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

Abstract— Researchers recognize the potential of affective, or emotional, features in enhancing learning systems, but many current systems use classifiers with limited emotions, limiting model flexibility. This paper introduces a novel educational assessment tool that leverages 52 localized facial expression features, 39 pose landmarks, and time-series algorithms to discern user-specific affect-performance associations, enabling real-time interventions in e-learning environments. Our *Assessment Unit* evaluates users' subject-specific proficiency, the *Affect Unit* predicts real-time performance based on affect features, and a simple *Intervention Unit* offers targeted recommendations. In future work, we aim to shift from rule-based interventions to advanced methods integrating reinforcement learning, investigate adaptive data fusion, and develop an affect-performance dataset to train and evaluate performance predictors. 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, educational assessment tool, personalized educational experiences, e-learning, learning system, real-time 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 —encompassing facial expressions, body language, and vocal tones—holds the greatest potential for enhancing learning systems [5]. 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 localized expressions, pose landmarks, 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. To enhance the flexibility of facial expression detection, we identify and extract 52 localized facial expressions, including individual eye squinting and brow expressions, as well as several pose features that capture the positions of various facial and shoulder markers. During initial interactions, we use time-varying affect features gathered as the user interacts with lessons or assessments, along with their corresponding intermittent lesson evaluations or assessment scores, to train time-series algorithms for forecasting user performance. Once the system has undergone sufficient tuning, the performance predictor becomes operational in subsequent sessions, allowing the system to provide more informed interventions to the user to accelerate learning and improve overall retention. The educational interventions, which include suggesting material reviews, breaks, or offering emotional support, aim to optimize the user's learning experience by providing real-time guidance. 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 personalized educational interventions within e-learning environments. In the sections that follow, we discuss leading approaches in sign language recognition, multimodal data in learning systems, and methods for selecting an optimal educational intervention.

ii. use case: learning american sign language

While initially developed and validated for ASL learning, the system's versatility extends beyond sign language education. Our system 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 effectively use multimodal information, data fusion is crucial. Chango et al. [1] categorized fusion into feature-level (early fusion), decision-level (late fusion), and hybrid fusion. Early fusion combines features into a single vector using dimensionality reduction techniques. Decision-level fusion combines class predictions from various classifiers. Hybrid fusion combines both early and late fusion. Chango et al. [14] explored hybrid configurations with feature-based classifier ensembles. Qu et al. [15] used similarity metrics, like the Spearman coefficient, with mathematical formulas for data source weighting. Lou et al. [16] developed a binomial tree model fusion algorithm for student interest classification. Presently, our system uses vector concatenation for data fusion. In future work, we will explore adaptive fusion methods to weigh our 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]. Xiao et al. [18] introduced a personalized recommendation system (PRS) that leverages association rules, content filtering, and collaborative filtering to recommend learning resources to online course participants. Similarly, Benhamdi et al. [19] developed a PRS 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. [17], reinforcement learning (RL) is gaining traction. For example, Gordon et al. [20] utilized RL to enhance a robot's ability to engage and emotionally support children during educational activities, enabling personalized affective responses across 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

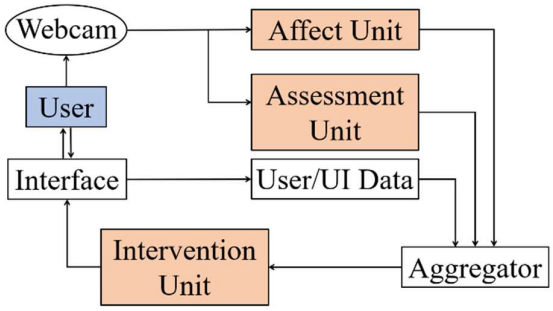

Figure 1: Overview of the proposed adaptive multimodal learning system architecture.

We introduce a novel learning system that utilizes individualized affect-performance patterns to guide educational interventions, with the goal of enhancing learning outcomes. Our method integrates computer vision and time-series algorithms, focusing on localized facial expressions for improved model adaptability and flexibility. Prior work often classifies emotions into a limited set, constraining the model's expressiveness and flexibility [2, 3, 4, 5]. Furthermore, our system is designed to seamlessly integrate into existing e-learning environments, requiring only a webcam for implementation.

The system's architecture, depicted in *Figure 1*, represents the user as a distinct blue square. At the start of their initial interaction, users establish a profile by responding to questions about demographics and relevant experience. This user profile/user interface data encompasses session durations and past session data. Users are recorded during their interaction with educational content, which can include lessons or assessments. The Affect unit extracts sequences of affect features from webcam footage, comprising 91 localized facial expressions and pose landmarks per frame. Performance scores are obtained either directly in the assessment or throughout the lessons. In our system, users' proficiency in signing ASL characters is evaluated. They are prompted to sign a character, which is then analyzed by a CNN to assess the accuracy of the sign. In contrast, other systems may employ traditional assessment methods.

During initial interactions, the *Affect Unit* utilizes time-series algorithms trained on affect sequences, user profiles, and assessment data to forecast user performance. Once the system undergoes sufficient tuning, the performance predictor becomes operational in subsequent

sessions. With the system now equipped to provide real-time performance forecasts using affect features, the Intervention Unit can make more informed recommendations to the user. This capability enables real-time, targeted interventions within e-learning environments, aimed at accelerating learning and enhancing overall retention.

a. assessment unit

The Assessment Unit's purpose is to evaluate users' performance in a given task, essential for both fine-tuning the performance forecaster within the *Affect Unit* and informing the *Intervention Unit*. While traditional assessment methods can be used, our system employs a CNN to assess users' ASL signing proficiency. When prompted, users capture an image of themselves signing a character. Utilizing MediaPipe's hand landmarks task [13], an isolated ASL sign is extracted and subsequently analyzed by a CNN to predict the accuracy of the sign.

The CNN architecture, inspired by a high-performing Kaggle submission [21], 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 has a stride of 1, while the subsequent 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.

This CNN was trained on a dataset consisting of 90,000 isolated ASL images spanning 29 classes [22]. The model was trained on 90% of the training data, 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

The affect unit serves a dual purpose: firstly, to extract localized facial expressions, and secondly, to predict real-time user performance based on these features. As the user interacts with the system, MediaPipe [13] is employed to extract 52 localized facial expressions and 39 pose features. Localized facial expressions are quantified based on their presence in a given frame, while pose attributes encompass (x, y) coordinates, along with a presence value. All features scaled to the [0,1] range, resulting in a total of 91 normalized affect features. To enable real-time prediction of user-specific performance, the system undergoes initial tuning. During initial interactions, sequences of affect features and corresponding scores are collected. If these interactions take the form of lessons, the student's video is segmented into sequences, delineated by in-lesson assessments. Affect sequences and scores are used to tune the performance predictor, and we plan to both train and test Transformers, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models. After adequate tuning, the

performance predictor becomes operational in subsequent sessions. Consequently, the performance predictor can alert the Intervention Unit of the presence of features associated with subpar performance, facilitating real-time monitoring.

Our next step involves constructing an affect-performance dataset. We have devised a method to gather video footage from users completing a brief online questionnaire, with the intention of using this data to train and evaluate the performance of our RNN and LSTM predictors. Our goal is to identify the algorithm that most efficiently learns affect-performance patterns and to investigate both transfer learning techniques for pretraining the performance predictor and user-specific fine-tuning methods.

c. intervention 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 (i.e., a low performance is predicted by the affect unit), 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 or not. If the user deems the system trigger unwarranted, we will proceed to re-tune the performance predictor in 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 introduces a novel educational assessment tool that leverages localized expressions, pose landmarks, and time-series algorithms to discern user-specific affect-performance associations, facilitating real-time interventions within e-learning environments. We developed an *Assessment Unit* to evaluate users' ASL signing proficiency, an *Affect Unit* to predict real-time performance based on affect features, and a simple *Intervention Unit* for targeted interventions. In future work, we aim to transition from our rule-based *Intervention Unit* to advanced methods that incorporate reinforcement learning or LLMs, explore adaptive data fusion, and develop an affect-performance dataset to train and evaluate performance predictors. While still a work in progress, this tool provides a promising avenue for exploring associations between a user's affect states and their preferences, needs, and overall learning outcomes. As such, it holds the potential to enhance the personalization of educational experiences.

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, issue. 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], Nov. 2021. Available: arXiv:2111.14262. (accessed Feb. 5, 2024).

[3] M. S. Hussain, O. AlZoubi, R. A. Calvo, and S. K. D’Mello, “Affect detection from multichannel physiology during learning sessions with AutoTutor,” in *International Conference on Artificial Intelligence in Education, AIED 2011, Auckland, New Zealand, June 28-July 1, 2011*, G. Biswas, S. Bull, J. Kay, A. Mitrovic, Eds. Springer, Berlin, Heidelberg, 2011. pp. 131-138. 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 Dev*, vol. 65, 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,” in *International Conference on Artificial Intelligence in Education, AIED 2017, Wuhan, China, June 28-July 1, 2017, E.* André, R. Baker, X. Hu, M. Rodrigo, B. du Boulay, Eds. Springer, Cham, 2017. pp. 544-547. doi:10.1007/978-3-319-61425-0_60.

[6] K. A. Weaver and T. Starner, “We Need to Communicate! Helping Hearing Parents of Deaf Children Learn American Sign Language,” in *Proc. of 13th International ACM SIGACCESS conference on Computers and accessibility, ASSETS 2011, Dundee Scotland, UK, Oct. 24-26. 2011*, pp. 91-98. doi:10.1145/2049536.2049554.

[7] C.-H. Chuan and C. A. Guardino, “Designing SmartSignPlay: An Interactive and Intelligent American Sign Language App for Children who are Deaf or Hard of Hearing and their Families,” Companion Publication of the *21st International Conference on Intelligent User Interfaces, IUI 2016, Sonoma, CA, USA, Mar. 7-10, 2016*, pp. 45-48. 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*, issue. 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,” in *2017 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2017, Sydney, NSW, Australia, Nov 29-Dec 1, 2017*, pp. 1-8. doi:10.1109/dicta.2017.8227451.

[12] A. Chow, G. Cameron, M. Sherwood, P. Culliton, S. Sepah, S. Dane, and T. Starner. 2023. “Google - Isolated Sign Language recognition,” *Kaggle.com*. https://www.kaggle.com/competitions/asl-signs (accessed Jan. 7, 2024).

[13] Google LLC. 2023. "MediaPipe." 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, 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," in *2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2017, Chengdu, China, Apr. 24-26, 2021*, pp. 121-125. 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, 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, issue, Oct. 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, pp. 667–677, Mar. 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, pp. 1455–1477, May 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," in *Proc. of the Thirteenth AAAI Conference on Artificial Intelligence, AAAI 2016, Phoenix, AZ, USA, Feb. 12-17, 2016*, pp. 3951-3957. doi:10.1609/aaai.v30i1.9914.
[21] Akash Nagaraj. 2018. "Asl alphabet," *Kaggle.com*. https://www.kaggle.com/datasets/grassknoted/asl-alphabet/data (accessed Nov. 14, 2023).
[22] Dan Becker. 2018. "Running Kaggle Kernels with a GPU", *Kaggle.com*. https://www.kaggle.com/code/dansbecker/running-kaggle-kernels-with-a-gpu/notebook (accessed Nov. 21, 2023).