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

Andres Gomez
University of Florida, College of
Electrical and Computer Engineering
Gainesville, FL
Andresggomez@ufl.edu

***Abstract* — This paper introduces an innovative affect-aware multimodal learning assessment system that utilizes features like facial expressions and head pose to inform educational interventions. Distinguishing itself from prior research, our approach excels in offering heightened flexibility for detecting affective states. The system establishes correlations between learners' diverse affective features and their performance. By learning these correlations during initial interactions with the learner, the system facilitates interventions during subsequent content engagements, triggering a system response when features linked to substandard performance are detected. The proposed educational interventions, encompassing material reviews, breaks, or emotional support, aim to enhance future learning experiences by intervening before evaluation, optimizing the learner's time. This research seeks to raise awareness of future directions in engineering education research, specifically emphasizing the utilization of affect-aware features for optimal educational interventions within e-learning environments.**

Keywords—Multimodal assessment, affect-aware learning, educational interventions, facial expressions, head pose estimation, engineering education research

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 learner'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 an educational assessment tool that establishes connections between a learner's diverse affective mannerisms and their corresponding performance, enabling targeted interventions that enrich the learning experience. The primary challenges involved in developing this system are (1) detecting the learner'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 learns the learner's behavioral patterns and their associations with performance, and an intervention recommendation phase, where it utilizes these patterns to suggest appropriate actions.

To enhance facial expression detection flexibility, we identify and extract fifty-two localized facial expressions, encompassing individual eye squinting and brow expression, alongside thirteen pose features capturing positions of various facial and shoulder markers. The features, gathered while the learner 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 learner's future learning experiences by intervening before evaluation, thereby optimizing time for the learner. While the current intervention module is rule-based, we aspire to explore advanced methods, such as reinforcement learning. The tool's versatility also extends to various subjects, with our specific focus on learning American Sign Language (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.

In the sections that follow, we discuss leading approaches in hand gesture recognition, multimodal data in learning systems, multimodal data fusion techniques, adaptable learning systems, and methods for selecting an optimal action. Finally, we introduce our proposed approach. 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.

II. HAND-GESTURE RECOGNITION

Various tools and applications have been developed to facilitate ASL learning for deaf children and other users.

SMARTSign [6] and SmartSignPlay [7] are web and mobile applications that provide sign language learning resources, including dictionaries, quizzes, and video recording functionality. Other initiatives like Bouzid et al.'s game-based mechanism for learning SignWriting, a system of writing sign languages [8], and Brashear's game-based interactive tool for developing deaf children's language skills offers interactive tutoring, real-time evaluation, and is equipped with a camera and sensors to recognize ASL [9].

Additionally, studies like Kumar et al. [10] focus on recognizing 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 was launched by Google to classify isolated ASL signs using TensorFlow Lite models trained from MediaPipe extracted features. Over 1,000 teams entered, resulting in a variety of solutions, with the top teams achieving nearly 90% accuracy on the provided dataset.

III. MULTIMODAL DATA IN LEARNING SYSTEMS

Many studies investigating multimodal learning systems integrate various sensory data sources, including visual, auditory, and electrophysiological data, to better understand and model student engagement and emotions. Chango et al. conducted a comprehensive review of multimodal learning analytics studies, categorizing them based on in-person, online, or hybrid classroom settings [1]. These studies encompassed data sources such as electrophysiological data, student heart rate, room temperature, questionnaires, student posture, teacher movements, and more.

For instance, Pourmrzaei et al. developed an intelligent tutoring system that leverages facial expressions and head pose estimation to adapt the learning environment [2]. Other research by Hussain et al., identified a wide range of affective states using features like facial expressions, skin conductance, heart rate, and respiration [3]. Research by Thompson and McGill [4] even 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. categorize data fusion into feature-level (early fusion), decision-level (later fusion), and hybrid fusion [6]. Early fusion often involves dimensionality reduction techniques and concatenation of features into a single vector. The next category is decision-level or later fusion, which consists of obtaining classes from various classifiers and then fusing the predictions. The last category is hybrid fusion and involves a combination of early and later fusion techniques.

Some studies employing later fusion techniques have relied on selecting the best-performing classifiers to make informed decisions [1]. Additionally, Chango, Cerezo, Sanchez-Santillan, et al., have ventured into hybrid configurations, commencing with the fusion of features from various sources to create classifiers, followed by the ensemble-based fusion of these classifiers [12]. A study by Qu et al., delves into fusion methods that combine similarity metrics, namely the Spearman coefficient, with various mathematical formulas to weigh data sources [13]. Research by Lou et al. introduced a binomial tree model fusion algorithm with guided weight allocation for student interest classification [14]. To summarize, multimodal information fusion can encompass an array of methodologies, including aggregation, ensembles, statistical operators, and filtering techniques. Among these, ensembles, which amalgamate results from multiple classifiers, stand out as a favored fusion approach.

IV. OPTIMAL EDUCATION INTERACTION

As previously mentioned, the proposed learning system will learn the optimal educational strategy for users over time. Some studies use rule-based systems to determine optimal decisions, while others use reinforcement learning. While some studies rely on rule-based systems to determine optimal decisions, others turn to reinforcement learning for this purpose. An illustrative example by Gordon et al. presents a social robot companion's behavior influenced by two primary components: game logic and an affective policy [15].

During educational games, the robot issues instructions, hints, and positive feedback for correct responses, utilizing gaze direction to reinforce interactions. The affective policy comes into play upon task completion or when interaction time expires, using a Q(s, a) matrix to manage various emotional states and actions, encompassing valence and engagement. Employing a reinforcement learning algorithm, specifically SARSA, this approach personalizes the robot's affective responses for individual learners, adapting to their unique emotional dynamics to enhance their educational experience

V. PROPOSED APPROACH

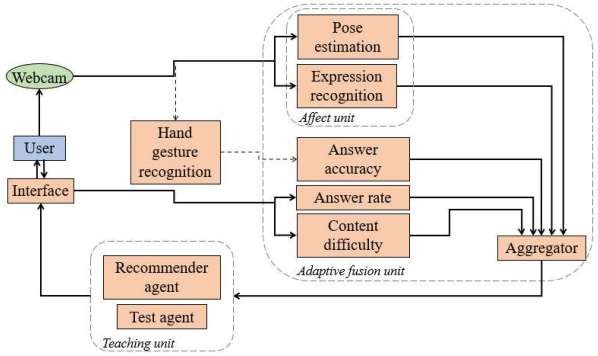

Figure 1: Overview of the adaptive multimodal learning system architecture.

As outlined earlier, we present a learning system designed to extract affect features from recorded footage as a learner 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*. Within this framework, the learner 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 a hand-gesture recognition system, pose estimation, and expression recognition system. The response accuracy is determined by the hand-gesture recognition system. These features are then concatenated and channeled to the teaching unit, which provides recommendations to the learner.

During the learner's interaction with educational videos or content, our system utilizes MediaPipe to extract fifty-two localized facial expressions and thirteen pose features. These localized facial expressions are quantified by their presence in a given frame, while pose attributes include x and y coordinates, along with a presence value. All features are normalized and constrained within the [0,1] range. In total, our system's facial expression recognition and head pose estimation modules yield 91 normalized affect features.

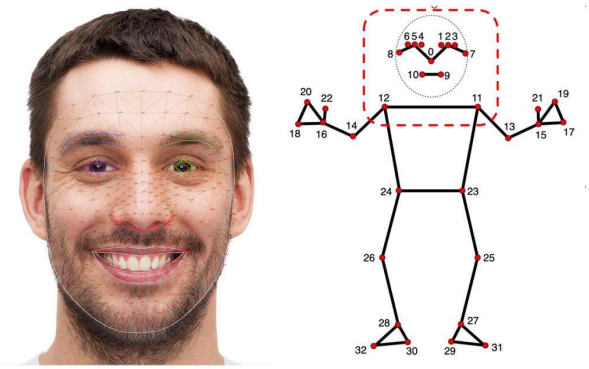

Figure 2: Facial and pose landmarks extracted by MediaPipe. The dashed red boundary on the right include the thirteen extraced pose features.

Upon completing the video or content review, the learner transitions to the assessment phase of the system. In this scenario, the learner is prompted to provide a sign corresponding to the letter under evaluation. Leveraging MediaPipe, 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 [16]. This Kaggle submission was developed for an American Sign Language (ASL) dataset comprising 90,000 isolated ASL images spanning 29 classes [17].

The CNN architecture, as illustrated in *Figure 3*, 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 underwent training on 90% of the training data, with the remaining 10% designated for the validation set. Various data augmentation techniques were explored, yielding optimal results through a random brightness adjustment ranging from 80% to 120%, coupled with rescaling by 1./255. During training, the accuracy reached 100%, indicating a potential concern of overfitting, while the validation accuracy stood at 87%. Impressively, the model achieved a perfect testing accuracy of 100% on a limited held-out test set.

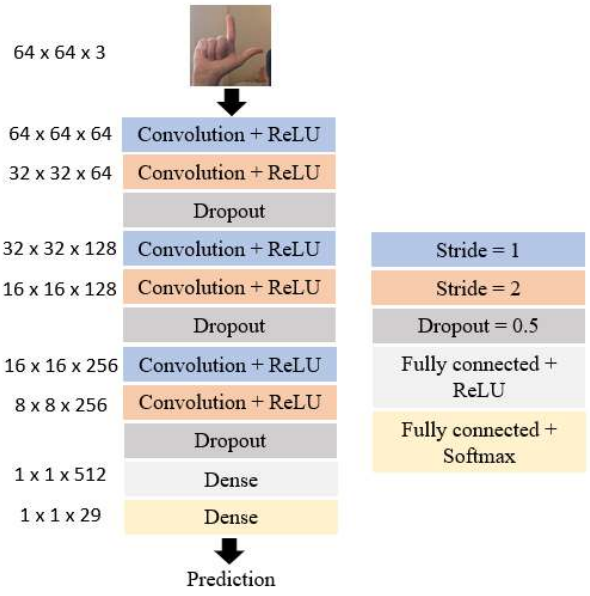

Figure 3: The CNN architecture used within the ASL recognition module.

Despite excelling on this specific test set, the model faced challenges in generalizing effectively to self-generated data. For future applications involving ASL recognition, it is imperative to conduct extensive experiments or employ image-preprocessing techniques to enhance the model's performance on self-generated data.

Upon generating the probability for each submitted sign, this information guides the action recommendation module. Currently, the action recommendation module operates on a rule-based system, relying solely on the probability of the true label generated by the ASL recognition module. 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 this work progresses, we will explore more sophisticated methods for optimal intervention selection. This may involve refining the rules or incorporating reinforcement learning to develop a policy that maximizes user performance.

VI. REFERENCES

[1] Wilson Chango, Juan A. Lara, Rebeca Cerezo, Cristóbal Romero A review on data fusion in multimodal learning analytics and educational data mining. WIREs Data Mining and Knowledge Discovery. 05 April 2022

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

[3] M. S. Hussain, Omar AlZoubi, Rafael A. Calvo & Sidney K. D’Mello, “Affect Detection from Multichannel

Physiology during Learning Sessions with AutoTutor', AIED 2011: Artificial Intelligence in Education pp 131–138
[4] Thompson, N., & McGill, T. J. (2017). Genetics with Jean: The design, development and evaluation of an affective tutoring system. Education Technology Research and Development, 65, 279–299.
[5] B. Nye et al., "Analyzing learner affect in a scenario-based intelligent tutoring system," 17, doi: 10.1007/978-3-319-61425-0_60.
[6] K. A. Weaver, "We need to communicate!: Helping hearing parents of deaf children learn american sign language", Proc. 13th Int. ACM SIGACCESS Conf. Comput. Accessibility, pp. 91-98, Oct. 2011.
[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", Proc. 21st Int. Conf. Intell. User Interfaces., pp. 45-48, 2016.
[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 Comput., vol. 89, no. 1, pp. 11-17, Sep. 2007.
[10] P. Kumar, H. Gauba, P. P. Roy and D. P. Dogra, "A multimodal framework for sensor based sign language recognition", Neurocomputing, vol. 259, pp. 21-38, Oct. 2017.
[11] M. Oliveira, H. Chatbri, S. Little, Y. Ferstl and N. E. O, "Connor and A. Sutherland "Irish sign language recognition using principal component analysis and convolutional neural networks", Proc. Int. Conf. Digital Image Comput. Techn. Appl. (DICTA), pp. 1-8, Dec. 2017.
[12] Chango, W., Cerezo, R., Sanchez-Santillan, M. et al. Improving prediction of students' performance in intelligent tutoring systems using attribute selection and ensembles of different multimodal data sources. J Comput High Educ 33, 614–634 (2021).
[13] Qu, J., Liu, A., & Liu, R. (2021, 2021). 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 (pp. 121– 125). IEEE Xplore.
[14] Z. Luo, C. Zheng, J. Gong, S. Chen, Y. Luo, Y. Yi1, "3DLIM: Intelligent analysis of students' learning interest by using multimodal fusion technology", Education and Information Technologies (2023) 28:7975–7995
[15] 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", Thirtieth AAAI Conference on Artificial Intelligence, Feb 2016
[16] Akash, (2017), ASL Alphabet, Kaggle. Available: https://www.kaggle.com/datasets/grassknoted/asl-alphabet/data, doi: 10.34740/KAGGLE/DSV/29550
[17] DanB (2017), Running Kaggle Kernels with a GPU, Kaggle. Available: https://www.kaggle.com/code/dansbecker/running-kaggle-kernels-with-a-gpu/notebook