Adaptive intelligent agent for e-learning: First report on enabling technology solutions

Dora Doljanin*, Luka Pranjić*, Ljudevit Jelečević* and Marko Horvat*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Applied Computing
marko.horvat3@fer.hr

Abstract – Because of the global COVID-19 pandemic, online learning has become the dominant teaching method. Moreover, a wide range of e-learning pedagogies are rapidly gaining importance, and in some cases emerging as the preferred approach in education over the traditional methods and techniques of classroom teaching. However much has to be done to efficiently assess student engagement and the learning curve. In this regard, we have proposed construction of an intelligent agent for personalized and adaptive assessment of learning performance based on methods for automated estimation of attention and emotion. We report on the first progress towards the development of the intelligent agent. Three classifiers were used in parallel to detect information about the progress of student engagement. Object detection in video is accomplished with YOLOv3, emotion detection from facial expressions using PAZ software library, and detection of head, arms, and upper-body orientation and position with OpenPose system. NimStim facial expression database, WIDER Attribute Dataset, and UPNA Head Pose Database were used for experimental validation of the individual classifiers. Our system attained the highest precision and recall of 79.13% and 94.15%, respectively, and the highest success rate of 59.56% in recognition of 6 discrete emotions from facial expressions.

Keywords – digital learning; adaptive learning; emotion recognition; pose estimation; object detection

I. INTRODUCTION

During the COVID-19 pandemic online learning has emerged as the predominant education model. Formal higher education institutions such as universities and colleges, but also professional education schools and others, have largely shifted their academic activities online and away from customary methods of teaching in a classroom. Most likely distance learning will remain a significant education paradigm after the COVID-19 crisis. This situation opens a possibility to envision new systems for automated intelligent assessment of students' vigilance, motivation, and attention during online classes, thus enabling customization of individual learning curves for each student.

Previously we have described a concept of one such system of an adaptive intelligent agent for e-learning, justified its development, and envisioned which tools could be employed to develop it in practice [1]. Such agent could reveal hidden information about the students' learning curves and help educators to better focus on outliers, either helping in certain indicated areas those who are less successful or brining additional and targeted course materials to talented students. Also, teaching personalization is another important goal for the proposed agent.

In this paper we are reporting our first findings on capabilities of three enabling technologies: human pose estimation, emotion recognition from facial expressions and person detection in images.

Our motivation is to establish if contemporary video conferencing tools and image recognition APIs are capable of resolving required features in sufficient detail regarding spatial, field-of-view, color-depth and temporal constraints. This first report on the three enabling technologies, and individual experimentation with each classifier, should be followed by a larger and more complex experiment of the integrated system in laboratory settings.

The remainder of this paper is organized as follows; Section 2 describes significance and application of body posture estimation, the OpenPose subsystem, the dataset that was developed for testing the subsystem and its experimental validation. Sections 3 illustrates facial expression analysis and emotion recognition with the PAZ software library and NimStim database. Section 4 provides information about the person detection experiment in pictures with YOLOv3 software kit. Finally, Section 6 concludes the paper and provides insight into the planned future development of the adaptive intelligent agent for e-learning.

II. POSE ESTIMATION

In order to efficiently assess student engagement and the learning curve, the intelligent agent should collect data from students' web cameras during an online class. Monitoring visual signals such as facial expressions, gaze, head and body posture, gestures and hand movements may provide an insight into students' involvement and attention during lessons, enabling to further assess their learning performances based on their behavioral patterns [2] [3].

Body language plays an important role in nonverbal communication, including between students themselves and between students and teachers. For example, leaning forward and taking notes signals a higher interest and engagement, whereas leaning back, yawning, supporting head and looking away are associated with low level of

attention [2]. Sustained attention has been recognized as an important factor of the learning success [4]. Hence, monitoring students' body language, position and posture using a pose estimation classifier may be of great value in the assessment of students' attention during online classes. Several papers describing similar systems exist in literature. A convolutional neural network architecture has been created for unobtrusive students' engagement analysis using non-verbal cues such as face expressions, hand gestures and body postures, trained and tested in the wild of more than 350 students present in a classroom environment [5][6]. Student's facial expressions and body postures were captured using iPad's web camera and classified students' engagement into four different engagement levels [7]. A system using artificial neural networks has been built in order to classify behavior among kindergarten students in e-learning environment using a spatio-temporal model from sequences of digital images [8]. Kinect One sensor was used for classifying the students' attention using facial expression, eye gaze, and body posture [2]. Information from multiple input modalities such as webcam and mouse were also used for effective human attention detection [3]. Students' attention has been successfully modelled using eye tracking sensors and machine learning [9].

We created new pose estimation dataset consisting of 16 pictures capturing different and characteristic human body positionings that may often be seen in online learning environments. The pictures present a variety of human torso poses within a web camera field of view that contains many realistic challenges, including various degrees of head rotation, body rotation, partly visible body parts, interfering body parts, poor lighting, varying distance from the camera etc. The dataset will be used in our subsequent research to test the robustness of the pose estimation method in critical scenarios that are highly likely to occur in online education.

A. *Pose estimation experiment*

Our pose estimation method uses the OpenPose 18 keypoint detection model [10]. Each keypoint was given a corresponding unique ID number, as shown in Table 1. The system input is an image and the output are the 2D locations of anatomical keypoints of each person recognized in the picture, appropriately connected into bodyparts [10]. The results of pose estimation classifier are shown in Fig. 1.

TABLE I. BODY POSTURES KEYPOINTS ENUMERATION.

Keypoint Name	ID
Nose	0
Neck	1
Right Shoulder	2
Right Elbow	3
Right Wrist	4
Left Shoulder	5
Left Elbow	6
Left Wrist	7
Right Hip	8
Right Knee	9
Right Ankle	10
Left Hip	11
Left Knee	12
Left Ankle	13
Right Eye	14
Left Eye	15
Right Ear	16
Left Ear	17

The evaluation procedure consists of processing each image from the previously described dataset using the OpenPose model. The detected keypoints are used for further validation. In order to evaluate the performances of the pose estimation method, we observed a set of kepoints detected in each picture. Moreover, we analyzed the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) keypoints detected in each processed image.

Using the collected data, we measured three metrics as indicators of pose estimation success rate: accuracy (ACC), True Positive Rate (TPR) and the F1 score [11]:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (3)$$

Figure 1. Estimated body positions in the developed dataset. Picture 1 is in top left, picture 4 top-right, picture 13 bottom-left, and picture 16 bottom-right.

B. *Pose estimation results*

We analyzed the performance of the pose estimation classifier and a variety of factors influencing the results. The results are shown in Fig. 2.

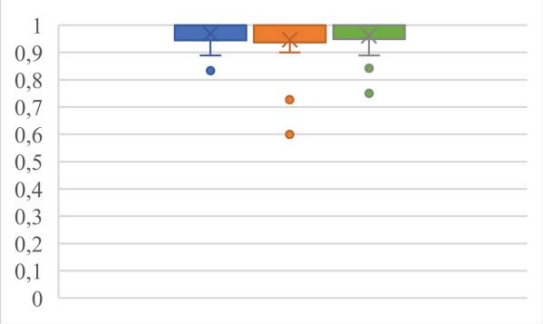

Figure 2. Box plots with accuracy, TPR, and F1 score metrics for estimation of body positions with OpenPose on the developed dataset.

The experiment demonstrated a mean accuracy of 96.88%, with a maximum of 100% and a minimum of 83.33%. Accuracy is found to be challenged the most in scenarios where the person is standing laterally opposite the camera, as the core and the limbs are positioned in an unusual way compared to the standard front view. Furthermore, pictures with partially visible parts of the body within a web camera field of view, such as half face, no forehead or partly visible arm (pictures 9, 12 and 14 in the dataset), have shown up to 11.12% lower accuracy levels than the ones with fully visible body parts.

The calculated TPR has a mean of 94.65%. There is a significant variability between a maximum (100%) and a minimum (60%). The most errors have occurred in images with partly visible body parts, as well. The greedy parsing algorithm used in the OpenPose model demonstrates efficient parses of body poses and produces high-quality matches. However, most false negatives are a result of missing adjacent keypoints, since the greedy algorithm estimates the position of limbs by connecting adjacent joints [9]. Thus, in pictures 12 and 14 (Fig. 1) some limbs failed to be recognized.

The value of F1 score mean is 96.16%, with the highest value of 100% and the lowest value of 75%. The lowest performance has been shown in case of partially visible face features (picture 9 in the dataset). Most false positives come from imprecise estimation of keypoints positioning.

Importantly, the results show that the model is highly robust to direction or levels of lighting in the picture, retaining high quality performances even in scenarios with poor lighting conditions (e.g. picture 4 in Fig. 1), since it is quite difficult to always attain perfect lighting conditions for each of the students during online classes.

Also, the results indicate a potential room for improvement of performances in scenarios with partially visible body parts, particularly face features, since those situations not only happen to be the biggest challenge of the model but are also most often seen in real-time online communication in learning environments.

III. EMOTION RECOGNITION

The evaluation of emotion recognition subsystem, based on the PAZ software library [ref], was carried out on a dataset extracted from the NimStim facial expressions database [ref]. In this test, a total of $N = 122$ pictures (out of the total of 519) were used representing 11 test subjects who acted specific discrete emotions and had a ground truth confidence interval ≥50%. Mean age of the individuals was 19.4 years, $sd = 1.2$. The ground-truth confidence level in the expressed discrete emotions was determined manually by domain experts [ref].

Emotions analyzed in this experiment were labelled Angy, Neutral, Disgust, Fear, Happy, Sad, and Surprise. Category Calm in the PAZ was identified with the emotion Neutral in the NimStim database.

Analysis of the emotion subsystem was split into two categories. In Fig. 3 we calculated the Sum of Squares Error (SSE) as in Eq. 4, average confidence of the emotion subsystem, and its standard deviation.

$$SSE = \sum_{i=1}^{n} (x_i - c_i)^2 \,;\, x_i \in [0,1], c_i \in [0,1] \qquad (4)$$

Here x_i represents the ground-truth confidence in expression of a discrete emotion for a picture i in the NimStim database and c_i confidence value obtained from the classifier for the same picture i.

Figure 3 shows every test example that the system guessed correctly in relation to the ground truth confidence interval was taken into account. Test examples that did not meet this criterion were ignored.

The results have shown that $SSE = 8.63$, *average confidence* = 0.63, $sd = 0.19$. This data represents the overall accuracy of the system. We can clearly see that the classifier works best for the emotion Happy ($SSE = 0.64$; *average confidence* = 0.84; $sd = 0.15$) if we take the number of correctly guessed examples of the specific emotion into account, while the worst performance was for emotion Disgust. ($SSE = 1.4$, *average confidence* = 0.47, $sd = 0.05$).

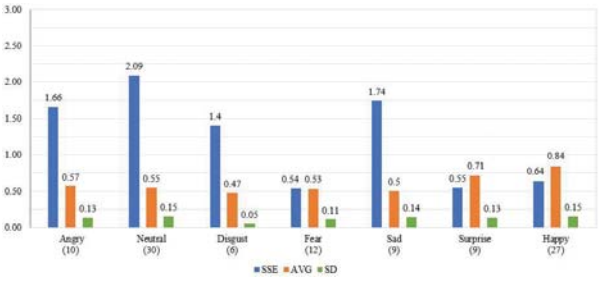

Figure 3. Estimation accuracy of individual discrete emotions for $N = 122$ pictures from the NimStim database with the PAZ software library.

Confusion matrix that shows the overall classification performance of the subsystem is shown in Fig. 4. Certain images of some test subjects who tried to mimic certain emotions and who got from experts a ground truth confidence interval of less than 50% were disregarded. A prerequisite was also that the system guessed the emotion correctly. Out of the total of 156 processed pictures 122 had confidence interval greater than 50%, and we selected those pictures for experimentation.

As can be seen in Fig. 4 the classification shows the best accuracy for discrete emotion Happy (90%) while the worst accuracy (40%) is for Disgust. The overall accuracy of emotion recognition in this experiment is 66%.

		GROUND TRUTH							CORRECT
		ANGRY	NEUTRAL	DISGUST	FEAR	HAPPY	SAD	SURPRISE	GUESSES
OUTPUT	ANGRY	10		4			1		10
	NEUTRAL	1	13	2	1	2	3		13
	DISGUST	2		6					6
	FEAR	7	4	1	8		5	1	8
	HAPPY			2	1	26			26
	SAD						8		8
	SURPRISE				4	1		9	9
	∑	20	17	15	14	29	17	10	80
	TOTAL	122							66%

Figure 4. Confusion matrix for classification of $N = 122$ pictures.

IV. Person Detection

In order to test object detection accuracy in a setting similar to the one it is going to be used in, we had to choose a fitting dataset. Such dataset should consist of two types of images: one image type showing multiple people relatively close together and the other containing a single person which should represent a web camera feed during an online lecture. Former image type is supposed to test accuracy of person detection inside a lecture hall where the number of detectable people is rather large.

Our dataset of choice was WIDER Attribute Dataset primarily because it contains a good mix of the two types while also coming with ground truth human bounding boxes. WIDER Dataset consists of 14,000 images divided into 30 different scene categories out of which we are going to use only five: Handshaking, Dancing, Meeting, Couple, and Surgeons [14].

A. Person detection experiment

Our object detection framework is based on You Only Look Once (YOLO) object detection algorithm [15]. More specifically, we used YOLOv3-608 pre-trained model with a non-maximum suppression (NMS) algorithm using various NMS thresholds.

During this experiment we are going to process 160 suitable photos from WIDER dataset and compare our results (i.e., resulting bounding boxes) with ground truth bounding boxes. We compare them using the Jaccard index with multiple thresholds to determine the values of TP, TN, and FP which are used to calculate previously mentioned TPR, F1 score, and SSE along with precision (PPV) [11]:

$$PPV = \frac{TP}{TP + FP} \tag{5}$$

These metrics are calculated once per Jaccard index threshold (abbreviated as JIT) for every image. Dataset averages for various JITs are also calculated.

B. Person detection results

As mentioned, results are divided in four distinct groups with differing Jaccard index thresholds ranging from 0.5 up to 0.9. Correlation between threshold and person detection metrics might provide us with interesting insight.

Intersect over union ratio lower than 0.5 means there is a relatively low overlap between the bounding boxes, 0.5 – 0.8 represents the medium and everything above could be considered a high degree of overlap.

The next table shows correlation between JIT increase and the following four metrics: average precision (Avg PPV), average recall (Avg TPR), average F1 score (Avg F1), and total SSE.

TABLE II. Person Detection Results in Relation to Five Selected Jaccard Index Values.

Metrics / JIT	Threshold				
	0.5	**0.6**	**0.7**	**0.8**	**0.9**
Avg PPV	0.83	0.78	0.69	0.53	0.23
Avg TPR	0.89	0.84	0.76	0.57	0.25
Avg F1	0.83	0.78	0.7	0.54	0.23
SSE	9.26	13.14	17.42	24.68	19.12

As expected, all metrics quantifying person detection's quality decrease as JIT increases which is clearly visible from the following chart in Fig. 5.

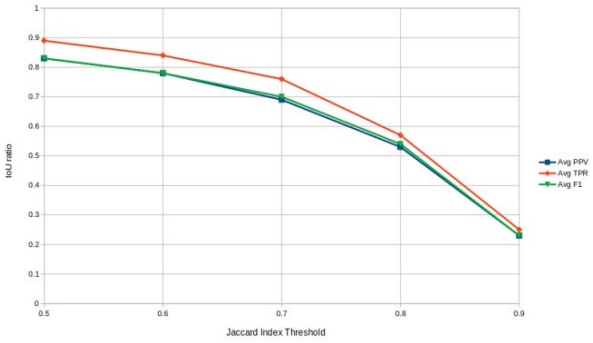

Figure 5. Correlation between JIT and dataset average metrics in person detection experiment.

The steep decrease between JIT = 0.8 and JIT = 0.9 noticeable in Fig. 5, along with observations made during manual analysis of processed images, indicates that person detection is not acting very pedantic, in a sense that it tends to enclose people in a box that is too large. This could be an indication of greater inaccuracy in more crowded images due to occlusion.

If we examine at results with respect to the fact our person detection evaluation dataset consists of two types of images, we can see that images emulating web camera feed during lectures get processed rather correctly, unlike lecture hall type images where person detection gets worse proportionally as number of people in the field of view increases.

V. Conclusion and Future Work

We have proposed a multifaceted intelligent agent for monitoring student behavior during online video lectures using pose estimation, facial expression emotion recognition and person detection in images. For this report we have selected three free and readily available classifiers: OpenPose, PEZ and YOLOv3, which we plan to upgrade and adapt to our needs [1]. Further, we have prepared three separate datasets for validation and testing of the classifiers. A limited dataset for pose estimation consisting of 16 distinctive video conference pose pictures has been prepared for this purpose. Custom software applications have been developed for experimentation with the classifiers. Each classifier was separately tested in a dedicated experiment and metrics such as accuracy, precision, TPR and F1-score have been measured. Individual performances of all three classifiers have been

described in separate sections and they give confidence that the proposed intelligent agent could be practically developed, and have its general performance tested at least as a prototype in laboratory settings.

Continuing with the implementation of the e-learning intelligent agent, it will be necessary to precisely define a rigorous experimentation protocol. Especially important will be to determine a set of rules that control cognition and behavior of homogenous groups and individuals in classrooms, in particular emotional responses to standardized images and video-clips [16]. We expect that establishing the ground-truth will greatly assist in personalization and a more effective learning. These multimedia documents are available in affective multimedia databases and can be used to establish individual baseline emotional responses [17].

Looking ahead, it would be interesting to see how other physical and physiological modalities that are accessible during video e-learning sessions contribute to the estimation accuracy of student attention and emotional responses. In this regard, it should be valuable to test additional information channels such as gaze direction, pupil dilation, eye blink rate, hand gestures, voice recognition, keyboard usage, and mouse dynamics. Even EEG brain signals – since they already have been well-researched regarding understanding of the complex human emotion mechanisms – with inexpensive consumer hardware [18]. Also, speech is known to be a potent source of information for estimation recognition and might provide useful metrics during video conferencing [19]. Finally, in long term we plan to create an ontology for formal description of knowledge about sequences of education materials and their causal relationship with emotion, attention, and student success, much the same as has already been accomplished with ontologies for description of multimedia sequences used for emotion elicitation [20] [21] [22]. We believe that the development of the adaptive intelligent agent for e-learning will significantly assist professors in personalization, participation, and productivity of education and thereby increase the common quality of teaching.

REFERENCES

[1] M. Horvat and T. Jagušt, "Emerging opportunities for education in the time of COVID-19: Adaptive e-learning intelligent agent based on assessment of emotion and attention,", In Proceedings of the 31st Central European Conference on Information and Intelligent Systems (CECIIS 2020), pp. 203–210, 2020.

[2] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features," EURASIP journal on image and video processing, 2017(1), pp. 1–12, 2017.

[3] J. Li, G. Ngai, H. V. Leong, and S. C. Chan, "Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics," ACM SIGAPP Applied Computing Review, 16(3), pp. 37–49, 2016.

[4] F. Al-Shargie, U. Tariq, H. Mir, H. Alawar, F. Babiloni, and H. Al-Nashash, "Vigilance decrement and enhancement techniques: a review," Brain sciences, 9(8), pp. 178, 2019.

[5] T. S. Ashwin and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," Education and Information Technologies, 25(2), pp. 1387–1415, 2020.

[6] T. S. Ashwin and R. M. R. Guddeti, "Unobtrusive Behavioral Analysis of Students in Classroom Environment Using Non-Verbal Cues," in *IEEE Access*, vol. 7, pp. 150693–150709, 2019, doi: 10.1109/ACCESS.2019.2947519.

[7] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions", *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, 2014.

[8] A. Jalal and M. Mahmood, "Students' behavior mining in e-learning environment using cognitive processes with information technologies," Education and Information Technologies, 24(5), pp. 2797–2821, 2019.

[9] N. Veliyath, P. De, A. A. Allen, C. B. Hodges, and A. Mitra, "Modeling students' attention in the classroom using eyetrackers," In Proceedings of the 2019 ACM Southeast Conference, pp. 2–9, 2019.

[10] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291–7299, 2017.

[11] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," Cambridge University Press, Cambridge, UK, 2008.

[12] O. Arriaga, M. Valdenegro-Toro, M. Muthuraja, S. Devaramani, and F. Kirchner, "Perception for Autonomous Systems (PAZ)", *arXiv preprint arXiv:2010.14541*, 2020.

[13] N. Tottenham, J. W. Tanaka, A. C. Leon, T, McCarry, M. Nurse, T. A. Hare, D. J. Marcus, A. Westerlund, B. J. Casey, and C. Nelson, "The NimStim set of facial expressions: Judgments from untrained research participants," Psychiatry Research, vol. 168:3, pp. 242–249, 2009.

[14] Y. Li, C. Huang, C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," In European Conference on Computer Vision, pp. 684–700, 2016.

[15] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," Computer Vision and Pattern Recognition, 2018.

[16] M. Horvat, D. Kukolja, and D. Ivanec, "Comparing affective responses to standardized pictures and videos: A study report," In Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2015), pp. 1394−1398, 2015.

[17] M. Horvat, "A brief overview of affective multimedia databases," In Proceedings of the 28th Central European Conference on Information and Intelligent Systems (CECIIS 2017), pp. 3–9, 2017.

[18] M. Horvat, M. Dobrinić, M. Novosel, M., and P. Jerčić, "Assessing emotional responses induced in virtual reality using a consumer EEG headset: A preliminary report," In Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018), pp. 1006−1010, 2018.

[19] S. Lugović, I. Dunđer, and M. Horvat, "Techniques and applications of emotion recognition in speech," In Proceedings of the 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2016), pp. 1278−1283, 2016.

[20] M. Horvat, N. Bogunović, and K. Ćosić, "STIMONT: A core ontology for multimedia stimuli description," Multimedia tools and applications, 73(3), 1103−1127, 2014.

[21] M. Horvat, "StimSeqOnt: An ontology for formal description of multimedia stimuli sequences," In Proceedings of the 43rd International Convention on Information, Communication and Electronic Technology (MIPRO 2020), pp. 1134−1139, 2020.

[22] M. Horvat, "Generation of multimedia stimuli based on ontological affective and semantic annotation," Doctoral thesis, University of Zagreb, 2013.