and (8).

$$\hat{S}_i^{(t+1)} = \hat{S}_i^t + r_t(S_{\pi_i}, \hat{S}_i^t) \tag{7}$$

$$\Delta S_i^{(t+1)} = S_{\pi_i} - \hat{S}_i^{(t+1)} \tag{8}$$

This process is repeated in a cascade of *T* regressors $r_0, r_1, \ldots, r_{T-1}$ in order to learn the best combination with a sufficient level of accuracy. Once the T regressor trained by n face images was obtained, a regression model for face landmark was defined. The pre-defined model was used to identify the face landmark.

Based on the face landmark, the eye location was identified and denoted by 37 - 42, and 43 - 48 on the left and right eye, respectively. The pupil is located in the middle of the eyes as the darkest object.

3.2.2. Face tracking

Face movement was used to detect the pose of the participant relative to the camera. It is three-dimensional, however, only horizontal and vertical movement were considered.

The video consisted of n frames with *i* = 1 to *n - 1*. The *x* and *y* positions were calculated by considering the displacement of the face position relative to the previous frame. The nose tips were considered as the centre of the face, while the displacement of the position of the nose tips was considered as face movement data. The value was set between -10 … 10, were recorded as 25 pixels for 1 unit.

$$f(x_i) = \begin{cases} round\left(\frac{x_{i-1}-x_i}{25}\right), & 0 < |x_{i-1} - x_i| < 250 \\ 10, & x_{i-1} - x_i \geq 250 \\ -10, & x_{i-1} - x_i \leq -250 \end{cases} \tag{9}$$

where x_i is the current frame, x_{i-1} is the previous frame face coordinates.

Face tracking produced two output variables which were horizontal (x) and vertical (y) movements. The horizontal and vertical movements, normalized in the range -10 to 10, and were calculated with respect to Eq. (9).

3.2.3. Eye-tracking

Eye-tracking shows user focus to the view. It provides a clue when the participant looks to another side of the test user interface focus. For example, when help is provided from outside the camera view, the user's eyes tends to leave the test sheet focus.

The normal eye condition is when the participant focuses on the testing area on the smart phone. The pupil position becomes in two-dimension space (x and y) with +/- relative to the center in the number of pixel displacement. For a single face, the left and right eye movements are assumed to be in the same direction.

The first task was to identify the face and eyes in the image. Followed by estimating the center of the eye by calculating x_2 - x_1 and y_2 - y_1, where x_2/ x_1 are it's left and right corner and y_1/y_2 are the top and bottom of the eye. Each frame sequence estimates the new position of the black node in the eye relative to the pupil in the previous frame. Figure 4 shows the face identification and pupil tracking during a particular frame in a cheating video. Once the position of the pupil is detected, the relative position is calculated against the normal with respect to Eq. (10).

$$f(x_i) = \begin{cases} x_{i-1} - x_i\,, & 0 < |x_{i-1} - x_i| < 10 \\ 10, & x_{i-1} - x_i \geq 10 \\ -10, & x_{i-1} - x_i \leq -10 \end{cases} \tag{10}$$

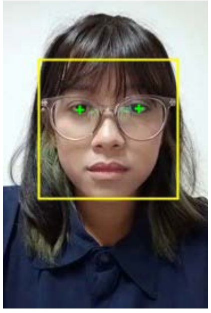

Fig. 4. Face and pupil identification in a particular frame.

3.2.4. Sound

During the simulation, the smart phone recorded the sound of the test environment, which is transformed in certain periods synchronous to the frame sampling. Furthermore, the sounds were normalized into 0 to 10 and put on the time-series data. The audio input amplitude represents the power of the sound and originally the peak of the loudest audio signal(peak). The peak signal usually achieve in certain point. The effective audible signal listened by human usually called the root mean square (RMS) amplitude. The original audio signal is distributed over positive and negative value; therefore, a positive and negative peak are existed. We can consider an audio signal as a composite of many sinusoidal signal in varied frequencies. Let's assume in certain small windows for example 1/12 second we sampled n=10 data {S_1, S_2, …. S_{10}}, RMS can be calculated using Eq. (11).

$$RMS = \sqrt{\frac{(S_1{}^2 + S_2{}^2 + S_3{}^2 + \ldots\ldots S_n{}^2\ )}{n}} \tag{11}$$

The RMS therefore all positive number, between 0 and 1. In our experiment this scale is normalized to 0 to 10.

3.3. Multilayer perceptron for activity recognition

Multilayer perceptron (MLP) is a class of artificial neural network (ANN) commonly referred to as "vanilla", especially when it consists of a single hidden layer [19]. The minimum architecture of MLP consists of three layers which are input, hidden, and output. Each node in the hidden and output layer is a neuron with a nonlinear activation function. The network is trained by a supervised learning technique called back-propagation to adjust the weight for each node [20, 21]. MLP is capable of handling non-linear separation of classes due to tits activation function [22]. This capability has become the main difference between MLP and linear perceptron.

It employs two common activation functions sigmoid, in Eq. (12) and rectifier linear unit (RELU) as described in Eq. (13):

$$S(x) = \frac{e^x}{1+e^{-x}} \tag{12}$$

$$R(x) = max(0, x) \tag{13}$$

where S(x) is the sigmoid activation function, x is the input variable and R(x) is the RELU function. Figure 5 shows the architecture of MLP in this research. It consists of input layer with 1290 nodes, varied number of hidden layer and single node in output layer.

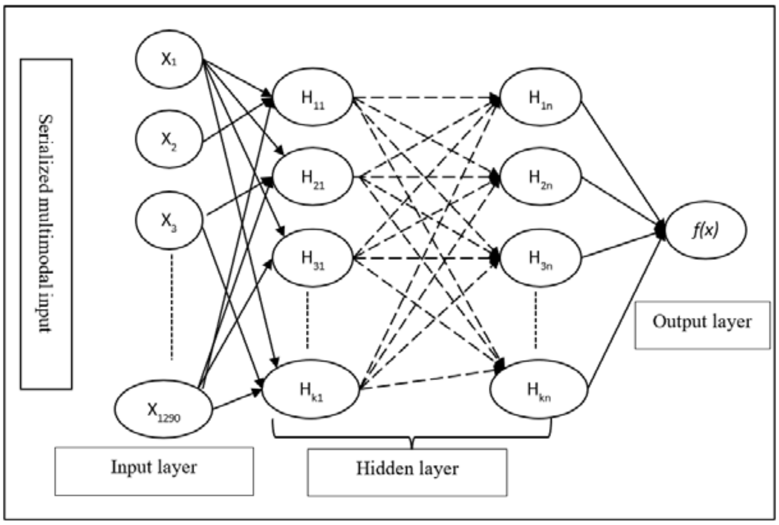

Fig. 5. Multi-layer perceptron architecture.

In the experiments conducted, the network was set with 10, 20, 30 and 40 hidden layers using the 'RELU' activation function as described in Eq. (13) and trained with 'Adam' optimizer. The input layer was set at 1290 nodes to fit with the number of input data which comes from the multi modal feature extraction of pupil tracking, face tracking and environment sound. The final output is a single node with binary class cheating and non-cheating.

3.4. Experimental design

This experimental research was carried out in a quiet testing environment away from the crowd to benefit the exam room. The research location was a 3x3 meter room which had bright lighting, tables, and chairs. Each participant sitting and working on exam questions all had a smart phone. Their profiles were quite diverse, with male and female, wearing/not wearing glasses, or veil. The data collection was carried out simultaneously among four participants in meeting rooms.

A simulation of suspicious activity was defined in seven possible cheating scenarios as follows:

- Opening a book activity.
- Reading a book next to the smart phone.
- Using a laptop to obtain information.
- Reading someone else's clue outside the camera.
- Sending and receiving messages from the second smart phone.
- Creating someone else's presence outside of the back camera.
- Asking questions from someone else through a voice simulation.

A total of twenty subjects were used to test all the scenarios. It involved recording seven activities for each subject in the cheating scenarios detailed above, as well as seven in the non-cheating categories. This led to a total of 280 activities.

Three low end smart phones with the specification are explained in Table 1.

Table 1. Smartphone specification.

Specs	Samsung A2 Core	Samsung A20	Vivio Y15
CPU	1.6 GHz Octacore	2x 1.6 GHz Octacore	2.0 GHz Octacore
RAM	1 GB	3 GB	4 GB
Display	5" 540 x 960 px	6.4" 720 x 1560 px	6.35"720 x 1544 px
Front Cam	5 MP	8 MP	16 MP
Back Cam	5 MP	13 Mp	13 MP
OS	Android 8.0	Android 9.0	Android 9.0

3.5. Evaluation

The data was divided into 5 folds and a cross-validation was carried out. The proportion of training and testing data were 80% and 20%, respectively. The quality metrics of the classification measured were accuracy, recall, precision, and f1 score.

4. Results and Discussion

The activity classification was carried out by a multilayer perceptron (MLP) on a standardized input parameter. In order to standardize the data, the duration of each activity in the listed scenarios were set to fifteen seconds. The shorter activity was padded with zero value in each variable.

A total of 12 frames per second were considered to be enough since the mobile exam scenario was controlled by ensuring the participants head movement while holding the smart phone was limited by the camera position.

The original audio-video signal was transformed into face, pupil and sampled audio tracking. The length of the output of the transformation was varied due to the duration of the video between 14 to 22 seconds. In order to ensure the equal length of the input vector to the network, zero padding was put at the end of the signal. The expected length of the signal was 258 to fit the number of input nodes in the MLP architecture. Figure 6 shows the snippet of sound signal representation of a particular activity. Meanwhile, Figs. 7 and 8 show the face and eye tracking diagrams, respectively. X and y-axis represent horizontal and vertical movement.

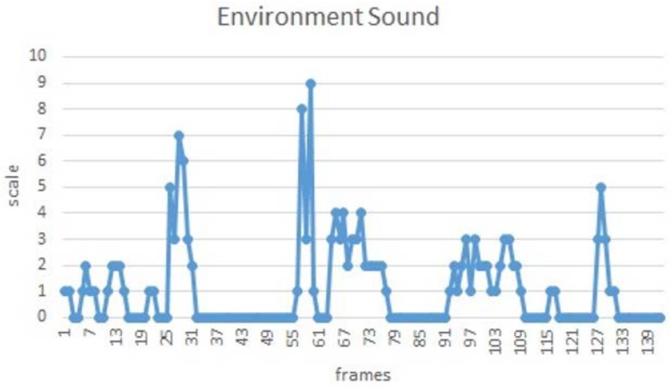

Fig. 6. Sound signal feature.

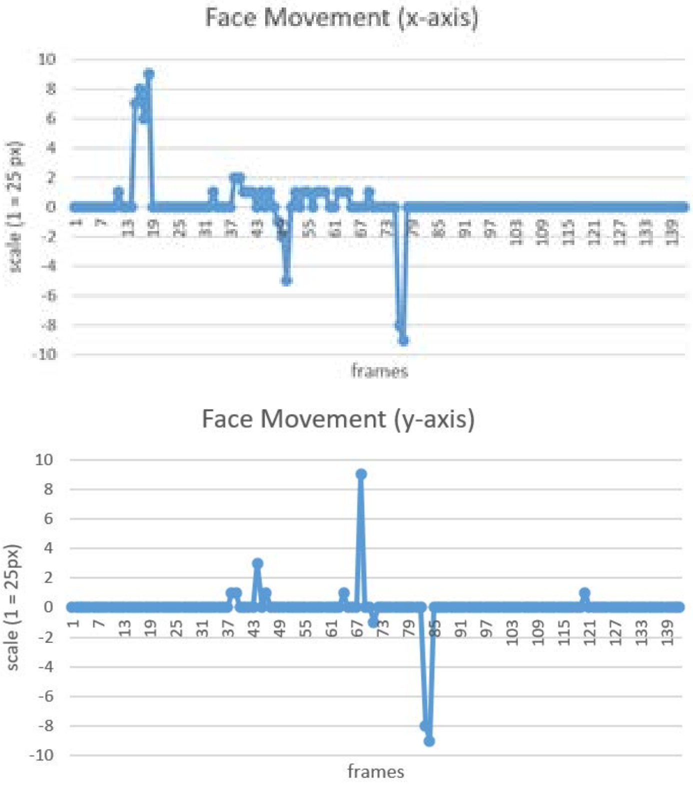

Fig. 7. Face tracking movement.

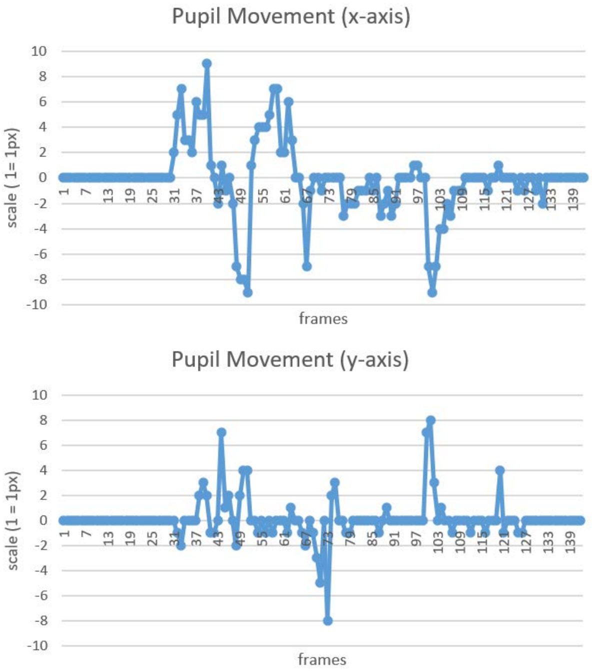

Fig. 8. Eye tracking signal.

The signal was concentrated into a single vector for each activity while the mid-level signal 5×258 was transformed into a 1 x 1290 size.

The entire mid-level signal for all the dataset consisted of 280 rows, with the data randomly divided into five folds and cross-validated using the classifier. Table 2 shows the average, accuracy, precision and f1score of the classification result using MLP with varied hidden layer. Table 2 showed that 20 hidden layers achieved the best classification metrics. Table 3 shows the classification metrics of 5 folds experiments.

Table 2. Average accuracy, recall, precision and F1 score for different hidden layer.

Hidden Layer	Accuracy	Recall	Precision	F1 Score
10	0.910325	0.910325	0.911421	0.910325
20	0.917338	0.917338	0.929832	0.916798
30	0.903312	0.903312	0.909436	0.902822
40	0.903182	0.903182	0.912396	0.902383
50	0.913896	0.913896	0.918950	0.913387

Table 3. Classification result for each fold in 20 hidden layers.

Fold	Accuracy	Recall	Precision	F1 Score
1	0.928571	0.928571	0.928571	0.928571
2	0.892857	0.892857	0.909774	0.889796
3	0.964286	0.964286	0.96659	0.964194
4	0.946429	0.946429	0.950893	0.945724
5	0.854545	0.854545	0.893333	0.855703
Average	0.917338	0.917338	0.929832	0.916798
Std Dev	0.039291	0.039291	0.026573	0.039201

According to the experiments, the best model was achieved on the third fold, with a successful 96% classification accuracy based on the available data. The worst experiment achieves 85% accuracy, while the balance weight between the recall and precision in calculating the F1 score was considered due to the balance portion of cheating and non-cheating classes in the dataset.

The conversion of the video data into face and pupil tracking were carried out successfully. The down sampling was reduced into 12 frames per second to produce a good result where the achieved accuracy was above 91.7% on average and 96% in the best model as shown Table 3. The simple mechanism to transform video and audio data into a simple representation still has enough discriminating power for a binary classifier.

An experiment of using multimodal input from webcam, gaze tracked and EEG devices [10]. Three scenarios were performed which are automatics cheating detector (ACD) which is comparable to our research. They also proposed a combination of the ACD with peer cheating detector (PCD) and Final review committee (FRC). PCD involved peer exam taker to review piece of suspicious exam video raised by automatic detection (ACD). Once PCD conclude that the piece of video is suspicious cheating, a final review committee work for final decision.

Table 4 shows the result of Le et al. [10]. Compared to single webcam ACD and Multimodal ACD, our result outperforms in recall, precision and accuracy. The fair comparison with single webcam only (86.1%) However, compare to peer cheating detection (PCD) in Table 4, our achievement slightly under their achievement. This is reasonable due to human involvement in their PCD experiment.

Table 4. The performance of previous research [10].

Cheating Classifier	Recall	Precision	Accuracy
Webcam-only ACD	76.60%	84.60%	83.30%
Multi-modal ACD	82.10%	60.00%	81.10%
PCD	96.40%	94.10%	95.60%
Webcam-only MOOP	78.60%	90.40%	86.10%
Multi-modal MOOP	90.50%	93.80%	92.70%

Atoum et al. [11] proposed a multimodal proctoring using double camera and sound recorder. They need a special camera (wearcam) installed on the eyeglasses to observe the view area of the participant. They achieve 87% of cheating recognition using SVM classifier. Our result was over 4% better at 91.73% compare to theirs'.

This research is limited by the pre-defined cheating action under specified requirements in the experimental setting. However, despite the unavailability of empirical data, and the specified resolution of the video, the result might be degraded. Secondly, other types of cheating are likely to be beyond the simulated activity pre-set in this research, with undefined outcomes. The lighting in this study was set in stable and visible condition, therefore, a darker situation is likely to lead to a significant decrease on recognition quality. The noisy sound is also a challenge as the dataset was taken in a controlled condition where the noise is at a minimum level. Another potential problem of this research is privacy issue, due to the need of recording the video and audio during the exam. In the future, a signal transformation on the recording device would be an alternative to overcome the privacy issue. Due to the limited possibility of cheating/non-cheating actions detected, the exact activity was not precisely recognized such as reading a book, using mobile phone etc. These limitations are a guide for future research.

5. Conclusion

This study presented an audio-video analysis for cheating action recognition to support a mobile exam. Therefore, in conclusion, the audio and video signals are necessary to recognize the cheating and non-cheating activity. The average accuracy, precision, recall, and f1 score were found to be at 91.73±3.9%, 91.73±3.9%, 92.99±2.6% and 91.67±2.3%, respectively. A high recognition rate was achieved, compared to the state of the art in online cheating detection. Compare to multimodal automatic detection and double camera plus audio from the literature, this research achieved better accuracy with almost a 4% margin. In addition, this research was designed for a smart phone exam environment with an implementation of standard accessories without any additional device. Further research needs to be conducted for multiclass classification in order to identify the action conducted. The possibility of using more sensors in the smart phone, such as accelerometer and the gyro tends to enrich the dataset and is expected to improve the accuracy.

Acknowledgment

The research activity leading to the publication has been funded by the Ministry of Research, Technology/National Research and Innovation Agency (Indonesia) Grant No. 227/SP2H/AMD/LT/DRPM/2020. Also, the authors would like to thank the research Directorate of Universitas AMIKOM Yogyakarta for their support.

References

1. Seaman, J.E.; and Seaman, J. (2017). *Distance education state almanac 2017*. Retieved September 5, 2019, from https://onlinelearningsurvey.com/reports/almanac/national_almanac2017.pdf
2. Allen, I.E.; and Seaman, J. (2014). *Grade change tracking online education in the United States*. Retrieved September 5, 2019, from https://www.onlinelearningsurvey.com/reports/gradechange.pdf
3. Fask, A.; Englander, F.; and Wang, Z. (2014). Do online exams facilitate cheating? An experiment designed to separate possible cheating from the effect of the online test taking environment. *Journal of Academic Ethics*, 12(2), 101-112.
4. King, C.G.; Guyette, R.W.; and Piotrowski, C. (2009) Online exams and cheating: An empirical analysis of business students' views. *Journal of Educators Online*, 6(1), 1-11.
5. King, D.L.; and Case, C.J. (2009). E-cheating: Incidence and trends among college students. *Issues in Information Systems*, 15(1), 20-27.
6. Corrigan-Gibbs, H.; Gupta, N.; Northcutt, C.; Cutrell, E.; and Thies. W. (2015). Deterring cheating in online environments. *ACM Transactions on Computer-Human Interaction*, 22(6).
7. Foster, D.; and Layman, H. (2013). Online proctoring system compared. *U.S. Pat. Appl. No. 12/723,666.*, 1-12.
8. Kaiiali, M.; Ozkaya, A.; Altun, H.; Haddad, H.; and Alier, M. (2016). Designing a secure exam management system (SEMS) for M-learning environments. *IEEE Transactions on Learning Technologies*, 9(3), 258-271.

9. Bawarith, R.; Basuhail, A.; Fattouh, A.; and Gamalel-Din, S. (2017). E-exam cheating detection system. *International Journal Advanced Computer Science and Application*, 8(4), 176-181.
10. Li, X.; Chang, K.; Yuan, Y.; and Hauptmann, A. (2015). Massive open online proctor: Protecting the credibility of MOOCs Certificates. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*. Vancouver BC, Canada, 1129-1137.
11. Atoum, Y.; Chen, L.; Liu, A.X.; Hsu, S.D.H.; and Liu, X. (2017). Automated online exam proctoring. *IEEE Transactions on Multimedia*, 19(7), 1609-1624.
12. Hung, H.; and Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 12(6), 563-575.
13. Nguyen, L.S.; Frauendorfer, D.; Mast, M.S.; and Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, 16(4), 1018-1031.
14. Zajdel, W.; Krijnders, J.D.; Andringa, T.; and Gavrila, D.M. (2007). CASSANDRA: Audio-video sensor fusion for aggression detection. Proceedings of the *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. London, United Kingdom, 200-205.
15. Lefter, I.; Burghouts, G.J.; and Rothkrantz, L.J.M. (2012). Automatic audio-visual fusion for aggression detection using meta-information. *Proceedings of the IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance*. Beijing, China, 19-24.
16. Wang, Z.; Yu, J.; and He, Y. (2017). Soccer video event annotation by synchronization of attack-defense clips and match reports with coarse-grained time information. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(5), 1104-1117.
17. Kolekar, M.H.; and Sengupta, S. (2015). Bayesian network-based customized highlight generation for broadcast soccer videos. *IEEE Transactions on Broadcasting*, 61(2), 195-209.
18. Cluskey, G.R.; Ehlen, C.R.; and Raiborn, M.H. (2011). Thwarting online exam cheating without proctor supervision. *Journal of Academic and Business Ethics*, 4, 1-7.
19. Wahid, A.; Sengoku, Y.; and Mambo, M. (2015). Toward constructing a secure online examination system. *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*. Bali, Indonesia, 1-8.
20. Clarke, N.L.; Dowland, P.; and Furnell, S.M. (2013). e-Invigilator: A biometric-based supervision system for e-assessments. *Proceedings of the International Conference on Information Society*. Toronto, Canada, 238-242.
21. Asha, S.; and Chellappan, C. (2008). Authentication of e-learners using multimodal biometric technology. *Proceedings of the International Symposium on Biometrics and Security Technologies*. Islamabad, Pakistan, 1-6.
22. Jalali, K.; and Noorbehbahani, F. (2017). An automatic method for cheating detection in online exams by processing the student's webcam images. *Proceedings of the 3rd Conference on Electrical and Computer Engineering Technologies*. Tehran, Iran, 1-6.

23. Kumar, A.V.S.; and Rathi, M. (2019). *Biometric Authentication in Online Learning Environments*. Chapter 8: Keystroke dynamics: A behavioral biometric model for user authentication in online exams. IGI Global, 183-207.
24. Bal, A.; and Acharya, A. (2011). Biometric authentication and tracking system for online examination system. *Proceedings of the International Conference on Recent Trends in Information Systems*. Kolkata, India, 209-213.
25. Fayyoumi, A.; and Zarrad, A. (2014). Novel solution based on face recognition to address identity theft and cheating in online examination systems. *Advances in Internet of Things*, 4(2), 5-12.
26. Xiao, B.; Georgiou, P.; Baucom, B.; and Narayanan, S.S. (2015). Head motion modeling for human behavior analysis in dyadic interaction. *IEEE Transactions on Multimedia*, 17(7), 1107-1119.
27. Liu, Y.; Xie, Z.; Yuan, X.; Chen, J.; and Song, W. (2017). Multi-level structured hybrid forest for joint head detection and pose estimation. *Neurocomputing*, 266, 206-215.
28. Dornaika, F.; and Raducanu, B. (2009). Three-dimensional face pose detection and tracking using monocular videos: tool and application. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(4), 935-944.
29. Qing, C.; Jiang, J.; and Yang, Z. (2010). Normalized co-occurrence mutual information for facial pose detection inside videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(12), 1898-1902.
30. Drouard, V.; Horaud, R.; Deleforge, A.; Ba, S.; and Evangelidis, G. (2017). Robust head-pose estimation based on partially-latent mixture of linear regressions. *IEEE Transactions on Image Processing*, 26(3), 1428-1440.
31. Cheung, Y.; and Peng, Q. (2015). Eye gaze tracking with a web camera in a desktop environment. *IEEE Transactions on Human-Machine Systems*, 45(4), 419-430.
32. Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. (2016). Eye tracking for everyone. *Proceedings of the IEEE Computer Vision and Pattern Recognition*. Las Vegas, Nevada, 2176-2184.
33. Meng, C.; and Zhao, X. (2017). Webcam-based eye movement analysis using CNN. *IEEE Access*, 5, 19581-19587.
34. Tran, H.D.; and Li, H. (2011). Sound event recognition with probabilistic distance SVMs. *IEEE Transactions on Audio, Speech and Language Processing*, 19(6), 1556-1568.
35. Wang, C.; Wang, J.; Santoso, A.; Chiang, C.; and Wu, C. (2018). Sound event recognition using auditory-receptive-field binary pattern and hierarchical-diving deep belief network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(8), 1336-1351.
36. Dalal, N.; and Triggs, B. (2004). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Diego, USA, 1, 886-893.
37. Kazemi, V.; and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, Ohio, 1867-1874.