DOI: https://doi.org/10.15276/aait.06.2023.14
UDC 004.932

Human action analysis models in artificial intelligence based proctoring systems and dataset for them

Svitlana G. Antoshchuk[1)]
ORCID: https://orcid.org/0000-0002-9346-145X; asg@op.edu.ua. Scopus Author ID: 8393582500
Anastasiia A. Breskina[1)]
ORCID: https://orcid.org/0000-0002-3165-6788; anastasia.breskina@gmail.com
[1)] Odessa Polytechnic National University, 1, Shevchenko Ave. Odessa, 65044, Ukraine

ABSTRACT

This paper describes the approach for building a specialized model for human action analysis in AI-based proctoring systems and proposes a prototype of dataset which contains data specific to the application area. Boosted development of machine learning technologies, the availability of devices and the access to the Internet are skyrocketing the development of the field of distance learning. And in parallel with distance learning systems the AI-based proctoring systems, that provide the functional analysis of student work by imitating the teacher's assessment, are developing as well. However, despite the development of image processing and machine learning technology, the functionality of modern proctoring systems is still at a primitive level. Within the image processing functionality, they focus entirely on tracking students' faces and do not track postures and actions. At the same time, assessment of physical activity is necessary not only as part of the learning process, but also to keep students healthy according to regulatory requirements, as they spend the entire duration of learning process in front of computers or other devices during the distance learning. In existing implementations, this process falls entirely on the shoulders of teachers or even the students themselves, who work through the lesson materials or tests on their own. Teachers, at the same time, have to either establish contact through video communication systems and social media (TikTok, Instagram) and/or analyse videos of students doing certain physical activities in order to organise physical activities evaluation. The lack of such functionality in AI-based proctoring systems slows down the learning process and potentially harms students' health in the long run. This paper presents additional functionality requirements for AI-based proctoring systems including human action analysis functionality to assess physical activity and to monitor hygiene rules for working with computers during the educational process. For this purpose, a foundation model called InternVideo was used for processing and analysis of student's actions. Based on it, the approach for building a specialized model for student action analysis was proposed. It includes two modes of student activity evaluation during the distance learning process: static and dynamic. The static mode (aka working phase) analyses and evaluates the student's behavior during the learning and examination process, where physical activity is not the main component of learning. The dynamic mode (aka physical education mode) analyses and assesses the student who purposefully performs physical activity (physical education lesson, exercises for children during the lesson, etc.). A prototype dataset designed specifically for this application area has also been proposed.

Keywords: Computer vision; neural networks; dataset, transformer; action understanding; video understanding; artificial intelligence based proctoring systems; online proctoring; proctoring system; distance learning; online learning

For citation*: Antoshchuk S. G., Breskina A. A. "Human action analysis models in artificial intelligence based proctoring systems and dataset for them".** ***Applied Aspects of Information Technology. **2023; Vol.6 No.2: 190–200. DOI: https://doi.org/10.15276/aait.06.2023.14**

INTRODUCTION

The widespread use of the Internet and the development of computers and portable devices have given a rapid boost to the field of distance learning. The COVID-19 pandemic and quarantines around the world forced most schools and universities to switch to online learning, which triggered an additional round of development of online proctoring systems (OPS). Online proctoring systems are online tools that are used to monitor the examination process.

These systems simulate the work of a teacher controlling students' behavior during the learning process [1, 2]. Online proctoring systems are divided into three types [1, 2], depending on the level of automation of the student assessment process and the level of teacher or external observer involvement in this process.

Live proctoring. In live proctoring systems, students are assessed in real time by a teacher or a third person (a hired observer). These systems are used during theory exams, which take about 2-3 hours.

Recorded proctoring. As with live proctoring, the whole process of assessing student behavior is the responsibility of the teacher. However, the process of analyzing student behavior takes place post-factum, based on stored video recordings of the learning process or exam.

Automated proctoring. The aim of these systems is to minimize the teacher's work in assessing student behavior. Such systems generate automatic assessment of student work according to different factors:
activity on the desktop, third-party programs and internet resources, assessment of audio and conversations during the learning process and assessment of student behavior based on information obtained from the video camera (absence of the student at the workplace, presence of third persons together with the student, reduced concentration, etc.).

The development of Machine learning (ML) and artificial intelligence has opened up a large area for the development and improvement of automated OPS. Online proctoring systems that use elements of artificial intelligence to do their work (Fig. 1) are called artificial intelligence based proctoring systems (AIPS)[2].

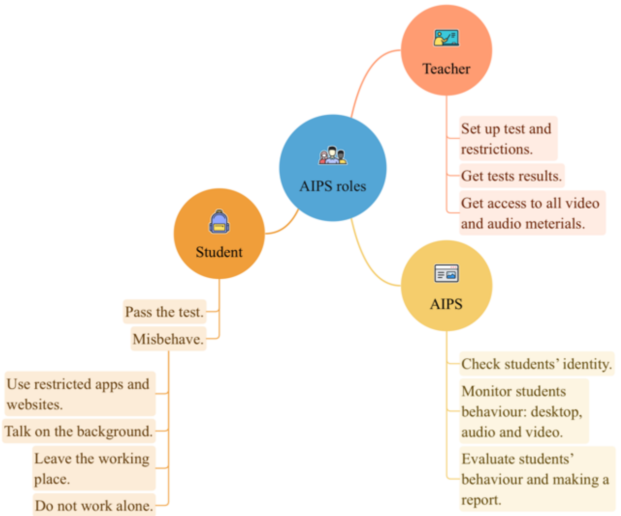

Teacher
Set up test and restrictions.
Get tests results.
Get access to all video and audio meterials.
AIPS roles
Student
Pass the test.
Misbehave.
Use restricted apps and websites.
Talk on the background.
Leave the working place.
Do not work alone.
AIPS
Check students' identity.
Monitor students behaviour: desktop, audio and video.
Evaluate students' behaviour and making a report.

Fig. 1. **Artificial intelligence based proctoring systems roles Mind map**
***Source:* compiled by the authors**

Modern ML methods make it possible to characterize students' behavior during classes with greater accuracy. And while with program and audio noise analysis the task is trivial (whether the restricted program or website is open or not, whether there is an audio signal containing suspicious noise patterns or not), video analysis already includes more complex Computer Vision algorithms.

This analysis includes such checks as:

– the presence or absence of the student on his working place;

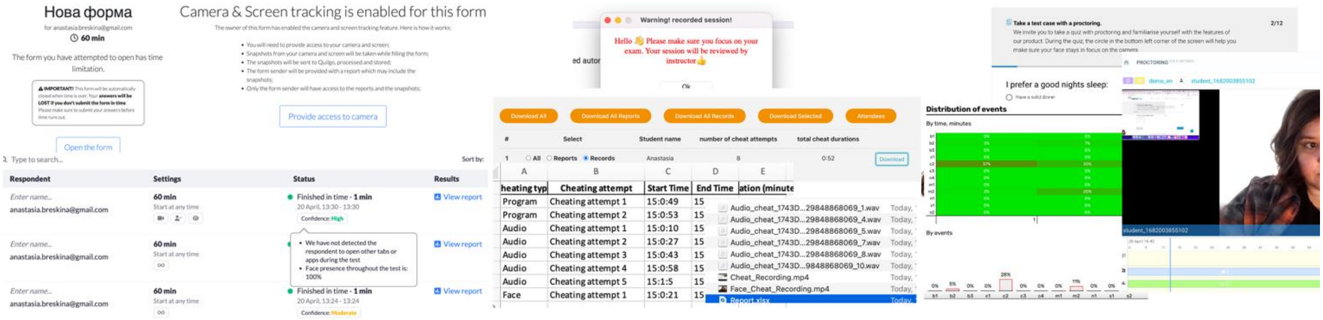
– presence or absence of other people around the student;

– analysis of the student's emotional state from facial expressions and body movements.

LITERATURE REVIEW

Three artificial intelligence based proctoring systems implementations were tested: Quilgo, AI Proctor and ProctorEdu. The testing was organized as follows: three exams were taken (except for ProctorEdu, as this system provided only one free test) and in each of them two "inappropriate" actions were performed while working with forbidden software and talking during the task, as well as four types of activity on video (changing the direction of looking, absence of the student at the desk, presence of third person at the desk, chaotic movements and waving of hands).

As a result, the systems performed well in analysis of the desktop and the programs used in all test cases, and an acceptable in audio noise analysis (about 80 % responses were detected correctly, there were a few False positives (FP), which is not critical). The most problematic was the CV module: all software products gave a True positive (TP) result of only 25-30 % test cases and False negative (FN) in others (Fig. 2).



Нова форма
for anastasia.breskina@gmail.com
60 min
The form you have attempted to open has time limitation.
IMPORTANT! This form will be automatically closed when time is over. Your **answers will be LOST if you don't submit the form in time**. Please make sure to submit your answers before time runs out.

Camera & Screen tracking is enabled for this form
The owner of this form has enabled the camera and screen tracking feature. Here is how it works:
- You will need to provide access to your camera and screen;
- Snapshots from your camera and screen will be taken while filling the form;
- The snapshots will be sent to Quilgo, processed and stored;
- The form sender will be provided with a report which may include the snapshots;
- Only the form sender will have access to the reports and the snapshots;

Respondent	Settings	Status	Results
Enter name... anastasia.breskina@gmail.com	60 min Start at any time	Finished in time - 1 min 20 April, 13:30 - 13:30 Confidence: High	View report
Enter name... anastasia.breskina@gmail.com	60 min Start at any time	• We have not detected the respondent to open other tabs or apps during the test • Face presence throughout the test is: 100%	View report
Enter name... anastasia.breskina@gmail.com	60 min Start at any time	Finished in time - 1 min 20 April, 13:24 - 13:24 Confidence: Moderate	View report

Warning! recorded session!
Hello 👋 Please make sure you focus on your exam. Your session will be reviewed by instructor 👍

#	Select	Student name	number of cheat attempts	total cheat durations
1	○ All ○ Reports ● Records	Anastasia	8	0:52

heating typ	Cheating attempt	Start Time	End Time	ation (minute
Program	Cheating attempt 1	15:0:49	15	
Program	Cheating attempt 2	15:0:53	15	
Audio	Cheating attempt 1	15:0:10	15	
Audio	Cheating attempt 2	15:0:27	15	
Audio	Cheating attempt 3	15:0:43	15	
Audio	Cheating attempt 4	15:0:58	15	
Audio	Cheating attempt 5	15:1:5	15	
Face	Cheating attempt 1	15:0:21	15	

Audio_cheat_1743D...29848868069_1.wav Today,
Audio_cheat_1743D...29848868069_4.wav Today,
Audio_cheat_1743D...29848868069_5.wav Today,
Audio_cheat_1743D...29848868069_7.wav Today,
Audio_cheat_1743D...29848868069_8.wav Today,
Audio_cheat_1743D...9848868069_10.wav Today,
Cheat_Recording.mp4 Today,
Face_Cheat_Recording.mp4 Today,
Report.xlsx Today,

Take a test case with a proctoring. 2/12
We invite you to take a quiz with proctoring and familiarise yourself with the features of our product. During the quiz, the circle in the bottom left corner of the screen will help you make sure your face stays in focus on the camera.
I prefer a good nights sleep:
○ Have a solid dinner

a b c

Fig. 2. **Comparison of artificial intelligence based proctoring systems implementations: a – Quilgo; b – AI Proctor; c – ProctorEdu**
***Source:* compiled by the authors**

As a result, the systems performed well in analysis of the desktop and the programs used in all test cases, and an acceptable in audio noise analysis (about 80 % responses were detected correctly, there were a few False positives (FP), which is not a critical). The most problematic was the computer vision module all software products gave a True positive (TP) result of only 25-30% test cases and False negative (FN) in others.

Unfortunately, the software is not open source, and we can only guess at how the video stream from the camera is handled. According to the developers' claims and since the teacher receives a full video or video clips as the output, these systems should process a video stream. However, the test results show that the video sequence is not fully processed, but rather a frame at a random moment in time is analyzed.

In total, the analysis (Table 1) of these software solutions revealed the following common problems:

– lack of transparency in the handling of personal customer data, which does not comply with the general data protection regulation (GDPR);

– unpredictable operation of image recognition modules;

– complete lack of control over the hygiene requirements for computer use, which can lead to health problems for students.

***Table 1.* Comparison of AI-based proctoring systems implementations**

Name/ Features	CV model TP %	Integration	Price
Quilgo	25	Google Forms	50 tests free, up to $35/month
AI Proctor	30	A wrapper for other apps.	$6 per exam, free trial.
Proctor Edu	25	No	Presentation, $6 per exam

***Source:* compiled by the authors**

THE PURPOSE OF THE ARTICLE

The purpose of this work is to develop additional functional requirements for artificial intelligence based proctoring systems and based on these requirements, describe a specialized model for student action analysis and implement a prototype dataset designed specifically for this application area.

MAIN PART. ARTIFICIAL INTELLIGENCE BASED PROCTORING SYSTEMS AND VIDEO FOUNDATION MODELS

Formation of requirements for artificial intelligence based proctoring systems. Based on the highlighted shortcomings, it is proposed to introduce new and clear rules for artificial intelligence based proctoring systems.

These requirements are divided by functional area and are as follows:

– regulation of handling users' personal data;

– monitoring of student behavior and engagement during the learning process to monitor the mental and physical health of the student during the long-distance learning process.

Regulation of handling users' personal data. This point implies controlling the handling of personal data of teachers and students in accordance with modern rules on the handling of personal data. For instance, for the European Union this is GDPR (Available from: https://commission.europa.eu/law /law-topic/data-protection/data-protection-eu_en). These rules include the description on how to implement full control over the account by user and the depersonalize the stored personal data of each user in the system as much as possible.

Monitoring of student behavior and engagement during the learning process. A comprehensive approach is needed to assess student engagement. In addition to the analysis of audio noise and the use of third-party software, a comprehensive analysis of student behavior on video should be implemented as well. This analysis should consist of both the analysis of the student's actions and their emotional state based on the video processing of the student's face (facial expression recognition machine learning task). This point is very important, as modern artificial intelligence based proctoring systems focus entirely on the comfort of the teacher but are not conducive to the health of the students. The lack of control of physical activity of students during the distance learning process leads to the fact that a person spends all the time in front of the monitor, which can severely affect both their physical and mental health. So, it is suggested to add a module for monitoring physical activity, emotional state and health norms while working on the computer.

It is proposed to not only count down the allowable time of working with the computer, but also to analyze the physical activity of the student and assess the quality of the sets of exercises he/she performs during the physical exercises.

For a complex analysis and assessment of student behavior, the following data should be taking into the account:

– the current mode of the system. It’s suggested to use two modes: static and dynamic. Static mode is when the student's behavior is analyzed during a lecture or examination when no active physical activity is expected.

And dynamic mode is when the student is actively engaged in physical activity and that physical activity should be evaluated:

– data on the student's desktop and the websites they are currently working with;

– audio, to analyze the presence or absence of noise on the background;

– description of the student's activities and psychological state, the result of the analysis of the video sequence.

While the first three input values are easy to obtain, in order to classify a student's performance by analyzing a video sequence, we need to preprocess the data, get a text description of the video sequence and then use it in the classification process. And since to evaluate and control engagement, it is necessary not only to analyze the student's actions, but also their emotional state by face recognition, it was suggested to integrate the fundamental model and use its output values for further classification.

Video foundation models. A new successful paradigm for building artificial intelligence systems has emerged in recent years [3, 4]. Its idea is to train one model on a vast amount of data and adapt it to many applications. Such models are called foundation models and their use allows reducing training time of other, specialized, models that are based on them, and decreases resource intensity of this process. In order to solve the problem of analyzing students' activities during the learning process, it was decided to use one of such models.

For this purpose, several video foundation models (InternVideo[6], mPLUG[7], UnMasked Teacher [8]) were analyzed within such machine learning tasks, which can be divided into two groups:

– human action classification, action recognition, spatio-temporal action localization,

– video question answering and visual question answering.

The action recognition and classification tasks are computer vision tasks, which involve classifying and assigning actions performed in a video to a predetermined set of action classes. Spatio-temporal action localization at the same time aims not only at identifying the category of action, but also at localizing it in time and space. The output of models depends on both the model itself and the data used to train it. And the complexity of creating video datasets for model learning is that the most popular benchmarks for action recognition are relatively small (about 10000 videos per benchmark).

In the context of these machine learning tasks, the Top-1 Accuracy [9] metric was used as a quantitative measure of the quality of the models reviewed. It represents how well the model performs across all classes and whether it is useful when all classes have equal importance. Top-1 means that the model's answer with the highest probability is assumed to be TP and only it will be taken into the account.

A large number of datasets and benchmarks are available to evaluate the performance of the models.

Since scientists have used different variety of benchmarks in their papers, those on which all three models were evaluated were used. The following datasets were used to compare the performance of the models: Kinetics-400, Kinetics-600 and Kinetics-700. In addition, the Something-Something V2 and Something-Something V1 datasets were selected to evaluate the performance of the InternVideo model (Fig. 3, Fig. 4, Fig. 5 and Fig. 6).

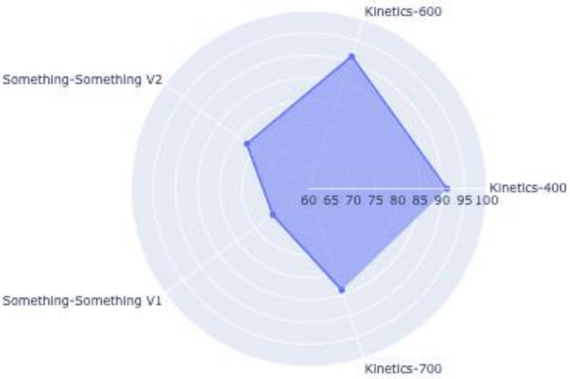


Fig. 3. **Radar diagram of InternVideo model performance**

Source: **compiled by the authors**

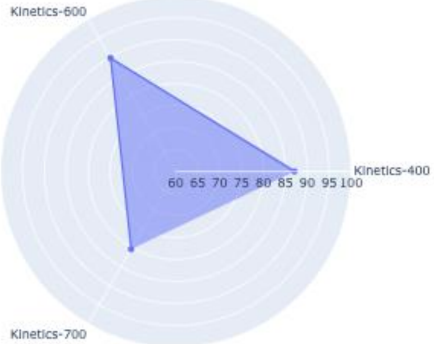


Fig. 4. **Radar diagram of mPLUG model performance**

Source: **compiled by the authors**

As can be seen in the diagrams, the difference in performance in the recognition and classification of actions is small for InternVideo and UnMasked Teacher models (InternVideo performed slightly better). mPLUG performed worse than the others.

Video question answering task aims to answer questions in natural language according to the given video. By receiving the video and the question in natural language, the model gives accurate answers according to the content of the video. Visual question answering task is pretty the same and involves answering questions about an image. The purpose of visual question answering is to teach machines to understand the content of an image and to answer questions about it in natural language.

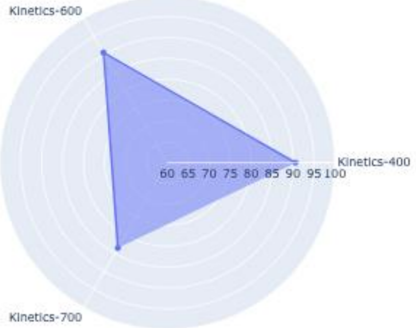


Fig. 5. **Radar diagram of UnMasked Teacher model performance**
***Source:* compiled by the authors**

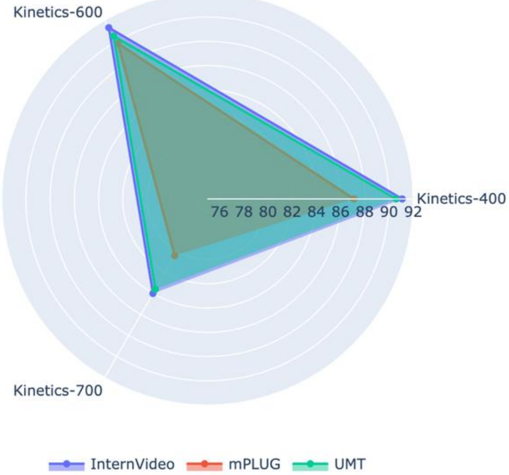


Fig. 6. **Radar diagram of all three models performance**
***Source:* compiled by the authors**

For these tasks, the common datasets for the three models were MSRVTT-QA and MSVD-QA. As a quantitative characteristic of the quality of the models considered, the Accuracy metric was used (unlike Top-1 Accuracy, the model responses are not filtered in any way).

As can be seen in the diagram, the difference in performance for the video question answering and visual question answering tasks is small for Intern-Video and UnMasked Teacher models, but at the same time mPLUG performed better than the other two.

Taking into account the slight difference in the results of the performance evaluation and the fact that InternVideo general video foundation models achieve state-of-the-art performance on 39 datasets [5], it was decided to use it for the solution of the students' action recognition task.

It was also chosen because it was the first to offer a new approach for learning based on a video material. Since the most popular benchmarks for action recognition are small and are oriented on feature analysis on images instead of video and do not take into account the temporal component of the action, the InternVideo approach is a game changer in this area (Fig. 7).

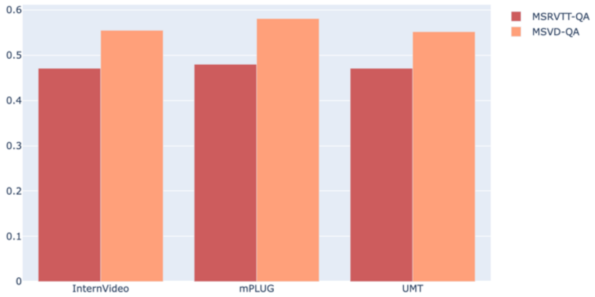


Fig. 7. **Bar chart of all three models performance**
***Source:* compiled by the authors**

InternVideo. This is a video general foundation model [4], one of the features of which is the extension of an image-pretrained vision transformer into video representation learning. Intern-Video adopts the vision transformer (ViT, used variant UniformerV2 [10]), as well as additional local spatiotemporal modelling modules for multilevel representation interaction. It progressively improves its representation by integrating both self-supervised mask-based modelling and multimodal learning and supervised learning.

The model inputs both image and video data and a test description of the events taking place. Three application tasks are handled (Action Understanding [11], Video-Langugae Alignment [12] and Video Open Understanding [5]) and used a set of corresponding datasets to work with each of them (Kinetics [13], ActivityNet [14], MSR-VTT [15], DiDeMo [16], UFC101-HMDB51 [17], Ego4d [18], etc.).

Below is a scheme of interaction using cross-model attention [19] during the model training (Fig. 8). These cross-model attention models are used to learn a unified video representation based on both video masked modeling and video-language contrastive learning. Snowflake means "Freezing a layer" (the same images are run through the same layers without updating the weights, it is a technique to accelerate neural network training [20]) and fire means that the weights will change.

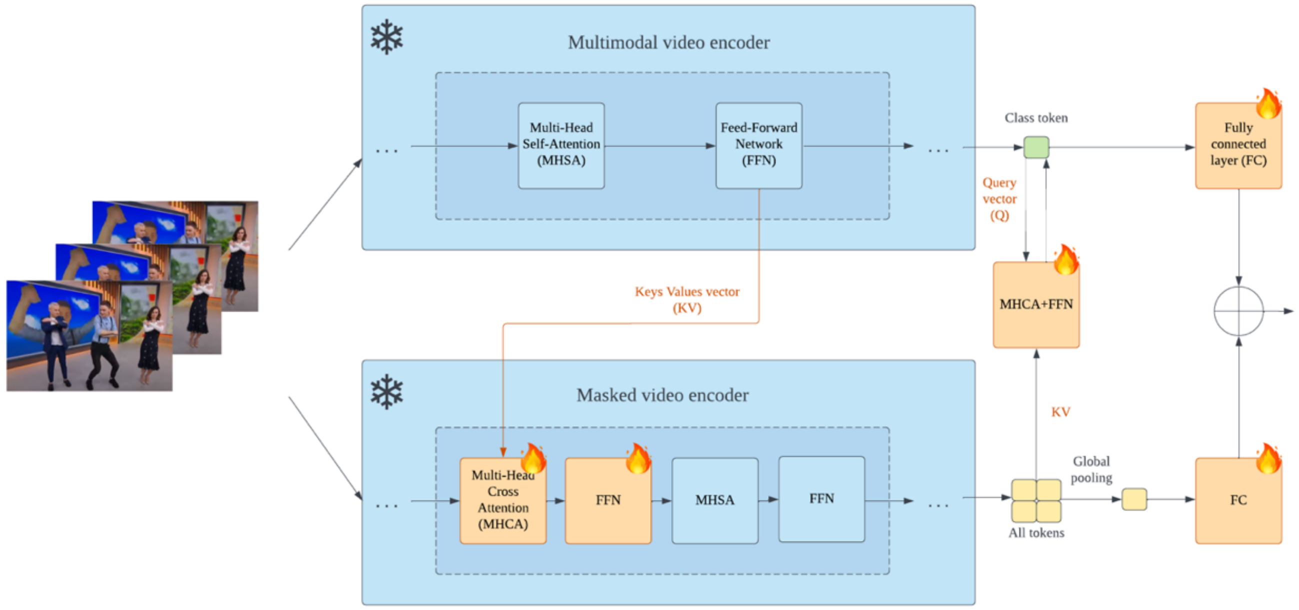


Fig. 8. **The InternVideo model interaction using cross-model attention**
Source: **compiled by the authors**

For specific cases, like artificial intelligence based proctoring systems, there are no datasets in the public domain (Available from: https://paperswithcode.com/datasets).

Common emotion analysis datasets consist of images rather than video sequences (Available from: https://paperswithcode.com/task/facial-expression-recognition). Even if there are datasets with video sequences, they are small (e.g., 242 [25] or 500[26] videos). Due to the lack of temporal analysis of emotion changes, programs can give false positive or false negative responses (the transition period between emotions can be falsely identified as another emotion).

At the same time, for exercise and sports analysis, the emphasis is on analyzing group sports competitions and matches (volleyball, basketball, football), whereas for artificial intelligence based proctoring systems it is the activity of a single person in the context of gymnastics and stress exercises that matters. This is why it is proposed to develop a prototype dataset specializing in the task of student behavior analysis in artificial intelligence based proctoring systems.

IMPLEMENTATION

InternVideo in artificial intelligence based proctoring systems. In the context of using the InternVideo model in distance learning systems, it is proposed to adapt the existing model and implement domain specialization which is a big research area nowadays [21, 22].

These are the points which are necessary to cover it with extra data and functionality (Fig. 9):

– expand its vocabulary with more information on types of physical activity (exercises) for its further suggestion and evaluation;

– expand the existing model with functionality for analyzing student behavior.

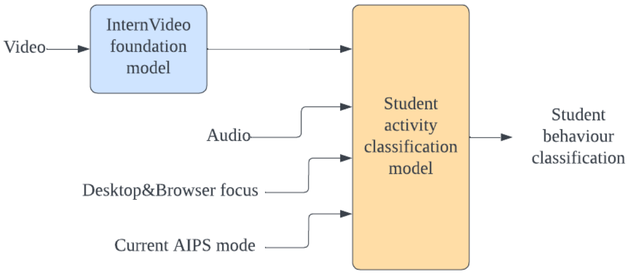

Video
InternVideo foundation model
Audio
Desktop&Browser focus
Current AIPS mode
Student activity classification model
Student behaviour classification

Fig. 9. **The architecture of the student productivity assessment model in artificial intelligence based proctoring systems**
Source: **compiled by the authors**

Physical activity data vocabulary. It is necessary to adapt the existing model by extending it with data on possible PE exercises. In Ukraine, there is an official document describing a set of exercises to maintain health while working with computers (Available from: https://zakon.rada.gov.ua/rada/show/v0007282-98#Text). Also, the Ministry of Education and Science of Ukraine has developed a set of exercises for schoolchildren to reduce stress (Available from: https://mon.gov.ua/ua/osvita/doshkilna-osvita/suchasne-doshkillya-pid-krilami-zahistu/vseukrayinskij-naukovo-metodichnij-zhurnal-doshkilne-vihovannya/ruhankita-vpravi-dlya-znyattya-napruzhennya). In comparison, in other countries (e.g., Portugal and the United Kingdom) the focus is on physical education (PE) classes, active games and outdoor activities.

Therefore, alternative recommendations have been chosen as additional data about possible physical activity indoor (Fig. 10), specifically:

– ten-minute shake up games from NHS (the UK) (Available from: https://www.nhs.uk/healthier-families/activities/10-minute-shake-up);

– instruction for exercises at home for adults (the UK) (Available from: https://www.sfh-tr.nhs.uk/media/9816/pil202207-02-heb-home-exercise-booklet.pdf);

– exercise during quarantine due to COVID-19 from Hospital da Luz (Portugal) (Available from: https://www.hospitaldaluz.pt/pt/saude-e-bem-estar/covid-19-exercicio-quarentena).

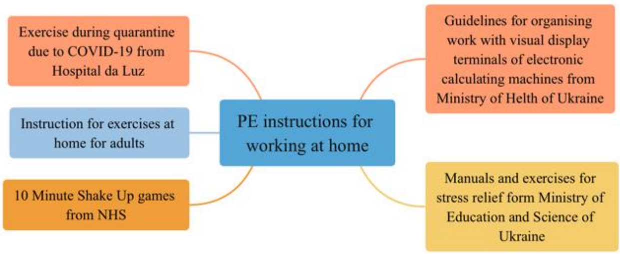

Exercise during quarantine due to COVID-19 from Hospital da Luz
Instruction for exercises at home for adults
10 Minute Shake Up games from NHS
PE instructions for working at home
Guidelines for organising work with visual display terminals of electronic calculating machines from Ministry of Helth of Ukraine
Manuals and exercises for stress relief form Ministry of Education and Science of Ukraine

Fig. 10. **Different sources for indoor physical education instructions**
Source: **compiled by the authors**

Behavior analysis. It is proposed to classify student behavior depending on which of two modes of human behavior analysis is turned on. The static mode (aka “working phase”) is the mode of student behavior analysis during a lesson, when emotions and movements of the student are analyzed and unnecessary changes in body position are considered as bad undesirable behavior.

The dynamic mode (aka “Physical Education (PE) mode”) is a mode of activity analysis during the mandatory exercise sessions that are added to maintain physical and mental health. In this case, the model also describes the student's behavior, but evaluates their performance according to a set of exercises specified in advance by the teacher.

To implement students’ behavior evaluation, it is suggested to extend the usage of the existing InternVideo model with extra training data, since it was originally used for action recognition and video question answering task and not specifically for facial expression recognition.

To reduce computational requirements, it was decided to use low-storage adaptation approach: the approach freezes all model parameters and places new MLP modules with teachable parameters between existing model layers. Additional datasets must be used for post-training. Since InternVideo uses video sequences for training, unlike existing analogues, in order to achieve better results in temporal analysis, a prototype of a dataset for the application area of proctoring systems has been developed.

The data from the proposed prototype dataset was divided into two large classes: video sequences for emotion assessment and video sequences of physical activity. In the context of emotion-assessment video sequences, three classes of emotion state of a student during training were identified based on Action unit classification for facial expression recognition [23, 24], [25, 26] (Fig. 11).

The division of the presented types of emotion (Action units) was carried out as follows:

– concentrated and neutral or positive (Fig. 11a). These sets of emotions are clearly recognizable. A student with these emotions is considered to be concentrated and not in need of emotional release;

– concentrated and negative (Fig. 11b). In this case the student is experiencing negative emotions and the system should offer to do some distracting physical activity, but this behavior does not give many penalty points in the context of the assessment as the student is still focused on work;

– negative (Fig. 11c). With this set of emotions, it is not possible to indicate reliably how well the student has performed, but the system should ask them to do physical exercise to relieve the emotional stress;

– distracted neutral (Fig. 11c). With this set of emotions, the student shows low interest in the learning process, so the system should also ask them to do appropriate physical exercises. The system should also give penalty points for poor performance.

The sources of the video sequences were videos of online conferences, reaction video, vloggers, online lectures, videos with facial expressions and micro expressions. Videos of recreational gymnastics and workouts for the Office were also used. The videos were downloaded from YouTube and divided into one-minute clips. Amazon SageMaker Data Labeling platform was used to implement data labeling.

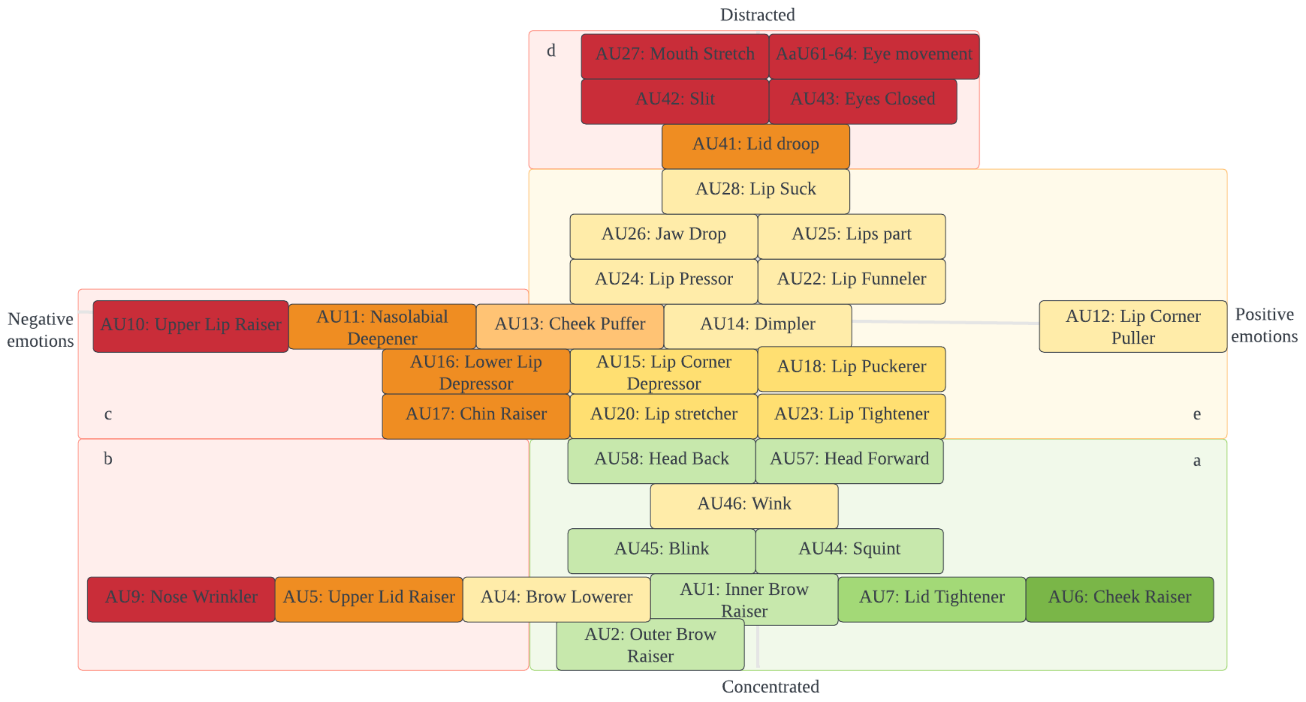


Fig. 11. **Division of Action unit in the context of distance learning:**
a – student is attentive and has positive emotions; b – attentive and negative;
c – not sufficient to understand how attentive a person is but is angry; d – inattentive;
e – not sufficient to assess level of attention
Source: **compiled by the authors**

CONCLUSIONS

The paper proposes a new functional requirement for artificial intelligence based proctoring systems, including the analysis of student's posture and the tracking of necessary student activity to maintain their health, and proposes an adaptation of the existing foundation model to implement these capabilities. For the first time, a dataset specialized for the application domain of proctoring systems has been prototyped. It uses a data set with video sequences, as opposed to standard image analysis, in the context of solving the facial expression recognition problem.

Several existing software solutions for distance learning have been analyzed. This analysis showed that despite the development of image processing and ML technologies, the functionality of current proctoring systems is still at a primitive level. They lacked a comprehensive approach to analyzing student behavior in class and did not include support for emotional and physical health of the student. Assessment of physical activity is necessary not only as part of the learning process, but also to keep students healthy in accordance with regulatory requirements, as they spend the entire learning process in front of computers or other devices during distance learning. In existing implementations, this process still falls entirely on the shoulders of teachers or even the students themselves to work through the lesson or test materials on their own.

Based on an analysis of several existing basic models for human action recognition, the InternVideo model was chosen, which has been shown to be state-of-the-art in 39 datasets in the tasks of action recognition and classification and video sequence characterization.

As a result, the InternVideo model was adopted for use in student action analysis. It includes two modes of student action assessment in distance learning: static and dynamic. The static mode (aka working stage) analyses and assesses student behavior during the teaching and examination process, where physical activity is not a major component of learning. The dynamic mode (aka physical education (PE) mode) analyses and assesses the student who performs physical activity purposefully (PE lesson, children's exercises during the lesson, etc.).

Based on an analysis of existing datasets, a lack of datasets for the application area under study was identified, as well as a lack of extensive datasets with videos for analyzing people's emotions. For this purpose, a prototype dataset specifically for proctoring systems was proposed, containing both a video for emotion analysis and a video for analysis of a person's activity.

REFERENCES

1. Nigam, A., Pasricha, R., Singh, T. & Churi, P. "A systematic review on AI-based proctoring systems: past, present and future". *Education and Information Technology*. 23 June 2021; Vol. 26 No 5: 6421–6445. DOI: https://doi.org/10.1007/s10639-021-10597-x.

2. Nigam, A., Pasricha, R., Singh, T. & Churi, P. "Correction to systematic review on AI-based proctoring systems: Past, present and future". *Education and Information Technologies*. June 2021; Vol. 27 No 5: 7377–7378. DOI: https://doi.org/10.1007/s10639-021-10846-z.

3. Motwani, S., Nagpal, C. & Motwani, M. "AI-Based proctoring system for online tests". *4th International Conference on Advances in Science & Technology (ICAST2021)*. 15 June 2021. DOI: https://dx.doi.org/10.2139/ssrn.3866446.

4. Yasunaga, M., Leskovec, J. & Liang, P. "LinkBERT: pretraining language models with document links". 29 March 2022. p. 8003–8016. DOI: https://doi.org/10.48550/arXiv.2203.15827.

5. Bommasani, R., Drew A. Hudson, Adeli, E., Altman, R., Arora, S. & others "On the opportunities and risks of foundation models", 2021. DOI: https://doi.org/10.48550/arXiv.2108.07258.

6. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L. & Qiao, Y. "Intern video: General video foundation models via generative and discriminative learning". 2022. DOI: https://doi.org/10.48550/arXiv.2212.03191.

7. Xu, H. Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F. & Zhou, J. "mPLUG-2: A modularized Multi-modal foundation model across Text, Image and Video". 2023. DOI: https://doi.org/10.48550/arXiv.2302.00402.

8. Li, K, Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., Qiao, Y. "Unmasked teacher: towards training-efficient video foundation models". 2023. DOI: https://doi.org/10.48550/arXiv.2303.16058.

9. Antoshchuk, S. & Breskina, A. "Evaluation metrics systematization for 2D human poses analysis models". *Herald of Advanced Information Technology*. April 2023; Vol. 6 No 1: 26–38. DOI: http://dx.doi.org/10.15276/hait.06.2023.2.

10. Li, K., Wang, Y., He, Y., Li, Y., Wang, L., Wang, L. & Qiao, Y. "UniFormerV2: Spatiotemporal Learning by arming image ViTs with Video UniFormer". 2022. DOI: https://doi.org/10.48550/arXiv.2211.09552.

11. Hutchinson, M. & Gadepally, V. "Video Action Understanding". *IEEE Access*. 24 September 2021; Vol. 2: 134611–134637. DOI: https://doi.org/10.1109/ACCESS.2021.3115476.

12. Huang, J., Li, Y., Feng, J., Wu, X., Sun, X. & Ji, R. "Clover: Towards A unified video-language alignment and fusion model". *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. p. 14856–14866. DOI: https://doi.org/10.48550/arXiv.2207.07885.

13. Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A. & Zisserman, A. "A Short Note on the Kinetics-700-2020 human action dataset". 2020. DOI: https://doi.org/10.48550/arXiv.2010.10864.

14. Heilbron, F. C., Escorcia, V., Ghanem, B. & Niebles, J. C. "ActivityNet: A large-scale video benchmark for human activity understanding". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 15 October 2015. p. 961–970. DOI: http://dx.doi.org/10.1109/CVPR.2015.7298698.

15. Tan, G. Liu, D. Wang, M. & Zha, Z. "Learning to discretely compose reasoning module networks for video captioning". *IJCAI'20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. 2021. p. 745–752. DOI: https://doi.org/10.48550/arXiv.2007.09049.

16. Mithun, N. C., Paul, S. & Roy-Chowdhury, A. K."Weakly supervised video moment retrieval from text queries". *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019. p. 11592–11601. DOI: https://doi.org/10.48550/arXiv.1904.03282.

17. Wang, L. Qiao, Y. & Tang, X. "Action recognition with trajectory-pooled deep-convolutional descriptors". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 15 October 2015. p. 4305–4314. DOI: https://doi.org/10.1109/CVPR.2015.7299059.

18. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J. and others "Ego4D: Around the world in 3,000 hours of egocentric video". *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 24 June 2022. p. 18995–19012. DOI: https://doi.org/10.48550/arXiv.2110.07058.

19. Chen, Y., Yuan, J., Zhao, L. Chen, T., Luo, R., Davis, L. & Metaxas, D. N. “More than just attention: Improving cross-modal attentions with contrastive constraints for image-text matching”. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* May 2023. p. 4432–4440. DOI: https://doi.org/10.48550/arXiv.2105.09597.

20. Yuan, G., Li, Y., Li, S., Kong, Z., Tulyakov, S., Tang, X., Wang, Y. & Ren, J. “Layer freezing & data sieving: missing pieces of a generic framework for sparse training”. September 2022. DOI: https://doi.org/10.48550/arXiv.2209.11204.

21. Deng, B. & Jia, K. “Universal domain adaptation from foundation models”. 2023. DOI: https://doi.org/10.48550/arXiv.2305.11092.

22. Li, B., Hwang, D., Huo, Z., Bai, J., Prakash, G., Sainath, T. N., Sim, K. C., Zhang, Y., Han, W., Strohman, T. & Beaufays, F. “Efficient domain adaptation for speech foundation models”. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 05 May 2023. DOI: https://doi.org/10.48550/arXiv.2302.01496.

23. Benitez-Quiroz, C. F., Srinivasan, R. & Martinez, A. M. “EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild”. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 12 December 2016. p. 5562–5570. DOI: https://doi.org/10.1109/CVPR.2016.600.

24. Mollahosseini, A. Hasani, B. & Mahoor , M. H.“AffectNet: A database for facial expression, valence, and arousal computing in the wild”. *IEEE Transactions on Affective Computing.* Jan,-March Jan.-March 2019; Vol. 10: 18–31. DOI: https://doi.org/10.1109/TAFFC.2017.2740923.

25. McDuff, D., Kaliouby, R. E., Senechal, T., Amr, M., Cohn, J. F. & Picard, R. “Affectiva-MIT facial expression dataset (AM-FED): naturalistic and spontaneous facial expressions collected in-the-wild”. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 12 September 2013. DOI: https://doi.org/10.1109/CVPRW.2013.130.

26. Zafeiriou, S., Papaioannou, A., Kotsia, I. Nicolaou, M. & Zhao, G. “Facial affect “In-the-Wild”: A survey and a new database”. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* 19 December 2016. 36–47. DOI: https://doi.org/10.1109/CVPRW.2016.186.

Conflicts of Interest: the authors declare no conflict of interest

Received 10.04.2023
Received after revision 08.06.2023
Accepted 19.06.2023

DOI: https://doi.org/10.15276/aait.06.2023.14

УДК 004.932

Моделі аналізу дій людини в системах прокторингу на основі штучного інтелекту та набори даних для них

Антощук Світлана Григорівна[1)]
ORCID: https://orcid.org/0000-0002-9346-145X; asg@op.edu.ua. Scopus Author ID: 8393582500
Бреськіна Анастасія Аршавирівна[1)]
ORCID: https://orcid.org/0000-0002-3165-6788; anastasia.breskina@gmail.com
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

У цій статті описано підхід до побудови спеціалізованої моделі для аналізу дій людей у системах прокторингу на основі штучного інтелекту та запропоновано прототип набору даних, який містить дані, специфічні для прикладної області. Стрімкий розвиток технологій машинного навчання, доступність пристроїв та доступ до Інтернету швидко розвивають сферу дистанційного навчання. Паралельно з системами дистанційного навчання розвиваються і системи автоматизованого навчання, які забезпечують функціональний аналіз роботи студентів, імітуючи оцінювання викладачем. Однак, незважаючи на розвиток технологій обробки зображень та машинного навчання, функціональність сучасних систем прокторингу все ще залишається на примітивному рівні. В рамках функціоналу обробки зображень вони повністю зосереджені на відстеженні обличь студентів і не відстежують пози та дії. У той же час, оцінка фізичної активності необхідна не тільки в рамках навчального процесу, але й для збереження здоров'я студентів згідно з нормативними вимогами, оскільки під час дистанційного навчання вони проводять весь навчальний процес за комп'ютером або іншими пристроями. В існуючих реалізаціях цей процес повністю лягає на плечі викладачів або навіть самих студентів, які самостійно опрацьовують матеріали уроку або тести. При цьому викладачам доводиться або встановлювати контакт через системи відеозв'язку та соціальні мережі (TikTok, Instagram) та/або аналізувати відеозаписи виконання студентами певних фізичних вправ для організації оцінювання фізичної активності. Відсутність такої функціональності в системах прокторингу на основі штучного інтелекту уповільнює навчальний процес та потенційно шкодить здоров'ю учнів у довгостроковій перспективі. У цій статті представлено додаткові вимоги до функціональності систем прокторингу на основі штучного інтелекту, зокрема функціональність аналізу дій людини для оцінки фізичної активності та контролю за дотриманням гігієнічних правил роботи з комп'ютером під час навчального процесу. Для цього було використано фундаментальну модель InternVideo для обробки та аналізу дій студента. На її основі було розроблено підхід до побудови спеціалізованої моделі для аналізу дій студента. Він включає в себе два режими оцінки діяльності студента в процесі дистанційного навчання: статичний і динамічний. Статичний режим (так звана фаза роботи) аналізує та оцінює поведінку студента під час навчального та екзаменаційного процесу, де фізична активність не є основним компонентом навчання. У динамічному режимі (так званий режим фізкультури) аналізується та оцінюється студент, який цілеспрямовано виконує фізичну активність (урок фізкультури, вправи для дітей під час уроку тощо). Також запропоновано прототип набору даних, розроблений спеціально для цієї прикладної області.

Ключові слова: Комп'ютерний зір; нейронні мережі; набір даних; трансформер; розуміння дій; розуміння відео; системи прокторингу на основі штучного інтелекту; онлайн-прокторинг; система прокторингу; дистанційне навчання; онлайн-навчання

ABOUT THE AUTHORS

Svitlana G. Antoshchuk – Doctor of Engineering Sciences, Professor, professor of Department Information Systems. Odessa Polytechnic National University, 1, Shevchenko Ave. Odessa, 65044, Ukraine
ORCID: https://orcid.org/0000-0002-9346-145X; asg@op.edu.ua. Scopus Author ID: 8393582500
Research field: Automated management systems and progressive information technologies; information technology; theoretical and applied aspects of multimedia data processing

Антощук Світлана Григорівна – доктор технічних наук, професор кафедри Інформаційних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1, Одеса, 65044, Україна

Anastasiia A. Breskina – PhD Student of Information Systems Department. Odessa Polytechnic National University. 1, Shevchenko Ave. Odessa, 65044, Ukraine
ORCID: https://orcid.org/0000-0002-3165-6788; anastasia.breskina@gmail.com
Research field: Information technology; computer vision; deep Learning

Брескіна Анастасія Аршавирівна – аспірант кафедри Інформаційних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1, Одеса, 65044, Україна