An Analysis of Emotion for Children and Elders

Xinyu Wang
Carnegie Mellon University
Pittsburgh, Pennsylvania
xinyuw3@andrew.cmu.edu

Kaixin Ma
Carnegie Mellon University
Pittsburgh, Pennsylvania
kaixinm@andrew.cmu.edu

Mingtong Zhang
Carnegie Mellon University
Pittsburgh, Pennsylvania
mingtonz@andrew.cmu.edu

Xinru Yang
Carnegie Mellon University
Pittsburgh, Pennsylvania
xyang@andrew.cmu.edu

ABSTRACT

Automatic emotion recognition plays a critical role in technologies such as intelligent agents and social robots and is increasingly being deployed in applied settings such as education and healthcare. Most research to date has focused on recognizing the emotional expressions of young and middle-aged adults and, to a lesser extent, children and adolescents. We present a thorough analysis of children and elders' emotions by building a new dataset of elders reacting to emotion elicitation stimuli and comparing it against an existing dataset for children's emotion. Specifically, we use the EmoReact dataset and create a new dataset containing 1323 video clips of 46 unique individuals with human annotations of six discrete emotions: anger, disgust, fear, happiness, sadness, and surprise. We conducted statistical analysis of the most indicative features for each emotion on both dataset and build several predictive models using unimodal and multimodal features. Our experiments and analysis indicate that children and elders express emotions quite differently and models trained on one age group do not generalize to the other.

KEYWORDS

affective computing, emotion recognition, multimodal analysis

ACM Reference Format:
Xinyu Wang, Kaixin Ma, Mingtong Zhang, and Xinru Yang. 2019. An Analysis of Emotion for Children and Elders. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 BACKGROUND AND MOTIVATION

Emotion is one of the most common ways we express ourselves, either consciously or unconsciously and it plays an important role in communication. Interpreting emotions helps us determine the true intention of the message we receive.[5] Thus understanding emotions helps us communicate effectively and build better relationship with others. Recently, a lot of research interests have focused on the automatic emotion recognition because this can be an vital step toward building more human-like machine system.

On the one hand, automatic emotion recognition plays a central role in various technologies such as virtual home assistant and intelligent learning system. Successful automatic emotion recognition serve as key foundation for adaptive systems which make more appropriate response based on users' emotions. On the other hand, a successful emotion recognition system can potentially open up a new domain for psychological research. For example, machine learning could discern the most salient features for recognizing emotions from high dimensional space, which may have been overlooked by traditional approach. Thus traditional psychological study may find a brand new method to solve those long debated propositions with the reference to how a well-functioning machine recognize certain emotions.

Previous studies on emotion recognition tend to pay little attention to age differences. However, we do know children and elders vary a lot when interpreting emotions,[10] and the neural system, which connects closely to the expression of emotions also differs a lot in children and elders. [9] These evidence motivate us to explore age-aware emotion recognition systems for socially vulnerable groups like children and elders so as to provide them more adaptive caring in automatic systems catering to their needs. For example, in an intelligent tutoring systems, the tutor can make adaptive instructions based on its recognition of the learners' emotions; since both the young and old learners are increasing, to help it make more accurate judgement (e.g. to use age-specific algorithm for different age groups of learners) will make the instruction for both groups more effective.[17]

Specifically, we plan to target on several different emotions and two age groups to explore the aforementioned domains. The emotions are: happiness, surprise, excitement, curiosity, uncertainty, disgust, fear and frustration for both children and happiness, fear, disgust, anger, sadness, surprise for elders. By this research, we will give particular contribution on the difference between emotion expression patterns for children and elders.

2 LITERATURE REVIEW

2.1 Emotion Recognition for Children and Elders

There are plenty of research about the different recognition patterns of emotions of children and elders i.e how they interpret/recognize

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

others emotions. [16]However, little research has been done about the different output (expression) patterns of emotions of children and elders, especially with combination with affect computing. Most research either takes it for granted that unlike the recognition of emotions, the expression of emotions will not vary in different age groups, or involve only issues in which the difference of expressing emotions in different age groups does not matter a lot. The novelty of our research lies mostly in stepping into this uncharted water. We made several hypothesis regarding the difference between children and elders in expressing their emotions and tested them with our data.

2.1.1 *ANS and autonomic actions.*

Our first hypothesis regarding the potential difference in patterns of emotion expression between children and elders originates from existing findings. Neuroscience shows the young and the old share the same pattern of autonomic nervous system (ANS) while the general magnitude of changes in ANS activity of the young is larger than which of the old. Some expressions, especially some movements of facial muscles to a certain extent rely on the autonomic nervous system. And they are all very useful in expressing one's emotions.

Take blinking as an example, despite we can blink deliberately, in most of time we blink unconsciously. Hence we can call blinking a semi-autonomic action and to be autonomic means it will be highly influenced by ANS system. Based on the difference of activity in ANS systems of children and elders, we hypothesize that elders will generally blink less frequently that children while expressing any kind of emotions. [9]

2.1.2 *Disgust as a Controversial Basic Emotion.*

The term emotion in our comparison between children and elders refers only to basic emotions. However, disgust is usually argued possessing two different basic subtypes: core disgust and social-moral disgust.[19] Core disgust is the disgust caused by concrete substances like faces, vomit or saliva. Social-moral disgust is the disgust caused by abstract concepts which are culturally believed disgusting, like racism, government corruption and massacre.

Based by the aforementioned theory, we hypothesized that while expressing disgust, the elders will express more in a manner of social disgust and the children more in a manner of core disgust. This is because more experience and developed morality system make the elders tend to feel disgusted more from social events. So what will this make them differ in expressing disgust?

Certain kinds of correlation are frequently reported in different kind of emotions,[19] like frustration with social emotion and fear with core emotion. So if our hypothesis is true, elders will express their disgust in a manner which shares more common features with frustration, and children in a manner which shares more common features with fear.

2.1.3 *Different Muscle Movements around Eyes while Happy.*

A big smile is definitely an indicative symbol of happiness. However, there are subtle differences in seemingly similar smiles. If we define fake smile as a smile not from pure delight in soul but from complex motivations to show happiness, then a subtle but distinctive feature of fake smile is the movement of zygomatic major without the movement of orbicularis oculi.[4]

Table 1: Number of videos and for each emotion and number of children who expressed the emotion in EmoReact

Emotion type	Number of Videos	Number of Children
Curiosity	385	51
Uncertainty	344	53
Excitement	355	49
Happiness	604	60
Surprise	298	49
Disgust	137	35
Fear	50	20
Frustration	131	31

Table 2: Annotation Reliability Measurements

Emotion	Observed	Sscore	AlphaK	ICCK	FINN
Anger	0.85	0.65	0.30	0.64	0.76
Disgust	0.88	0.71	0.42	0.78	0.79
Fear	0.92	0.81	0.32	0.68	0.87
Happiness	0.82	0.57	0.52	0.85	0.69
Sadness	0.87	0.68	0.25	0.60	0.77
Surprise	0.79	0.50	0.28	0.66	0.64
Valence	0.83	0.56	0.45	0.83	0.76

Based on the above finding, we hypothesized that there might be a difference in the general patterns of children and elders to express happiness. If the elders, due to more social experience and social concerns, tend to have a more complex motivation to show others their happiness, then their most significant expression of happiness, the smile, tend be more like a fake smile, with similar activity of zygomatic major with that of a similarly happy child but significantly lower activity of orbicularis oculi than that of a similarly happy child.

2.1.4 *Age-Related Changes in the Face.*

decoding emotional faces is also likely to be influenced by stimulus features, and age-related changes in the face such as wrinkles and folds may render facial expressions of elders. [8] Generally, due to the aging of facial muscles, the controllability and flexibility of an elder's will decrease compared to which of a child. This is also supported by an experiment asking children and elders to pose expressions following muscle to muscle instruction. The performance of the aged in the the experiment is worse than the young. [14]

We hypothesized some general differences in patterns for expressing all kinds of emotion of children and elders. Generally, because the aging of their facial muscles, the expression of emotions in elders will be less harder to decode compared to that of children, so an expression actually denotes a more intense feeling in elders.

2.2 Recognition Technology

With the advanced techniques in computational methods, there have been a great amount of research work on quantitatively modeling affective states. Early work as as [22] uses visual signals to recognize human affective behaviors while [12] takes the advantages of acoustic signals and uses decision tree method and random forest ensemble. Besides, [2] presents a direct and obvious way to detect human emotions by semantic labels and attributes that represents the verbal text. Meanwhile, as the development in deep learning, there have been many powerful models that demonstrate huge potential compared to Support Vector Machine (SVM). For example,[7] applied Long Short Term Memory (LSTM) and Deep Neural Network (DNN) for aggregation and representation of face features on eight-class emotion recognition from VGGFace dataset.

2.3 Existing Datasets

Previous studies have explored the field of children emotions. [11] introduced the MMDB dataset which contains video recordings of children of 1 to 2 years old interacting with adults instructors. Each stage of interaction session and children's attention are annotated in this dataset. However, its main goal is to study social interaction for children. There are other children emotion datasets which only contains images such as the Child AïňĂective Facial Expression (CAFE) [15] which study children of age 2 to 8 and the Dartmouth dataset [3] which study children of age 6 to 16.

There is also previous work that explored emotions of elders. [20] introduced a dataset to study emotion interaction for elders. They collected audio and video data from TV-series and annotated 7 major emotions including anger, anxiety, boredom, disgust, happiness, neutrality and sadness. They also proposed a emotion interaction model to analyze the emotion transitions. [21] studies the emotion recognition task for elders, with an emphasis on speech features. They collected their dataset from actor performed emotional speeches, which contain 7 emotions same as [20].

3 DATA DESCRIPTION

3.1 EmoReact

The EmoReact[1] dataset was created mainly to study children's emotion. The original data source was the Youtube React channel. There are in total 63 children (32 female and 31 male) appear in the dataset. These children age from 4 to 14 and they perform a series of tasks during the video including 1. Getting to know the subject. 2. Being asked a question about it. 3. Answer the question. 4. Being told a fact about it and react. 5. Talking about their opinion about it. The videos are segmented into short clips so that each clip only contain one child react to one subject. There are 1102 clips in total and the average duration of the clips are 5 seconds. The released version of dataset contains annotation of 8 emotions including: curiosity, uncertainty, excitement, happiness, surprise, disgust, fear and frustration as well as valence. The labels for emotions are in binary to indicate whether the emotion exist. The statistics about the dataset is shown in the table 1. The distribution of emotion types is unbalanced toward positive types. It is also worth noting that each clip may contain zero to many emotions.

3.2 ElderReact Data Collection

Following a similar pipeline as EmoReact, we created our own dataset which we call ElderReact. We first collected 43 videos from the YouTube Elder Reaction channel in which elders react to different subjects. These videos' subjects cover a wide range of topics including video games, social events and online challenges etc. These videos typically have 3 stages: (1) The elders are first told the subject they will be reacting to and asked what they know about the subject. (2) Elders are presented the subjects and they interact with the subjects. (3) Elders are asked their opinions about the subjects after getting to know more. The original videos downloaded from YouTube contain multiple people reacting to the subject. We used the open source tool PySceneDetect to segment videos into short clips so that each clip only have one person's reaction. We only kept clips that are at least 3 seconds and manually filtered out clips with low quality, i.e. some scene change may be omitted by PySceneDetect so that more than one person would end up in one clip. We generated 1,323 clips in total after this step and these clips are typically 3 to 10 seconds long. Additionally, we manually annotated the identity of the elder in each of these 1,323 clips. In total we have 46 people in our dataset from which 26 are female and 20 are male.

3.3 Emotion Annotations

From our pilot study, we noticed that some of the more fine-grained emotions such as curiosity rarely appear in our videos. Thus We decided to only focus on 6 basic emotions: Anger, Disgust, Fear, Happiness, Sadness and Surprise. We used the crowd-source platform - Amazon Mechanical Turk to collect annotations. We hired 3 workers to annotate each of the clips. For each emotion type, we asked a compound question about the existence of that emotion. For example, we asked "How fearful/scared/panicked was the individual in the video?" for emotion fear. Workers would answer each of the questions on a 1-4 scale where 1, 2, 3, 4 represent Not at all, Slightly, Moderately and Very much respectively. We also asked workers to judge the overall valence of he video on a 1-7 scale where 1 mean very negative and 7 means very positive. The definitions of each of the emotions and valences are also shown in the Mturk interface. Since we observed that multiple emotions may appear in one video clip, i.e transition of emotions, we asked workers to identify all emotions that exist. Additionally, we asked the worker to annotate the gender of the individual in the video to test the vigilance of the workers. We required all workers to have U.S. location, lifetime approval rate to be greater than 97% and number of hits approve to be greater than 500. All hits in gender was answered incorrectly were rejected and new workers were asked to provide substitute annotations.

3.4 Annotation Statistics

After collecting the annotations, we compute observed agreement, SScore and KripendorffâĂŹs alpha to evaluate the agreement among workers. The agreement scores are shown in table 1. As we can see that the observed agreement for all emotions and valence are relatively high. However, due the emotion imbalance, the Sscore and KripendorffâĂŹs alpha dropped by a large margin. Besides inter-annotator agreement, we also compute intra-class correlation scores to estimate how consistent the annotations are among videos. The scores are shown in the last column. Using the ICC estimates, workers' rating are good for disgust, happiness, and valence and acceptable for all the others. Using the Finn estimates, workers' rating are good for anger, disgust, fear, sadness, and valence and

Table 3: Number of videos and for each emotion and number of people who expressed the emotion in the ElderReact

Emotion type	Number of Videos	Number of People
Anger	350	39
Disgust	278	36
Fear	152	28
Happiness	742	44
Sadness	241	38
Surprise	566	45

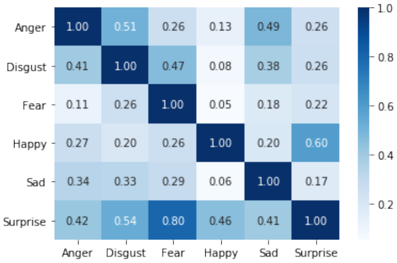

Figure 1: Co-occurrence between different affective states in elder people

acceptable for happiness and surprise. Finally, we convert the raw ratings of workers into binary values to indicate whether the emotion exist in the videos. We consider an emotion to be present if at least 2 workers agree that the emotion exists, i.e ratings to be greater or equal to 2. The number of videos for each emotion and the number of people who have shown the emotion are presented in table 2. Although there is still imbalance among emotions, we get a good number of samples for each of the emotion and each emotion is expressed by many people, which would allow researcher to build more complicated computational models and study the generalization ability to unseen individuals. Since each clip may contains many emotions, we also plot the co-occurrence patterns of emotions in figure 1. We normalize each cell by the number of videos of emotions on the x-axis. It's interesting to note that happiness co-occurred with surprise very often. We think the main reason is that in many of our videos, the elders are asked to react to things that they're unfamiliar with such as video games but they also enjoy them a lot. On the other hand, surprise also co-occurred with fear quite often. When elders are asked to react to online challenges, the contents are often unexpected for them and they are afraid of what they saw sometimes, resulting the relative high co-occurrence rate of fear with surprise.

4 STATISTICAL ANALYSIS

4.1 Problem Conceptualization

Traditionally, the attention has been mostly focused on visual and audio features themselves for emotion recognition task. In other words, previous effort had the assumption that emotion is independent of the age. However, we believe that age plays an important role in emotion recognition and different age groups need to be treated differently. The figure 2 illustrates the construct we developed for our task. Basically we think that age would affect how the emotion is recognized besides various features from visual and audio modalities. Since the exact age information for individuals

Table 4: Most indicative visual features for each emotion in EmoReact

Emotion type	Feature	Stat	Hedge	Direction
Curiosity	H Gaze	μ	0.86	↑
	Z Gaze	μ	0.95	↑
Uncertainty	Lip corner depressor	μ	0.73	↑
Excitement	Blink	σ	0.82	↑
Happiness	Lip corner puller	σ	0.97	↑
	Cheek Raiser	σ	0.97	↑
Surprise	Jaw drop	σ	1.15	↑
	Blink	μ	1.62	↑
Disgust	P3	σ	0.86	↑
	P6	σ	0.89	↑
Frustration	H Gaze	μ	2.06	↑
	Z Gaze	μ	2.53	↑

Table 5: Most indicative audio features for each emotion in EmoReact

Emotion type	Feature	Stat	Hedge	Direction
Curiosity	VUV	σ	0.99	↑
	NAQ	σ	0.71	↑
Uncertainty	MDQ	μ	0.48	↑
Excitement	MFCC 20	μ	0.74	↑
Happiness	MFCC 4	σ	0.45	↑
Surprise	Creak	σ	1.23	↑
	MFCC 19	σ	1.41	↑
Disgust	F0	μ	1.38	↑
	peakSlope	μ	1.25	↑
Frustration	VUV	μ	-2.42	↓
	MDP	σ	1.74	↑

Table 6: Most indicative visual features for each emotion in ElderReact

Emotion	Feature	Stat	Hedge	Trend
Anger	Lip corner depressor	σ	0.97	↑
	Lip stretcher	σ	0.96	↑
Disgust	H Pose Shift	σ	1.01	↑
	H Gaze Shift	σ	0.88	↑
Fear	Blink	σ	-2	↓
	Chin raiser	σ	-2.12	↓
Happiness	Cheek raiser	σ	1.23	↑
	Lip corner puller	σ	1.49	↑
Sadness	P 16	σ	0.88	↑
Surprise	H Gaze	μ	0.58	↑
	V Gaze angle	μ	0.54	↑

Table 7: Most indicative audio features for each emotion in ElderReact

Emotion	Feature	Stat	Hedge	Direction
Disgust	MFCC 0	μ	1.17	↑
	MFCC 2	μ	-1	↓
Fear	F0	μ	1.47	↑
Happiness	F0	μ	0.98	↑
	MFCC 2	μ	-1.08	↓
Sadness	NAQ	σ	0.60	↑
Surprise	MFCC 0	μ	0.86	↑

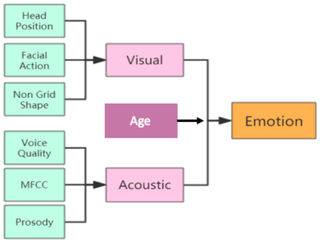

Figure 2: Age is an important factor to recognize emotions

in the videos is not available in both EmoReact and ElderReact dataset, we could not add age as an input to the predictive model to test out if our construct makes sense. Instead, we make comparison between these 2 dataset on several aspects. Specifically, we developed a series of hypothesis to compare children and elders' emotional response. Then we conducted thorough feature analysis to compare the features extracted on the same emotion between 2 age groups. Finally, we train our predictive models using one dataset and test on the other to see if the age difference would affect models' performance.

4.2 Features Extraction

As for feature extraction, we use the same feature templates and open source tools suggested in the EmoReact paper. We briefly describe feature templates here. For visual features, we extract the following 3 types:

(1) Head position: The feature of head position provides information on some basic behaviors such as nodding or shaking head, which have been shown to be indicative to sense of agreements and disagreements[13].
(2) Facial action: The feature of facial action such as brow rising, eye blinking and lip corner puller are proved to be strongly indicative for emotion presentation[6].
(3) Non grid shape: The feature of non grid shape comes from PCA algorithm on facial landmarks that can be useful in some specific emotion expression such as fear.

For audio features, we extract the following 3 types:

(1) Voice quality: The feature of voice quality is extracted by signal process functions such as normalized amplitude quatient (NAQ) and it is important in measuring speech characteristics such as tenseness, creakiness and breathiness.
(2) MFCC: The feature of Mel-Frequency Cepstral Coefficient is widely accepted and successfully applied in emotion recognition [18].
(3) Prosody: The feature of prosody indicates the information of pitch in speech.

We extract same set of features for both EmoReact and ElderReact datasets. We selected frames where the faces are successfully detected. In total 97.9% of the frames in our dataset are successfully processed by OpenFace. Audio features are computed every 10 milliseconds. After extracting the raw features, we computed the mean and standard deviation for each video. Then we concatenated them as summaries for the video clips. The resulting visual feature is a 178-dimensional vector for each video. The audio features 72-dimensional vectors.

4.3 Hypothesis Testing

To test the hypothesis regarding difference between children and elders regarding expressing the same emotion, we did significance test to visual features from both children and elders with the same emotion. For one emotion E, we did an independent 2 sample group t-test with one group containing clips of children who exclusively express E and another group containing clips of elders who exclusively express E. To eliminate the bias from individually unique patterns to express E, within every group, we combined all different clips with the same person to one clip by using the average of the feature from those clips for every feature. We selected particular interesting and inspiring features regarding our hypothesis and interpreted them based on our hypothesis.

4.3.1 *Blinking.*

Here we have found that for the same surprise, the younger tend to blink significantly more frequently while expressing it compared to the elders. This is also the pattern for most of other emotions, so based on our data and statistic analysis, we can say generally the young blink more to express even the similar emotion compared to the old. Please refer to figure 14 in the appendix.

4.3.2 *Fake Smile.*

Here we did find there are significant difference between the movement of orbicularis oculi of happy elders and chilren, while the result is the contrary of our hypothesis, that the movement of elders' orbicularis oculi is more active. So this hypothesis is not proved and we will further research this interesting finding. Please refer to figure 11 in the appendix.

4.3.3 *Core and Social Disgust.*

Between groups of elders and children expressing similar disgust, the feature of upper lid raiser is significantly different. Elders have a larger upper lid raiser index while the index has negative effect in predicting the existence of sadness, which is a major emotional element for social disgust. So the previous hypothesis is partially proved, that the disgust of elders have a bigger social part than that of children. Please refer to figure 15 in the appendix.

4.3.4 *Expression Intensity.*

We chose happiness to test this hypothesis for that is the only emotion in which children and elders share the same most predictive features, cheeck raiser and lip corner puller. The result shows both features are significantly more intense on happy elders than happy children. So it seems the hypothesis is

Table 8: Results for each emotion in ElderReact (F1 score)

Model	Anger	Disgust	Fear	Happy	Sad	Surprise
Random	0.30	0.26	0.14	0.51	0.27	0.41
Naive Bayes	0.36	0.30	0.16	0.53	0.32	0.41
RBF SVM	0.39	**0.35**	0.16	0.70	0.35	0.54
XGBoost	**0.40**	0.26	0.17	**0.71**	**0.37**	**0.57**
TFN	0.34	0.30	**0.18**	0.61	0.32	0.51

Table 9: Results for each emotion in ElderReact (Accuracy)

Model	Anger	Disgust	Fear	Happy	Sad	Surprise
Random	0.50	0.50	0.50	0.50	0.50	0.50
Naive Bayes	0.58	0.18	0.31	0.60	0.20	**0.70**
RBF SVM	0.58	0.57	0.65	0.71	0.50	0.63
XGBoost	0.59	0.60	0.67	0.69	**0.63**	0.64
TFN	**0.61**	**0.66**	**0.70**	0.60	0.59	0.34

Table 10: Results for each emotion in ElderReact using only visual features (F1 score)

Model	Anger	Disgust	Fear	Happy	Sad	Surprise
Random	0.30	0.26	0.14	0.51	0.27	0.41
Naive Bayes	0.37	0.30	0.17	0.57	0.32	0.32
RBF SVM	**0.41**	**0.35**	**0.18**	**0.70**	**0.35**	0.50
XGBoost	0.39	0.30	0.17	**0.70**	0.32	**0.51**
TFN	0.04	0.26	0.16	0.65	0.32	**0.51**

proved for generally the elders neeed to express more intensely to be recognized as having a certain emotion. Please refer to figure 10 and figure 12 in the appendix.

4.4 Feature Analysis

For each emotion state in both EmoReact and ElderReact dataset, we performed t-test to identify the most indicative features. To reduce any potential bias, we treat the group of videos in which the none of the emotion exist as the control group and against the group of videos that only contain the emotion we try to test. Also since multiple videos may contain the same individual, which could potentially bias the t-test, we first average all instances of the same individual and then perform the independent t-test. We used p-value of 0.05 to be our threshold. Additionally, we also compute the Hedge's g to indicate the effective size of the significant features. The most indicative visual and audio features for children's emotion are shown in table 4 and 5 and those for elders' emotion can be found in table 6 and 7. Most of these results are intuitive. For example, the Lip Corner Puller and the Cheek Raiser are considered important for happiness. We can also observe that every emotion tend to have different most salient features.

Table 11: Results for each emotion in ElderReact using only visual features (Accuracy)

Model	Anger	Disgust	Fear	Happy	Sad	Surprise
Random	0.50	0.50	0.50	0.50	0.50	0.50
Naive Bayes	0.54	0.18	0.59	0.64	0.19	**0.70**
RBF SVM	**0.59**	0.56	0.66	**0.71**	0.53	0.62
XGBoost	0.58	0.54	0.64	0.69	**0.59**	0.60
TFN	0.58	**0.58**	**0.69**	0.62	0.57	0.34

Table 12: Results for each emotion in ElderReact using only audio features (F1 score)

Model	Anger	Disgust	Fear	Happy	Sad	Surprise
Random	0.30	0.26	0.14	0.51	0.27	0.41
Naive Bayes	0.27	0.34	0.17	0.54	**0.36**	0.51
RBF SVM	0.28	0.28	**0.18**	0.64	0.34	0.50
XGBoost	0.33	**0.35**	0.17	0.65	**0.36**	**0.57**
TFN	**0.35**	0.30	0.15	**0.69**	0.32	0.51

Table 13: Results for each emotion in ElderReact using only audio features (Accuracy)

Model	Anger	Disgust	Fear	Happy	Sad	Surprise
Random	0.50	0.50	0.50	0.50	0.50	0.50
Naive Bayes	0.58	0.62	0.62	0.54	0.66	**0.63**
RBF SVM	0.58	0.59	0.65	0.54	0.52	0.60
XGBoost	0.54	0.63	0.65	**0.57**	0.56	0.60
TFN	**0.78**	**0.82**	**0.92**	0.53	**0.81**	0.34

5 PREDICTIVE MODELING AND EXPERIMENTS

5.1 Predictive Modeling

5.1.1 Baseline Models. We have built several baselines including Gaussian Naive Bayes, radial basis function kernel SVM, XGBoost as our baseline models. We also include random guess to calculate our F1 scores so that we can have a better understanding of how well those baseline models perform.

5.1.2 Tensor Fusion Network. Besides traditional machine learning approaches, we have also done experiments with deep learning methods to further improve the performances. In particular, we finetuned the Tensor Fusion Network[23] on our dataset including children and elder people. Specifically, the previous network provides with various embedding functions that embeds the raw input from different modalities into numerical representations and fuses them all by using outer product operation. However, the number of modalities in their work was three (text, video, audio) while we only consider video and audio input in our case. Thus, instead of fusing from three modalities into a three-dimensional tensor, we re-factorize the code that fuses from two modalities into a two-dimensional matrix eventually. Meanwhile, rather than using embedding layers for raw input, we have used public and popular

Table 14: Feature abalation Study on ElderReact

Emotion	Feature Removed	F1	F1 After
Surprise	Blink	0.502	0.5
Frustration	VUV	0.36	0.346
Happy	Lip Corner Puller	0.702	0.588
Happy	Cheek Raiser	0.702	0.654
Happy	MFCC 2	0.702	0.551

tools to extract features that serve as representations for both visual and acoustic modalities.

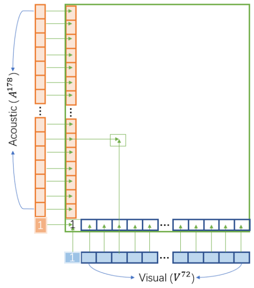

Figure 3: Tensor Fusion

5.2 Experiments

We separated ElderReact into three subsets: Training set, which contains 615 videos; validation set, which has 355 videos; and test set, which has 353 videos. These sets are defined in a person independent manner. Specifically, training set has 8 people, dev set has 8 people and test set has 30 people. We hope this split could help better generalizability of our models.

We have conducted following experiments in this work. For EmoReact dataset, since previous work has established several non-neural baselines. We only experimented TFN on it using visual/audio/ early-fusion features (Fig 5). For ElderReact dataset, we experimented all of our baseline models as well as TFN using visual/audio/early-fusion features. We report F1 scores and accuracy on these experiments (Table 8-13)

To explore the differences between EmoReact and ElderReact, we used training data from EmoReact and tested on elder data. We also did the other way around. Figure 8 and Figure 9 shows the performance differences using RBF SVM classifier. Note that there are only four overlapping emotions among these datasets so F1 scores on only four affect states are reported.

We have dropped two test videos where there are NaN values in the feature input. We did min-max feature normalization as the range of values of raw features can vary a lot. To handle the problem of imbalanced emotion labels, we performed under-sampling and used ensemble classifiers. Specifically, we first selected a subset of negative examples that has the same size as positive examples and used it as training data to train a classifier. We used the classifier to predict labels. Then we repeated the process for 100 times and used the majority voting method to decide the final labels. We used the same approach for all the baseline models.

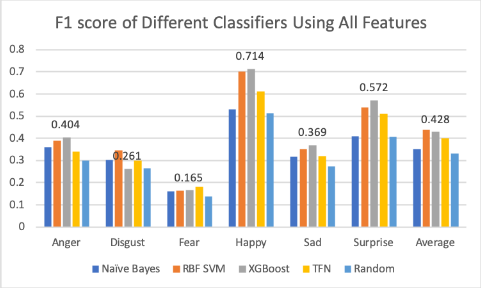

Figure 4: F1 scores using different classifiers on ElderReact

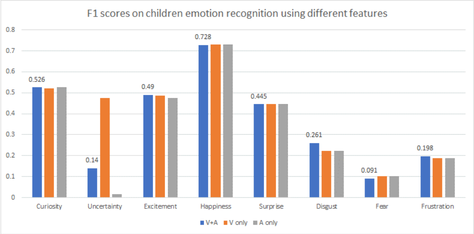

Figure 5: F1 scores under different configurations of visual and acoustic features on EmoReact

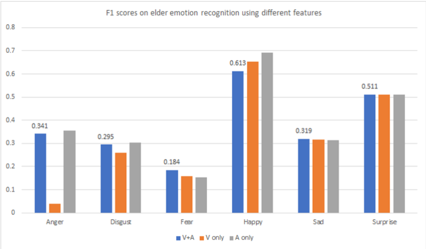

Figure 6: F1 scores under different configurations of visual and acoustic features on ElderReact

For settings of experiments on tensor fusion network, the division of dataset, the input of features and the metrics to measure remain the same as baseline models. However, to tackle with the imbalance in dataset, we have used biased weights on different classes in cross entropy loss function on PyTorch. To achieve unimodal setting, the features of the other modality would be set to zero before being fused into the final tensor. In this way, the model is only able to exploit the information from the rest modality.

For the ablation studies, we selected 5 features considered as significant from our analysis. We removed these features and compared the results with and without these features.

5.2.1 Discussion. After tuning the hyper parameters of networks to avoid overfitting and get a reasonable result, we have observed some interesting findings both on children and elder dataset. On children dataset (figure 5), generally there is no huge difference with different combinations of feature input in recognizing some emotions such as curiosity, excitement, happiness and surprise. This probably indicates that both of visual features and acoustic features

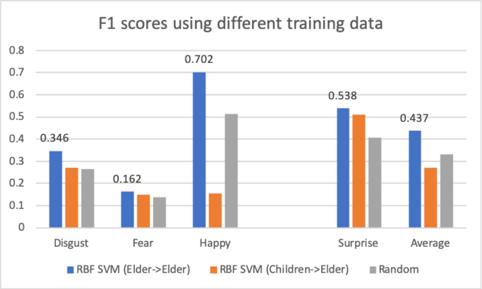

Figure 7: F1 scores using training data from EmoReact or ElderReact, test on ElderReact, use all the features

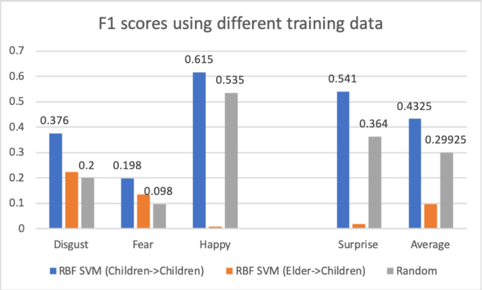

Figure 8: F1 scores using training data from EmoReact or ElderReact, test on EmoReact, use all the features

are of equal importance in detecting such emotions. However, for some emotions such as uncertainty, the performance of using visual features only far outperforms it of using acoustic features only. It might result from the fact that when people or children are uncertain, they usually make little sound but show some actions on face. Therefore, the information is very limited from acoustic source and could be useless in recognizing such emotions while visual features play a crucial part in it.

On elder dataset, the overall F1 scores are slightly lower than that of children dataset (see figure 5 and tabel 8). A common explanation would be that elder people are calm and children are more expressive. It is also interesting to compare the results of the same emotion on these two different age groups. For instance, happiness is the simplest emotion for all ages. While the indication power of acoustic features and visual features are of similar level for recognizing happiness from children, they show different abilities on elder people. Essentially, acoustic features seem to be more powerful in indicating happiness than visual features (figure 6) for elder people and this might come from the biological evidence that the emotion arousal from the face of elder people is much less than that of children, making acoustic features pivotal in recognizing happiness. Another explanation might come from the external interference that in elder dataset, where the background music is light and cheerful, elder people tend to be happy in the same video. This might also adds to the high performance in acoustic features since it is hard to filter the background music from the real sound made by elder people.

From Figure 8 and Figure 9, we can see that a model trained on one dataset does not generalize to the other. The performance of the model trained on EmoReact dropped a lot when testing on ElderReact, so did the other way around. This can be attributed to several reasons. First, children and elder people look differently. Also, there are voice changes during the aging process. In addition, children and elder may express emotions differently so that models may "pay attention" to different features when training on different datasets. All these together may explain why the model performance dropped a lot.

We have also done ablation studies (Table 14) RBF SVM to investigate the important features. F1 scores trained with either audio features or visual features are reported (whether to use audio/visual features depends on which feature we are ablating). The F1 scores dropped a lot when ablating features on ElderReact dataset, supporting that these features are important for recognizing the corresponding emotions. The F1 scores dropped a bit when ablating the audio feature VUV on EmoReact dataset. It didn't change the classifier performance when ablating blink feature on EmoReact, which does not align with our expectation.

Another thing that is noteworthy is that when features are limited, neural networks might have the trend of over exploiting the distribution of data that it simply learns to output a single answer due to the sharp imbalance. For example, the accuracy of tensor fusion network under only audio features (table 13) is generally much higher than that of other models. With a closer look to other metrics at the same time, we found out that it actually returns negative results and because of many of the samples are negative, it achieves high accuracy. To deal with this issue still needs further development of the model itself as well as more balanced data.

6 CONCLUSION AND FUTURE DIRECTIONS

6.1 Take Home Message

Based on our hypothesis testing, we did find out that children and elders express certain emotions differently, either with different muscle movement or with different intensity. Our transfer learning experiments, i.e training the model on one age group and test on the other also shows that the model can not generalize well across between children and elders. We believe that our work prove the need for research effort on the area of emotional response for elders.

6.2 Future Work

Regarding the future work, we believe that 2 directions are worth trying. First, to better leverage the models of neural network, more data can be collected. Right now the deep learning model does not outperform traditional machine learning models, which can be due to the small data size. With more annotated data, the neural network model may start showing its power.

Besides gathering more data, attention mechanism can be utilized on neural network model to visualized the most important features so that we can better understands the results. This can also be done with traditional machine learning models by printing out the weights it assigned to different features. Gaining better understanding of the models would help us find more effective ways to improve them.

7 APPENDIX: TEAM COLLABORATION

Xinyu: Annotated around 35 clips for emotion annotation and around 110 clips for distinguishing whether the disgust is core disgust or social disgust. Designed and ran statistical analyses to see if the hypotheses hold. Conducted feature extraction on AWS and implemented various machine learning models for recognizing different emotions among different age groups. Conducted ablation studies and cross testing.

Xinru: Filtered and categorized around 700 clips of elder people for annotation. Processed and converted features format for predictive models. Finetuned and experimented tensor fusion network, a deep learning model, under various configurations on emotion recognition across different age groups.

Kaixin: Segmented raw videos into short clips and manually filtered out clips with low quality. Manually annotated names of individuals in the dataset. Design and implemented MTurk interface for annotation. Collected annotation results, computed agreement scores for raw annotation and converted raw annotation into labels. Conducted statistical analysis on visual and audio features on both datasets.

Mingtong: Generated instructions for annotation, made standards for annotation and selected annotation materials based on research topic. Gathered relevant psychology background research to delineate the concepts of this research. Annotated all initial 762 clips for emotions, valence etc. Distributed data to other annotators (outside of class) to gather 3rd round of annotation. Raised hypothesis and made hypothesis test.

8 APPENDIX: SUPPLEMENTARY MATERIAL

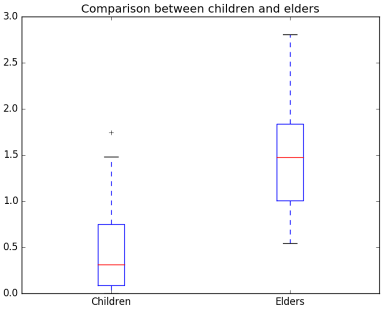

Figure 9: Children and Elders's Upper lip raiser for Happiness

REFERENCES

[1] Charles E Hughes Behnaz Nojavanasghari, Tadas BaltruÅąaitis and Louis-Philippe Morency. 2016. EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (2016).

[2] Z.-J. Chuang C.-H. Wu and Y.-C. Lin. 2006. Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)* (2006).

[3] Duchaine B Dalrymple KA, Gomez J. 2013. The Dartmouth Database of ChildrenâĂŹs Faces: Acquisition and Validation of a New Face Stimulus Set. *PLoS ONE 8(11): e79131, https://doi.org/10.1371/journal.pone.0079131* (2013).

[4] Paul Ekman. 1992. Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 335, 1273 (1992), 63–69.

[5] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384.

[6] P. Ekman. 1994. Strong evidence for universals in facial expressions: a reply to russell's mistaken critique. (1994).

[7] Taorui Ren et al. 2018. Video-based Emotion Recognition Using Multi-dichotomy RNN-DNN. *Asian Conference on Affective Computing and Intelligent Interaction* (2018).

[8] Mara Fölster, Ursula Hess, and Katja Werheid. 2014. Facial age affects emotional expression decoding. *Frontiers in psychology* 5 (2014), 30.

[9] Wallace V Friesen and Paul Ekman. 1991. Emotion, Physiology, and Expression in Old Age. *Psychology and Aging* 6, 1 (1991), 28–35.

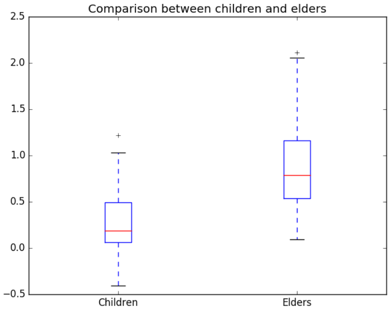

Figure 10: Children and Elders's Cheek raiser for Happiness

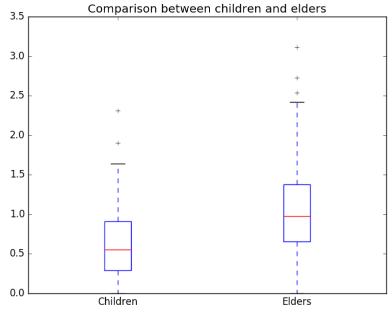

Figure 11: Children and Elders's Lid tightener for Happiness

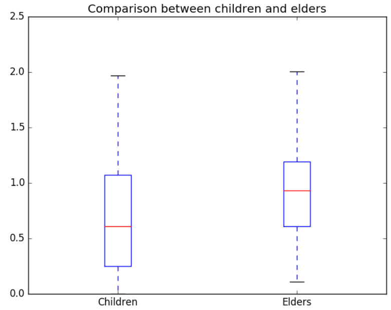

Figure 12: Children and Elders's Lip Corner Puller for Happiness

[10] Derek M Isaacowitz, Corinna E Löckenhoff, Richard D Lane, Ron Wright, Lee Sechrest, Robert Riedel, and Paul T Costa. 2007. Age differences in recognition of emotion in lexical stimuli and facial expressions. *Psychology and aging* 22, 1 (2007), 147.

[11] A. Rozga M. Romero M. Clements S. SclaroïňĂ I. Essa O. Ousley Y. Li C. Kim J. Rehg, G. Abowd. 2013. Decoding childrenâĂŹs social behavior. *InProceedings of the IEEE Conference on ComputerVision and Pattern Recognition* (2013).

[12] G. Li J. Rong and Y.-P. P. Chen. 2009. Acoustic feature selection for automatic emotion recognition from speech. *Information processing management* (2009).

[13] L.-P. Morency K. Bousmalis and M. Pantic. 2011. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. (2011).

[14] Robert W Levenson, Laura L Carstensen, Wallace V Friesen, and Paul Ekman. 1991. Emotion, physiology, and expression in old age. *Psychology and aging* 6, 1 (1991), 28.

[15] V. LoBue and C. Thrasher. 2015. The child aïňĂective facialexpression (cafe) set: Validity and reliability fromuntrained adults. *Frontiers in psychology* (2015).

[16] Aire Mill, Jüri Allik, Anu Realo, and Raivo Valk. 2009. Age-related differences in emotion recognition ability: A cross-sectional study. *Emotion* 9, 5 (2009), 619.

[17] Bunmi O Olatunji, Craig N Sawchuk, Peter J De Jong, and Jeffrey M Lohr. 2007. Disgust sensitivity and anxiety disorder symptoms: Psychometric properties of the disgust emotion scale. *Journal of Psychopathology and Behavioral Assessment* 29, 2 (2007), 115–124.

[18] N. Sato and Y. Obuchi. 2007. Emotion recognition using mel-frequency cepstral coefficients. (2007).

[19] Jane Simpson, Sarah Carter, Susan H Anthony, and Paul G Overton. 2006. Is disgust a homogeneous emotion? *Motivation and emotion* 30, 1 (2006), 31–41.

[20] Kunxia Wang, ZongBao Zhu, Shidong Wang, Xiao Sun, and Lian Li. 2016. A database for emotional interactions of the elderly. *10.1109/ICIS.2016.7550902* (2016).

[21] Kunxia Wang, Zongbao Zhu, Jian Zhang, and Lifei Chen. 2018. Speech emotion recognition of Chinese elderly people. *Web Intelligence* 16 (08 2018), 149–157. https://doi.org/10.3233/WEB-180382

[22] G. I. Roisman Z. Zeng, M. Pantic and T. S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* (2009).

[23] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Empirical Methods in Natural Language Processing, EMNLP.*

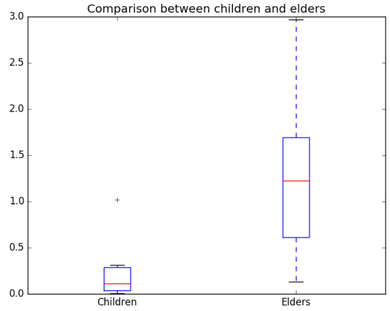

Figure 13: Children and Elders's Upper lip raiser for Surprise

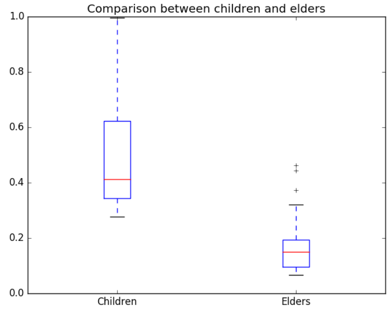

Figure 14: Children and Elders's Blink for Surprise

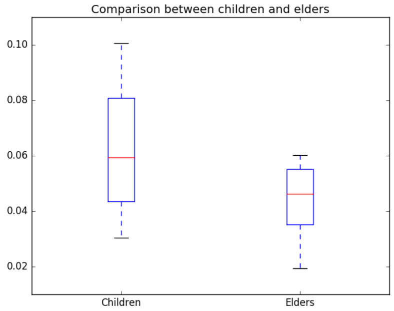

Figure 15: Children and Elders's Upper lid raiser for Disgust