Affect Recognition for Individuals with Developmental Disabilities

Justin Le, *Student Member, IEEE,* and Pushkin Kachroo, *Senior Member, IEEE,*

I. Introduction

INDIVIDUALS with developmental disabilities (IDD) tend to have unique mannerisms in communication. They may, for example, speak with words or phrases that belong entirely to languages of their own making, languages that only a long-time personal caretaker would understand. Thus, the task of inferring their emotions from their expressions presents a novel difficulty when compared to previous problems in affect recognition: the lack of a large database of expressions on which to base our inference. Previous works in affect recognition aimed to infer emotions that tend to be common among a large population, and these methods could therefore leverage the available photo and video recordings of other individuals. Emotions such as joy, fear, anger, and disgust are universal, and a model trained with examples of these emotions can generalize well to unseen individuals, something that cannot be done for IDDs for whom very little training data is available due to the individuality of their expressions. Even spontaneous affect, which is often subtle and ambiguous, can be recognized with reasonable accuracy by a method when training with appropriate data.

In this paper, we consider the problem of automatically recognizing a person's emotions given constraints on the amount, quality, and type of training data, as well as the efficiency of computation. These constraints seek to accomodate the requirements of a practical system that can aid communication between IDDs and their caretakers, which we further discuss below.

For IDDs, the difficulty of affect recognition lies in the uniqueness of representation—the expressions that we commonly associate with a certain emotion may not have the same association in an IDD's "language". That is, any two IDDs may represent the same emotion with two vastly different expressions. Furthermore, these expressions tend to be unfamiliar to individuals who have no disabilities, making it difficult to capture, segment, and label training data for each IDD, or to enforce a prior on appearance or geometry that would be applicable across multiple IDDs.

Furthermore, any practical system that aims to recognize emotions of IDDs for the sake of communication must be considerate of time-efficiency and the position of sensor hardware. We require inference to be computationally efficient, as we intend for our methods to be used in real-time scenarios in which the results of inference may become part of routine communication between IDDs and their caretakers. We also require that the data be taken from sensors placed at an angular displacement from the frontal plane of the face and at a reasonable distance, so as not to cause discomfort for the IDD.

Because we expect an IDD to have limited control over muscles in any given part of the body, we account for a wide variety of occlusions caused by uncontrollable hand-to-face contact and for large variations in head pose caused by limited strength of the neck. In localizing the face, the current state-of-the-art is sufficiently robust to this type of occlusion and large pose variation. We use the recent approach proposed by Kazemi, et al., as it exhibits robustness to these conditions, and its efficiency makes it suitable for real-time use, although other current facial registration techniques can be substituted.

We also account for the possibility that changes in head pose are themselves expressions that potentially indicate significant emotions of the IDD, as movements of the head may be a primary mode of communication for an IDD who has limited control of other modes such as speech or facial expression. Hence, constructing features for affect recognition will require that we fuse the measurements taken of the head pose with those of the facial expressions. Similar approaches fuse either the extracted features or the results from separate classifiers. Fusing features allows for exploiting correlations between different modalities when they are temporally synchronous but can potentially lead to a feature space whose high dimensionality might degrade computational speed. Fusing separate results of inference removes the need for synchronicity between modalities and presents many interesting avenues for experimentation due to the abundance of ensemble methods available in state-of-the-art machine learning. We study both approaches to fusion with the goal of determining the benefits and shortcomings of each.

Our contribution is a system for automatically recognizing a person's emotions given the following constraints:

- Limited, noisy, or mislabelled training data
- Only RGB training data taken from a non-frontal view of a person's upper body
- No domain-specific priors for learning
- At least near-real time inference on streaming RGB data without a compromise in accuracy
- Theoretical guarantees for bounds on generalization error

This work was supported in part by the University of Nevada, School of Medicine.

J. Le is with the Department of Electrical and Computer Engineering, University of Nevada, Las Vegas, Las Vegas, NV 89154-4026 USA (lej6@unlv.nevada.edu; http://justin-le.github.io).

P. Kachroo is with the Department of Electrical and Computer Engineering, University of Nevada, Las Vegas, Las Vegas, NV 89154-4026 USA and also with the University of California Berkeley, Berkeley, CA 94720 USA (pushkin@unlv.edu; http://faculty.unlv.edu/pushkin/).