Abstract

In the current digital era, technology-enhanced learning is evolving rapidly, setting new trends in educational environments and enabling students to learn more efficiently than ever before. Conventional learning course content typically involves only two sensory modalities—audio and video—which limits its ability to engage learners fully. In contrast, immersive learning course content and environments incorporate multiple senses, allowing learners to interact with multimedia content in ways that go beyond sight and sound. This approach, known as *mulsemedia* (multiple sensorial media), posits that engaging sensory channels—such as audio, visual, haptic, olfactory, temperature, gustatory, and even airflow—can significantly reinforce the learning process. Furthermore, measuring learner engagement is essential to ensuring that learners remain actively involved in learning. Various detection methods can assess engagement levels; in this study, we focus on analyzing engagement through facial expressions, particularly in a mulsemedia-synchronized learning environment.

Modern Facial Expression Recognition (FER) systems have achieved significant results through deep learning techniques. However, existing FER systems face two primary challenges: overfitting due to limited training datasets, and additional complications unrelated to expressions, such as occlusion, pose variations, and illumination changes. To improve the performance of FER in analyzing learners' engagement within a mulsemedia-based learning environment and to address some of these challenges, we propose three key aspects.

In our first study, face detection is a crucial step for identifying and cropping faces to train FER models. We observed that the conventional Viola-Jones face detection algorithm often produced false positives, particularly in complex images containing multiple faces or cluttered backgrounds. To address this issue, we enhanced the Viola-Jones algorithm by integrating particle swarm optimization to improve prediction accuracy in challenging images. The integration optimizes threshold selection and refines feature selection, enabling AdaBoost within the Viola-Jones framework to focus on the most relevant features for constructing a robust classifier. This enhancement significantly reduces false positives by fine-tuning feature selection and cascade thresholds, thereby improving prediction accuracy in complex environments.

In our second study, we observed that existing supervised FER approaches are inadequate for analyzing spatiotemporal features in real-time environments involving

dynamic facial movements. To overcome this limitation, we introduced a fusion of convolutional neural networks and Bidirectional Long Short-Term Memory (Bi-LSTM) networks to recognize emotions from facial expressions and capture relationships between sequences of expressions. Our approach employs a VGG-19 architecture with optimized hyperparameters and TimeDistributed layers to independently extract spatial features from each frame within a sequence. These spatial features are subsequently fed into a Bi-LSTM, which captures temporal relationships across frames in both forward and backward directions. This fusion enhances the model's ability to recognize emotions from expression sequences. The proposed method achieves significant accuracy in FER analysis, with results compared against baseline techniques.

In our third study, we introduced a Deep Semi-Supervised Convolutional Sparse Autoencoder to address the limitations of supervised FER approaches, particularly their reliance on extensive datasets and the challenges posed by imbalanced facial expression distributions, which can adversely affect model performance. This approach consists of two main stages. In the first stage, a deep convolutional sparse autoencoder is trained on unlabeled facial expression samples. Sparsity is introduced in the convolutional block through penalty terms, encouraging the model to focus on extracting the most relevant features for latent space representation. In the second stage, the trained encoder's feature map is connected to a fully connected layer with a softmax activation function for fine-tuning, forming a semi-supervised learning framework. This approach enhances FER accuracy in real-time environments. Furthermore, these two approaches were conducted using the Extended Cohn-Kanade+, Japanese Female Facial Expression, and an In-house dataset. Model performances are evaluated using metrics including accuracy, precision, recall, F1-score, the confusion matrix, and the receiver operating characteristic curve.

Finally, all proposed methods were integrated to effectively analyze learner engagement levels in mulsemedia-synchronized learning environments. To achieve this, a mulsemedia-synchronized web portal was developed, incorporating olfactory, vibration, and airflow effects. The FER system mapped eight facial expressions to three engagement levels—highly engaged, engaged, and disengaged—based on the system's predicted probability scores and predefined threshold values. The final results demonstrate that mulsemedia-based learning significantly improved learning outcomes and memory retention compared to conventional methods.