FEELS: a full-spectrum enhanced emotion learning system for assisting individuals with autism spectrum disorder

Amir-Hossein Karimi* University of Waterloo, ON, Canada
Ameneh Boroomand* University of Waterloo, ON, Canada
Kaylen J. Pfisterer University of Waterloo, ON, Canada
Alexander Wong University of Waterloo, ON, Canada
* co-first authors

Abstract

Autism Spectrum Disorder (ASD) is a developmental disorder that can lead to a variety of social and communication challenges, and individuals with ASD are at a higher risk of loneliness and depression as a result of the disconnect and isolation they may feel from the rest of society as a result of their ASD. Interventions targeting improved emotional detection has been clinically shown to be quite promising; however, there are considerable barriers that make it challenging to incorporate emotion detection within daily life scenarios. Motivated by the need to fill this gap, we introduce the concept of FEELS, a full-spectrum enhanced emotion learning system which could be useful as a tool to assist individuals with ASD. FEELS facilitates enhanced emotion detection by capturing a live video stream of individuals in real-time, then leveraging deep convolutional neural networks to detect facial landmarks and a custom hybrid neural network consisting of a time distributed feed-forward neural network and a LTSM neural network to determine the emotional state of the individuals based on a sequence of facial landmarks over time. The feasibility of such an approach was explored through the construction of a proof-of-concept FEELS system that can detect between five different basic emotional states: neutral, sad, happy, surprise, and anger. Future work will include extending the proof-of-concept FEELS system to detect more emotional states and evaluate the system in more natural settings.

1 Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder characterized by difficulties communicating and interacting with other individuals in society, as well as restrictive and repetitive behaviour that can be viewed by others as unusual, and as such can lead to a variety of social deficits. Individuals with ASD often experience higher prevalence of loneliness, depression, and lower degrees of social support than their neurotypical counterparts [1, 2]. A 2012 review by Chevallier and colleagues [3] describe research regarding individuals with ASD within the context of social motivation theory. Social motivation theory is based on theory of mind or the capacity to read emotions both in others as well as one's own in order to explain and predict behaviour and understand that others' mental states may differ from one's own [3]. For those with ASD, this manifests itself in several ways including diminished capacity for: social orienting (e.g., decreased eye contact for those with ASD), engaging in social rewards like looking for and taking pleasure in social interactions (e.g., social anhedonia has been correlated with degree of severity of ASD), and social maintenance or working to foster social bonds (e.g., those with ASD "place less emphasis on preserving their reputation and managing their self image") [3].

However, for many individuals with ASD, these social deficits do not imply a lack of social interest [1]. To close the gap between deficit and desire, social skills training, including group-based social skills training, has been used as an intervention to address these social deficits [1, 4]. White and colleagues [1] summarize that many of these interventions have been successful at improving: theory of mind skills [5], peer relations [6], greeting and play skills [7], frequency of positive social behaviour [8], and emotion recognition [9, 2]. For example, high-functioning children saw significant improvements in ability to detect both basic (i.e., happiness, sadness, anger fear) and complex emotions (i.e., embarrassment, empathy, loneliness, surprise, pride) after the intervention [2]. However, there has been some speculation of whether these skills gained and observed in clinical settings may translate or transfer to improved skills in daily life [1, 2](i.e., the real-world and in real-time). To further complicate the situation, there are considerable barriers that make it challenging to incorporate emotion detection within daily life scenarios. Motivated by the need to fill this gap, we introduce the concept of FEELS, a novel full-spectrum enhanced emotion learning system that can potentially be an accessible technological solution to support real-time emotional cues "in the wild".

One of the key motivations behind FEELS is that the face plays an important role in visual communication. By looking at an individual's face, humans can automatically extract many nonverbal messages, such as the individual's identity, intent, and emotion [10]. Seminal research points to facial expressions as the biggest contributor (in fact, over 50% contribution) to conveying the meaning of a message [11]. In this work, we set out to design and build an integrated system to assist with the detection of a wide range of emotional states through real-time analysis of facial expressions.

As mentioned in their seminal review [12] Pantic and Rothkrantz show that the development of an automated system for facial expression analysis involves a three-stage pipeline: 1) face detection, 2) facial expression data extraction, and 3) facial expression classification. In the proof-of-concept FEELS system developed in this study, we overcome the challenges of the first two stages by employing deep convolutional neural networks made available through open-source libraries that has been demonstrated to achieve state-of-the-art performance for tracking the relationship between points of significance on an individual's face, known as facial landmarks. We tackle the third stage by leverage a custom neural network of our design that uses a series of detected facial landmarks over time to classify emotional states, thus providing valuable information for human and computer interaction, and can be directly used in this setting to help individuals with ASD, as described above. In the proof-of-concept FEELS system, live streams of subjects are captured using a video camera, and the detected emotional states of the captured subjects are overlaid on the live stream itself to enable individuals with ASD to visualize and help them understand the emotional states of others and themselves.

The paper is organized as follows. In Section 2, we provide a detailed description of the presented FEELS framework. In Section 3, we present an example of the visualizations that an individual with ASD would see through the proof-of-concept FEELS system. Finally, conclusions are drawn and future work discussed in Section 4.

2 Proposed FEELS framework

The proposed FEELS framework for assisting individuals with ASD leverages real-time visual perception as a means for detecting the emotional states of subjects captured via video and visualizing the detected emotional states back to individuals with ASD so they can better understand the emotional states expressed not just by those they are interacting with but also themselves. The FEELS framework leverages a combination of state-of-the-art deep convolutional neural networks made available through open-source libraries for facial landmark detection, but also custom neural networks of our own design to infer the emotional states of subjects captured in the video stream, which are then overlaid back on the live stream in real time to enable real-time visualization of the emotional states of different individuals to enable those with ASD to get a better understanding and insights into those they are interacting with as well as the way they convey their emotions to others. In what follows, we describe each of the four steps of the FEELS framework in detail: (1) data acquisition, (2) facial landmark detection, (3) emotional state classification, and (4) emotional state visualization.

Data acquisition. The first step of the FEELS framework involves data acquisition. To perform real-time emotion recognition, a video camera is used in FEELS to capture a live stream of subjects that an individual with ASD is interacting with or of themselves. For the proof-of-concept, we limit the set of emotional states that can be identified by FEELS based on the live video stream to i) neutral, ii) happy, iii) sad, iv) surprise, and v) angry.

Facial landmark detection. The second step of the FEELs framework involves the automatic detection of facial landmarks from the subjects captured in the live video stream. Traditionally, facial landmark extraction typically involves modeling (explicitly or implicitly) both the global features of the face shape (i.e., boundary around face) and the local features of facial appearance (i.e., neighbourhood around the desired landmarks such as moth, nose and eyes). Traditional facial landmark detection algorithms typically fall under three categories: the holistic methods, the constrained local methods, and the regression-based methods. While many variants exist for each category, generally algorithms are trained using a labeled data set with annotations marking key points on a training subject's face. The number of anthropological landmarks vary between 5 and 200, and can be identified on still images or a sequence of frames. These landmarks identify either dominant points describing unique locations on the face (e.g., corner of the eye), or mark an interpolated point connecting those dominant points around the facial component or contour. Formally, given an image $\mathcal{I}$, a landmark extraction algorithm will output the locations of d landmarks (as was specified on each and every image in the training set) as $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{d}$.

Facial landmark detection is challenging for several reasons, including: i) the variability in facial appearance across subjects or due to facial expressions, ii) environmental factors such as camera resolution and varying lighting conditions, iii) and partial visibility or self-occlusions due to the subjects hand or hair covering the individual's face. Over the years, algorithms have gone from detecting facial landmarks in "controlled" environments (e.g., faces of certain size and orientation under specific lighting conditions) to becoming more robust to the variations above and extracting landmarks for faces "in-the-wild" (e.g., extreme head poses and facial occlusions under intense illumination).

While tasks such as face recognition or face detection may produce adequate results by only using primary landmarks (e.g., mouth corners, eye corners, and nose tip), higher level tasks such as face animation or facial expression understanding require greater number of landmarks e.g., from 20-30 to 60-80 for higher accuracy [13]. For the proposed FEELS framework, we require a real-time yet accurate solution that can detect and extract a large number of facial landmarks without the need for calibration, which would not be feasible for natural conversational and interaction scenarios. To satisfy these requirements, the proposed FEELS framework leverages the state-of-the-art deep convolutional neural network for facial landmark detection made available through OpenPose [14], an open-source library that facilitates for the joint detection of human body, hand, and facial keypoints and landmarks (in total, 135 different keypoints and landmarks) given an image capture. A key advantage of the deep convolutional neural network provided via OpenPose is that it is not only highly accurate and provides a large number of facial landmarks, but also is very fast and robust and thus makes it ideal for the proposed FEELS framework. An example of detected facial landmarks is shown in Figure 1.

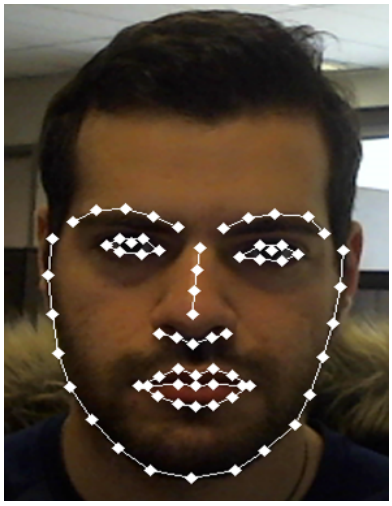

Figure 1: An example of detected facial landmarks.

Emotional state classification. The third step in the proposed FEELS framework is the classification of emotional states based on the extracted landmarks from the facial landmark extraction step. To achieve reliable and robust emotional state classification, we are motivated to move beyond the single-frame emotional state classification approaches that are commonly employed, and introduce a custom neural network architecture of our own design that leverages spatial-temporal facial landmark behaviour. In particular, we introduce a hybrid neural network architecture that consists of a time distributed feed-forward neural network and a LSTM neural network that analyzes the series of facial landmarks over time in order to infer one of the five emotional states.

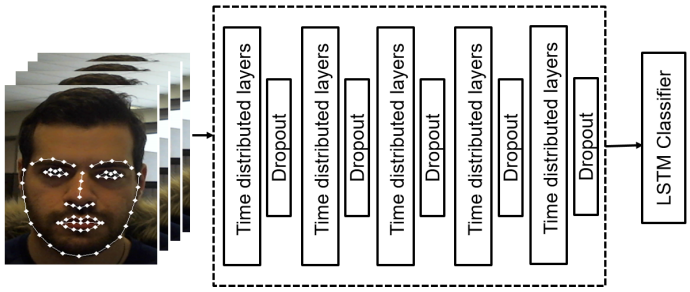

Figure 2: Overall design of the proposed neural network architecture for inferring emotional state using a series of facial landmarks over time.

The overall design of the proposed neural network architecture is shown in Figure 2. More specifically, the network consists of five time distributed layers, each followed by a tanh nonlinearity and a 10% dropout layer, followed by an LSTM classification network. The network was trained in batches of 32 samples (five frames each) over 1000 epochs with the Adadelta [15] optimizer. Each frame of each sample contains 96 features: 70 landmark estimates extracted using OpenPose, and an additional 26 features that capture a combination of L2 and cosine pair distance between the tip of the nose and other landmarks, to provide better classification.

Emotional state visualization. The fourth and final step of the proposed FEELS framework involves taking the detected emotional states of the individual subjects captured in the live video stream, and overlaying this information directly onto the live stream itself to enable an individual with ASD to not only observe the emotional states of those they are interacting with, but also of themselves to learn how they are conveying their emotions to others. The visualization results of the proposed FEELS framework will be shown and discussed in the next section.

3 Results and Discussion

To have a better understanding about the proposed FEELS framework, we constructed a proof-of-concept system that is capable of detecting five different emotional states (i.e., neutral, sad, happy, surprise, and anger) of subjects captured within a live video stream, and creating visualizations of the detected emotional states overlaid on top of the live stream so that an individual with ASD can see and understand in real-time the emotions being conveyed by those they are interacting with, as well as the facial expression they themselves are expressing. An example of emotional state visualizations created by FEELS overlaid on top of the live stream for the five emotional states is shown in Figure 3. As such, the proposed FEELS system can potentially help overcome some of the considerable barriers that make it challenging to incorporate emotion detection within daily life scenarios. Furthermore, this type of information will hopefully give individuals with ASD better feedback to help them better communicate and interact with others in society.

4 Conclusions

In this paper, we introduced FEELS, a full-spectrum enhanced emotion learning system which could be useful as a tool to assist individuals with ASD. FEELS is designed to facilitate for enhanced emotion detection by capturing a live video stream of individuals in real-time, leveraging deep convolutional neural networks to detect facial landmarks, and introducing a custom hybrid neural network consisting of a time distributed feed-forward neural network and a LTSM neural network to determine the emotional state of the individuals based on a sequence of facial landmarks over time. Future work involves exploring the use of a series of video frames directly