Impact of Age on Emotion Recognition

Sk Rahatul Jannat
Department of Computer Science and Engineering, University of South Florida, Tampa, FL

***Abstract*— In this paper I detail my current research on the impact of age on emotion recognition. I detail and introduce related work, including why the problem of investigating age, as it relates to emotion recognition, is important. New approaches for generalizing emotion recognition across age are proposed, as well as a new approach to help diagnose Autism Spectrum Disorder in children, through analysis of emotion-related features. This approach will also give an explanation as to why the diagnosis occurred. I detail my overall research vision, give a summary of work to date, and conclude with future work and challenges of the proposed research.**

I. Introduction to Work

Affective Computing has been an exciting and growing field in the past two decades and has important applications in artificial intelligence (AI), as being able to recognize emotion is an important part of human intelligence [22]. The ability to recognize emotion has broad impacts for real-world applications in fields as diverse as medicine, defense, entertainment, and retail. Some of these applications include pain recognition [26], customer feedback [3], and educational video games [13]. To move forward with developing these applications, we need to understand the foundation of autonomy, as well as advance interfaces between human and machines. To do this, we must first understand emotions role in autonomy, including what exactly emotion is. This is a difficult problem as there are currently around 100 definitions of what emotion is [21]. In recent years, there has been many encouraging works for recognizing emotion [4], [7], [24], [27], however, one limitation of these works has been on the range of ages that have been tested. The majority of subjects in these studies have largely focused on adults with an approximate age range of 18-60 years old. Fewer works have investigated emotion recognition from children and elders. Without investigating these age ranges, it is difficult to determine if the proposed algorithms generalize well to different age populations [16].

Although there are less works on these age ranges, there are some recent encouraging works that have begun to investigate this problem. Ma et al. [16] developed the EmoReact multimodal dataset which consists of 1323 video clips of 46 elder subjects reacting to Youtube videos. In this work they report baseline results of recognizing emotion on elders, as well as cross-age generalization by investigating children as well. To do this, they extracted hand-crafted features and trained classifiers that include support vector machines and naïve Bayes. This was done on 6 target emotions that include anger, disgust fear, happy, sad, and surprise. In this work to investigate children, they used EmoReact [19] dataset from Nojavanashari et al. This dataset is similar to ElderReact, as the subjects are recorded watching YouTube videos. The main difference being the subjects consist of 63 children. In this work, they investigated facial action units [6], head pose and orientation, and non rigid shape parameters. Nagaragan et al. [18] conducted cross-domain transfer learning on the EmoReact dataset. Using a pre-trained AlexNet [12], they investigated transfer learning from the visual domain to the acoustic domain, as well as from simple emotions to complex emotion recognition. Interesting work from Rincon [15], involved the investigation into emotion recognition in children using the NAO robot. They used a convolutional neural network (CNN) to recognize emotion in children using facial expressions. Using a network pre-trained with the AffectNet dataset [17], they recognized the 6 standard emotions.

Motivated by these works, I propose to investigate the impact of age on emotion recognition to facilitate the development of emotion recognition systems that can generalize across a large range of age ranges (i.e. children, adult, and elder). My research has the following major aims and innovations.

1) Investigation into emotion recognition across children, adults, and elders is proposed. This investigation will result in insight for further development of emotion recognition algorithms that generalize across age.
2) New approach to mitigate the difficulty of generalizing emotion recognition across age ranges [16], is proposed.
3) A multimodal approach to help diagnose Autism Spectrum Disorder (ASD) in children, by analyzing their emotions, is proposed. This approach will also provide an explanation for the diagnosis.

II. Research Vision

A. Impact of Age on Emotion Recognition

I propose to analyze how emotion recognition varies across age including children, adults, and elders. More specifically, I propose to develop models that generalize emotion recognition results across age. Models that better generalize can facilitate real-time applications including automatic detection of disorders such as ASD, PTSD, and down syndrome. This can be a difficult problem, as can be seen in Fig. 2 where there are marked visual differences between the same emotions across age (child vs elder). To facilitate my proposed research, I will make use of multiple large-scale multimodal datasets that include the age ranges necessary to conduct this investigation. These datasets include, but are not limited to, EmoReact [19], ElderReact [16], and BP4D+ [27]. To analyze the impact of age across emotion recognition, I

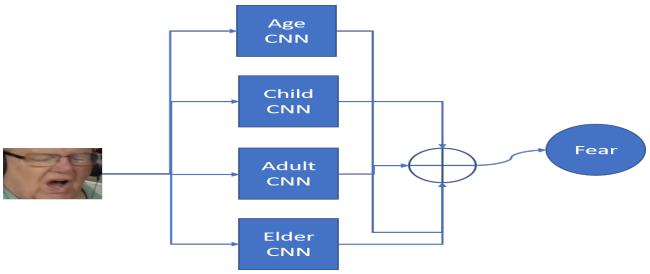

Fig. 1: Proposed approach for generalizing emotion recognition across age. The age CNN is trained on all age ranges with a class label of child, adult, or elder. The other 3 CNNs are trained on the specific age range with emotion labels (e.g. Child CNN is only trained on child data). The final output is weighted by the probability of the Age CNN. In the example shown here, the Elder CNN probabilities are more heavily weighted due to age range of elder.

Fig. 2: Comparison of children emotions [19] and elder emotions [16]. Top row: child; bottom row: elder. From left to right: happy, disgust, surprise.

propose to conduct statistical analysis across data (e.g. face) that includes children, adults, and elders. I will investigate the correlation between emotions displayed, to learn how often different emotions are felt across the age ranges, as well as what stimuli elicited said emotion (e.g. what video was watched, or which task was performed). This type of analysis will lead to further insight into developing models that can recognize emotion across large age ranges.

To generalize across age, I propose an age-aware ensemble-based approach to emotion recognition. The proposed approach will first recognize the general age of the subject (i.e. child, adult, elder), which will then be used weight the fusion of an ensemble of deep neural networks. This ensemble includes 1 network for each age range (child, adult, and elder). To recognize the final emotion of the subject, I propose to average the pseudo probability from each of the networks where the probability from the network that matches the recognized age will be more heavily weighted. This ensemble-based approach has been successful in medical imaging [20], and I hypothesize that it will also positively impact the accuracy of emotion recognition across large age ranges. See Fig. 1 for an overview of the proposed approach.

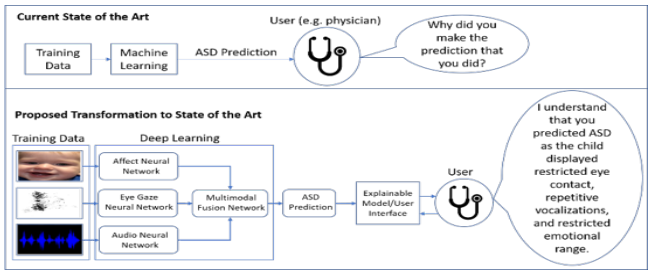

Fig. 3: Proposed approach for helping diagnose ASD in children, compared to current state of the art.

B. Diagnosing ASD in Children

To diagnose ASD in children, I propose a multimodal fusion-based approach that incorporates facial expression, audio, and gaze information. This information will be used to train deep neural networks for diagnosis. See Fig. 3 for an overview of my proposed system. The training of deep neural networks can require a large amount of data to achieve state of the art results in many domains. Considering this, I propose to collect over 20 million frames of multimodal data (face, gaze, and audio), resulting in the necessary data to train the many weights of effective deep neural networks. This will ensure the proposed dataset is state of the art and will contribute to future advancements in the diagnosis of ASD using multimodal data. This data will be collected at the University of South Florida in the department of Health, through the children (with ASD and typically developing peers) performing tasks meant to elicit emotion.

To fuse multimodal data, I propose a hybrid approach where 3 separate convolutional neural networks are used to extract features from each of the modalities (Fig. 3). The extracted features from the deep networks (e.g. from fully connected layers), will then be used as input to a fusion network. I propose to investigate multiple fusion networks that include deep belief networks [10] and other late fusion models [5] to learn which fusion-based approach to diagnose ASD is most accurate. There are multiple advantages to this hybrid approach. First, the fusion of deep features from multiple neural networks has been found to improve emotion recognition when fusing audio and video data [2]. I hypothesize that a similar hybrid approach will result in more accurate diagnosis of ASD, when fusing gaze, audio, and facial affect. Second, I will investigate the diagnosis of ASD using a single modality, which allows us to evaluate the efficacy of the proposed fusion approach compared to a single modality. The features from the single modalities will skip the fusion network and be directly used as input to the diagnosis algorithm. This design will further validate the need for a multimodal approach to diagnosing ASD.

I propose to explain the diagnosis based on a model that uses the Autism Diagnostic Observation System 2nd Edition (ADOS-2). I will use a similar scoring mechanism as ADOS-2 to provide an explanation of the diagnosis. I will model each task that can be scored using facial affect, gaze, and

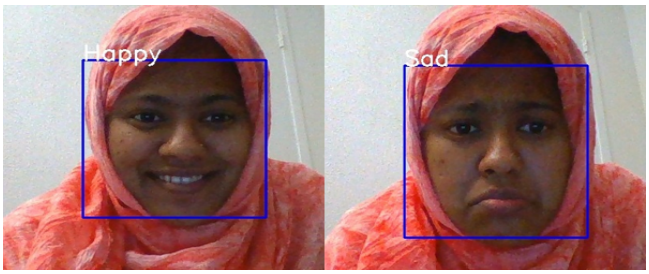

Fig. 4: Real-time expression recognition.

audio. During data collection, the children should give a range of expressions directed at an interviewer. The proposed system will be able to analyze the images to determine if the child has given a range of expressions (i.e. happy, surprised, etc.), and it will analyze the gaze to determine that the expressions are directed towards the interviewer (i.e. they are looking at the interviewer). This approach also allows us to model the interactions between modalities.

III. Summary of Work to Date

A. Ubiquitous Emotion Recognition

In this work, my ultimate goal is to fuse audio and video data for the task of emotion recognition in a ubiquitous environment (e.g. mobile phone). To do this we investigated the use of 2 publicly available datasets. For video (image) data, we use the BP4D+ multimodal emotion corpus [27]. For our experiments we use approximately 13,000 RGB images from this dataset. The next dataset we use is the Ryerson Audio-Visual Database of Emotional Speech and Song [14]. For our experiments, we use approximately 700 recordings from this dataset. Before we can efficiently fuse this multimodal data, we must first pre-process the data. For the image data, we first detect the face in each image using Haar features (Fig. 4) [23]. Once the face is detected, we crop the image to include the face region and scale it to 256x256. To pre-process the audio data, we plot the raw audio signal onto the 2D image plane. The final waveform image is also scaled 256x256, to be consistent with the face data.

To conduct my experiments on emotion recognition, I used Convolutional Neural Networks (CNN) for recognition. I used an Inception V3 CNN with 3 convolutional layers of size 32, 64, and 128, each followed by max-pooling, with a final fully connected layer used in the output. The Adam optimizer was used with a learning rate of 0.0003. Given pre-processed image data that includes faces and audio waveforms, we then train our deep network to recognize emotion. We conducted 3 experiments, with 3 separately trained networks, to do this. We trained one network only on image data, another only on the plotted audio waveforms, and the third on both image and waveform data. This was done to test the accuracy of our method using a single modality compared to a multimodal approach. To test our networks, we trained on 2 emotions (happy and sad) for both audio and video data. We employed a 90/10 split of training and testing data (trained on 90% of the data and tested on 10%).To train our network that contains both audio and image data, each of the signals were used as a single instance of emotion (i.e. one face had a class of happy, one separate waveform had a class of happy). When using audio and video data, we achieved an accuracy of 96.09%.

The developed ubiquitous application was developed to run on PCs (e.g. laptops), and Android devices. The application captures face and audio data. Like the offline pre-processing, the face was detected using Haar features, and cropped to a size of 256x256. To be able convert the captured raw audio signal to a plotted waveform, we created a lightweight Node Js server. The audio file is captured from the application, sent to the server, which is then transformed to the 2D image plane (waveform plot), and finally sent back to the phone. More details can be found in my previous work [11]. This work motivates my current investigation into the impact of age, as a model that generalizes across age is important to develop an accurate ubiquitous emotion recognition system that can be used in wild settings.

B. Fusion of Hand-crafted and Deep Features

We proposed the fusion of hand-crafted and deep features extracted from both actor and listener data from the One-Minute Graduate Empathy Prediction Challenge (OMG-Empathy) dataset. This dataset contains 80 videos that involved a semi-scripted talk with an actor, telling 8 stories, and a listener. After the conversation between the actor and listener was finished, the listener watched the video and rated how they felt giving the ground-truth valence levels in the range of negative one to positive one.

For our hand-crafted features, we used 68 facial landmarks to represent the face, and spectrogram features to represent the audio. For our deep features, we used a CNN with 8 convolutional layers, and 3 fully connected layers with the Adadelta optimizer. From this network, 256 deep features were extracted from the fully connected layer. Given this data, we then fused the actor and listener images, and actor and listener facial landmarks. We then performed a weighted fusion of valence levels from the hand-crafted and deep features. A random forest was used to predict empathy from this data. See Fig. 5 for an overview of our approach.

For this challenge, there were 2 tracks: (1) personalized (model trained on each person); and (2) generalized (1 model for all subjects). We calculated the Concordance Correlatiosn Coefficient (CCC) for each track. We achieved an average CCC of 0.03 across all subjects for the personalized track, and a CCC score of 0.03 for the generalized track. More details can be found in my previous work [8]. This work motivated my current investigation through encouraging results with a multimodal approach.

C. Impact of Age on Emotion Recognition

I have begun my initial investigation into the impact of age on emotion recognition, using the ElderReact [16] and EmoReact datasets [19] to test generalizing across age. Within these two datasets, there are four similar emotions (”Happiness”, ”Surprise”,”Disgust”, ”Fear”). For my initial

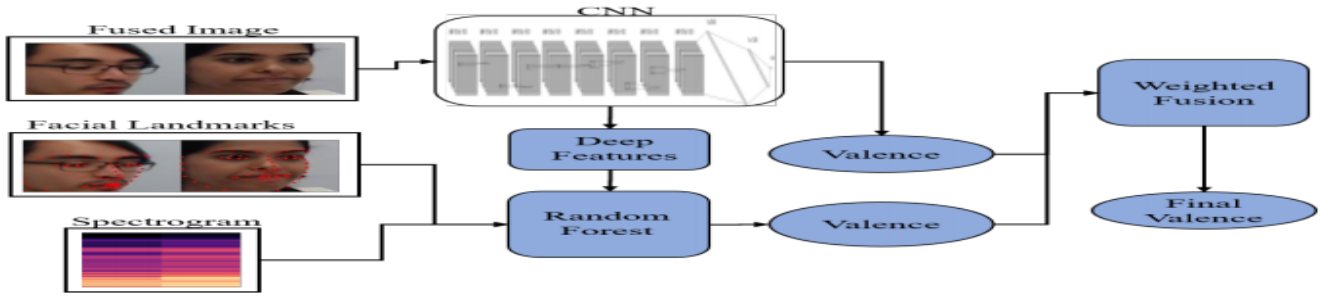

Fig. 5: Fusion of hand-crafted and deep features for recognizing empathy.

experiments, I trained a sequential Deep Network, on three of these emotions which are Fear, Happiness and Surprise. To test generalization, I have trained the network on the child dataset (EmoReact) and validated on Elderly dataset (ElderReact) and also the vise versa. Initial results from this sequential deep network are a low accuracy ($< 30\%$), showing the difficulty of generalizing across age. To account for this difficulty, I am currently evaluating the usefulness of a multi-tail network [9] to recognize multiple emotions across age. After this, I will then begin to implement my proposed ensemble-based approach (Section II-A).

IV. Plans and Challenges

I am currently finishing up my last semester of course-work. At the end of this semester (Spring 2020), I will give my major research area presentation, which consists of a major literature review of my field, my current work to date, and my future plans for research. This is my last required task before being a PhD candidate. I anticipate I will graduate in 3 years (Spring 2023).

After finalizing and submitting my current work, I will begin an extension to a top ranked journal such as the IEEE Transactions on Affective Computing. This work will further improve upon my current work, by also comparing to a third class of age (i.e. adult). Along with this third class, I will also investigate the impact of age across gender and ethnicity. It has been shown that emotions are felt different across gender [9], and I hypothesize that this difference will be increased when age is factored in. This work will be conducted along with my proposed ensemble-based approach. I will also conduct my research into early diagnosis of ASD.

There are a number of challenges with my research. First, there are a small number of datasets that have age ranges that include children and elders. Secondly, my aim is to investigate the impact of age on emotion recognition, however, expression does not equal emotion [1]. Can multimodal data help with this (e.g. audio, thermal, physiological)? A challenge with this, is to the best of my knowledge there are not datasets that contain multimodal age-related data outside of audio and video. Lastly, while deep network have shown to be successful in analyzing facial images for expression [25], age was not investigated. The impact of using deep neural networks on investigating the impact of age on emotion recognition is not clear. This may result in the need for more advanced techniques to generalize across age.

References

[1] L. Barrett et al. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019.

[2] P. Barros, E. Barakova, and S. Wermter. A deep neural model of emotion appraisal. *arXiv preprint arXiv:1808.00252*, 2018.

[3] E. Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[4] W. Chu et al. Modeling spatial and temporal cues for multi-label facial action unit detection. *arXiv preprint arXiv:1608.00911*, 2016.

[5] W. Chu et al. Learning spatial and temporal cues for multi-label facial action unit detection. In *FG*, 2017.

[6] P. Ekman. What the face reveals: Basic and app studies of spon exp using the facial action coding system (facs). *Ox Uni Press*, 1997.

[7] D. Fabiano and S. Canavan. Emotion recognition using fused physiological signals. In *ACII)*, 2019.

[8] S. Hinduja et al. Fusion of hand-crafted and deep features for empathy prediction. In *FG*, 2019.

[9] S. Hinduja et al. Recognizing perceived emotions using facial expressions. *FG (To Appear)*, 2020.

[10] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[11] R. Jannat et al. Ubiquitous emotion recognition using audio and video data. In *UbiComp*, 2018.

[12] A. Krizhevsky et al. Imagenet class with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[13] C. A. Lara, H. Mitre-Hernandez, J. Flores, and H. Perez. Induction of emotional states in educational video games through a fuzzy control system. *IEEE Transactions on Affective Computing*, 2018.

[14] S. Livingstone and F. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dyn, multimodal set of fac and voc expressions in north american english. *PloS one*, 13(5), 2018.

[15] A. Lopez-Rincon. Emotion recognition using facial expressions in children using the nao robot. In *CONIELECOMP*, 2019.

[16] K. Ma et al. Elderreact: A multimodal dataset for recognizing emotional response in aging adults. In *ICMI*, 2019.

[17] A. Mollahosseini et al. Affectnet: A db for facial exp, valence, and arousal computing in the wild. *IEEE Trans on AC*, 10(1):18–31, 2017.

[18] B. Nagarajan and V. Oruganti. Cross-domain transfer learning for complex emotion recognition. In *TENSYMP*.

[19] B. Nojavanasghari et al. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *ICMI*, 2016.

[20] R. Paul et al. Mitigating adverserial attacks on medical image understanding systems. *ISBI*, 2020.

[21] R. W. Picard. *Affective computing*. MIT press, 1995.

[22] R. W. Picard et al. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[23] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[24] C. Wang et al. Learning bodily and temporal attention in protective movement behavior detection. *arXiv preprint arXiv:1904.10824*, 2019.

[25] H. Yang et al. Facial expression recognition by de-expression residue learning. In *CVPR*, 2018.

[26] G. Zamzmi et al. Machine-based multimodal pain assessment tool for infants: a review. *arXiv preprint arXiv:1607.00331*, 2016.

[27] Z. Zheng et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, 2016.