HCC: An explainable framework for classifying discomfort from video

William Valentine[1*], Megan Webb[2*], Christopher Collum[2], David Feil-Seifer[2], and Emily Hand[2]

[1] Rose-Hulman Institute of Technology, Terre Haute, IN 47803, USA
valentwa@rose-hulman.edu
[2] University of Nevada, Reno, NV 89557, USA
{meganwebb, ccollum, dave, emhand}@unr.edu

Abstract. We present Human Comfort Classifier (HCC): A framework for classifying human discomfort from video. Recognizing comfort and discomfort in social interactions is something that many of us do without having to think about it. However, identifying discomfort in others can be a challenge for individuals with social skills deficits, who often become socially isolated. Social isolation can lead to many negative outcomes for individuals and is recognized by the CDC and WHO as a priority public health problem. In this work, we propose HCC to detect discomfort in videos. This can be utilized for training for individuals with social skills deficits. HCC utilizes a multi-modal approach of pose estimation, facial landmarks, and natural language processing to determine comfort in real time. We utilize an explainable rule-based model to categorize behavior and achieve approximately 78% prediction accuracy on an interview dataset.

Keywords: Emotion perception · Explainable computer vision · Video understanding

1 Introduction

State-of-the-art emotion detection technology has predominantly focused on discerning the six basic emotions of happiness, sadness, anger, surprise, fear, and disgust which are considered to be clearly recognizable emotions [12,25]. We seek to expand on the work in this field to include more subtle emotions for a particular application. We introduce Human Comfort Classifier (HCC), a practical emotion classifier for the subtle emotions of comfort and discomfort. HCC is the first of its kind and will open up a new research direction into the detection of more subtle emotions.

HCC could be particularly valuable for individuals with social skills deficits (SSDs) in social skills training (SST). Individuals with SSDs include those with Autism Spectrum Disorder (ASD), Attention-Deficit Hyperactivity Disorder (ADHD),

* equal contribution

and anxiety [7,13,21]. These individuals struggle with personal and professional relationships and face higher rates of social isolation and loneliness [6,28]. Social isolation and loneliness are serious public health risks that are recognized by the CDC and WHO as a priority public health problem [5,31]. Individuals experiencing social isolation and loneliness can incur long-term physical damage and higher rates of depression [5,24]. These negative effects have become increasingly important in recent years with 1 in 44 children having a diagnosis of ASD in 2018, which is over three times the rate of just eighteen years prior [23]. SST can assist this increasing population of individuals through different approaches, including behavioral approaches and relationship-based approaches [2]. Cognitive-behavioral approaches focus on expanding the perspective of the individual to others, and video methods are already used in this approach [2]. Particular SST approaches vary between geographical areas, and can be expensive in terms of the cost and travel required. An online SST tool – such as the proposed HCC – that classifies comfort and discomfort in others can be used to expand the perspective of individuals with SSDs in a behavioral approach, and it can help make SST more accessible. For example, the tool can: 1) supplement professional social skills trainers by allowing for at-home practice, 2) be part of the development of an entirely at-home training, or 3) be made available to those who cannot afford more personalized SST.

Three main obstacles exist when applying emotion perception technology to a particular application. First, current datasets and models face particular limitations when applied to real-world scenarios. Second, expressions of specific emotions can be inconsistent between individuals and contexts. Third, emotion perception methods often rely on deep learning, rather than explainable models, which can reduce explainability and limit downstream applications. These obstacles are discussed below.

Numerous datasets and accurate models have been developed for the classification of the basic emotions of happiness, sadness, anger, surprise, fear, and disgust [11], although obtaining reliable data remains a challenge. These datasets have largely been based on static images of individuals rather than videos [25]. Using videos, rather than static images, would allow the model to perceive transitory states of emotion expression within social situations [9]. Additionally, these images of emotional expressions are often preformed by actors in a lab rather than obtained from natural emotions expressed in the wild [25]. There are systematic differences with acted performances that lead to atypical features compared to natural expressions [10]. HCC uses video data from public interviews in order to best reflect real emotional expressions.

Another obstacle involves the variability of physical expressions of emotion in this research. Individuals may express emotions through facial expression, body language, speech content, and speech style (volume, pitch, pauses). Emotion detection by facial expression has been most frequently studied [25]. However, different emotions are expressed to varying degrees across parts of the body [4,17]. This variability makes it important to incorporate multiple types of physical expression when detecting emotions. Limited work has been done with mod-

els based on expressions through body language, but they tend to yield worse results than studies on facial expressions [25]. In this work, we incorporate multiple modalities into our model – facial expressions, body language, and verbal expressions – for a more holistic approach to emotion perception.

Lastly, an emotion perception system that is opaque to its user would be limited in its informative capacity for SST applications. Because it is important for assistive technologies to be designed specifically for the individuals who will use them, we favor the use of explainable systems for our results rather than deep learning techniques. Deep learning techniques may be able to achieve a higher level of accuracy when classifying comfort and discomfort, but are not able to explain the results. We utilize deep learning based models to process the data that is fed into our explainable model, thereby attaining the benefits of deep learning without sacrificing explainability where it is needed most: in what features are expressing comfort and discomfort. It is crucial when assisting these individuals in learning to recognize these emotions themselves that it is understandable what cues result in a classification of comfort of discomfort. Our method produces a set of features for comfort and discomfort that are applied to each video, which informs the viewer of the cues that resulted in a comfort or discomfort classification.

Our Human Comfort Classifier (HCC) responds to constraints of applying emotion detection and perception to a subtle emotion. We utilize existing models crucial for emotion expression [17,22] and data from videos recorded in real-world environments to construct a rule-based system that prioritizes explainability. Our primary contributions are summarized as follows:

- We present HCC: A framework for classifying expressions of comfort and discomfort from video.
- We utilize facial landmarks, pose estimation, and sentiment analysis as inputs for our model.
- We develop a rule-based framework that operates in real time.

2 Background

The field of affective computing – the study of recognizing and processing human emotion using computer vision – has been active for decades. Our approach draws on emotion detection, pose estimation, and sentiment analysis to provide rule-based, explainable, real-time capabilities for discerning discomfort for the assistance of those with social skills deficits. We discuss the areas related to our work below.

2.1 Emotion Detection and Perception

Emotions are internal experiences within a person that cannot be objectively determined by others. Expressions, whether facial, body, or verbal, indicate that a person is experiencing an emotion (although a lack of expression does not mean

that an emotion is not occurring) [25]. The process of emotion detection consists of predicting emotion labels through the use of expressed cues [27]. Since humans cannot objectively determine the emotional states of others without their confirmation, we seek to contribute a model that predicts human perception of emotions, rather than the internal emotional state of an individual. For example, if Person A and Person B perceive Person C to be frustrated when they squint, then HCC will perceive what Person A, Person B, and others collectively perceive, whether or not Person C is truly experiencing that emotion. We seek to contribute a model that can detect cues of particular emotions, not one that determines internal emotional states [29].

Emotions can be detected from visual cues of the face and body and from verbal cues of words and voice [19]. Most commonly, cues from facial expressions are studied since they are considered reliable [25]. Facial cues are less frequently studied in conjunction with other important cues from language and context, additionally, only the six basic emotions are typically studied [12,25]. There has been work that includes additional emotions [27], although to our knowledge comfort and discomfort have not yet been studied. More commonalities among affective computing research include the utilization of datasets containing images, rather than video, and posed or acted emotions rather than natural expressions that occur in real-world scenarios and interactions [10,25]. However, recent research has found that video-based methods of affective computing are more robust and effective than image-based methods because of the dynamic nature of facial expressions [9]. In response, the proposed HCC is a video-based methodology that includes only natural expressions. Beyond facial expressions, HCC also factors in pose, word sentiment, and analysis of the audio's pitch and volume.

2.2 Pose Estimation

Previous studies have not always incorporated body language, or the way of measuring it which is pose estimation, despite its value in perceiving emotion. Similar to facial expressions, body language is dynamic and video methods capture this data better than still images [19]. Human interpretation of emotion can be greatly influenced by body language. For example, prior work has found that participants will perceive a face expressing anger as fearful when a fearful body pose is displayed underneath that facial expression [1]. When distinguishing between subtle emotions, visual cues beyond facial expression must be included in the model.

2.3 Sentiment Analysis

The majority of human communication is conveyed through non-verbal elements including facial expression and body language [16]. However, non-verbal elements make up 55% of human communication and the remaining 45% consists of an individual's vocal tone and words [16]. Although verbal elements are not the

majority, they are still a sizable component worth incorporating [4,27]. Additionally, the expression of particular emotions may not be equally distributed across modalities: sadness is best interpreted through audio, and anger is best interpreted from video [8]. The proposed HCC incorporates video and audio modalities in order to achieve greater accuracy in predicting discomfort.

Previous emotion perception studies that incorporate sentiment analysis tend to utilize methodologies that inhibit real-time applications. For example, one study hired professionals to write transcriptions of audio data for the model, which means that the model cannot incorporate new data without the use of these professionals [4]. The proposed HCC utilizes a transformer to produce transcripts of speech as well as a rule-based system that contributes to model explainability. With the ability to produce transcriptions automatically, HCC has real-time capabilities. Real-time capabilities allow for flexibility in assisting individuals participating in social skills training.

3 Methodology

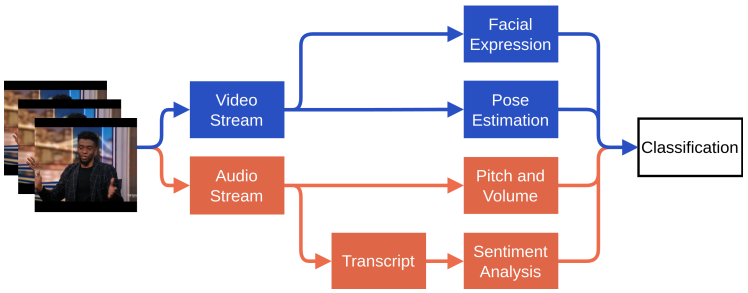

Fig. 1: Pipeline for video processing. Initially, processed video frames are split into video (the upper half of the diagram in blue) and audio (the lower half of the diagram in orange) streams. From the video stream, facial expressions and pose estimation are extracted. From the audio stream, pitch and volume, as well as transcripts are extracted. Sentiment analysis is performed on the generated transcripts. Finally, the rule-based system evaluates the data from both video and audio streams and classifies the perceived emotion as comfort or discomfort. The separation of audio and video data allows for simultaneous video and audio processing, so that predictions are made in real time. Real-time predictions permit this model to have SST applications to benefit individuals who are learning to better understand comfort and discomfort in social situations.

In this study, we utilize pose estimation, facial landmark generation, audio and text analysis as inputs to our rule-based model. Therefore, we must briefly discuss each of the primary stages in our framework's pipeline. HCC takes a real-time video feed and splits it into simultaneously occurring audio and video

streams, as illustrated in Figure 1. Then, two CNN models, BlazePose [3] and FaceMesh [18], are used to generate pose and facial landmarks for our rules to analyze. Although BlazePose includes 11 landmarks for the face, the landmarks from FaceMesh are more extensive at 468 landmarks for the face. We utilize FaceMesh for facial expressions, rather than only using BlazePose. This is because FaceMesh has higher precision for facial landmark locations and BlazePose lacks enough landmarks to capture smaller motions of the face [3,18]. Both of these models are capable of running in real time. In the audio stream, an additional transformer-based model, Whisper [26], generates transcripts which are analyzed for positive and negative sentiment. We also generate arrays of pitch and volume corresponding to each video frame. Finally, this information is fed into our rules which provide a binary classification of the subject's comfort. We emphasize that all aspects of this process are real-time allowing for HCC to be useful in SST applications. HCC's binary classification is explainable, hence it does not use deep learning methods to perform the final comfort/discomfort classification, so that it can always provide its exact rationale for why a classification was made. The exact rationale for a classification is essential for at home trainings for social skills and other HCI applications. We elaborate on each step of the pipeline in the following subsections.

3.1 Pose Estimation

We estimate pose with a CNN that generates 33 pose-based landmark points. We build off of prior work that incorporates heat and offset maps which allow for greater precision in extracting pose landmarks [3]. To optimize this network for our purposes, we do not utilize pose landmarks below the torso. We find that cameras positioned for conversation often leave the lower body and arms below the elbow out of view. This allows for a reduction in errors of the model. This CNN can support 256x256 resolution images, so the input frames are scaled down to that resolution. This is the only pre-processing performed on the input frames. Ultimately, the performance of this model is beyond real time, since it can inference 30 to 100 frames per second. Real-time capabilities are important for this model to be applied to collaborative situations, such as a training program.

3.2 Facial Expression

In a similar fashion to the pose estimation segment, we utilize FaceMesh to generate 468 facial landmarks for use in our HCC [18]. This model supports 256x256 resolution images, which allows for the previously preprocessed frames to be reused for efficiency. The performance of this model is beyond real time at 100 to 1000 frames per second when optimized by a GPU.

3.3 Audio Stream

Volume and Pitch We utilize Fast Fourier Transforms (FFTs) to extract our pitch data from the raw audio stream. We then extract the intensity of the sound

to represent our volume and create an array of the volume and pitch for each frame.

Sentiment Analysis We use Whisper [26], a transformer-based model, to generate a transcript in five second chunks. Our audio is resampled to 16,000 Hz as that is the expected input frequency for the Whisper model [26]. We perform a simple sentiment analysis on the transcript using a dataset of positive and negative sentiment words [15]. Each word is given a value according to its determined sentiment that is either 0 (neutral), greater than 0 (positive), or less than 0 (negative). All of these values are summed together for each sentence to determine a basic sentiment for the sentence.

3.4 Rules

We utilize the visual and auditory sources of data to generate our binary classification. We use a rule-based approach where we identify rules based on psychology research and four videos from our annotated dataset, CID. First, we have rules that dictate discomfort from rapid twitching movement, or from complete stillness in the body landmarks. These rules determine the displacement of the body between frames. Second, we have rules that examine the negative sentiment of a statement and if combined with any other positive discomfort rule, will classify the person as uncomfortable. Finally, we have rules dictating the length of pauses and the number of filler words in a conversation. If both of these pass a certain threshold, we classify the person as uncomfortable. We utilize this rules-based model instead of a deep-learning approach to enable the extraction of the exact rationale for why a classification is made. The exact rationale for a classification is essential for at home trainings for social skills and other HCI applications. In order to properly train an individual to understand comfort and discomfort in social settings, it is imperative that a natural language explanation can be provided. The following is a list of rule descriptions.

3.5 Summary of Rules

We present a summary of our rules.

Rule **1**: Increased movement of the arms indicates comfort [16].
Rule **2**: The lack of movement of the torso indicates discomfort [16].
Rule **3**: Long pauses in conversation indicate discomfort [20,30]
Rule **4**: A high number of words spoken with a rapid response time indicate comfort [20].
Rule **5**: A high amount of time an individual spends smiling indicates comfort.

$$score = 0.2R_1 - 0.1R_2 - 0.1R_3 - 0.2R_4 + 0.2R_5 \quad (1)$$

Fig. 2: One example of an interview in the CID dataset. The individual being interviewed displays discomfort according to our three human reviewers. Her face and upper torso are clearly visible to the camera for about 30 seconds.

Fig. 3: A second example of an interview in the CID dataset. The individual being interviewed displays comfort according to our three human reviewers. His face and upper torso are clearly visible to the camera for about 30 seconds.

Equation 1 shows the formula for calculating the comfort score from the rules listed above. The coefficients are determined experimentally. We run our data points through our rule-based system which provides a binary classification of true (for comfortable) or false (for uncomfortable). Each of these rules modifies a score. The range of this score is -0.4 to 0.4. If the threshold is passed for each of these rules then their weight is added to the score. If this score is negative, then the individual is classified as uncomfortable, and if it is non-negative, then the classification is comfortable.

3.6 Dataset

We introduce a dataset, **Celebrity Interview Dataset (CID)**, of 18 YouTube celebrity interview videos that take place in the United States of America. These interviews were annotated with respect to the discomfort or comfort of the interviewee lasting for a duration of approximately 30 seconds from a culturally American perspective. Each clip had a front facing view of the interviewee so that the face and majority of the torso are clearly visible. We process each of the videos to pad or scale the video to a 1080x1080 resolution. Additionally, the frame rate for each of the videos is set to 30fps. Of the videos, 44% display discomfort and the other 56% display comfort resulting in a relatively even split between classes. We introduce CID because, to the best of our knowledge, there are no other datasets annotated for comfort and discomfort. CID will be made publicly available after publication. We believe 18 videos to be a number sufficient for a proof-of-concept test of our novel HCC framework.

Annotation A group of three annotators annotated each of the video clips with either a comfortable or uncomfortable tag based on a group vote. Videos that could not reach a unanimous decision were removed from the dataset.

Sample Frames We provide two sample frames (Figure 2 and Figure 3) from our dataset to help to illustrate the common attributes of the videos we select. The individual's torso and most of their arms must be visible. Additionally, they must have the majority (over 90%) of their face visible along with being able to be accurately heard by the microphone.

Licenses The videos we use are generally under thirty seconds and are used for research purposes. Therefore, they fall under the category of Fair Use by YouTube's guidelines and do not need a formal license.

Demographics Table 1 provides context to the subjects within the videos. We found that white men are over-represented in available interview videos of individuals experiencing discomfort.

	Male	Female
Hispanic	1	0
White	9	1
Black	2	2
Asian	1	2

Table 1: Demographic Distribution

4 Experimental Results

Approach	Accuracy
Only audio	76.91%
Only video	64.93%
Combined	**77.96**%

Table 2: HCC Results

We utilize a simple average of correct predictions divided by total frames to determine the accuracy of our rules. We run our model on every frame in every video in our dataset, and then average the accuracy from each video. Four videos were used from the dataset for creating the rules which represents 22% of our total videos. All of the videos were then used to validate the rules.

Table 2 shows the accuracy of using only audio or only video inputs to show the increased accuracy of using a multi-modal approach. Audio alone has a high accuracy, but the video rules combined with audio rules reduce false positives from HCC. Figure 4 shows the confusion matrix for our combined audio and video HCC. We see that there is no major imbalance between false positives or false negatives, and therefore the model is doing well on predicting both comfort and discomfort.

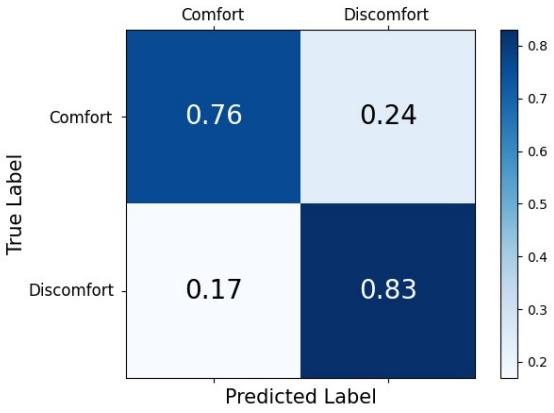

Fig. 4: This confusion matrix shows the predicted amounts of false comfort (bottom left) and false discomfort (top right) that occur when HCC runs on CID. We see that no unexpected results are found from the matrix. There is no high distribution of false positives or false negatives.

Performance and Frame Rate We simultaneously generate landmarks and process audio in order to reduce computational idleness and optimize the pipeline's performance. We use a GeForce RTX 2080 to run HCC which achieved an average frame rate of 44.27. We assume real-time to be 30 frames per second, thus this demonstrates that HCC is 47.57% faster than real time. Prioritizing real-time capabilities permits this model to have beneficial applications in social skills training.

5 Discussion

This research both supports the study of subtle emotions [14] and demonstrates a novel method for designing emotion detection models for an assistive application. We demonstrate the feasibility of studying subtle emotions, specifically comfort and discomfort, in affective computing.

5.1 Results

We found that the audio rules tended to produce high overall accuracy, with the primary issue being that these rules performed poorly with the display of subtle comfort or discomfort. However, the video rules did well when confronted by subtle displays of feeling, but suffered when the presentation of these feelings was blatant. The combination of both audio and video rules demonstrates a clear increase of the accuracy and usefulness of the prediction. If we had a high false discomfort score, the model's usefulness would be reduced. Due to the

inclusion of both video and audio rules, this false discomfort score is minimized. Our accuracy is strong considering that the typical accuracy of state-of-the-art emotion detection on videos rests within 56.64%-84.39% by one standard deviation [32]. HCC is within an ideal range of accuracy while being a proof-of-concept that is uniquely centered around subtle emotions, real-time capabilities, and explainable rules rather than deep learning.

5.2 Limitations

The following limitations should be considered while interpreting these results. First, the novel rules for comfort and discomfort are based broadly on psychology research and intuition. HCC is not able to adjust or create rules based on the input data itself. This allows for HCC to be explainable, but it means that expressions of (dis)comfort that humans are not yet aware of may not yet be incorporated into the model. To accommodate for this limitation, we include hyper-parameters that can be adjusted as signals of discomfort become better understood. Second, CID is biased toward individuals who have attained celebrity status, and CID is limited in size (18 videos) and demographics. Although the six basic emotions are considered to be universal among different cultures, we can not validate that the more subtle emotion of comfort is expressed the same in different contexts by different groups, which may be the cause of error in our model. Third, the context in which our data takes place is predominantly one-on-one interview settings, and our model has not been tested on individuals within group contexts. We seek to expand CID in future work. Nonetheless, our data consists of individuals experiencing natural emotions in real-world contexts, which is an advantage over datasets that consist of acted emotions in a lab environment.

5.3 Future Work

The numerous capabilities and practicality of this novel system leave many areas to be expanded upon. We foresee experimenting with a more extensive verbal expression system (more robust sentiment analysis, expansion to other languages, calibration to different individuals), including more features (eye contact detection, distinction between speakers, rules that incorporate differences in expression by gender), and expanding beyond the binary classification to include additional emotions. Importantly, the hyper-parameters can also be easily adjusted in response to continuing emotion detection and psychology research into displays of discomfort.

HCC is designed to be applied to social skills training environments, which work to reduce social isolation and related negative health effects. Other applications for the multi-modal, real-time, rule-based model can be found in retail and service industries, virtual and augmented reality experiences, and the field of human-robot interaction.

6 Conclusion

We present Human Comfort Classifier: A novel framework for classifying human discomfort in real time from videos. HCC utilizes a multi-modal approach of pose estimation, facial landmarks, and natural language processing to determine comfort. We opt for HCC to be explainable and thereby do not utilize deep learning for its binary classification. The reason why we need to keep HCC explainable is because the exact rationale for a classification is essential in many HCI situations such as at home trainings for social skills. We find that rules related to silence in conversations have a strong accuracy in predicting the discomfort of an individual. We find that our rules about stillness of the torso and arms tend to accurately predict comfort. Additionally, HCC inferences in super real-time: about 47.57% faster than real-time. Finally, we introduce CID, a novel dataset for classifying comfort and discomfort. We have utilized a rule-based model to categorize human behavior and achieve approximately 78% prediction accuracy on CID, successfully demonstrating proof-of-concept of our novel work.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant #IIS-2150394. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Aviezer, H., Hassin, R.R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., Bentin, S.: Angry, disgusted, or afraid? studies on the malleability of emotion perception. Psychological science **19**(7), 724–732 (2008)
2. Baker, J.: Key components of social skills training. Teaching Social Skills to People with Autism: Best Practices in Individualizing Interventions. Bethesda, MD: Woodbine House, Inc (2013)
3. Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M.: Blazepose: On-device real-time body pose tracking. CoRR **abs/2006.10204** (2020), `https://arxiv.org/abs/2006.10204`
4. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**, 335–359 (2008)
5. CDC: Loneliness and social isolation linked to serious health conditions. Alzheimer's Disease and Healthy Aging (2021)
6. Conger, J.C., Keane, S.P.: Social skills intervention in the treatment of isolated or withdrawn children. Psychological Bulletin **90**(3), 478 (1981)
7. De Boo, G.M., Prins, P.J.: Social incompetence in children with adhd: Possible moderators and mediators in social-skills training. Clinical psychology review **27**(1), 78–97 (2007)
8. De Silva, L.C., Miyasato, T., Nakatsu, R.: Facial emotion recognition using multimodal information. In: Proceedings of ICICS, 1997 International Conference on

Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. vol. 1, pp. 397–401. IEEE (1997)
9. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops). pp. 2106–2112. IEEE (2011)
10. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Towards a new generation of databases. Speech communication **40**(1-2), 33–60 (2003)
11. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of personality and social psychology **17**(2), 124 (1971)
12. Gendron, M., Mesquita, B., Barrett, L.F.: 538539 Emotion Perception: Putting the Face in Context. In: The Oxford Handbook of Cognitive Psychology. Oxford University Press (03 2013). https://doi.org/10.1093/oxfordhb/9780195376746.013.0034, https://doi.org/10.1093/oxfordhb/9780195376746.013.0034
13. Hagopian, L.P., Kuhn, D.E., Strother, G.E., Van Houten, R.: Targeting social skills deficits in an adolescent with pervasive developmental disorder (2009)
14. Harrigan, J.A., O'Connell, D.M.: How do you look when feeling anxious? facial displays of anxiety. Personality and individual differences **21**(2), 205–212 (1996)
15. Hu, M., Liu, B.: Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (2004), https://api.semanticscholar.org/CorpusID:207155218
16. Ilyas, C.M.A., Nunes, R., Nasrollahi, K., Rehm, M., Moeslund, T.B.: Deep emotion recognition through upper body movements and facial expression. In: VISIGRAPP (5: VISAPP). pp. 669–679 (2021)
17. Kadambi, A., Ichien, N., Qiu, S., Lu, H.: Understanding the visual perception of awkward body movements: How interactions go awry. Attention, Perception, & Psychophysics **82**(5), 2544–2557 (Jul 2020). https://doi.org/10.3758/s13414-019-01948-5, https://doi.org/10.3758/s13414-019-01948-5
18. Kartynnik, Y., Ablavatski, A., Grishchenko, I., Grundmann, M.: Real-time facial surface geometry from monocular video on mobile gpus (2019)
19. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Context based emotion recognition using emotic dataset. IEEE transactions on pattern analysis and machine intelligence **42**(11), 2755–2766 (2019)
20. Koudenburg, N., Postmes, T., Gordijn, E.: Disrupting the flow: How brief silences in group conversations affect social needs. Journal of Experimental Social Psychology **47**, 512–515 (03 2011). https://doi.org/10.1016/j.jesp.2010.12.006
21. Liberman, R.P.: Assessment of social skills. Schizophrenia Bulletin **8**(1), 62 (1982)
22. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines. ArXiv **abs/1906.08172** (2019), https://api.semanticscholar.org/CorpusID:195069430
23. Maenner, M.J., Warren, Z., Williams, A.R., Amoakohene, E., Bakian, A.V., Bilder, D.A., Durkin, M.S., Fitzgerald, R.T., Furnier, S.M., Hughes, M.M., Ladd-Acosta, C.M., McArthur, D., Pas, E.T., Salinas, A., Vehorn, A., William, S., Esler, A., Grzybowski, A., Hall-Lande, J., Nguyen, R.H., Pierce, K., Zahorodny, W., Hudson, A., Hallas, L., Mancilla, K.C., Patrick, M., Shenouda, J., Sidwell, K., DiRienzo, M., Gutierrez, J., Spivey, M.H., Lopez, M., Pettygrove, S., Schwenk, Y.D., Washington, A., Shaw, K.A.: Prevalence and characteristics of autism spectrum disorder among

children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2020. MMWR Surveillance Summaries **72**(2) (2023)
24. Novotney, A.: The risks of social isolation. Monitor on Psychology **50**(5), 32–37 (2019), https://www.apa.org/monitor/2019/05/ce-corner-isolation
25. Pereira, R., Mendes, C., Ribeiro, J., Ribeiro, R., Miragaia, R., Rodrigues, N., Costa, N., Pereira, A.: Systematic review of emotion detection with computer vision and deep learning. Sensors **24**(11) (2024). https://doi.org/10.3390/s24113484, https://www.mdpi.com/1424-8220/24/11/3484
26. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022)
27. Ranganathan, H., Chakraborty, S., Panchanathan, S.: Multimodal emotion recognition using deep learning architectures. In: 2016 IEEE winter conference on applications of computer vision (WACV). pp. 1–9. IEEE (2016)
28. Segrin, C., Kinney, T.: Social skills deficits among the socially anxious: Rejection from others and loneliness. Motivation and emotion **19**, 1–24 (1995)
29. Stratton, D., Hand, E.: Bridging the gap between automated and human facial emotion perception. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2400–2410 (2022). https://doi.org/10.1109/CVPRW56347.2022.00268
30. Templeton, E.M., Chang, L.J., Reynolds, E.A., Cone LeBeaumont, M.D., Wheatley, T.: Fast response times signal social connection in conversation. Proceedings of the National Academy of Sciences **119**(4), e2116915119 (2022)
31. WHO: Social isolation and loneliness (2024)
32. Xue, J., Wang, J., Wu, X., Zhang, Q.: Affective video content analysis: Decade review and new perspectives (2024), https://arxiv.org/abs/2310.17212