xLSTM-FER: Enhancing Student Expression Recognition with Extended Vision Long Short-Term Memory Network

Qionghao Huang(✉) and Jili Chen

Zhejiang Key Laboratory of Intelligent Education Technology and Application,
Zhejiang Normal University, Jinhua, Zhejiang, China
{qhhuang,irelia}@zjnu.edu.cn

Abstract. Student expression recognition has become an essential tool for assessing learning experiences and emotional states. This paper introduces xLSTM-FER, a novel architecture derived from the Extended Long Short-Term Memory (xLSTM), designed to enhance the accuracy and efficiency of expression recognition through advanced sequence processing capabilities for student facial expression recognition. xLSTM-FER processes input images by segmenting them into a series of patches and leveraging a stack of xLSTM blocks to handle these patches. xLSTM-FER can capture subtle changes in real-world students' facial expressions and improve recognition accuracy by learning spatial-temporal relationships within the sequence. Experiments on CK+, RAF-DF, and FERplus demonstrate the potential of xLSTM-FER in expression recognition tasks, showing better performance compared to state-of-the-art methods on standard datasets. The linear computational and memory complexity of xLSTM-FER make it particularly suitable for handling high-resolution images. Moreover, the design of xLSTM-FER allows for efficient processing of non-sequential inputs such as images without additional computation.

Keywords: Facial Expression Recognition · Student Academic Performance · Memory Network · Vision xLSTM

1 Introduction

Student facial expression recognition is a burgeoning field with significant implications for educational technology. By analyzing students' facial cues, educators can gain insights into their emotional states, engagement levels [25], cognitive load [12], and academic performance [8,11] during learning activities [9]. The current student face recognition systems primarily include those based on traditional CNN-based and Vision Transformer [5] (ViT)-based approaches. The lightweight and efficient characteristics of CNNs have attracted the attention of early education researchers, leading to the development of a series of face recognition systems and teaching environments based on CNNs [22]. The ViT has

W. Zhang et al. (Eds.): APWeb-WAIM 2024, CCIS 2246, pp. 249–259, 2025.
https://doi.org/10.1007/978-981-96-0055-7_21