EMERSK -Explainable Multimodal Emotion Recognition With Situational Knowledge

Mijanur Palash and Bharat Bhargava, *Fellow, IEEE*

***Abstract*—Automatic emotion recognition has recently gained significant attention due to the growing popularity of deep learning algorithms. One of the primary challenges in emotion recognition is effectively utilizing the various cues (modalities) available in the data. Another challenge is providing a proper explanation of the outcome of the learning. To address these challenges, we present Explainable Multimodal Emotion Recognition with Situational Knowledge (EMERSK), a generalized and modular system for human emotion recognition and explanation using visual information. Our system can handle multiple modalities, including facial expressions, posture, and gait, in a flexible and modular manner. The network consists of different modules that can be added or removed depending on the available data. We utilize a two-stream network architecture with convolutional neural networks (CNNs) and encoder-decoder style attention mechanisms to extract deep features from face images. Similarly, CNNs and recurrent neural networks (RNNs) with Long Short-term Memory (LSTM) are employed to extract features from posture and gait data. We also incorporate deep features from the background as contextual information for the learning process. The deep features from each module are fused using an early fusion network. Furthermore, we leverage situational knowledge derived from the location type and adjective-noun pair (ANP) extracted from the scene, as well as the spatio-temporal average distribution of emotions, to generate explanations. Ablation studies demonstrate that each sub-network can independently perform emotion recognition, and combining them in a multimodal approach significantly improves overall recognition performance. Extensive experiments conducted on various benchmark datasets, including GroupWalk, validate the superior performance of our approach compared to other state-of-the-art methods.**

***Index Terms*—Emotion Recognition, Deep Learning, Multimodal, Convolutional neural network (CNN), LSTM.**

I. Introduction

EMOTION shapes our social life by influencing our communications with others. Therefore, automatic human emotion recognition (ER) holds significant potential in various aspects of our lives. In the current era of online learning, which has become prevalent due to the Covid-19 pandemic, an integrated ER system can help teachers maintain an effective learning environment by providing insights into the emotional state of students. Similarly, a car equipped with driver emotion recognition capability can proactively prevent road rage or accidents by alerting the driver when they are tired, frustrated, or angry. Furthermore, the implementation of an ER system in CCTV cameras can enable the detection of individuals displaying anger near sensitive locations such as schools or children's playgrounds, triggering timely alarms to prevent potential harm, including deadly school shootings. Human-computer interactions, law enforcement and surveillance, interactive games, consumer behavior analysis, customer service enhancement, and healthcare are just a few examples of the diverse fields where emotion recognition technology can significantly impact outcomes and experiences.

Moreover, we perceive other people's emotions using both visual and non-visual cues. Visual cues include facial expressions, posture, gestures, eye movement, and walking gait, to name a few. Non-visual cues encompass speech, text, brain signals, and EEG signals, among others. Working with visual cues is more common than working with non-visual cues, as visual cues are more easily obtainable. Facial expressions and postures can be observed directly by looking at a person or analyzing images or videos captured by regular cellphones, CCTV cameras, or similar devices. However, the same convenience does not apply to non-visual cues. For instance, obtaining a brain scan requires specialized instruments to be attached to the individual, and obtaining permission for such procedures can be challenging. Moreover, the knowledge of them being recorded can potentially influence the subject's emotional state. Therefore, this work primarily focuses on visual cues, specifically facial expressions, postures, and gaits, for the purpose of emotion recognition.

Many of the existing works use only one type of cue (unimodal) such as facial expression [1], [2] or gait [3] etc. However, solely depending on a single mode can make the model less reliable in wild deployment. For example, a facial expression model trained on many of the existing benchmark datasets with no masked sample will perform poorly when encountering a subject wearing mask, which is very common during the Covid-19 pandemic. Similarly, a person's body may be blocked from the camera view by an obstacle and we may not have posture or gait information. So considering multiple modes at the same time can make the model more reliable. Moreover, several prior works show that combining multiple cues (multimodal) can results in higher accuracy in automatic emotion recognition [4], [5]. A person with a smiley face (mode 1) is possibly happy, but if we know they have open arms and stand straight (mode 2) we can be much more confident in our deduction.

Manuscript received 5 October 2022; revised 24 March 2023 and 14 June 2023; accepted 5 August 2023. Date of publication 10 August 2023; date of current version 2 February 2024. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yong Luo. *(Corresponding author: Mijanur Palash.)*

The authors are with the Department of Computer science, Purdue University, West Lafayette, IN 47906 USA (e-mail: mpalash@purdue; bbshail@purdue.edu).

Digital Object Identifier 10.1109/TMM.2023.3304015