# Facial expression interpretation in ASD using Deep Learning

Pablo Salgado[1][0000−0002−1750−2627], Oresti Banos[2][0000−0001−5434−4253], and Claudia Villalonga[2][0000−0003−4253−7909]

[1] iungo Education SAS, Bogota, Colombia.
pablo.salgado@iungoeducation.com
[2] Research Centre for Information and Communications Technology,
University of Granada, Granada, Spain.
{oresti,cvillalonga}@ugr.es

**Abstract.** People with autism spectrum disorder (ASD) are known to show difficulties in the interpretation of human conversational facial expressions. With the recent advent of artificial intelligence, and more specifically, deep learning techniques, new possibilities arise in this context to support people with autism in the recognition of such expressions. This work aims at developing a deep neural network model capable of recognizing conversational facial expressions which are prone to misinterpretation in ASD. To that end, a publicly available dataset of conversational facial expressions is used to train various CNN-LSTM architectures. Training results are promising; however, the model shows limited generalization. Therefore, better conversational facial expressions datasets are required before attempting to build a full-fledged ASD-oriented support system.

**Keywords:** Autism · AI · Deep learning · Emotions · Facial expression.

## 1 Introduction

The last decade has witnessed an explosive growth in the field of deep learning in AI. Since the presentation of AlexNet [1] in 2012, a convolutional neural network (CNN) developed for image recognition, deep learning has achieved an impressive record in cognitive tasks. One such task is human facial expression recognition, with a prominent relevance in the affective computing area. In other application areas of deep learning, such as natural language processing, automatic translation, or speech recognition, recurrent neural networks (RNN) have been proven to learn how to solve complex problems requiring sequences of data over time. Soon enough it was envisioned that the combination of these architectures, i.e., CNN-RNN architectures, such as convolutional neural network - long-short term memory (CNN-LSTM) can learn from video sequences, for example to remove commercials or detect human activity.

The success in applying such AI techniques in complex cognitive tasks poses the question of whether AI can also assist people with ASD to improve their quality of life. One main challenge for people with autism is to be able to interpret social cues that humans normally use in their daily nonverbal communication. More specifically, facial expressions most often convey essential information to interlocutors which is usually misinterpreted by people with autism. Hence, having the capacity of decoding such information for them is found of much relevance to facilitate the communication among people with and without autism.

In the light of this issue, this work aims to develop a deep neural network model, namely a CNN-LSTM, to classify conversational facial expressions in order to support people with ASD. As it is usual in AI research, data must be collected and preprocessed before it is fed to AI algorithms for learning. Searching a suitable dataset was therefore the first step. Before feature extraction, a preprocessing of the dataset was required. For the facial feature extraction, some pretrained CNNs were used. In the final step of classification, a LSTM was used. After ten trials with different CNN-LSTM settings, it was found that available publicly datasets seem to be insufficient to solve the problem at hand. Lack of enough data for training and testing results in difficulties in finding appropriate settings to improve the proposed deep learning models.

## 2 State of the art

There exist some previous works that have applied AI techniques to detect human facial expressions in ASD. At least two projects are active by the time this document is written. The first project is a commercial one, known as Brain Power™ [2]. Among other provided applications, this project offers Emotion Charades™ [3], which uses deep learning for human emotion recognition. The mobile app is designed to run on Google Smart Glasses as an augmented reality game where children with ASD try to guess the emotion on faces of surrounding people. If the child interprets emotions correctly, she is awarded with points or encouraged to try again otherwise. This system can only recognize six basic human emotions: sadness, happiness, fear, anger, surprise and disgust.

The second project is driven by the University of Stanford in California. The solution, known as The Autism Glass Project [5], consists of an Android app that works with a pair of Google Smart Glasses™which in turn feeds the video to the mobile app for deep learning emotion recognition. In response, the detected emotion is sent to the glasses, which informs the child of the detected emotion. Additionally, the child is encouraged to interpret captured faces in the app with an emotion. Several papers [6–10] have been published about this project in which the improvements experienced by children using the smart glasses and the app are reported. This system can only recognize seven basic human emotions: happiness, surprise, anger, disgust, sadness, fear and indifference.

There exist commercial cognitive systems designed to recognize human facial emotions not necessarily related to ASD. Table 1 presents a non-comprehensive list of the available systems, where the common denominator is that all these systems can recognize a few human basic emotions. Apart from that, several image datasets are publicly available that can be used for model training in order to recognize human facial emotions in pictures. As it is observed with available commercial cognitive systems, these datasets are restricted to a few human emotions. In [11] three potential datasets for basic emotion recognition training can be found: JAFFE, UMD and Cohn-Kanade. JAFFE and UMD datasets provide pictures for basic emotions (happiness, surprise, disgust, anger, sadness and fear). Cohn-Kanade, instead, provides combinations of facial action units [12] of those basic emotions to build some complex facial expressions.

**Table 1.** Commercial Human Emotion Recognition Systems.

| Provider | Web Site | Emotions |
|---|---|---|
| Eyesee | https://eyesee-research.com/facial-coding/ | Happiness, surprise, confusion, disgust, fear, sadness, neutral. |
| Emotion Research Lab | https://emotionresearchlab.com/online-platform/ | Happiness, surprise, anger, disgust, fear, sadness, neutral. |
| iMotions | https://imotions.com/biosensor/fea-facial-expression-analysis/ | Happiness, anger, fear, disgust, contempt, sadness, surprise. |
| Kairos | https://www.kairos.com/ | Anger, disgust, fear, happiness, sadness, surprise. |
| Microsoft Azure | https://azure.microsoft.com/en-us/services/cognitive-services/face/ | Anger, contempt, disgust, fear, happiness, neutral, sadness, surprise. |
| MoodMe | https://www.mood-me.com/insights/ | Happiness, surprise, sadness, anger, fear, disgust. |
| Noldus | https://www.noldus.com/facereader | Happiness, sadness, anger, surprise, fear, disgust. |
| NVISO | https://www.nviso.ai/en | Happiness, surprise, sadness, disgust, fear, anger, neutral. |
| Realeyes | https://www.realeyesit.com/technology/emotion-recognition/ | Happiness, surprise, confusion, sadness, disgust, fear. |
| RefineAI | https://www.refineai.com/ | Happiness, surprise, sadness, fear, anger, disgust, contempt. |
| Sightcorp | https://sightcorp.com/ | Happiness, surprise, sadness, disgust, anger, fear. |

Since datasets like JAFFE and Cohn-Kanade, where emotions are posed by actors in laboratory-controlled environments, the AffectNet dataset has been collected from internet with 1,000,000 images [13]. Again, this dataset only tags basic human emotions: happiness, sadness, surprise, fear, contempt and uncertainty. FER-2013 [14] is a very well-known dataset for facial analysis but, like the others, is limited to the basic emotions: anger, disgust, fear, happiness, sadness and surprise. Regarding video datasets, the AFEW [15] has been collected from movies with near real conditions but only for six basic human emotions: anger, disgust, fear, happiness, sadness and surprise.

Conversely to the above datasets, the Large MPI Facial Expression dataset [16] comprises 51 facial expressions commonly used by people in conversations. This dataset consists of 510 videos (10 videos per facial expression) posed by ten non-professional actors and actresses and delivered as tagged and numbered image sequences totaling 88823 photographs. The large array of facial expressions covered in this dataset makes it particularly representative of potential affective social situations that people with ASD may encounter regularly. Yet, as for most datasets, it is important to account for bias. The MPI dataset is posed by German males and females in their 20's. According to [17] facial expressions are subject to cultural differences even for the most basic and universal human emotions. The MPI dataset is hence biased, thus the systems developed out of this data might not be applied in a context different than German. Nevertheless, this type of bias can be a good characteristic since it may be expected that a person with ASD normally lives in a homogeneous cultural environment.

In regard to AI techniques for recognizing human facial expressions, the classic approximation is based on the facial action coding system (FACS) [12]. This system describes the human expression in terms of actions units (AU) in which the "geometry" of the face can be represented according to the relative position of the face muscles. Since the 80s, statistical analysis-based AI algorithms such as Gabor filters were used along with FACS to extract facial features from pictures. Statistical machine learning algorithms such as support vector machines, bayesian networks, or Markov hidden models were used to learn how to classify human emotions. By the year 2012, with the introduction of AlexNet [1], CNNs proved to excel in extracting features from images so efficiently that in the ICML 2013 [14] one workshop was dedicated to facial emotion interpretation. The top three teams, out of 56, attained accuracy scores of 71.162%, 69.256% and 68.821% using CNNs for classification of emotions in static pictures of human facial expressions.

Training a CNN is a costly process, so the transfer learning technique provides an excellent approach to transfer what a neural network has learned to solve a specific problem to a new, potentially unrelated domain. Several CNNs are pretrained with the ImageNet[3] dataset and are available for transfer learning in popular frameworks such as Keras[4]. Three of them (MobileNet, MobileNetV2 and NasNetMobile) will be used in this work applying the transfer learning technique in attempt to resolve the problem at hand. These CNNs are specifically designed for the limited resources of the mobile devices, hence the main reason for selecting them, as we intend to deploy the neural network on such devices. Although these CNNs have relatively few training parameters, for instance MobileNetV2 which only has 3504872, it can achieve an accuracy of 90.1 % in object classification [24].

[3] http://www.image-net.org

[4] https://keras.io/

Despite CNNs are the way to go for classification of static images, this project requires to analyze the temporal dimension of image sequences to find the pattern that composes a conversational facial expression. RNN are called to solve this type of problems. Specifically, the LSTM type of RNN is defined to remember start and end of a sequence. A combination of both is therefore quite practical for facial expression recognition problems. Some work has been done in this direction [20], where the authors achieved an accuracy of 88.02% in classifying six basic emotions (anger, disgust, fear, happy, sad and surprise) from the extended Cohn-Kanade dataset [21]. Also, in [18] the authors proposed the use of CNN-LSTM neural networks, attaining an accuracy of 41.67% while classifying six basic emotions (happy, sadness, anger, fear, surprise, disgust) available in the AFEW dataset [15].

## 3 Methodology

The overall objective of this work is to develop a classifier for the 51 facial expressions available in the MPI dataset, commonly used in conversations, as a base for a support cognitive artificial system for people with ASD. Ten different trials are performed in order to obtain the best neural network architecture. The first three trials are designed to find the most appropriate training mini-batch size, video sequence size and dataset using MobileNetV2 (the Keras mobile CNN with the fewest parameters). Trial four tries training MobileNetV2 without transfer learning. Trial five attempts to train MobileNet (the Keras mobile CNN with the fewest layers) while trial six attempts to train NasNetMobile (the Keras mobile CNN with the most layers). At this point, some conclusions are drawn to guide the last four trials. Trials seven, eight and nine test some RNN configurations based on LSTM using the best CNN model found in earlier trials. These first nine trials are conducted on models trained with just three (bored, confused, contempt) of the 51 classes available in the MPI dataset. This approach reduces the training time required while searching the best hyper-parameters for mini-batch size, sequence size, neural network deepness and LSTM units. Based on the conclusions of the first nine trials, the final CNN-RNN neural network to classify the 51 facial expressions is trained in Trial 10.

Since MPI dataset is relatively small, just 10 videos for each expression, a sliding window technique is used to generate multiple sequences from each video. Such a technique is shown in Fig. 1, where a given 11 frames video turns out in three sequences of three frames each. The sliding window algorithm receives as parameters the video itself and the desired sequence size. In response it generates all possible sequences of the given size from the video. This technique is applied across all neural networks trained in the ten trials to increase the quantity of videos sequences available.

**Trial 1**. A MobileNetV2 CNN is instantiated without the classification layer. The last convolutional layer is trained again, while all remaining layers are configured for non-training in order to apply the transfer learning technique. One LSTM layer is added, just before the new classification layer with 3 outputs. The

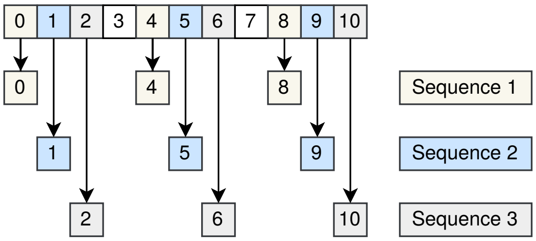

**Fig. 1.** Sliding window algorithm.

training data is not preprocessed, simply 30 videos are created from the photograph frames provided with the MPI dataset for the three selected classes. MPI dataset pictures are at a resolution of 768x576 so this will be the final resolution for the videos. Despite that, Keras pretrained networks receive a 224x224x3 tensor as input so each video is resized in runtime to match that requirement.

Fifteen models are trained by combining a mini-batch size from (2, 4, 8, 16, 32) and sequences size from (6, 12, 24). An early stopping technique is used to stop the training when validation loss completes ten epochs of continuous increasing. In order to establish the best model a linear regression is calculated for validation accuracy and loss. The slope of the line is used as an indicator of how fast the precision increases and the loss decreases. The model with the biggest slope for precision (0.0099) and the smallest slope for loss (0.0126) is considered the best model. This model is trained with a mini-batch size of 8 and a sequence size of 12. The model fails to converge consistently and is over-fitted as it is trained with just 312 sequences and validated with 78 sequences. So, next trial tries a Data Augmentation technique to increase the number of sequences for training.

**Trial 2**: This trial focus on Data Augmentation by preprocessing the same 30 videos of trial 1. The objective is to create 20 videos from each original video totaling 600 videos for training. No more changes are made from trial 1. For augmentation, 600 random transformations are generated making each video slightly different from the previous one. One transformation is created from nine parameters (rotation, displacement x and y, shear, x and y zoom, horizontal flip, brightness and grayscale) with values randomly selected from a given range. Once the transformation is instantiated it is applied to each frame of the original video and saved as a new video. Fig. 2 shows an example of some transformations applied to a video frame.

Fifteen models are trained by combining a mini-batch size from (2, 4, 8, 16, 32) and sequences size from (6, 12, 24). The model trained with a mini-batch size of 32 and a sequence size of 12 turns out to be the best model for this trial as the linear regression slope is 0.0169 for accuracy and it is 0.0056 for loss. The model converges consistently on the validation accuracy chart achieving more than 91.73% by the ninth epoch, but the model is over-fitted since loss starts to increase consistently starting at 0.5 from epoch seven. Anyway, the

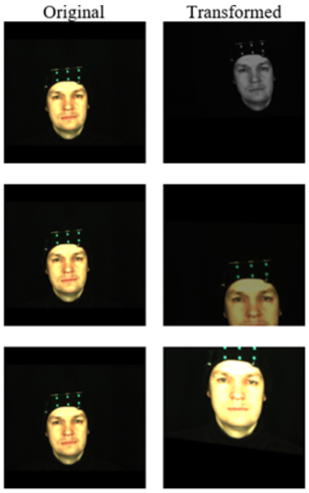

**Fig. 2.** Example of image transformations applied to a video frame.

Data Augmentation technique shows better results on training as this model was trained with 12000 sequences and validated with 3000 sequences. Next trial focuses in making a close-up of the face in each video.

**Trial 3**: This trial focuses in preprocessing the 30 videos from trial 1 creating 600 videos with a close-up of the face in the video. The same 600 transformations from trial 2 are applied to videos after the face close-up is preprocessed. No more changes are made from trial 2. Each video is scanned twice, the first scan gets the area where the face is located and the second one, cuts the face in all frames saving a new video. Fig. 3 shows an example resulting of a face frame close-up.

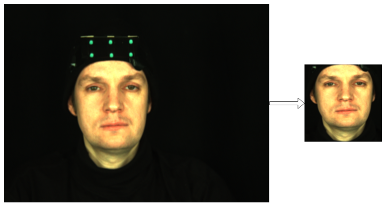

**Fig. 3.** Frame close-up.

Fifteen models are trained by combining a mini-batch size from (2, 4, 8, 16, 32) and sequences size from (6, 12, 24). The model trained with a mini-batch size of 32 and a sequence size of 6 turns out as the best model for this trial as the linear regression slope is 0.0247 for accuracy and it is 0.0391 for loss. This model failed to converge consistently on the validation accuracy despite it seems to show an upward trend without exceeding 80%. Anyway, the model is highly over-fitted. What these results seems to show is that a high-resolution picture of face close-up turns out in over-fitting the neural network. In this regard, Goodfellow et al. [14] have found that a 48x48 resolution is just enough for facial expression recognition on FER-2013 dataset. Same way, Cunningham et al. [19] have found that conversational facial expressions from MPI dataset are recognizable by humans at resolutions as lower as 64x48. Based in these studies and obtained results, the closed-up face dataset is abandoned in this work as well as the best model for this trial. As trial 3 is abandoned, the best model from trial 2 is by far the best model. So, for the following trials, the mini-batch size will be 32, sequence size will be 12 and the preprocessed augmented video dataset from trial 2 will be used for training. The MobileNetV2 was used in all these trials.

**Trial 4**: This trial focuses on fully training the CNN used in the first three trials (MobileNetV2). Same neural network architecture from trials 1, 2, and 3 are used. The mini-batch size and sequence size hyper-parameters are 32 and 12 respectively. The same 600 preprocessed videos from trial 2 are used to fully train the CNN, but they are resized in runtime to 48x48 (see [14] and [19]) since emotion recognition can be successfully made at this lower resolution. The model virtually learns nothing since the validation accuracy stays constantly at around 33%. The conclusion is that transfer learning is very useful for this project as there is not enough data to fully train the CNN portion of the neural network.

**Trial 5**: This trial focuses in training a neural network with a CNN portion deeper than MobileNetV2 which is 157 layers deep. The NasNetMobile is selected as it is the deepest mobile CNN available in Keras. Neural network architecture is the same from trial 4, but MobileNetV2 was switched for NasNetMobile. Hyper-parameters are the same as trial 4. Training dataset is the same from trial 2. It is observed that the validation accuracy hardly reaches 67% while loss indicates that the model is very over-tuned as it increases in value rapidly from 1.0 since epoch 1. The best model from trial 2 seems to be better than this one. So, the conclusion is that a deeper CNN network does not improve the training results for the problem at hand.

**Trial 6**: This trial, focuses on trying a less deep CNN portion. The selected CNN is MobileNet as this CNN is the less deep mobile network. This trial is conducted with same hyper-parameters and network architecture from trial 5 but NasNetMobile is switched to MobileNet. By far, the best model is from this trial as it is the only one with a negative slope for loss at -0.1491, while it has the highest value for the slope of accuracy 0.0352. The conclusion is that a less deep CNN is the most appropriate for the problem at hand. So, MobileNet is selected as the final CNN portion for the neural network.

**Trial 7**: Now that a CNN portion has been selected, this trial focuses on the RNN portion. The RNN is made deeper by adding three additional LSTM layers of 64 units. In order to control the over-fitting of the neural network, each LSTM layer is added between dropouts at 50%. Remaining architecture is the same as the trial 6 as well as hyper-parameters and training data. The model of this trial start to converge to a validation accuracy around 90% while loss shows a similar pattern of convergence below 1. The over-fitting of the network has been controlled and the model looks appropriate to solve the problem at hand.

**Trial 8**: This trial aims to evaluate if an RNN with more LSTM units outperforms the model from trial 7. The only change from trial 7 is that the four LSTM layers are expanded to 128 units. Expanding the LSTM layers with more units makes the neural network to lose its ability to converge and to start overfitting as accuracy varies widely from 37% and 94% over 33 training epochs. The loss behaves in a similar way varying between 0.2452 and 4.7848.

**Trial 9**: This trial focuses on training a bidirectional RNN. From trial 7 architecture, each LSTM 64 units layer is embedded into a bidirectional RNN. No more changes are performed. The neural network loose even more its ability to converge as accuracy varies widely from 50% and 95% over 43 training epochs and the loss behaves in a similar way varying between 0.2872 and 3.4150. So, trials 8 and 9 are abandoned and the architecture from the trial 7 is taken for the final trial.

**Trial 10**: In the first three trials it is found that the hyper-parameters of mini-batch size and sequence with the best results are 32 and 12 respectively, so is the augmented and preprocessed video dataset from trial two. From trial four, it is observed that transfer learning is more than appropriate for this project. From trials five and six, it is concluded that a network with fewer layers may be more effective, so the MobileNet CNN is selected. From trials seven, eight, and nine, it is found that the neural network from trial seven presents the best accuracy and loss curves, compared to experiment eight and nine, so the RNN portion from trial seven is selected.

This trial is intended to train the final neural network to classify the 51 classes from the MPI dataset. 51 videos (all from one actor) are held apart for testing. From each video of the remaining 459 videos, 20 videos are generated using the augmentation procedure from trial 2. Thus, the training dataset for this trial is composed of 9180 preprocessed and augmented videos. The sliding window algorithm finally produces 119340 sequences for training. The neural network architecture trained is the same as for trial 7.

## 4 Results and Discussion

The final neural network to classify the 51 classes from the MPI dataset obtained in Trial 10 is validated and tested. The results from training this model are show in Fig. 4. As seen in the figure, the neural network converges consistently and by the epoch 33 hits an 88.6% of accuracy with a loss of 0.954. Trained is stopped by epoch 40 since no improvement is shown for accuracy neither loss. Kaulard et al. [16], performed a validation of the MPI dataset with three human evaluators for each facial expression in the dataset. They obtained an accuracy of 60% when the three evaluators classify a video with same class. The neural network seems to be as good as humans to classify the MPI facial expressions.

Testing a neural network requires independent data that has not been used to train the model. The 51 videos held apart are used to perform this test. The

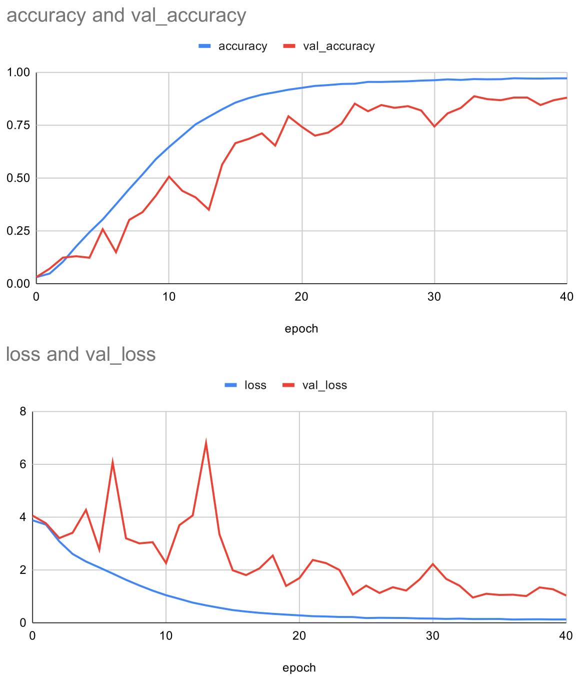

**Fig. 4.** Evaluation charts for the generated model (trial 10).

sliding window algorithm is used to generate the required sequences of size 12 frames. Finally, 663 test sequences are generated from the 51 videos. Once the network is evaluated just an accuracy of 7.97% is achieved with a loss greater than 10. These results suggest that the neural network is over-fitted despite the efforts made to avoid such a situation. It seems to be memorizing the faces of the actors/actresses of the MPI dataset. The results are not entirely unexpected as MPI dataset provides only 10 different faces and 51 classes. However, it is important to note that the final neural network achieves an accuracy of 7.97% which is slightly over four times the accuracy of 1.96% of a given random classifier for 51 classes.

Either way, the low accuracy achieved may be due, in part, to the inclusion of the neutral facial expression frames in the videos. In the same way, another issue may be that the data augmentation preprocessing produced some frames with the face displaced such that a slightly part of the face is missing. Another set of trials addressing these issues should be run in a future work to see if greater accuracy is achieved. Anyway, it seems that the main issue is the MPI dataset itself as this dataset has a limited number of people posing in videos. There is not enough data for the models to start generalizing, despite the augmentation of the videos and the sliding window technique.

## 5 Conclusions

The final CNN-LSTM neural network model developed in this project has the potential to develop an artificial cognitive system aimed to provide people with ASD with a tool that can improve their lives. The main issue to achieve such an artificial cognitive system is found in the lack of enough training data. There is plenty of room in this area to develop such datasets before to attempt to build this kind of systems.

Facial expression recognition is a complex cognitive task that even humans find difficult, and people with ASD may found impossible to perform. AI in general and artificial cognitive deep learning models have the potential to develop applications that somehow can improve the quality of life for this people. Beyond applications for people with ASD, further research in this field may find application in areas like human-robot interactions or facial sentiment analysis.

## References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton: "ImageNet Classification with Deep Convolutional Neural Networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, https://doi.org/10.1145/3065386
2. Brain Power — Autism Education - Empowering Every Brain, https://brain-power.com/. Last accessed Mar. 01, 2021
3. Learn to Play Emotion Charades, https://youtu.be/lGoxUd2nTDc. Last accessed Apr. 01, 2021
4. "Glass – Glass.", https://www.google.com/glass/start/. Last accessed Apr. 01, 2021
5. Autism Glass Project, http://autismglass.stanford.edu/. Last accessed Apr. 01, 2021
6. J. Daniels et al., "Feasibility Testing of a Wearable Behavioral Aid for Social Learning in Children with Autism," Appl. Clin. Inform., vol. 9, no. 1, pp. 129–140, Jan. 2018, https://doi.org/10.1055/s-0038-1626727
7. J. Daniels et al., "Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism," npj Digit. Med., vol. 1, no. 1, p. 32, 2018, https://doi.org/10.1038/s41746-018-0035-3
8. C. Voss et al., "Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder: A Randomized Clinical Trial," JAMA Pediatr., vol. 173, no. 5, pp. 446–454, 2019, https://doi.org/10.1001/jamapediatrics.2019.0285
9. C. Voss et al., "Superpower Glass: Delivering unobtrusive eal-Time social cues in wearable systems," in UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Sep. 2016, pp. 1218–1226, https://doi.org/10.1145/2968219.2968310
10. P. Washington et al., "SuperpowerGlass: A Wearable Aid for the At-Home Therapy of Children with Autism," Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol., vol. 1, no. 3, pp. 1–22, Sep. 2017, https://doi.org/10.1145/3130977
11. R. Gross, "Face Databases," in Handbook of Face Recognition, S. Z. Li and A. K. Jain, Eds. New York, NY: Springer, 2005, pp. 301–327.
12. P. Ekman et al., What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press, 1997.

13. A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,” IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, Jan. 2019, https://doi.org/10.1109/TAFFC.2017.2740923
14. I. J. Goodfellow et al., “Challenges in Representation Learning: A Report on Three Machine Learning Contests,” in Neural Information Processing, Berlin, Heidelberg, 2013, pp. 117–124, https://doi.org/10.1007/978-3-642-42051-1_16
15. A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, “Emotion Recognition in the Wild Challenge 2013,” in Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 509–516, https://doi.org/10.1145/2522848.2531739
16. K. Kaulard, D. W. Cunningham, H. H. Bülthoff, and C. Wallraven, “The MPI Facial Expression Database — A Validated Database of Emotional and Conversational Facial Expressions,” PLoS One, vol. 7, no. 3, p. e32321, Mar. 2012, https://doi.org/10.1371/journal.pone.0032321
17. H. A. Elfenbein, M. Beaupré, M. Lévesque, and U. Hess, “Toward a Dialect Theory: Cultural Differences in the Expression and Recognition of Posed Facial Expressions,” psycnet.apa.org, 2007, https://doi.org/10.1037/1528-3542.7.1.131
18. Y. Li, “Deep Learning of Human Emotion Recognition in Videos”, https://uu.diva-portal.org/smash/get/diva2:1174434/FULLTEXT01.pdf. Last accessed Mar. 01, 2021
19. D. W. Cunningham, M. Nusseck, C. Wallraven, and H. H. Bülthoff, “The role of image size in the recognition of conversational facial expressions,” Computer Animation and Virtual Worlds, vol. 15, no. 3–4, pp. 305–310, 2004, https://doi.org/10.1002/cav.33
20. S. Rajan, P. Chenniappan, S. Devaraj, and N. Madian, “Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM,” IET Image Processing, vol. 14, no. 7, pp. 1373–1381, Feb. 2020, doi: 10.1049/iet-ipr.2019.1188.
21. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.
22. P. Shaver, J. Schwartz, D. Kirson, and C. O’Connor, “Emotion knowledge: Further exploration of a prototype approach.,” Journal of Personality and Social Psychology, vol. 52, no. 6, pp. 1061–1086, 1987, doi: 10.1037/0022-3514.52.6.1061.
23. C. Pelachaud and I. Poggi, “Subtleties of facial expressions in embodied agents,” The Journal of Visualization and Computer Animation, vol. 13, no. 5, pp. 301–312, Dec. 2002, doi: 10.1002/vis.299.
24. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.