Adaptive Affect-Aware Multimodal Learning Assessment System for Optimal Educational Interventions

Abstract—This paper introduces an innovative method for enhanced personalized learning by establishing associations between users' emotional expressions and task performance. Our approach addresses inherent biases in affect recognition by leveraging facial and pose landmarks, integrating them with performance data, and employing time-series algorithms to uncover user-specific affect-performance relationships. Through real-time intervention capabilities, the system facilitates targeted educational interventions during subsequent learning sessions, aiming to improve learning outcomes. Currently rule-based, we aim to explore more advanced methods for optimal intervention selection. This system, while still in development, points towards future research directions in engineering education, exploring users' affect-performance associations to improve educational interventions, thereby offering more tailored and refined educational experiences.

Keywords—Affect-aware, educational interventions, education technology

I. INTRODUCTION

Educational landscapes are evolving to be more flexible, adaptable, and technological [1]. In these tech-driven learning environments, the capacity to gather and analyze novel data enables the optimization of the learning experience. Researchers have explored diverse indicators such as facial expressions, head pose estimation, skin conductance, heart rate, respiration, eye-tracking, and more to inform adjustments in learning systems or guide virtual tutor interactions [2, 3, 4]. Amongst these attributes, researchers believe that integrating a user's affective features holds the greatest potential for enhancing learning systems [5]. Affect refers to the outward expression of emotions, encompassing facial expressions, body language, and vocal tones. Despite compelling evidence, many learning systems currently overlook the incorporation of affect features, and those that do often utilize facial expression classifiers with a limited set of emotions, constraining the adaptability of the models.

In this work, we propose a novel educational assessment tool that utilizes facial landmarks, localized expressions, and time-series algorithms to learn associations between a user's diverse affective mannerisms and their corresponding performance, enabling real-time, targeted interventions within e-learning environments. The primary challenges involved in developing this system are (1) extracting the user's affective attributes, (2) discerning patterns between affect states and performance, and (3) selecting an optimal intervention. To accomplish these objectives, the system must include an affect-adaptation phase, during which it discerns the user's behavioral patterns and their associations with performance, and an intervention recommendation phase, where the system utilizes these patterns to suggest actions that aim to enrich the learning experience.

To enhance facial expression detection flexibility, we identify and extract localized facial expressions, encompassing individual eye squinting and brow expression, alongside pose features capturing positions of various facial and shoulder markers. The features, gathered while the user watches educational content, are subsequently correlated with their post-content evaluation performance. These correlations facilitate interventions during future content engagements, prompting a system response when the model identifies features linked to substandard performance.

The educational interventions, which include suggesting material reviews, breaks, or offering emotional support, seek to improve the user's future learning experiences by intervening before the end-of-lesson evaluation, thereby optimizing the learning experience for the user. While the current intervention module is rule-based, we aspire to explore advanced methods, such as reinforcement learning or the integration of large language models (LLM). The tool's versatility also extends to various subjects, with our specific focus on learning American Sign Language (ASL).

This work aims to raise awareness of future directions of engineering education research, with a particular focus on harnessing affect features to inform educational interventions within e-learning environments. In the sections that follow, we discuss leading approaches in sign language recognition, multimodal data in learning systems, data fusion techniques, adaptable learning systems, and methods for selecting an optimal educational intervention. Finally, we introduce our proposed approach.

II. USE CASE: LEARNING ASL

The motivation behind designing this system for ASL is to address the significant language deprivation faced by deaf children born to hearing parents, providing a crucial tool for parents to minimize the potential developmental delays. While initially developed and validated for ASL learning, the system's versatility extends beyond sign language education. It is being crafted as a plug-in for eLearning platforms, accommodating various learning interests and needs.

Many tools and applications facilitate ASL learning, like SMARTSign [6] and SmartSignPlay [7], offering sign language learning resources, including dictionaries, quizzes, and video recording functionality. Bouzid et al.'s SignWriting game [8] and Brashear's interactive tool [9] offer interactive tutoring, real-time evaluation, and ASL recognition [9]. Studies by Kumar et al. [10] focus on Indian Sign Language using a Kinect and Leap motion sensor, where each sensor collects data for sign recognition and facial data, respectively. Oliveira et al. [11] compared machine and deep learning techniques for Irish Sign Language fingerspelled letter recognition. More recently, a Kaggle competition [12], launched by Google, challenged teams to classify isolated ASL signs using TensorFlow Lite models trained from MediaPipe [13] features, with top teams achieving nearly 90% accuracy.

III. MULTIMODAL DATA IN LEARNING SYSTEMS

Numerous studies in multimodal learning systems integrate diverse sensory data sources—visual, auditory, and electrophysiological—to understand and model student engagement and emotions [1]. Chango et al. [1] conducted a comprehensive review of such studies, categorized by classroom setting (in-person, online, or hybrid). The studies therein encompassed data sources such as electrophysiological data, student heart rate, room temperature, questionnaires, student posture, teacher movements, and more. For instance, Pourmrzaei et al. [2] developed an intelligent tutoring system that leverages facial expressions and head pose estimation to adapt the learning environment. Research by Hussain et al. [3], identified affective states using features like facial expressions, skin conductance, heart rate, and respiration. Thompson and McGill [4] employed eye-tracking technology to assess student affective states and guide a virtual tutor's actions.

To make effective use of multimodal information, data fusion is essential. Chango et al. [1] categorized data fusion into feature-level (early fusion), decision-level (later fusion), and hybrid fusion. Early fusion often involves dimensionality reduction techniques and

concatenation of features into a single vector. Decision-level or late fusion consists of obtaining classes from various classifiers and then fusing the predictions. Hybrid fusion involves a combination of early and late fusion techniques. Chango et al. [14] explored hybrid configurations by fusing features to create classifiers, followed by ensemble-based fusion. Qu et al. [15] investigated fusion methods combining similarity metrics, such as the Spearman coefficient, with mathematical formulas to weigh data sources. Lou et al. [16] introduced a binomial tree model fusion algorithm for student interest classification. Presently, our system adopts vector concatenation of extracted features for data fusion. In future work, we will explore ways to adaptively fuse or weight features.

IV. Optimal Education Interactions

Numerous studies have explored strategies to improve learning outcomes by personalizing various aspects of the learning process, including content, paths, interfaces, suggestions, prompts and feedback [17]. Diverse methodologies exist for enriching the overall learning experience, with Xiao et al. [18] introducing a personalized recommendation system for online learning. This system recommends learning resources to online course participants by leveraging association rules, content filtering, and collaborative filtering. Similarly, Benhamdi et al. [19] developed a personalized recommendation system that tailors learning resource suggestions based on user preferences, background knowledge, and the user's capacity to retain knowledge.

While rule-based systems are prevalent in the studies highlighted by Xie et al., an alternative approach gaining traction is reinforcement learning (RL). For example, Gordon et al. [20] utilizes RL to enhance a robot's ability to engage and emotionally support children during educational activities. Their approach enables the robot to provide personalized affective responses tailored to each child across multiple sessions. Educational systems aiming to optimize interactions face a choice between rule-based and reinforcement learning approaches. Rule-based systems offer transparency, control, and ease of implementation but lack adaptability and scalability. In contrast, reinforcement learning provides personalization, flexibility, and scalability but faces challenges such as data efficiency, training complexity, and ethical considerations.

V. Proposed Approach

We present a learning system designed to extract affect features from video footage as a user interacts with educational content. This approach seamlessly integrates into existing educational setups, requiring only a webcam. The architecture of the proposed system is shown in *Figure 1*.

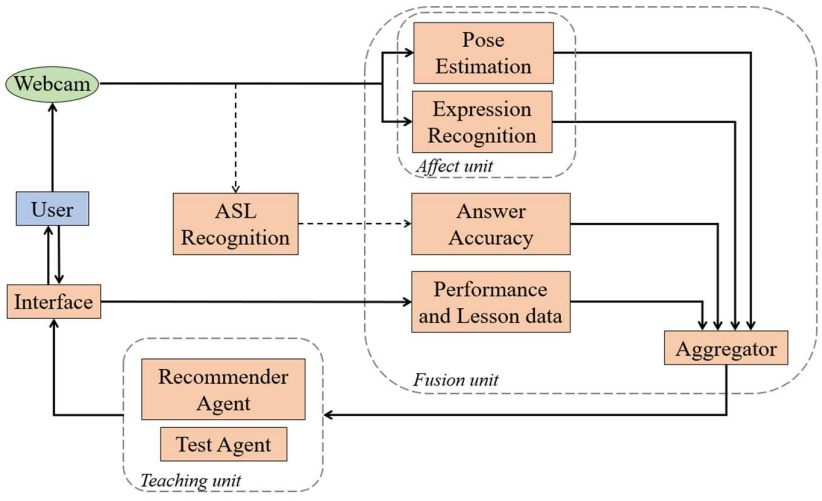

Figure 1: Overview of the proposed adaptive multimodal learning system architecture.

Within this framework, the user is symbolized as a distinctive blue square and is continuously monitored during their learning cycle via a webcam. The visual input is processed through key modules, including one for ASL recognition, pose estimation, and expression recognition. The response accuracy is determined by the ASL recognition system. These features are then concatenated and channeled to the teaching unit, which provides recommendations to the user.

A. ASL Recognition

During the assessment phase, the user is prompted to provide a sign corresponding to the letter under evaluation. Leveraging MediaPipe [13], hand landmarks from the submitted image are precisely identified, serving as reference points to crop an isolated sign with dimensions 64x64. This isolated sign is then input into the ASL recognition module, which features a convolutional neural network (CNN) inspired by a high-performing Kaggle submission [21]. This Kaggle submission was developed for an American Sign Language (ASL) dataset comprising 90,000 isolated ASL images spanning 29 classes [22].

The CNN architecture is composed of two sequential Conv2D layers with 64 filters, kernel sizes of 4, and a Rectified Linear Unit (ReLU) activation function. The initial layer, represented by a blue block, has a stride of 1, while the subsequent orange layer has a stride of 2. Following this, two additional convolutional layers are introduced, each with 128 filters, along with a dropout layer. Subsequently, the model includes two more convolutional layers with 256 filters, followed by a final dropout layer. In the dense layers, the first one comprises 512 units with a ReLU activation function, and the ultimate dense layer has units corresponding to the number of classes. This layer utilizes the SoftMax activation function to achieve a probability distribution across the classes.

The model was trained on 90% of the training data (sample size of 87k images), with the remaining 10% reserved for validation. Various data augmentation techniques were explored, yielding optimal results with random brightness adjustments between 80% to 120% and rescaling by 1./255. Training accuracy peaked at 100%, indicating potential overfitting concerns. Validation accuracy, obtained from 10k bootstrap iterations, is around 86.9% ± 0.4%. The 95% confidence interval for accuracy ranges from 86.9% to 87.0%, indicating high confidence in the estimate. The model achieved a perfect testing accuracy of 100% on a limited held-out test set but struggled to generalize effectively to self-generated data. To improve performance on self-generated data, further experiments or image preprocessing techniques are essential for future ASL recognition applications.

B. Affect Unit

Throughout the user's interaction with the system, MediaPipe [13] is employed to extract fifty-two localized facial expressions and thirteen pose features. The quantification of localized facial expressions is based on their presence in a given frame, while pose attributes encompass x and y coordinates, along with a presence value. All features are normalized and constrained within the [0,1] range. In total, the facial expression recognition and head pose estimation modules of our system generate 91 normalized affect features.

The identification of personalized associations between the 91 features and user performance is initiated during the early interactions. To initially tune the affect unit, users engage in playing a game or performing tasks with varying difficulties, during which affect

features are extracted. At the conclusion of each task, user evaluation takes place. Following multiple tasks of ranging difficulties, time-series algorithms such as recurrent neural network (RNN) or Long Short-Term Memory (LSTM) will be employed to learn the individualized affect-performance associations.

Our next step involves the creation of synthetic data featuring a spectrum of facial expressions and performance scores, linking expressions to scores. Subsequently, we will test RNN and LSTM algorithms to identify which best recognizes these patterns. Once the superior method is identified, we will tailor the affect unit for individual users and use it to guide the teaching unit.

C. *Teaching Unit*

Once the affect unit has undergone initial tuning, it becomes instrumental in intervening during subsequent learning cycles, facilitating real-time, targeted interventions. For instance, if affect features associated with substandard performance are detected, the system triggers a set of questions. We are contemplating the implementation of an initial query to the user, seeking feedback on the appropriateness of the system trigger—whether they comprehend the content recently reviewed when features associated with substandard performance were identified. If the user deems the system trigger unwarranted, we will proceed to re-tune the affect unit, integrating the new data. Conversely, if they acknowledge the trigger's relevance, an evaluation on the content covered during the system-triggered phase follows. Depending on their performance in this evaluation, we can propose an appropriate course of action.

Currently, the action recommendation module operates on a rule-based system, relying solely on the score of the assessment questions. If the score for the evaluated sign exceeds 80% accuracy, the module advises to 'continue to the next module'; for scores between 50% and 80%, the recommendation is to 'take a break'; and if it falls below 50%, the module suggests to 'review the module.'

As our work progresses, we will explore more sophisticated methods for optimal intervention selection. This exploration may encompass refining the existing rules or integrating reinforcement learning to formulate a policy that maximizes user performance and adapts to content. Additionally, we are considering the incorporation of a fixed set of domain/lesson-specific prompts for feeding into a language model, enabling us to discern which prompts contribute most effectively to accelerated student learning.

VI. CONCLUSION

Our approach presents a novel method aimed at mitigating biases in affect recognition. By extracting landmark features and integrating them into a time-series algorithm, we seek to learn user-specific associations. Through the utilization of a camera and performance data, our method is being developed as an assessment tool with the potential to predict user performance based on personalized affect-performance patterns. While still a work in progress, this tool provides a promising avenue for exploring associations between a user's emotional states and their preferences, needs, and overall learning outcomes. As such, it holds the potential to enhance the personalization of educational experiences in a more unbiased and individualized manner.

VII. REFERENCES

[1] W. Chango, J. A. Lara, R. Cerezo, and C. Romero, “A review on Data Fusion in multimodal learning analytics and educational data mining,” *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 4, Apr. 2022. doi:10.1002/widm.1458.

[2] M. Pourmirzaei, G. A. Montazer, E. Mousavi, “Customizing an Affective Tutoring System Based on Facial Expression and Head Pose Estimation”, arXiv preprint arXiv:2111.14262, Nov 2021.

[3] M. S. Hussain, O. AlZoubi, R. A. Calvo, and S. K. D’Mello, “Affect detection from multichannel physiology during learning sessions with AutoTutor,” Lecture Notes in Computer Science, pp. 131–138, 2011. doi:10.1007/978-3-642-21869-9_19

[4] N. Thompson and T. J. McGill, “Genetics with Jean: The design, development and evaluation of an affective tutoring system,” Educational Technology Research and Development, vol. 65, no. 2, pp. 279–299, Aug. 2016. doi:10.1007/s11423-016-9470-5

[5] B. Nye, S. Karumbaiah, S.T. Tokel, M. G. Core, G. Stratou, D. Auerbach, and K. Georgila, “Analyzing learner affect in a scenario-based intelligent tutoring system,” Lecture Notes in Computer Science, pp. 544–547, 2017. doi:10.1007/978-3-319-61425-0_60

[6] K. A. Weaver and T. Starner, “We need to communicate!,” The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility, Oct. 2011. doi:10.1145/2049536.2049554

[7] C.-H. Chuan and C. A. Guardino, “Designing smartsignplay,” Companion Publication of the 21st International Conference on Intelligent User Interfaces, Mar. 2016. doi:10.1145/2876456.2879483

[8] Y. Bouzid, M. A. Khenissi, F. Essalmi and M. Jemni, "Using educational games for sign language learning—A signwriting learning game: Case study", Educ. Technol. Soc., vol. 19, no. 1, pp. 129-141, Jan. 2016.

[9] H. Brashear, “Improving the efficacy of Automated Sign Language Practice Tools,” ACM SIGACCESS Accessibility and Computing, no. 89, pp. 11–17, Sep. 2007. doi:10.1145/1328567.1328570

[10] P. Kumar, H. Gauba, P. Pratim Roy, and D. Prosad Dogra, “A multimodal framework for Sensor Based Sign Language recognition,” Neurocomputing, vol. 259, pp. 21–38, Oct. 2017. doi:10.1016/j.neucom.2016.08.132

[11] M. Oliveira, H. Chatbri, S. Little, Y. Ferstl, N. O’Connor, and A. Sutherland, “Irish sign language recognition using principal component analysis and Convolutional Neural Networks,” 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Nov. 2017. doi:10.1109/dicta.2017.8227451

[12] A. Chow, G. Cameron, M. Sherwood, P. Culliton, S. Sepah, S. Dane, and T. Starner, “Google - Isolated Sign Language recognition,” Kaggle. https://www.kaggle.com/competitions/asl-signs (accessed Jan. 7, 2024)

[13] Google LLC, "MediaPipe," Online. https://developers.google.com/mediapipe (accessed Feb. 5, 2024)

[14] W. Chango, R. Cerezo, M. Sanchez-Santillan, R. Azevedo, and C. Romero, “Improving prediction of students’ performance in intelligent tutoring systems using attribute selection and ensembles of different Multimodal Data Sources,” Journal of Computing in Higher Education, vol. 33, no. 3, pp. 614–634, Oct. 2021. doi:10.1007/s12528-021-09298-8

[15] J. Qu, A. Liu, and R. Liu, “Research on evaluation and confirmation of college students’ learning behavior based on comprehensive weighted fusion algorithm,” 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Apr. 2021. doi:10.1109/icccbda51879.2021.9442584

[16] Z. Luo, C. Zheng, J. Gong, S. Chen, Y. Luo, and Y. Yi1, “3DLIM: Intelligent Analysis of Students’ learning interest by using Multimodal Fusion Technology,” Education and Information Technologies, vol. 28, no. 7, pp. 7975–7995, Dec. 2022. doi:10.1007/s10639-022-11485-8
[17] H. Xie, H.-C. Chu, G.-J. Hwang, and C.-C. Wang, “Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of Journal Publications from 2007 to 2017,” Computers & Education, vol. 140, p. 103599, 2019. doi:10.1016/j.compedu.2019.103599
[18] J. Xiao, M. Wang, B. Jiang, and J. Li, “A personalized recommendation system with combinational algorithm for online learning,” Journal of Ambient Intelligence and Humanized Computing, vol. 9, no. 3, pp. 667–677, 2017. doi:10.1007/s12652-017-0466-8
[19] S. Benhamdi, A. Babouri, and R. Chiky, “Personalized recommender system for e-learning environment,” Education and Information Technologies, vol. 22, no. 4, pp. 1455–1477, 2016. doi:10.1007/s10639-016-9504-y
[20] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, C. Breazeala, “Affective personalization of a social robot tutor for children’s Second language skills,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, no. 1, Mar. 2016. doi:10.1609/aaai.v30i1.9914
[21] Akash, “Asl alphabet,” Kaggle, https://www.kaggle.com/datasets/grassknoted/asl-alphabet/data (accessed Nov. 14, 2023)
[22] DanB, “Running Kaggle Kernels with a GPU”, Kaggle. https://www.kaggle.com/code/dansbecker/running-kaggle-kernels-with-a-gpu/notebook (accessed Nov. 21, 2023)