Expression Recognition Across Age

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

***Abstract*— Expression recognition is an important and growing field in AI. It has applications in fields including, but not limited to, medicine, security, and entertainment. A large portion of research, in this area, has focused on recognizing expressions of young and middle-age adults with less focus on children and elderly subjects. This focus can lead to unintentional bias across age, resulting in less accurate models. Considering this, we investigate the impact of age on expression recognition. To facilitate this investigation, we evaluate two state-of-the-art datasets, that focus on different age ranges (children and elderly), namely EmoReact and ElderReact. We propose a Siamese-network based approach that learns the semantic similarity of expressions relative to each age. We show that the proposed approach, to expression recognition, is able to generalize across age. We show the proposed approach is comparable to or outperforms current state-of-the-art on the EmoReact and ElderReact datasets.**

I. Introduction

Facial expressions are an important form of communication between humans, however, there is no consensus as to what they convey [8]. Investigating the role of expression is an important and challenging topic [5], as automatically interpreting expression is an important tool in human-human interactions [19], and human-machine interactions [24]. Systems that can automatically recognize expression have a wide range of applications in fields as diverse and education [20], medicine [21], and entertainment [2]. Although encouraging results have been achieved in these fields and more [18], many works focus on expression recognition in adults with fewer works focusing on different age ranges such as elderly subjects [14]. This lack of focus on a larger range of ages can be explained, in part, by many facial expression datasets being collected in university labs [30], which largely has college aged students as the subjects. Although data collection can be a challenging problem when a larger age range is considered, the lack of this data can cause bias in facial recognition systems [25]. This bias can cause difficulties in expression recognition as it has been shown that different age ranges have different representations of facial expressions. For example, children have different representations and they develop adult-like representations slowly over childhood [6], and elderly subjects display less intense expressions compared to younger subjects [10]. These difficulties can result in ethical concerns such as lack of trust in affective systems, from the general public [3].

While there has been encouraging results in facial expression recognition [13], [22], [27], [28], age was not considered in their experimental design. Although less, there are some interesting works that do take age into account when recognizing facial expression. Ma et al. [14] focused on elderly subjects for their experimental design. They extracted audio and video features from videos of elderly subjects reacting to different tasks. The audio features included voice quality, prosody, and Mel-frequency Cepstral Coefficient features. For video, they extracted facial landmarks, gaze, head pose, and facial action units [4]. They showed these features worked well on some expressions (e.g. happy), but not as well on others (e.g. fear). They also conducted cross-age experiments by training and testing on children data, as well as elderly. Nojavanasghar et al. [16] focused on children and expression recognition. Using similar audio and video features as Ma et al., they evaluated the accuracy of the different modalities separately, as well as with different fusion techniques such as late and hybrid fusion. Their work suggests that a multimodal approach to recognizing facial expressions, in children, results in more accurate recognition. Xu et al. [25] show that a disentangled approach is best to mitigate demographic bias, including age. Their approach disentangles the demographic information and concurrently learns the expression, as well as other information, such as age. They found that this disentanglement increased facial expression recognition accuracy, suggesting that demographic information, such as age, has a negative impact on recognition.

(a) Happy expression from elder and child. (b) Fear expression from elder and child.

Fig. 1: Example differences in expression across age. Children data take from EmoReact [16], and elderly data taken from ElderReact [14].

Motivated by these works, we investigate the problem of the impact of age on facial expression recognition. More specifically, we investigate two contrasting age ranges in children and elderly subjects. To facilitate this investigation, we evaluate the EmoReact (children) [16] and ElderReact (elderly) [14] facial expression datasets. To facilitate this investigation we conduct multiple experiments. First, we evaluate intra-age range expression recognition by training and testing on children and training and testing on elderly. Secondly, we evaluate inter-age range expression by training on children and testing on elderly, as well as training on elderly and testing on children. To conduct these experiments, we propose a Siamese-network architecture that is able to

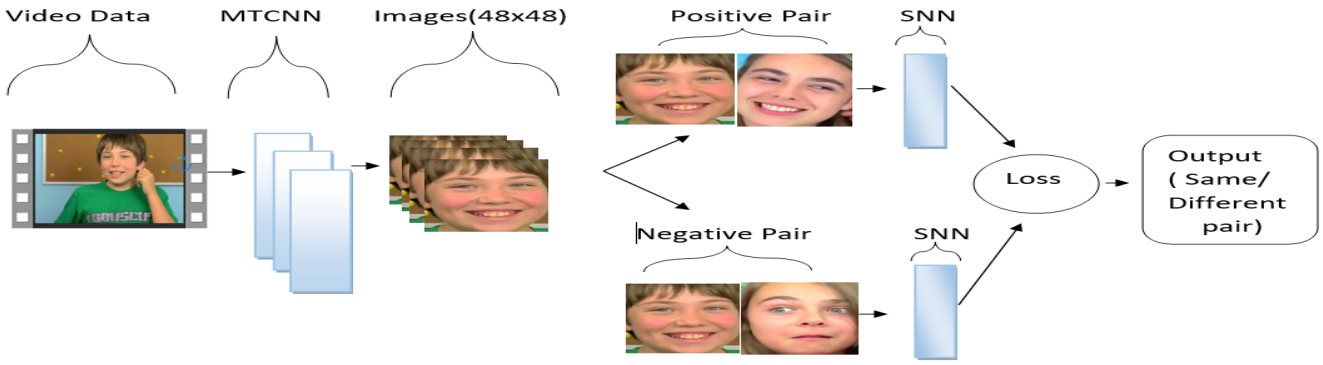


Fig. 2: Overview of end-to-end architecture. Face detection and landmark detection are first performed to crop the face, using MTCNN [29]. Cropped facial images are then resized to 48×48, which are then reshaped to size $[N, M]$, where N is the total number of images and M is the image size. Positive (similar) and Negative (dissimilar) pairs of images are then sent to the 2 sub-networks of the Siamese network (SNN). Contrastive loss is then used to find the distance to the positive and negative classes, which gives us our final classification.

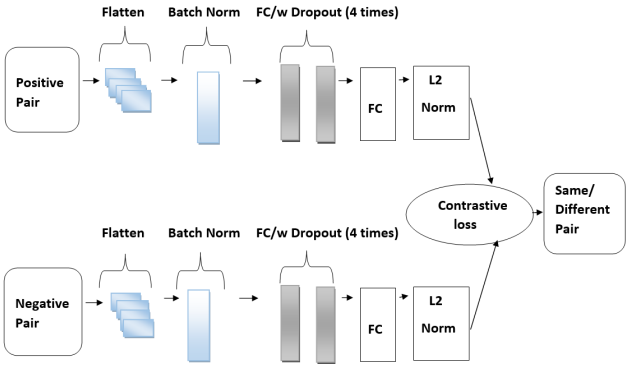


Fig. 3: Proposed Siamese network (SNN).

learn the semantic similarity of the two classes (children versus elderly). See Fig. 2 for an overview of our proposed end-to-end architecture for evaluating the impact of age on expression recognition. Our results suggest that the intra-age range accuracy is higher with elderly subjects, however, the inter-class accuracy is higher when training on children and testing on elderly. The main contributions of this work can be summarized as follows.

1) A Siamese network is proposed to learn the semantic similarity of children and elderly facial expressions. We show that the approach generalizes across age (e.g. children to elderly).
2) Insight into how age impacts expression recognition, including generalization, is detailed.
3) Proposed method, for expression recognition, is comparable to or outperforms state of the art on the EmoReact [16] and ElderReact [14] datasets.

II. Expression Recognition Across Age

We propose an end-to-end system for expression recognition across age, which makes use of Multi-task Cascaded Convolutional Networks (MTCNN) [29] and Siamese networks [17]. We are motivated to use these as MTCNNs have shown encouraging results when used for face detection and expression recognition [23] and Siamese networks have also shown encouraging results for expression recognition [1], [11]. As can be seen in Fig. 2, videos of facial expressions from either children or elderly subjects are used as input to the MTCNN, which then crops the faces. The cropped faces are then used as input to the Siamese network, which classifies the given expression. More details on the proposed approach are given in the following subsections.

A. Multi-task Cascaded Convolutional Network

We used an MTCNN for face detection and face alignment. This architecture consists of three separate convolutional neural networks. The first gets candidates for the bounding box. The second rejects a large number of false candidates, and the third exploits more supervision to find the final bounding box of the face. The MTCNN produces face/non-face classification, a bounding box, and facial landmarks as output. This information is used to crop the face, which we then resize to 48×48 for input to our proposed Siamese network. We refer the reader to the work from Zhang et al. [29] for more details on the MTCNN.

B. Siamese Network

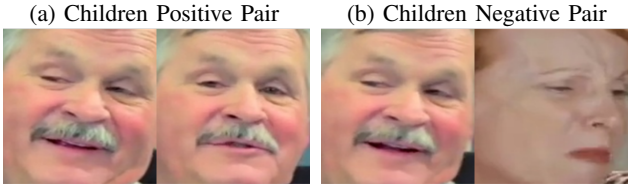
Siamese networks have the same configuration with the same parameters and weights. They find the similarity of the inputs by comparing its feature vectors, which in this case is the RGB data of size 48x48 of the input images. To train the network, anchor, positive, and negative images are used. The distance between the anchor and positive images, as well as the distance between the anchor and negative image is then calculated. The main idea is that the distance from anchor to positive is less than the distance between anchor to negative. For this purpose we have used Contrastive loss function which is defined as

$$loss = (1 - Y)\frac{1}{2}(D_w)^2 + Y\frac{1}{2}\max(0, m - D_w)^2 \quad (1)$$

where $Y = 0$ for similar pairs and $Y = 1$ for dissimilar pairs, and m is a predefined margin. D_w is the Euclidean distance defined as

$$D_w = ||G_w(X_1) - G_w(X_2)||^2 \quad (2)$$

(a) Children Positive Pair (b) Children Negative Pair



(c) Elderly Positive Pair (d) Elderly Negative Pair

Fig. 4: Positive and negative pair from EmoReact [16] (top), and ElderReact [14] (bottom).

where G_w is the output of our network for an image, and X_1 and X_2 are the given input images.

Our proposed architecture consists of a flatten layer, batch normalization, four fully connected layers of 2048 neurons each, dropout, then the final one more fully connected layer followed by L2 normalization (Fig. 3). Given this architecture, we feed the input image pairs to the network and use contrastive loss to minimize the distance for positive and negative pairs. This leads to a binary classification output as to whether the input pair is closer to positive or negative. As the Siamese network produces a binary classification we needed to do further calculations to get the final accuracy of the target expression. Given a test pair and the resulting distance from our proposed architecture, we then calculate whether the distance is positive (i.e. target expression) or negative (i.e. not the target expression) based on a threshold t. If $pred_{dist} < 0.5$, the test pair is the same expression (positive), otherwise it is a different expression (negative). Note, that we only classify the expression as something else, we don't classify it as a specific discrete expression. Finally, we match the ground truth for each class for each classification to get the total number of true positives and true negatives.

Lastly, to create the positive and negative pairs (Fig. 4) for the proposed Siamese network, we consider faces from the same expression as positive, and different expressions as negative. For example, given a Happy expression as the anchor image, another Happy expression is used as the positive, and another expression is randomly chosen as the negative (e.g. Sad). In this way, the negative sample can be any number of expressions that are available for training or testing (e.g. surprise, fear, anger).

III. Experimental Design and Results

As the main focus of this work is on the impact of age on expression recognition, we chose to evaluate two publicly available datasets with large differences in age. Namely one with children under 14 years of age, and another with elderly subjects. Details on these two datasets is given below.

1) EmoReact [16]: A multimodal dataset for recognizing emotional responses in children. It contains 1102 videos of 63 children ages 4-14 years. The dataset is relatively balanced in terms of gender with 51% of the subjects being female. The videos were downloaded from YouTube, where the children are reacting to context such as food, technology, other YouTube videos, and video games. The video were segmented into smaller clips (approximately 5 seconds each), where each contained one child reacting. For each of these, these children perform 5 tasks: (1) being shown the context; (2) being asked a question about it; (3) answering a question about it; (4) being told a fact about it; and (5) explaining their opinion about it. Crowd-sourced workers were used, from Amazon Mechanical Turk, to label the following discrete emotion categories: neutral, disgust, fear, happiness, sadness, surprise, curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment, frustration, and the continuous rating of valence.

2) ElderReact [14]: A multimodal dataset for recognizing emotional responses in elderly subjects. Similar to EmoReact, it contains videos downloaded from YouTube. It contains 43 videos of elderly subjects reacting to context such as video games, social events, and online challenges. These videos were also segmented into shorter clips of approximately 3-8 seconds in length. In total there are 46 subjects (46 female and 20 male). Similar to EmoReact, crowd-sourced workers, from Amazon Mechanical Turk, were also used to annotate the discrete emotions. Comparatively though, there are fewer discrete emotions; anger, disgust, fear, happiness, sadness, and surprise, along with continuous valence.

Anchor Expression	Positive Expression	Negative Expression
Happy	Happy	Disgust Surprise Fear
Disgust	Disgust	Happy Surprise Fear
Surprise	Surprise	Disgust Happy Fear
Fear	Fear	Disgust Surprise Happy

TABLE I: Summary of anchor, positive, and negative pairs.

Training Data	Validation Data	Testing Data
Child	Child	Child
Elder	Elder	Elder
Child	Child	Elder
Elder	Elder	Child

TABLE II: Experiments run on EmoReact and ElderReact.

A. Datasets

B. Experiments

Considering there are different discrete emotions between the EmoReact and ElderReact datasets, we only evaluated the subset that are found in both: disgust, fear, happiness, and surprise. Due to this, we create positive and negative pairs from this subset. See Table I for a summary of each of the possible positive and negative pair expression types. Using

Experiment	Disgust	Fear	Happy	Surprise	Average	Average F1
$Elder \rightarrow Elder$	**.84**	.82	.81	.85	.83	.86
$Elder \rightarrow Child$	.63	.64	.69	.67	.65	.73
$Child \rightarrow Child$	.73	.77	.82	.84	.79	.81
$Child \rightarrow Elder$	.80	**.85**	**.87**	**.88**	**.85**	**.87**

TABLE III: Experimental results across EmoReact and ElderReact. Accuracy is shown except for last column, which is the average F1-score across the four evaluated expressions. First column: $train_{set} \rightarrow test_{set}$.

this subset of data, we conducted four different experiments including within and cross dataset. Specifically, we evaluated expression recognition on elderly face images only, children only, and between elderly and children subjects (see Table II). Each of the datasets come with presorted training, validation, and testing sets. To conduct our experiments, we used these corresponding sets of data for further comparisons to state of the art. To conduct these experiments, we fed the cropped face images into the proposed Siamese network. As a note, the MTCNN preserved the pose of the faces, as well occlusion resulting in challenge data in regards to pose variation and occlusion.

C. Results

As can be seen in Table III, our proposed Siamese network shows encouraging results for expression recognition both within the same age range, as well as across. When the same age was used for training, validation, and testing, the proposed network was able to accurately recognize the expressions. The proposed network achieved an average accuracy of .83 and .79 for elder and child data, respectively. When training on elder data and testing on child data, the average accuracy dropped to .65 from .83. This can be explained, in part, due to the differences in expression across the age ranges, as can be seen in Fig. 1. Although $Elder \rightarrow Child$ recognition decreased, it is interesting to note that the opposite happened with $Child \rightarrow Elder$ recognition, as it achieved the highest average accuracy and F1 score, with .85 and .87, respectively. This is an interesting finding, as it suggests the proposed Siamese network was able to accurately learn the semantic similarities between children and elderly subjects, when children expressions were used to train the network. The decrease in $Elder \rightarrow Child$ recognition, and increase in $Child \rightarrow Elder$ recognition can be explained, in part, by the differences in expression between the two age groups. It can be seen in Fig. 1 that children subjects are visually more expressive compared to elderly subjects. It has also been found that intensity of expression can result in higher recognition accuracies for a variety of tasks [7]. Similar results have also been shown for classification of autism spectrum disorder, across age, where training on children boosted adult testing accuracy [9].

These results suggest that intensity of expression is an important factor for generalizing facial expression recognition across age. More specifically, when training neural networks (e.g. Siamese networks), having a larger range of expression may contribute to improved recognition when age is a considered variable. Finally, these results suggest that age does have an impact on expression recognition, however, the results do not directly support only a negative or positive impact. The impact of age is relevant to the type of data that is used for training.

Method	F1 Score
Proposed	**.81**
Nagarajan et al. [15]	**.81**
Nojavanasghari et al. [16]	.69

TABLE IV: Compare to state of the art on EmoReact [16]. Avg. F1 score across disgust, fear, happy, and surprise shown.

Method	Training/Testing Data	F1 Score
Proposed	$\boldsymbol{Elder \rightarrow Elder}$	**.86**
Ma et al. [14]	$Elder \rightarrow Elder$	.45
Proposed	$\boldsymbol{Elder \rightarrow Child}$	**.73**
Ma et al. [14]	$Elder \rightarrow Child$	.33
Proposed	$\boldsymbol{Child \rightarrow Elder}$	**.87**
Ma et al. [14]	$Child \rightarrow Elder$	.27

TABLE V: Compare to state of the art on ElderReact [14]. Avg. F1 score across disgust, fear, happy, and surprise shown.

1) Comparisons to State of the Art: We also compare our results to current state of the art on both EmoReact and ElderReact. To the best of our knowledge, there are two works that use EmoReact for recognizing discrete expressions, the rest focus on levels of valence and arousal [12]. As can be seen in Table IV, the proposed approach outperforms the baseline (Nojavanasghari et al. [16]), and we are comparable to the work from Nagarajan et al. [15].

To the best of our knowledge, the baseline work from Ma et al. [14], is the only work that has used ElderReact for recognizing discrete expressions. They have conducted the same experiments using a subset of expressions for cross-dataset evaluation. Considering this, the comparisons for each of the evaluated experiments can be seen in Table V. It can be seen that the proposed approach outperforms the baseline for all experiments: training and testing on elder data, training on elder and testing on child, and training on child and testing on elder. Most notable, is the result when training on child and testing on elder. Ma et al. [14] achieved an average F1 score of .27, while the proposed approach increased this by .6. This can be attributed, at least partially, to the proposed Siamese network, whereas the baseline work used hand-crafted features and an RBF SVM [26].

IV. Conclusion

We have presented an approach to expression recognition across age that makes use of a Siamese network to learn the semantic similarity between children and elderly subjects. We conducted experiments on the publicly available EmoReact and ElderReact datasets for both within dataset and cross dataset. The proposed approach outperforms the baseline and is comparable to other work on EmoReact, while the proposed approach outperforms the baseline on ElderReact for all experiments. This work is applicable to fields including medicine, security, and education.

References

[1] W. J. Baddar, D. H. Kim, and Y. M. Ro. Learning features robust to image variations with siamese networks for facial expression

recognition. In *International Conference on Multimedia Modeling*, pages 189–200. Springer, 2017.

[2] P. M. Blom, S. Bakkes, C. Tan, S. Whiteson, D. Roijers, R. Valenti, and T. Gevers. Towards personalised gaming via facial expression recognition. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 10(1), 2014.

[3] R. Cowie. Ethical issues in affective computing. In *The Oxford handbook of affective computing*, pages 334–348. Oxford University Press, 2015.

[4] P. Ekman and W. V. Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[5] C. Frith. Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3453–3458, 2009.

[6] X. Gao, D. Maurer, and M. Nishimura. Similarities and differences in the perceptual structure of facial expressions of children and adults. *Journal of Experimental Child Psychology*, 105(1-2):98–115, 2010.

[7] S. Hinduja. Analysis of emotions using multimodal data. *Ph.D. Dissertation, College of Engineering, University of South Florida*, 2021.

[8] G. Horstmann. What do facial expressions convey: Feeling states, behavioral intentions, or actions requests? *Emotion*, 3(2):150, 2003.

[9] S. R. Jannat and S. Canavan. Classification of autism spectrum disorder across age using questionnaire and demographic information. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 52–61. Springer International Publishing, 2021.

[10] R. W. Levenson, L. L. Carstensen, W. V. Friesen, and P. Ekman. Emotion, physiology, and expression in old age. *Psychology and aging*, 6(1):28, 1991.

[11] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, and S. Wen. Saanet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing*, 413:145–157, 2020.

[12] A. Lopez-Rincon. Emotion recognition using facial expressions in children using the nao robot. In *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 146–153. IEEE, 2019.

[13] Y. Lv, Z. Feng, and C. Xu. Facial expression recognition via deep learning. In *2014 international conference on smart computing*, pages 303–308. IEEE, 2014.

[14] K. Ma, X. Wang, X. Yang, M. Zhang, J. M. Girard, and L.-P. Morency. Elderreact: a multimodal dataset for recognizing emotional response in aging adults. In *2019 International Conference on Multimodal Interaction*, pages 349–357, 2019.

[15] B. Nagarajan and V. R. M. Oruganti. Cross-domain transfer learning for complex emotion recognition. In *2019 IEEE Region 10 Symposium (TENSYMP)*, pages 649–653. IEEE, 2019.

[16] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th acm international conference on multimodal interaction*, pages 137–144, 2016.

[17] S. K. Roy, M. Harandi, R. Nock, and R. Hartley. Siamese networks: The tale of two manifolds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3046–3055, 2019.

[18] M. Sajjad, M. Nasir, F. U. M. Ullah, K. Muhammad, A. K. Sangaiah, and S. W. Baik. Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services. *Information Sciences*, 479:416–431, 2019.

[19] M. Schlögl and C. A. Jones. Maintaining our humanity through the mask: Mindful communication during covid-19. *Journal of the American Geriatrics Society*, 68(5):E12, 2020.

[20] X.-Y. Tang, W.-Y. Peng, S.-R. Liu, and J.-W. Xiong. Classroom teaching evaluation based on facial expression recognition. In *Proceedings of the 2020 9th International Conference on Educational and Information Technology*, pages 62–67, 2020.

[21] P. Thiam, H. A. Kestler, and F. Schwenker. Two-stream attention network for pain recognition from video sequences. *Sensors*, 20(3):839, 2020.

[22] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020.

[23] J. Xiang and G. Zhu. Joint face detection and facial expression recognition with mtcnn. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 424–427. IEEE, 2017.

[24] G. Xiao, Y. Ma, C. Liu, and D. Jiang. A machine emotion transfer model for intelligent human-machine interaction based on group division. *Mechanical Systems and Signal Processing*, 142:106736, 2020.

[25] T. Xu, J. White, S. Kalkan, and H. Gunes. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, pages 506–523. Springer, 2020.

[26] X.-f. Yan, H.-w. Ge, and Q.-s. Yan. Svm with rbf kernel and its application research. *Computer Engineering and Design*, 27(11):1996–1997, 2006.

[27] M. Yu, H. Zheng, Z. Peng, J. Dong, and H. Du. Facial expression recognition based on a multi-task global-local network. *Pattern Recognition Letters*, 131:166–171, 2020.

[28] F. Zhang, T. Zhang, Q. Mao, and C. Xu. Geometry guided pose-invariant facial expression recognition. *IEEE Transactions on Image Processing*, 29:4445–4460, 2020.

[29] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[30] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016.