Chapter 1
Introduction

In recent years, advancements in digital media have significantly enhanced education, providing flexible and interactive learning environments for students. Moreover, educational leaders are continually exploring innovative ways to integrate technology into physical classrooms, aiming to improve student learning outcomes and better prepare students for a high-tech future.

Technology-Enhanced Learning (TEL) refers to any technology, such as laptops, tablets, and virtual learning environments, used to enrich learners' educational experiences [1-2]. According to a 2022 report, the use of educational technology (EdTech), including TEL tools, increased by 99% following the COVID-19 pandemic [3]. This surge has transformed education, making learning more accessible, flexible, and engaging. Over the past decade, advances in digital tools and internet connectivity have enabled TEL to thrive in both formal and informal educational settings. Today, online courses, virtual classrooms, adaptive learning software, and immersive technologies like virtual reality (VR) support a variety of learning styles and needs [4].

This growth is fueled by the increasing demand for personalized, on-demand education that meets the needs of diverse student populations, including remote learners, working professionals, and those with unique learning preferences. TEL platforms, such as learning management systems (LMS) and mobile apps, allow students to learn at their own pace, access a vast array of resources, and engage in interactive and collaborative learning activities.

As a research field, TEL investigates how technologies can enhance learning and teaching processes to achieve better outcomes and seeks to develop new technologies that improve these processes. TEL researchers ask a wide range of research questions—explorative, descriptive, analytical, predictive, interventionist, design-oriented, normative, and artistic—and employ diverse data collection and analysis methods, including quantitative and qualitative, as well as inductive and deductive approaches [5].

Furthermore, educational institutions and policymakers recognize the importance of integrating TEL to enhance learning outcomes and prepare students for a technology-driven world. The rapid expansion of TEL is expected to continue as emerging technologies, like artificial intelligence (AI) and multi-sensory media, create even more innovative, personalized, and immersive learning experiences.

1.1 Mulsemedia in the Learning Environment

1.1.1 Mulsemedia

Mulsemedia, short for "Multiple Sensorial Media," refers to multimedia systems that engage multiple senses beyond just sight and sound. This technology aims to create more immersive and engaging experiences by incorporating various sensory stimuli—such as taste, smell, touch, and even temperature—into a variety of applications, as mentioned below. The term "mulsemedia" was coined by Prof. George in his article [6]. Figure 1.1 shows the components of mulsemedia.

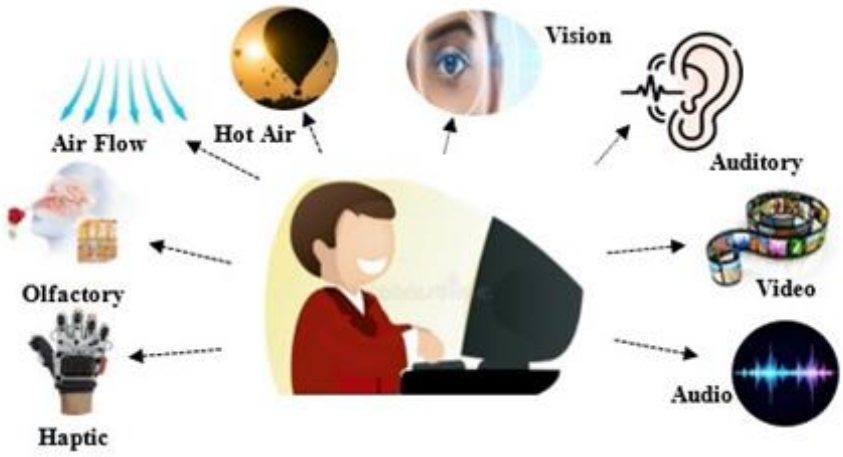

Figure 1.1. Mulsemedia Components

1.1.2 Applications of Mulsemedia Environment

The following are potential applications of Mulsemedia, as outlined in [7]

- **Entertainment:** Enhancing movies, video games, and virtual reality experiences by adding sensory elements that align with the on-screen action.
- **Education and Training:** Simulations for medical training, flight training, or historical recreations that engage multiple senses to enhance learning and retention.

- **Marketing and Advertising:** Creating memorable and impactful campaigns by engaging multiple senses to leave a stronger impression.
- **Healthcare:** Therapeutic uses, such as virtual reality for pain management, rehabilitation, and mental health treatments.
- **Retail:** Enhancing online shopping experiences by allowing customers to feel and even smell products virtually.

1.1.3 Components of Mulsemedia

The following compounds have been commonly included in the Mulsemedia environment, as outlined in [8]:

- **Visual:** Visual elements are the foundation of traditional multimedia and include images, videos, animations, and augmented reality (AR) or VR content. Visual stimuli in mulsemedia aim to captivate users by presenting realistic or highly interactive graphics that simulate real-world or imaginary environments. In VR settings, for example, visual components are carefully designed to provide a 360-degree perspective, immersing users in lifelike scenes that enhance the overall sensory experience.
- **Auditory:** Auditory components encompass sound effects, background music, and spoken language, which enhance engagement and emotional response. By integrating auditory cues, mulsemedia can simulate realistic soundscapes, including spatial audio that changes based on user interaction or positioning within a virtual environment. For example, in a VR tour of a forest, users might hear birds chirping or leaves rustling around them, heightening the sense of immersion and making the experience feel more authentic.
- **Haptic:** Haptic feedback engages the sense of touch, typically through devices like vibration motors, force feedback systems, or wearable technologies that provide sensations directly on the skin or muscles. These devices can simulate textures, resistance, and other tactile sensations, making digital interactions feel more physically engaging. For example, in a VR scenario where users "pick up" virtual objects, haptic gloves or controllers can provide feedback that mimics the weight and texture of the object, helping users feel more connected to the digital world.

- **Olfactory:** Olfactory components deliver scent-based stimuli, adding an extra layer of realism by engaging the sense of smell. Scent dispensers or olfactory devices release controlled amounts of fragrance at specific moments in an experience. This can enhance immersion in applications like VR or AR. For instance, a culinary learning module might emit the aroma of freshly baked bread, or a nature documentary might release forest scents, helping users to feel as if they are present in these environments.
- **Gustatory:** Gustatory stimuli are still in the experimental stages of mulsemedia and involve creating taste-based sensations. This sensory component faces technological challenges but has significant potential for applications in food-related experiences or culinary training. While more difficult to implement, experimental gustatory devices aim to simulate basic tastes (such as sweet, sour, bitter, salty, and umami) through electrodes or other methods that stimulate the tongue. These devices could one day allow users to "taste" foods virtually, opening new possibilities for sensory-rich experiences.
- **Thermal:** Thermal feedback involves changes in temperature to simulate environmental conditions or specific sensations, adding another layer of realism. For instance, a VR experience set in a desert might gradually increase the temperature around the user, while a winter-themed scene might provide cool air to mimic a snowy landscape. Thermal feedback devices, such as heated or cooled wearables, help create realistic sensations that make digital experiences feel more lifelike, engaging users' bodies in the simulation as well as their minds.

1.1.4 Mulsemedia in a Learning Context

In recent years, interactive multimedia has enabled dynamic and immersive learning experiences by integrating various media such as audio, text, video, images, and animation. VR and AR have significantly transformed the educational landscape by offering immersive and engaging learning experiences [9]. Multimedia makes the learning environment more interactive, aiding learners in retaining information for longer periods. The role of multimedia in education is to develop content in an interactive format using videos, images, audio, and animation to enhance learner interest and the overall learning process [9]. However, human perception encompasses more than just two senses in learning. Mulsemedia, or multiple sensorial media, combines more than one sense (e.g., visual, auditory, haptic, gustatory,

airflow, water jet, and olfactory) to provide a more interactive and immersive learning experience compared to conventional multimedia learning [10]. Conventional multimedia learning presents information interactively, enhancing cognitive skills by making it easier to understand and recall information quickly [10].

Recent research aims to incorporate multisensory effects into conventional multimedia learning environments, exploring multisensory integration to create immersive and enhanced learning experiences. This field examines the combination of visual, auditory, haptic, olfactory, and gustatory stimuli to deliver cohesive and engaging media content. The human brain has evolved to develop, learn, and function optimally in multisensory environments, suggesting that multisensory-based learning better mirrors natural settings and, as a result, is more effective for learning. By providing students with multiple ways to engage with course material, educators can create a more dynamic and interactive learning experience that caters to diverse learning styles. Mulsemedia research seeks to understand cross-modal perception, develop inclusive technologies, and leverage neuroscientific insights to create more intuitive and impactful media experiences.

1.1.4.1 Limitations in Conventional Multimedia Content

Conventional multimedia content, while widely used and beneficial in many ways, has several limitations and challenges that can impact the effectiveness of the learning experience. Here are some common problems associated with conventional multimedia content:

- **Limited Engagement:**
  - ***Passive Learning:*** Conventional learning content often involves passive consumption of content through reading or watching videos, which can lead to disengagement and lower retention.
  - ***Monotony:*** Repetitive and monotonous content can fail to capture and maintain students' interest.
- **Lack of Interactivity:**
  - ***Minimal Interaction:*** Conventional learning platforms may lack interactive elements such as quizzes, simulations, or interactive discussions, which are crucial for reinforcing learning.

- ***Limited Hands-on Practice:*** Subjects that require hands-on practice, such as lab-based sciences or technical skills, are challenging to teach effectively without interactive simulations.

- **Insufficient Sensory Engagement**:
  - ***Limited Sensory Stimulation:*** Conventional learning content primarily engages sight and sound, neglecting other senses which can enhance learning experiences.
  - ***Poor Real-World Simulation:*** Without engaging multiple senses, it's difficult to create realistic simulations that replicate real-world scenarios.

Integrating mulsemedia into conventional multimedia content can address many of these issues by making learning experiences more engaging, interactive, and realistic. By engaging multiple senses and providing more immersive content, mulsemedia can enhance student motivation, retention, and overall learning outcomes. Additionally, learning enjoyment is also an important factor in the learning context [11].

1.1.5 Measuring the Effectiveness of Mulsemedia by Quality of Experience

Quality of Experience (QoE) refers to a measure used to evaluate the satisfaction of users with a particular service or product [12]. It is commonly applied in telecommunications, multimedia services, and other IT-related fields. Unlike Quality of Service (QoS), which focuses on the technical aspects and performance metrics of a service, QoE emphasizes the end-user's overall experience and perception.

In the context of Mulsemedia learning, QoE refers to the learner's overall satisfaction and the effectiveness of the learning experience. It includes various factors that influence how learners perceive, interact with, and benefit from multimedia learning content and environments, often gathered through direct responses from users regarding their satisfaction.

When assessing the quality and effectiveness of mulsemedia research, a combination of subjective and objective measures is commonly employed [13]. Subjective measures involve gathering opinions and perceptions from human observers, typically through methods such as surveys and focus groups. These measures provide valuable insights into the overall effectiveness of mulsemedia, particularly regarding user experience and feedback. Objective measures, on the other hand, rely on quantitative data, such as physiological signals, to

evaluate technical aspects, including efficiency, accuracy, and speed. By integrating both subjective and objective measures, researchers can gain a more comprehensive understanding of the strengths and weaknesses of mulsemedia.

1.2 Emotion Recognition in the Learning Environment

Emotion recognition in the learning environment is an emerging technology designed to enhance the learning experience by identifying and responding to students' emotional states [14]. Recognizing students' emotions offers several benefits, such as determining which teaching styles or materials are most likely to foster positive feelings. As a result, learning materials or teaching methods can be tailored to spark students' interest in learning. Studies suggest that positive emotions in students promote engagement and a stronger intention to learn. Conversely, negative emotions can lead to poor learning outcomes, influence students' educational trajectories, and diminish their interest in learning. An enthusiastic student is significantly more likely to succeed in learning compared to a bored one, highlighting the importance of adapting teaching strategies to students' emotional states.

Emotion recognition technology employs various methods, including facial expression analysis, speech analysis, and physiological signals such as heart rate variability. The overarching goal is to enhance the learning process by understanding learners' emotional states and responding in ways that optimize their engagement and outcomes.

1.2.1 Benefits of Incorporating Emotion into Learning Environments

- **Personalized Learning**:
    - ***Adaptive Content*:** By recognizing a student’s emotional state, learning platforms can adapt the content delivery in real time. For example, if a student appears frustrated, the system can offer additional support or simpler explanations.
    - ***Customized Feedback*:** Educators can receive insights into students' emotional responses, enabling them to provide more tailored feedback and support.
- **Enhanced Engagement**:
    - ***Interactive Learning*:** Emotionally aware systems can make learning more interactive and engaging by adjusting difficulty levels, incorporating game elements, or providing encouraging feedback when needed.

- ***Real-Time Interventions:*** Instructors can be alerted to emotional distress or disengagement, allowing for timely interventions to keep students on track.

- **Improved Learning Outcomes**:
  - ***Emotional Support:*** Recognizing emotions helps in addressing not only academic needs but also emotional well-being, creating a more supportive learning environment.
  - ***Motivation and Satisfaction:*** Understanding and responding to students' emotions can boost motivation and satisfaction, leading to better retention and academic performance.

1.2.2 Emotion Recognition from Different Sources

In the context of learning environments, recognizing emotions is crucial to tailoring educational experiences that align with learners' emotional states. Emotion can be recognized from different sources such as facial expression, voice, hand signals, body motion, and physiological signals as shown in Figure 1.2, as outlined in [15].

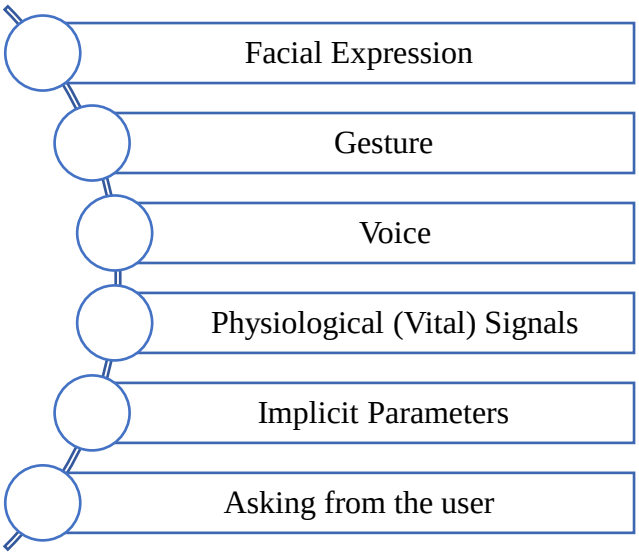

Figure 1.2 Emotion Recognition from Different Sources [15]

1.2.2.1 Facial Expression Recognition

Facial expression recognition (FER) involves analyzing facial movements to identify emotions by employing various techniques. Feature-based methods focus on detecting key facial features such as the eyes, mouth, and eyebrows. Appearance-based methods analyze

the overall texture and appearance of the face. Machine learning approaches involve training models on labeled facial expression data to classify emotions. Deep learning, particularly convolutional neural networks (CNNs), offers more accurate recognition by learning complex patterns from large datasets. These methods collectively enhance the accuracy of emotion recognition from facial expressions.

1.2.2.2 Gesture

Emotion recognition from gestures involves interpreting human movements to understand emotional states. This includes hand gestures, which analyze the movement and position of the hands; body posture, which focuses on overall stance and movements; and kinetic signals, which assess the speed and fluidity of motion. Sensor-based methods use accelerometers and gyroscopes to capture movement data, while vision-based methods employ cameras and computer vision algorithms to recognize gestures. These approaches work together to provide a comprehensive understanding of emotions through physical expression.

1.2.2.3 Voice

Voice recognition for emotion detection focuses on analyzing vocal attributes to identify emotional states. Key aspects include prosody, which examines the rhythm, pitch, and intonation of speech; speech rate, referring to how quickly a person speaks; volume, or the loudness of the voice; timbre, which is the quality and tone of the voice; and speech content analysis, which looks at the words and phrases used. These elements together help to assess emotions by interpreting how speech sounds and what it conveys.

1.2.2.4 Physiological Signals

Physiological signals offer valuable insights into emotions by measuring bodily responses. Key signals include heart rate, which looks at variations and changes in heart rate; skin conductance, indicating sweat gland activity associated with arousal; electroencephalography (EEG), which tracks brainwave patterns to detect emotional states; electromyography (EMG), which measures muscle activity, especially in facial expressions; and respiration rate, which examines changes in breathing patterns that correlate with emotional responses. These physiological signals collectively help in understanding and identifying emotions based on physical reactions.

1.2.2.5 Implicit Parameters

Implicit parameters are subtle indicators that reveal emotions indirectly. These include eye movement, which looks at patterns and focus of gaze to infer attention or emotion; pupil dilation, where changes in pupil size can signal emotional arousal; micro-expressions, which are brief, involuntary facial expressions that reflect true feelings; body temperature, where variations in skin temperature can indicate emotional states; and contextual information, which takes into account the surrounding environment and situation that may influence emotional reactions. These parameters help to uncover emotions that may not be overtly expressed.

1.2.2.6 Asking the user

Directly querying users involves methods where individuals explicitly share their emotional states. These include self-reporting, where users describe their emotions in their own words; questionnaires, which provide structured forms for detailed emotional feedback; interviews, involving verbal discussions to understand emotions more deeply; experience sampling, which collects real-time emotion data through mobile or digital devices; and diaries, where users regularly record their emotional experiences. These methods offer direct insight into how individuals feel, providing valuable data for emotion recognition.

Among these emotion recognition approaches; FER is identified as an effective method for detecting students' emotions through their facial expressions [15].

1.3 Emotion Recognition using FER in Learning Environment

In the context of learning environments, FER can play a significant role in understanding students' emotions during their educational experiences, enabling a more adaptive and personalized approach to teaching and learning. By analyzing students' facial expressions, educators and systems can gain insights into emotional states such as frustration, confusion, engagement, or joy, and adjust the learning process accordingly. This is achieved by capturing students' facial expressions in the learning environment using cameras, processing the sequences of images or frames, and extracting facial data through detection methods. Subsequently, students' learning statuses are assessed using expression recognition techniques.

Facial emotion is one of the most potent, natural, and universal ways for individuals to convey their emotions and thoughts, transcending differences in gender, ethnicity, and nationality. It can be recognized from static images or sequences of images and videos. Research indicates that spoken words contribute only 7% to a message's overall impact, while vocal aspects, such as tone and pitch, account for 38%, and facial expressions make up 55% [16]. This highlights the critical importance of facial expressions in human communication.

Ekman and Friesen identified six basic emotions—happiness, sadness, surprise, fear, anger, and disgust—which are often referred to as universal expressions or primary emotions. Some researchers also include neutral and contempt in this category [17]. However, recent advances in neuroscience and psychology suggest that the six basic emotions model may be culture-specific rather than universally applicable.

Facial expressions are generated by the complex interplay of facial muscles, each muscle contributing to the nuanced movements that convey emotions. The process begins when the brain sends signals via the facial nerve to specific muscles. For instance, the frontalis muscle raises the eyebrows to indicate surprise, while the orbicularis oculi muscles around the eyes contract to create genuine smiles, known as Duchenne smiles. Muscles like the corrugator supercilii pull the eyebrows together, expressing concentration or worry. Meanwhile, the zygomaticus major and minor muscles lift the corners of the mouth to form a smile.

The coordinated actions of muscles such as the orbicularis oris, which controls lip movements, and the buccinator, which tenses the cheeks, enable a wide range of expressions. Each muscle's contraction and relaxation pattern are finely tuned to produce the micro-expressions that reflect human emotions, from happiness and sadness to anger and surprise, making facial expressions a powerful non-verbal communication tool. Figure 1.3 shows the facial muscles and examples of how these expressions are coded numerically to identify emotions.

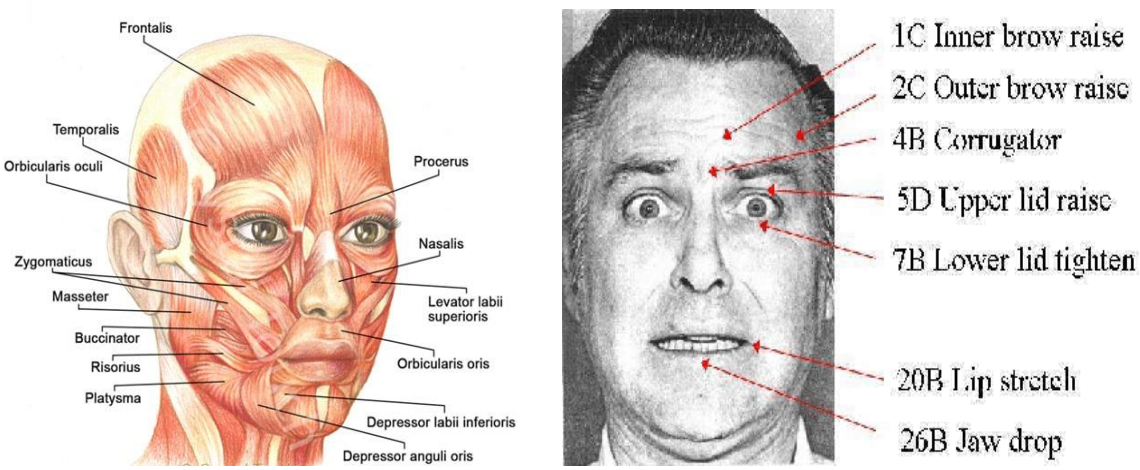

Figure 1.3. (a) Facial Muscles (b) Disgust facial expression muscle Movement [18]

1.3.1 Facial Action Coding System

The Facial Action Coding System (FACS), created by Paul Ekman and Wallace V. Friesen in the 1970s, is a method used to analyze facial expressions by breaking them down into basic Action Units (AUs) [18]. Each AU corresponds to the activation of specific facial muscles, which are associated with different emotional expressions. The system categorizes facial movements based on anatomical changes, allowing for a detailed and objective way to measure and study emotional expressions.

In the context of the disgusted facial expression, certain muscles are involved in creating the distinct features that characterize this emotion. Figure 1.3 (b) likely highlights these muscle movements, which typically include the contraction of the levator labii (raising the upper lip), the depressor anguli oris (pulling the corners of the mouth downward), and the nasalis (narrowing the nostrils), among others. These facial muscle actions combine to convey the emotion of disgust.

The following components are involved in FACS

- **Action Units (AUs)**
    - Each AU represents the movement of a specific set of facial muscles.
    - Examples include AU1 (Inner Brow Raiser), AU6 (Cheek Raiser), and AU12 (Lip Corner Puller).

- **Intensity Scores**
    - Each AU is scored on a five-point scale from A to E, representing the intensity of the movement, with A being the slightest trace and E being the maximum intensity.
- **Combinations of Aus**
    - Complex expressions are formed by combining multiple AUs. For instance, a genuine smile involves the combination of AU6 (Cheek Raiser) and AU12 (Lip Corner Puller).
- **Temporal Aspects**
    - FACS also considers the timing of facial movements, including their onset, apex, and offset, to capture the dynamics of facial expressions.

1.3.2 Conventional FER Process

The FER system consists of three primary steps: face detection, pre-processing, feature extraction, and emotion classification as shown in Figure 1.4.

Figure 1.4 Conventional FER Process

Initially, face detection is carried out to ascertain the presence of faces within images or video frames. Following this, pre-processing techniques are applied to enhance the image features, making them more pronounced and easier to analyze. In the third step, feature extraction, the system identifies key features from the detected faces to determine the emotions reflected in the facial expressions. Finally, the emotion classification step involves using a trained classifier to categorize the identified emotions based on the extracted features. This systematic approach ensures accurate and efficient recognition of facial emotions.

1.3.3 FER using Machine Learning

Conventional FER systems often use handcrafted feature extraction techniques such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and

Local Binary Patterns (LBP) to extract features from facial images [19]. Subsequently, geometric-based and appearance-based feature extraction methods were introduced to automate the FER process. According to various studies, geometric-based approaches retrieve geometric characteristics related to facial action units. Methods like the Active Appearance Model (AAM) and the Active Shape Model (ASM) extract geometric features based on the shape and location of facial expressions [20]. On the other hand, appearance-based techniques are more resilient to noise and more effective in feature extraction compared to geometric-based methods. In some cases, hybrid-based feature extraction approaches have been employed, leading to improved detection performance, though these are typically suitable only in clinical settings. Real-time emotion detection faces several challenges, including complex backgrounds, occlusions, varying illumination, and spontaneous expressions. Cultural differences can also lead to subtle variations in spontaneous expressions compared to typical expressions. Traditional feature extraction algorithms often result in high computational costs, extended learning times, and poor real-time performance under these conditions [20]. Additionally, complex images require substantial memory and the identification of discriminative visual features to accurately distinguish facial emotions and correlate them with the corresponding emotional state.

1.3.4 FER using Deep Learning

The advancement of deep learning has significantly impacted the computer vision sector, driven by increased computational power from graphics processing units (GPUs) and tensor processing units (TPUs) that support large-scale data training. This has led to substantial improvements in image recognition and detection. Convolutional neural networks (CNN), in particular, have achieved remarkable success in FER due to their robust feature extraction capabilities. Convolutional networks are extensively used with static images to extract appearance-based features. However, these features alone are insufficient to fully capture emotions, as static images lack the dynamic sequences associated with facial expressions [21]. Many two-dimensional (2D) CNN models fail to recognize temporal features in images [22]. Consequently, numerous studies have combined CNNs with Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNNs) to address this limitation. These integrated approaches extract the dynamic sequence of features from images, enhancing the accuracy and effectiveness of FER systems.

1.3.5 Limitation in Existing FER Systems

Facial expression recognition involves identifying and interpreting human facial expressions using computational methods. Despite significant advancements, several challenges remain in this field.

Here are some key challenges as identified in a systematic review.

- **Variability in Expressions**
  - **a.** *Subtle Differences*: Some emotions have very subtle differences that can be hard to distinguish even for humans, such as the difference between a polite smile and a genuine smile.
  - **b.** *Intensity Levels:* Expressions can vary in intensity, making it difficult to detect less intense emotions.

- **Inter-individual Differences**
  - a. *Cultural Variations:* Different cultures may express and interpret emotions differently.
  - b. *Personal Differences:* Individual differences in expressing emotions, such as unique facial features or personal habits, add to the complexity.

- **Environmental Factors**
  - a. *Lighting Conditions:* Variations in lighting can obscure or distort facial features.
  - b. *Occlusions:* Accessories like glasses, hats, or masks, and elements like hair can cover parts of the face, complicating detection.
  - c. *Camera Angles and Distances:* Different angles and distances from the camera can affect the visibility and clarity of facial expressions.

- **Dataset Limitations**
  - a. *Lack of Diversity:* Many FER datasets lack diversity in terms of ethnicity, age, and gender, leading to biased models.
  - b. *Size of Dataset:* High-quality annotated datasets are limited, hindering the training of robust models.

c. *Synthetic vs. Real-world Data:* Real-world data is more challenging due to uncontrolled environments, while synthetic or lab-generated data might not generalize well.

- **Dynamic Expressions**
  a. *Temporal Dynamics:* Emotions are dynamic and evolve. Capturing the temporal aspects of expressions requires sophisticated models that can process sequences of images or videos.
  b. *Micro-expressions:* These are brief, involuntary facial expressions that occur in response to emotions and are difficult to detect and interpret.
- **Technical and Computational Challenges**
  a. *High Computational Costs:* Real-time FER requires significant computational power and efficient algorithms.
  b. *Algorithmic Bias:* Machine learning models can inherit biases from training data, leading to inaccurate or unfair recognition.
  c. *Feature Extraction:* Identifying the most relevant features from facial images that correlate with different emotions is challenging.

1.4 Taxonomy of Learner Engagement Detection Methods

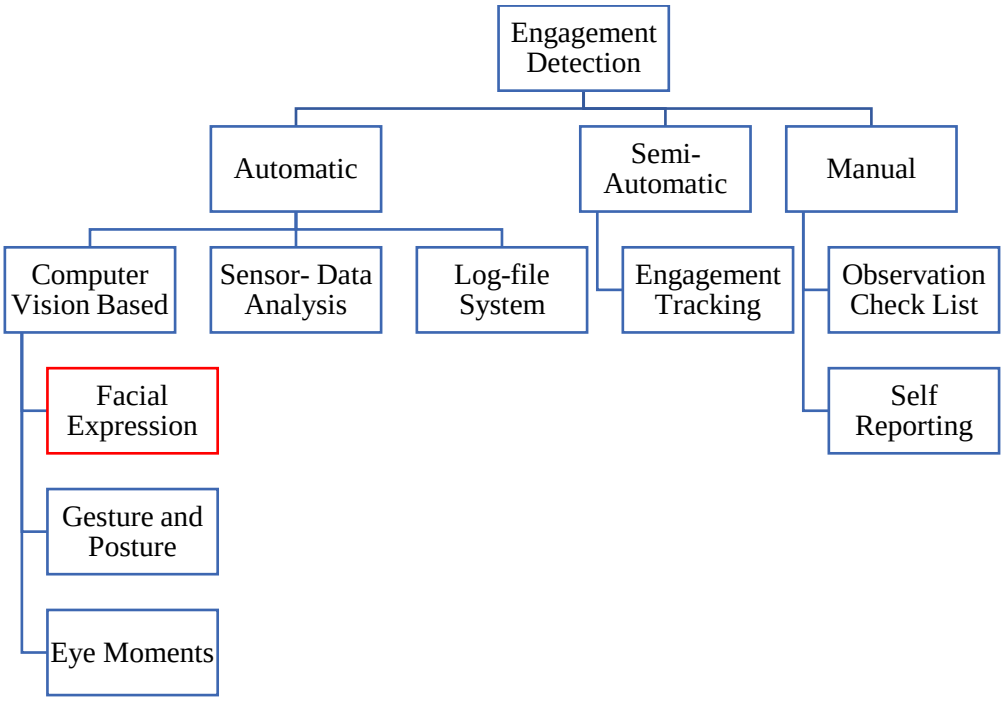

Figure 1.5. Learner Engagement Detection Methods

Learner engagement detection methods aim to measure and understand how engaged students are during learning activities. These methods integrate various technological and analytical approaches to capture, analyze, and interpret engagement levels [23]. Figure 1.5 illustrates the different types of learner engagement detection methods.

1.4.1 Automatic Methods

Automatic methods require minimal to no direct involvement from learners. They use various technologies to detect engagement levels.

- ***Computer Vision-Based:*** Utilizes visual data to analyze facial expressions, gestures, posture, and eye movements. This approach is particularly promising in online learning environments.
    - ***Facial Expression:*** Analyzes facial movements to infer emotional states and engagement levels.
    - ***Gesture and Posture:*** Examines body movements and positions to assess attention and engagement.
    - ***Eye Movements:*** Tracks eye movements and gaze patterns to determine focus and engagement.
- ***Sensor-Data Analysis:*** Uses data from sensors (e.g., wearable devices) to monitor physiological signals like heart rate and skin conductance, which are indicators of engagement.
- ***Log-File System:*** Analyzes interaction logs from learning platforms, such as click patterns, time spent on tasks, and navigation behaviors, to infer engagement.

1.4.2 Semi-Automatic Methods

Semi-automatic methods combine automated data collection with some level of learner participation.

- ***Engagement Tracking***: Involves monitoring and tracking learner interactions and behaviors within a learning environment, often requiring minimal input from the learner to validate or supplement the data.

1.4.3 Manual Methods

Manual methods rely heavily on direct involvement from learners or observers.

- ***Observation Checklist***: Instructors or observers use structured checklists to manually record and assess engagement behaviors during learning activities.
- ***Self-reporting:*** Learners self-report their engagement levels through surveys, questionnaires, or diaries, providing subjective data on their own engagement experiences.

Among various learner engagement detection methods, FER stands out as particularly effective because it offers real-time, non-intrusive insights into students' emotional states and engagement levels. Numerous studies have explored the use of facial expressions to estimate the emotions experienced by learners. FER leverages advanced computer vision techniques to analyze subtle facial movements, capturing a range of emotions such as interest, confusion, or boredom. This method's ability to operate continuously without requiring direct input from learners makes it especially valuable in learning environments, where traditional observational or self-reporting methods may be impractical. Additionally, FER's automatic and scalable nature allows for consistent monitoring of engagement across large groups of students, providing educators with immediate and actionable feedback to enhance the learning experience.

1.5 Motivation for Research

The rapid growth of TEL platforms has transformed the educational landscape, making learning more interactive and effective than ever before. Yet, conventional learning environments often lack immersive, multisensory elements, limiting student engagement and interaction with the content. Mulsemedia-based learning environments offer a solution by integrating multiple sensory channels, such as sight, sound, and even touch, to create a more immersive and engaging experience.

While numerous research efforts have been conducted on FER over the past decades, recognizing learners' emotional states through facial expressions remains limited, with little to no focus on deploying such models in real-time environments. Therefore, an effective FER system is needed to analyze learner engagement and dynamically adjust learning content

based on the learners' emotional states. Moreover, emotion recognition for learners in a mulsemedia-synchronized learning environment has not yet been explored.

1.6 Research Problem Statement

Given the limitations in automatically recognizing learner engagement in real-time environments, the primary aim is to design an effective automatic FER approach to identify learner engagement levels in mulsemedia-synchronized learning environments.

1.7 Research Objective

1.7.1 Primary Objective

- To develop an effective FER model for predicting learners' engagement through universal expressions using a deep learning approach in a mulsemedia-enhanced learning environment.

1.7.2 Secondary Objective

- To experiment with various deep learning approaches (3D-CNN, LSTM, and various types of Autoencoder) for building effective way of FER systems in mulsemedia learning environments using a universal facial expression dataset.
- To investigate the effectiveness of learners' satisfaction with mulsemedia-synchronised content, fostering improved engagement and knowledge retention among learners.
- Maximize classification accuracy and detection rate in FER.
- Minimize prediction time for FER.
- Maximize precision and recall scores in emotion classification.

1.8 Significant Contribution of Thesis

Despite the numerous FER methods available in the literature, no universal solution addresses all the identified issues. To address some of these concerns, this thesis aims to improve FER performance by considering specific characteristics highlighted in Sections 1.3.5. The thesis makes five contributions by exploring various aspects of FER, categorized as follows:

- **Understanding Deep Learning Techniques for FER:** An extensive review of FER techniques has been conducted using the Preferred Reporting Items for Systematic

Reviews and Meta-Analyses (PRISMA) framework to highlight significant advancements in the field. Additionally, this review examines the role of mulsemedia in learning, particularly it's potential to enhance the learning process and identify learner engagement. The literature review also presents taxonomies of several FER procedures, and the proposed classification system facilitates the selection of the most appropriate techniques for different FER-related scenarios. The accuracy, advantages, and disadvantages of several state-of-the-art, static-based FER approaches are also investigated experimentally.

- **Enhancing Viola-Jones Face Detection Algorithm Prediction Accuracy**: The conventional Viola-Jones algorithm employs AdaBoost for classifying faces in images and videos. The challenge lies in working with cluttered real-time facial images. AdaBoost needs to search through all possible thresholds for all samples to find the minimum training error when receiving features from Haar-like detectors. This exhaustive search consumes significant time to discover the best threshold values and optimize feature selection to build an efficient classifier for face detection. To address this, we proposed enhancing the conventional Viola-Jones algorithm by incorporating Particle Swarm Optimization (PSO) to improve its predictive accuracy, particularly in complex face images. We leverage PSO in two key areas within the Viola-Jones framework. Firstly, PSO is employed to dynamically select optimal threshold values for feature selection, thereby improving computational efficiency. Secondly, we adapt the feature selection process using AdaBoost within the Viola-Jones algorithm, integrating PSO to identify the most discriminative features for constructing a robust classifier. Our approach significantly reduces the feature selection process time and search complexity compared to the traditional algorithm, particularly in challenging environments. We evaluated our proposed method on a comprehensive face detection benchmark dataset, achieving impressive results, including an average true positive rate of 98.73% and a 2.1% higher average prediction accuracy compared to both the conventional Viola-Jones approach and contemporary state-of-the-art methods.

- **FER using CNN-BiLSTM architecture**: Existing FER systems primarily focus on spatial features for emotion recognition, but they struggle to accurately identify emotions from dynamic sequences of facial expressions in real-time. We propose deep

learning techniques that fuse CNN and Bidirectional LSTM (BiLSTM) to recognize emotions by leveraging spatiotemporal features, enabling the identification of relationships between sequences of facial expressions. This approach employs a hyperparameter-tuned VGG-19 model with time-distributed layers to automatically extract spatial features from a sequence of images, addressing the limitations of conventional feature extraction methods. Next, these feature sequences are fed into a BiLSTM network to analyze temporal features in both directions, enabling the recognition of emotions from a sequence of expressions. Experimental results demonstrate that the proposed method outperforms both baseline methods and state-of-the-art approaches.

- **FER using Semi-supervised Convolutional Sparse Autoencoder:** Most deep learning approaches in supervised FER systems heavily rely on large, labeled datasets. Implementing FER in CNNs often requires many layers, leading to extended training times and difficulties in finding optimal parameters. This can result in challenges in creating distinct facial expression patterns for classification, leading to poor real-time emotion classification. Therefore, we introduce a new approach called the Deep Semi-supervised Convolutional Sparse Autoencoder (DSCSA) to address these issues and enhance FER performance and prediction accuracy. This approach comprises two parts: Initially, a deep convolutional sparse autoencoder is trained with unlabeled facial expression samples. Here, sparsity is introduced in the convolutional block to enforce penalties, focusing on more relevant features for feature representation in the latent space. A trained encoder with a feature map is connected to a fully connected layer with softmax for final fine-tuning with learned weights and labeled facial expression samples in a semi-supervised approach for emotion classification. The results were analyzed using established state-of-the-art techniques.

- **Designing a Mulsemedia-enabled Web portal and analysis of learner engagement:** We designed an IoT-based mulsemedia-synchronized learning platform that uses affordable components, such as cooling fans, humidifiers, and haptics, to create a multisensory learning environment. This system provides learners with an immersive experience by incorporating sensory effects, such as the aroma of

rosemary, along with vibrotactile and airflow effects associated with thunder and lightning. These effects are synchronized with traditional audiovisual content. The study involved 70 participants pursuing science degrees, divided into two equal-sized groups: an experimental group and a control group, to analyze the impact of mulsemedia on their learning experience. The results showed that mulsemedia-based learning significantly improved learning outcomes and increased enjoyment levels. Additionally, it enhanced the sense of reality in the conventional learning environment. Finally, the proposed FER model was integrated to analyze learner engagement levels in the mulsemedia-enhanced learning environment.

1.9 Organization of Thesis

The thesis is mainly divided into eight chapters and framed based on research objectives.

Chapter 1: Describes the importance of emotion recognition in learning, introduces a mulsemedia-based learning environment, and discusses the research objectives and significant contributions to the research problem.

Chapter 2: Briefly discusses existing research based on systematic review approaches and explains current FER datasets, methods, and mulsemedia-based approaches in learning.

Chapter 3: Presents an overview of the proposed methodology based on the research objectives, highlighting four key contributions.

Chapter 4: Provides an overview of the enhanced Viola-Jones face detection algorithm, incorporating optimized techniques such as PSO, and presents experimental results in real-time environments.

Chapter 5: Describes the proposed CNN and Bi-LSTM architecture for FER, utilizing spatiotemporal features to identify the relationships within a sequence of images. It also presents experimental results using both in-house and benchmark datasets.

Chapter 6: Presents the introduced deep semi-supervised convolutional sparse autoencoder to enhance FER performance, addressing issues in supervised FER approaches, and showcases the results using both in-house and benchmark datasets.

Chapter 7: Briefly analyzes the mulsemedia-synchronized learning environment and the results of learners' engagement levels through FER.

Chapter 8: Summarizes the four contributions, discusses limitations, and suggests future research directions.

1.10 Chapter Summary

Emotion recognition in learning environments plays a crucial role in enhancing learning experiences and motivating continuous learning for students. Various methods are used to recognize emotions, with FER being particularly important for understanding learners' emotions in an educational context. Additionally, mulsemedia concepts have been introduced in conventional learning environments to provide an immersive experience, surpassing traditional multimedia approaches. This chapter discusses the existing challenges in emotion recognition, particularly in FER, and explores approaches to effectively analyze learner engagement. It also presents the research problems, core objectives, significant contributions, and an outline of the remaining chapters.