EmoStudent: Developing a Dataset to Analyse Students' Emotional Well-Being

Svitlana Antoshchuk and Anastasiia Breskina
Department of Information Systems, Institute of Computer Systems, Odesa Polytechnic National University, Shevchenko Avenue 1, Odesa, Ukraine
asg@op.edu.ua, anastasiia.breskina@op.edu.ua

Keywords: Dataset, Computer Vision, Emotion Understanding, Artificial Intelligence-Based Proctoring Systems.

Abstract: This article introduces an initial version of a dataset designed to educate and assess models that concentrate on studying of the emotional condition of students throughout the remote learning process. This dataset comprises short video clips showing the faces of individuals from diverse ethnic backgrounds and age groups in front of the computer screen. No dataset specialising in the proctoring systems problem solving task was found (process of working in front of the computer, emotions during the educational process). As a result, existing datasets for solving problems in related fields were analysed: emotion classification, emotion recognition, and face recognition. Building on this analysis and the specifics of the chosen data source (YouTube videos with Creative Commons license), the previously established criteria for creating the dataset were modified and expanded. A more adaptable approach was introduced concerning the categorization based on age and ethnicity. A path for future endeavors was also delineated, proposing an enhancement of the current implementation to encompass a broader spectrum of emotions and individuals with various forms of disabilities in subsequent iterations.

1 INTRODUCTION

Throughout the history of distance learning technologies there has always been a challenge in evaluating students' behavior and integrity during exams and individual assignments. To tackle this issue, proctoring systems were introduced. These proctoring systems serve as information systems aimed at supervising test or exam completion and monitoring and assessing students' honesty. Essentially, they take on the role of a teacher by observing and evaluating student conduct.

Initially, synchronous proctoring systems relied on human observers, such as teachers or hired staff, to monitor students in real-time. However, in recent years, with the advancement of artificial intelligence (AI) methods and models, there is a possibility to automate the assessment of student integrity [1, 2].

However, students encountered a lot of problems with these systems [1, 3]. Especially the computer vision modules were the subject of many complaints.

The problem was that the systems both had excessive requirements for students' behavior (e.g., the requirement not to take their eyes off the monitor for more than a certain number of seconds) and incorrectly detected the activity (e.g., students of different ethnicities). To address these problems, new rules for online proctoring systems and requirements for student behavior during online testing were proposed [3] to make the systems more humanoriented. More specifically, it was proposed to implement a complex analysis of student behavior, part of which is the analysis of the student's emotional state.

This paper focuses on datasets that are used to train and evaluate machine learning models that used for tasks such as emotion classification, emotion recognition and face recognition. Particularly, it delves into the implementation of an initial implementation of a dataset designed to train and evaluate models that are focused on analysing the emotional state of students during the distance learning process.

2 LITERATURE REVIEW

The task of face recognition is not a new one. There are many datasets created to train and evaluate models that solve these problems. These datasets can be divided into two broad types: image datasets and video datasets.

The most numerous of these are datasets consisting of a set of images of people or just their faces. They are mainly used for face recognition tasks and are characterized by large and high-quality sampling. Some of the well-known datasets relevant to face recognition task are as follows:

- COCO-WholeBody;
- WIDER FACE;
- FDDB (the Face Detection Dataset and Benchmark);
- AFW (Annotated Faces in the Wild);
- PASCAL FACE.

COCO-WholeBody [4] is a modified version of the COCO dataset, enriched with whole body annotations. This enhanced dataset includes four distinct types of bounding boxes for each person depicted in the images: face frame, upper body frame, left arm frame, and right arm frame. Additionally, it provides a substantial number of key points – 133 in total-comprising 17 points for the body, 6 points for the legs, 68 points for the face, and 42 points for the arms.

The WIDER FACE dataset [5] is used for face recognition is derived from the publicly accessible WIDER dataset. It comprises a total of 32,203 images and includes a large number of dummy 393,703 faces. The dataset exhibits significant diversity in terms of scale, pose, and occlusion, presenting a wide range of challenges for face recognition algorithms.

FDDB [6] is a dataset comprising labeled faces extracted from the Faces in the Wild dataset. It contains a total of 5171 face annotations, with images having various resolutions such as 363x450 and 229x410. The dataset encompasses a wide spectrum of challenges, including faces captured at difficult angles, defocused faces, and low-resolution images. Notably, both grayscale and color images are present in this dataset, offering a diverse and comprehensive set of face detection scenarios for evaluation and benchmarking purposes.

AFW [7] is a specific face detection dataset comprising 205 images that collectively contain 468 annotated faces.

The PASCAL FACE dataset [8] is designed for face detection and recognition purposes. It includes 851 images, which are extracted from the larger PASCAL VOC dataset, and encompasses a total of 1341 annotations for faces. However, it's important to note that this dataset is relatively small, containing only a few hundred images, and lacks diverse variations in the appearance of faces. As a result, it may offer limited coverage of real-world scenarios and may not fully represent the wide range of challenges faced in face detection and recognition tasks.

A limitation of these datasets is the lack of a temporal component, which limits the process of analyzing video sequences for human emotional state. To address this problem, video sequence datasets have been created. These include, for example, such sets and benchmarks as:

- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song);
- AM-FED; CK+ (the Cohn-Kanade database);
- MMI;
- UNBC-McMaster Pain;
- EngageNet.

RAVDESS [9] is a comprehensive dataset containing 7356 files. Within this database, there are recordings from 24 professional actors, equally divided between 12 females and 12 males. The actors provide two lexically comparable utterances in a neutral North American accent.

AM-FED [10] dataset consists of 42 webcam videos that have been recorded in authentic realworld settings. This dataset captures various scenarios and conditions that people commonly encounter when using webcams, making it an ideal resource for researchers and developers interested in real-life webcam-related applications and analysis.

CK+ [11] is a highly popular for advancing facial action and expression recognition systems. Within the CK+ database, there are 593 records of both posed and unposed sequences, meticulously encoded using the Facial Action Coding System (FACS) [12]. Moreover, the sequences are also encoded to represent the six fundamental emotions.

The MMI [13] dataset encompasses a vast collection of face videos, each encoded using the FACS. Specifically, it comprises a total of 1,395 video sequences, with each one manually encoded by identifying the Action Units (AUs) present in the facial expressions;

UNBC-McMaster Pain [14] is considered one of the most extensive collections of AU-encoded video sequences showcasing natural and spontaneous facial expressions. With a total of 10 action units, this

dataset offers a wealth of valuable information. However, it's important to note that even though the expressions are natural and spontaneous, the videos were recorded under controlled conditions..

EngageNet [15] is a dataset comprising a total of 31 hours of recorded data from 127 participants. This dataset captures various individuals under different illumination conditions, providing a diverse and comprehensive representation of facial expressions and interactions.

When it comes to datasets created specifically for the task of face recognition and emotional state recognition of students, there are even fewer of them. Such datasets usually consist of image sets. For example: Face dataset by Generated Photos (Face dataset for Academics by Generated Photos); the CUHK Face Sketch (CUFS).

Face dataset by Generated Photos [16] consists of 10,000 synthetic photos, each carefully balanced with diverse representations of race and gender. It also includes essential metadata and facial landmarks for research purposes. It's important to note that all the photos in this dataset are entirely synthetic, generated based on models, and not real photographs.

CUFS [17] database encompasses a diverse collection of 188 faces from the Chinese University of Hong Kong (CUHK) student database, along with 123 faces from the AR database and 295 faces from the XM2VTS database.

The analysis of existing solutions showed that existing datasets for face recognition and emotion analysis are not suitable for proctoring task for several reasons. Firstly, the video datasets represented people of limited age and ethnicity (usually young adults of Northern European or Asian appearance). Secondly, datasets with only images lose the temporal component. Some of the datasets consist of recordings of professional actors' reactions rather than emotions in the wild;

3 EMOSTUDENT DATASET

The aim of the paper is to develop the first iteration of a dataset that contains videos with emotions in the wild for training and evaluating a model of analyzing students' emotional state during distance learning.

During the conceptualization of the architecture of the automated online proctoring system [3], a prototype dataset was implemented, which is oriented specifically for solving problems within distance learning and online proctoring systems. Further work in this area revealed a number of limitations that led to the refinement of the dataset requirements.

The dataset was sourced from YouTube, specifically from videos that possess a Creative Commons license. The use of this license does not guarantee that there will be non-professional actors in the video. However, in order to get as close to the emotions "in the wild" as possible, it was decided to choose videos with interviews and amateur bloggers. To simplify the labelling process and facilitate dataset creation, the Amazon SageMaker platform was used.

3.1 Requirements for Data

To address the challenge of accurately determining the age of individuals in video data, a different approach was proposed. Instead of categorizing the sample based on specific age intervals, the division was made into broader groups: teenagers, young adults, and adults. This decision was taken due to the difficulty of precisely identifying individual ages from video footage.

The minimum age of children requirement remains the same due to the regulations and policies of various Internet platforms, which often impose a minimum age restriction of 13 years to ensure child safety and compliance with legal guidelines.

To formulate the proportions by age, it is proposed to take as a benchmark the data from the Education Data Initiative project, which collects data and statistics on the education system in the United States. Base on that data 66% of students are under 24 years old, 22.9% are young adults and 10.5% are adults.

The problem of obtaining data depending on people's ethnicity has also arisen. This is because people from different parts of the world have different levels of accessibility to Internet connection and portable devices and cameras. This caused a problem with detailed representation of different nationalities and ethnicities.

For simplicity, it was chosen to divide the data by ethnicity and region of residence: Africa, Caribbean, Asia-East, Asia-South, Europe-North, EuropeCentral, Europe-South, Latin America, Caucasus, Middle East, and Natives.

In the context of separation by region of residence, it is not recommended to form any dependency on the distribution of real data, as it is irrelevant in the context of solving the proctoring task. Instead, it is suggested to form a sample with an equal number of videos to represent each region.

Representation of people of different sex is also suggested to be done with equal representation.

To evaluate student engagement during the learning process, modified 0-3 grading scale designed for assessing student mindfulness [18] was used.

A score of 0 on the engagement scale denotes complete disengagement, where the subject exhibits minimal interest, frequently diverts their attention away from the screen, and may engage in active physical activities. Represented by emotions such as fear, overexcitement, anger. This rating indicates a lack of involvement or attentiveness during the observed activity.

A rating of 1 suggests minimal engagement, indicating that the subject shows only slight interest or involvement. They may barely open their eyes, look at the screen and display restlessness in their chair, suggesting a low level of. Represented by emotions such as bored.

It is proposed to change the essence of the score 2. It will indicate that the subjects are engaged in the content, show interest and interact with the proctoring system, but are in a stressful state. This includes negative emotions (anger, fear, sadness, disgust) combined with direct gaze at the monitor;

A rating of 3 on the engagement scale represents a high level of engagement. This indicates a strong sense of attentiveness and active involvement in the observed activity, showing a clear indication of the subject's deep interest and concentration. Represented by emotions such as neutral and joyful.

3.2 Characteristics of the Developed Dataset

Based on that final requirements for dataset include people from 11 regions; people of different age groups: teenagers, youth and adults. different type of engagement assessment, each of which is respectively described by a different type of student's emotional state.

In this dataset, there were originally 334 video files, each with a duration ranging from one minute to two and a half hours (Table 1). All of them have been split into small video sequences of maximum 1 minute length and 740p size. Average video sample length is 20 seconds; minimum number of the original video splits is from 1 to 50. This resulted in a total of 2816 videos.

The video sequences in this dataset capture individuals' facial expressions, with each recording featuring one person on stage. These individuals exhibit a range of emotional states, including basic emotions such as joy, neutral emotions, anger, sadness, and overexcitement. The dataset provides valuable insight into how these various emotions are expressed through facial cues and expressions, making it a valuable resource for studying and analyzing human emotions in different contexts “in the wild”.

For the labeling process, a group of three annotators was tasked with viewing and assessing the videos, paying particular attention to the level of engagement depicted through facial expressions. This approach allowed for a comprehensive evaluation of the subjects' attentiveness and involvement during the videos, as reflected in their facial expressions and movements. In the created dataset, the distribution of scores is approximately 30% zero mark, 10% one mark, 25% and 35% are two and three marks respectively.

Table 1: Original set of video sequences, categorized by ethnicity and age of people.

Ethnicity (location)/Age	Teen	Young Adult	Adult
Africa	9	10	11
Caribbean	3	10	12
Asia-East	20	25	35
Asia-South	7	4	7
EuropeanNorth	12	9	17
EuropeanSouth	10	10	13
EuropeanCentral	5	2	10
Latin America	5	5	13
Caucasus	7	10	12
Middle East	7	7	14
Natives	1	2	10
Total	86	94	154

The video sequences in this dataset capture individuals' facial expressions, with each recording featuring one person on stage. These individuals exhibit a range of emotional states, including basic emotions such as joy, neutral emotions, anger, sadness, and overexcitement. The dataset provides valuable insight into how these various emotions are expressed through facial cues and expressions, making it a valuable resource for studying and analyzing human emotions in different contexts “in the wild”.

For the labeling process, a group of three annotators was tasked with viewing and assessing the videos, paying particular attention to the level of engagement depicted through facial expressions. This approach allowed for a comprehensive evaluation of

the subjects' attentiveness and involvement during the videos, as reflected in their facial expressions and movements. In the created dataset, the distribution of scores is approximately 30% zero mark, 10% one mark, 25% and 35% are two and three marks respectively.

The audio for the videos was disabled to avoid any auditory influence since in the obtained sample the video sometime contains ambient sounds rather than the speech of the person on the video themselves.

3.3 Current Challenges and Further Plans

The existing dataset implementation faces a few challenges that need to be addressed. One of the issues is the heterogeneous nature of the data, especially concerning shot quality. The videos have been rescaled to a standard 740p size, but some of them still have lower quality compared to others.

Additionally, although efforts were made to include people from various ethnic backgrounds, there is currently an underrepresentation of certain ethnic groups, specifically "Natives" and "European" (especially "European-Central"). There is also a lack of representation for individuals with different types of disabilities.

To address these shortcomings and enhance the dataset's comprehensiveness, the next versions of the dataset implementation are planned to rectify these issues. Steps will be taken to increase the representation of individuals with different types of disabilities and expand the coverage of different ethnicities. These improvements aim to create a more inclusive and comprehensive dataset that better reflects real-world scenarios.

4 CONCLUSIONS

In this research paper, an initial dataset implementation designed to train and assess models with a focus on analyzing the emotional state of students during distance learning was introduced, also several most widely known datasets were reviewed. A notable aspect of this dataset is its inclusion of short video sequences capturing the faces of individuals from diverse ethnic backgrounds and various age groups, all within natural environments.

To create this dataset, a thorough analysis of various existing datasets commonly used for tasks like emotion classification, emotion recognition, and face recognition was made. Given the specific of the chosen data source (YouTube videos with Creative Commons licenses) and its limitations, certain adaptations and expansions were implemented to the existing criteria for dataset creation. Notably, a more adaptable approach was introduced when categorizing data based on age and ethnicity.

This study also presents potential avenues for future research. Specifically, efforts should be made to ensure a balanced representation of individuals with different ethnic background. Additionally, in forthcoming iterations of the dataset, it is planned to include data about individuals with different types of disabilities that can impact their emotional well-being evaluation. By including a wider range of people from different backgrounds, the dataset can contribute to a more accurate and in-depth understanding of emotional reactions in distance learning.

REFERENCES

[1] A. Nigam, R. Pasricha, T. Singh, and P. Churi, "A systematic review on AI-based proctoring systems: Past, present and future," Education and Information Technologies, vol. 26, pp. 6421-6445, Dec. 2021, doi: 10.1007/s10639-021-10597-x.

[2] S. Motwani, C. Nagpal, M. Motwani, N. Nagdev, and A. Yeole, "AI-Based proctoring system for online tests," Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021), Jun. 2021, doi: 10.2139/ssrn.3866446.

[3] A. Breskina, "Development of an automated online proctoring system," Herald of Advanced Information Technology, vol. 6, no. 2, pp. 163-173, 2023, doi: 10.15276/hait.06.2023.11.

[4] S. Jin et al., "Whole-Body human pose estimation in the wild," ECCV 2020: Computer Vision – ECCV, 2020, pp. 196-214, 2020, doi: 10.48550/arXiv.2007.11858.

[5] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 5525-5533, 2016, doi: 10.1109/CVPR.2016.596.

[6] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings.," Technical Report UM-CS-2010-009, Dept. Of Computer Science, University of Massachusetts, Amherst, 2010.

[7] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, pp. 2879-2886, 2012, doi: 10.1109/CVPR.2012.6248014.

[8] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," Image and Vision Computing, vol. 32, no. 10, pp. 790-799, 2014, doi: 10.1016/j.imavis.2013.12.004.

[9] S. R. Livingstone and F. A. Russo, "The ryerson audiovisual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," PLoS ONE, vol. 13, no. 5, May 2018, doi: 10.1371/journal.pone.0196391.

[10] D. McDuff, R. E. Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected "in-thewild"," 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, OR, USA,, pp. 881-888, 2013, doi: 10.1109/CVPRW.2013.130.

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended CohnKanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, pp. 94-101, 2010, doi: 10.1109/CVPRW.2010.5543262.

[12] E. A. Clark et al., "The facial action coding system for characterization of human affective response to consumer product-based stimuli: A systematic review," Front Psychol, May 2020, doi: 10.3389/fpsyg.2020.00920.

[13] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands, pp. 5, 2005, doi: 10.1109/ICME.2005.1521424.

[14] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA, pp. 57-64, 2011, doi: 10.1109/FG.2011.5771462.

[15] M. Singh, X. Hoque, D. Zeng, Y. Wang, K. Ikeda, and A. Dhall, "Do I have your attention: A large scale engagement prediction dataset and baselines," 2023. doi: 10.48550/arXiv.2302.00431.

[16] "Face dataset by generated photos (face dataset for academics by generated photos)," [Online]. Available: https://generated.photos/ datasets#research-dataset, [Accessed Jul. 3, 2023].

[17] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," CVPR 2011, Colorado Springs, CO, USA, pp. 513-520, 2011, doi: 10.1109/CVPR.2011.5995324.

[18] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, ACT, Australia, p. 1-8, 2018, doi: 10.1109/DICTA.2018.8615851.