Automatic Detection of Gaze and Smile in Children's Video Calls

Dhia-Elhak Goumri, Thomas Janssoone, Leonor Becerra-Bonache, Abdellah Fourtassi

Automatic Detection of Gaze and Smile in Children's Video Calls

Dhia-Elhak Goumri
dhia-elhak.goumri@univ-amu.fr
Aix Marseille Univ, CNRS, LIS
Marseille, France

Thomas Janssoone
thomas@enchanted.tools
Enchanted Tools
Paris, France

Leonor Becerra-Bonache
leonor.becerra@lis-lab.fr
Aix Marseille Univ, CNRS, LIS
Marseille, France

Abdellah Fourtassi
abdellah.fourtassi@gmail.com
Aix Marseille Univ, CNRS, LIS
Marseille, France

ABSTRACT

With the increasing use of video chats by children, the need for tools that facilitate the scientific study of their communicative behavior becomes more pressing. This paper investigates the automatic detection – from video calls – of two major signals in children's social coordination: smiles and gaze. While there has been significant advancement in the field of computer vision to model such signals, very little work has been done to put these techniques to the test in the noisy, variable context of video calls, and even fewer studies (if any) have investigated children's video calls specifically. In this paper, we provide a first exploration into this question, testing and comparing two modeling approaches: a) a feature-based approach that relies on state-of-the-art software like OpenFace for feature extraction, and b) an end-to-end approach where models are directly optimized to classify the behavior of interest from raw data. We found that using features generated by OpenFace provides a better solution in the case of smiles, whereas using simple end-to-end architectures proved to be much more helpful in the case of looking behavior. A broader goal of this preliminary work is to provide the basis for a public, comprehensive toolkit for the automatic processing of children's communicative signals from video chat, facilitating research in children's online multimodal interaction.

CCS CONCEPTS

• **Human-centered computing** → *Interaction design*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

smile, gaze, child-caregiver interactions, video chats, machine learning

ACM Reference Format:
Dhia-Elhak Goumri, Thomas Janssoone, Leonor Becerra-Bonache, and Abdellah Fourtassi. 2023. Automatic Detection of Gaze and Smile in Children's Video Calls. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3610661.3616241

ICMI '23 Companion, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0321-8/23/10…$15.00
https://doi.org/10.1145/3610661.3616241

1 INTRODUCTION

The role of online video calls in children's daily lives has significantly risen in the last few years following the COVID-19 pandemic. Understanding how children communicate via this new communicative medium is, therefore, critical to optimizing its use for children's social interactions and education. That said, studies about children's communication via video calls are still scarce: Although it is a priori not very challenging to collect children's video call data from home [5], there is a lack of cost-effective tools that allow researchers in child development to process such variable and naturalistic data, especially for multimodal communicative behavior.

Data annotation by experts is – and will remain – essential. Still, it is too resource-intensive, requiring specific training and familiarity with coding schemes that are often relatively complex in naturalistic social interaction. Thus, hand annotation, alone, makes it difficult to study a large sample that would adequately capture this phenomenon's high contextual and cultural variability. Automatizing the annotation process (at least partly) can speed up the research process and facilitate systematic, large-scale studies of children's computer-mediated communication, allowing researchers to draw robust scientific conclusions and policy-makers to make informed decisions about the pros and cons of using this technology in childhood.

Here, we focus on automatizing the detection of two signals: smile and gazing behavior, both known to be predominant in adult social and conversational coordination [12, 18], with precursors observed in infants' early interactions [8, 15, 20, 34]. Their development from infancy to adulthood is not well-understood, and how children perceive/produce them in video-call conversations is still largely unknown. For smiles, we are interested in the ability of models to reliably classify children's facial expressions into smiling vs. not-smiling. As for gazing behavior, we are interested in classification into "gazing at interlocutor" vs. "averting gaze," a signal that plays an important role in regulating multimodal conversational dynamics in adults and children [2, 21, 23]. In the context of video calls, gazing at an interlocutor is roughly operationalized as "looking at the screen," whereas gaze aversion is operationalized as "looking away from the screen."[1]

[1]Note that "looking at the screen" vs. "looking away from the screen" does not map exactly to gaze vs. gaze aversion in face-to-face conversations, mainly because the

We focus on middle childhood data, that is, school-age children (i.e., 6 to 12 years old) since this is the age when children start to be able to use this medium of communication without heavy supervision from adults to help them stay focused and engaged. This is also the age when video calls become an option for distance learning. Finally, from a cognitive point of view, school-age children witness significant developments in many socio-cognitive competencies (e.g., executive functions and theory of mind) that are understood to underlay coordinated communications[25]. The study of the extent to which these early competencies are hindered or facilitated by computer-mediated (as opposed to face-to-face communication) is still an open but pressing topic of research.

Our goal here is to provide a first investigation into the possibility of automatizing the detection of smiling and gazing behaviors in video calls involving school-age children. Using a corpus of child-caregiver conversations via Zoom, we tested and compared two main modeling approaches. The first approach is *Feature-based*: Following common practice in the fields of human-human and human-computer interaction, we first use the "OpenFace" library [3] to derive the relevant (continuous) features (e.g., Action Units values [13] and gaze/head direction coordinates [32]). Second, we train a binary classifier to learn mapping these continuous features to target categories (e.g., AU to detect smiles, and gaze coordinates + head pose estimates to detect gazing patterns).

The second approach is *End-to-End*, using Convolutional Neural Networks (CNN) to learn a direct mapping from images to the categories of interest, e.g., smiling vs. non-smiling or looking at screen vs. looking away. We follow state-of-the-art models in the automatic detection of facial expressions by using CNNs [6, 7, 9, 14, 16, 17, 29–31, 35]. CNN-based architecture hierarchically extracts features: the lower layers extract low-level features such as edges and lines; whereas higher layers can learn higher-level features (using features from the lower layers) to perform the classification.

While very large CNN models such as deep residual networks with dozens of layers (i.e., ResNet-50) have been used recently for various aspects of facial expression recognition ([7, 22]), such models typically require large datasets and computing resources. Here, we chose to use a simpler CNN architecture (see Figure 1) which provides a good compromise between practical use (given our relatively small data) and effectiveness. Indeed, an architecture of similar nature/size was shown to be enough to achieve near-perfect scores for the specific tasks we are interested in, that is, smile detection [6] and gazing behavior classification [14].

2 NOVELTY AND RELATED WORK

The current study is – to our knowledge – the first investigation of the automatic detection of smiling and gazing behaviors in children's video calls. Most current models are trained and/or tested exclusively (or disproportionately) on adult data. In addition, video call data is in itself a highly unconstrained context, with a significant degree of between-subject variability in terms of lighting, head positioning, the relative location of the person to the screen, the size of the screen, and webcam setup. These unique challenges require a thorough investigation and, perhaps, the development of new models.

We can find related work on the broad question of the automatic detection of smiling and gazing behaviors. Regarding smile detection, research has mainly focused on adults [9, 17, 29, 30, 35], with very few exceptions such as [33] who, however, focused on a much younger age (infants). Regarding the automatic coding of looking behavior (that is, gazing at interlocutor vs. gaze aversion), we found no directly related work. While most existing literature in computer vision focuses on deriving measures for gaze from a camera (for review, see [7, 16]), our task depends not only on the ability to estimate gaze but *also* the *relative* position of a third object (i.e., screen), especially in a completely unconstrained context where such relative location varies between subjects and is hard to estimate a priori. The closest work to our goal that we could find is the one by [14], who built a model of infants' looking behavior (looking at left vs. right stimuli). They, however, investigated gaze data in a much younger age (infants aged one year or younger) using a semi-controlled experimental setup that reduced variability between subjects.

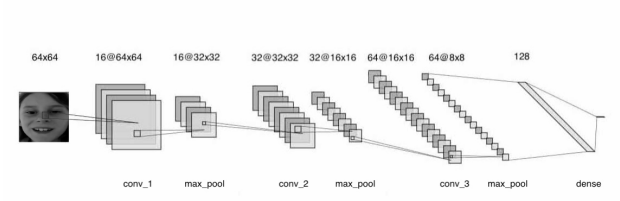

Figure 1: The CNN architecture we used for both smile and gaze classification. It takes as input the pre-processed grayscaled images of 64-by-64 pixels and passes them to three convolutional layers (conv_1, conv_2 and conv_3) followed by a max pooling layer each (max_pool). Information is then passed through a fully connected layer of size 128 (dense) to the output layer, ending with a sigmoid activation function to classify the image in a binary fashion.

3 METHODS

3.1 Data and pre-processing

3.1.1 Children's Video Call Data. For our experiments, we use videos from the ChiCo corpus [4]. It consists of 8 video call recordings at home (using Zoom software) between children (aged 6 to 12 years old) interacting with their caregivers. To elicit a balanced exchange between children and caregivers, the conversation involved an intuitive and weakly structured game where interlocutors guessed each other's words for 10 to 15 minutes. The setup required that children and caregivers use different computers and communicate from different rooms (if they record from the same house). The corpus was manually annotated for several facial expressions. Here we use the annotation for smile (smile vs. no-smile) and gaze (looking at screen vs. looking away). For further details about the corpus and the procedure for manual annotation and

position of the webcam is not aligned with the face. However, this is a constraint inherent to video calls that people have to adapt to – more specifically, to the fact that when an interlocutor is looking at them (as they appear on the interlocutor's screen), the interlocutor's gaze will appear on their screen as slightly misaligned because of the webcam angle.

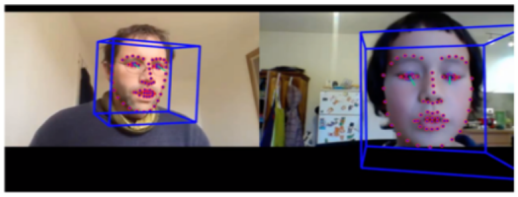

Figure 2: A snapshot from a child-parent zoom call. The figure also shows the OpenFace features we used in the current work: AUs for smile detection (illustrated in red) and eye gaze angle/direction (illustrated with the straight line in green).

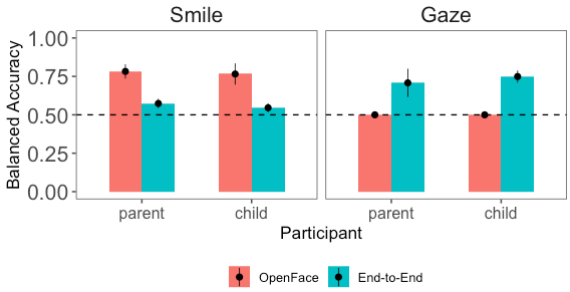

Figure 3: The (balanced) accuracy score for Smile and Gaze comparing Feature-based models to End-to-End models. Ranges indicate 95% confidence interval over 8 unique participants in ChiCo corpus, both parents and children. Scores for each individual were obtained by models that have not seen this individual in training, indicating the ability to generalize to unseen videos. The dashed lines represent chance level.

inter-annotation agreements, we refer the reader to the original paper [4].

3.1.2 Pre-processing. ChiCo dataset was pre-processed for face detection using RetinaFace [11]. We picked RetinaFace as it was reported by the authors to achieve state-of-the-art results in accurately detecting faces even in sub-optimal conditions with low lighting and contrast. Faces were detected frame by frame, extracted from the background, grayscaled, and resized to 64-by-64 pixels.[2] We used these "cropped" faces as input data for all the models investigated in this paper.

3.2 Smile detection

The common procedure is to train and test on data acquired under the same conditions. However, because of severe data imbalance in ChiCo (we had far fewer frames with a smiling face — according to manual annotation – than frames with a non-smiling face; see Table 1), we decided to perform two additional experiments involving transfer learning with or without fine-tuning:

Experiment 1: standard train-test. We train and test on ChiCo data.

Experiment 2: training on public data with direct testing. First, we trained the model on a publicly available dataset with more balanced and varied training examples called SMILE [19]. Second, we tested this trained model directly on ChiCo.

Experiment 3: fine-tuning. First, we trained the model on SMILE. Second, we fine-tuned it on videos from ChiCO. Finally, we test the fine-tuned models on videos from ChiCo (unseen during fine-tuning, see results section for details about evaluation).

The SMILE dataset consists of 13,165 faces, labeled as smiling (3,690 examples, or 28% of the data) or not smiling (9476 images, i.e., 72% of the data). The images were already cropped, grayscaled, and resized to a resolution of 64-by-64 pixels.

3.2.1 Feature-based model. We used OpenFace and the FEA tool [3] to extract AUs from faces in both SMILE (training data) and ChiCo (test data). OpenFace could only recognize and output interpretable AU values for about two-thirds of the faces in SMILE. That said, the resulting data was still fairly balanced (25% of smiles). As for ChiCo, OpenFace processed the videos frame by frame (where faces were already cropped, see pre-processing step above). It succeeded in processing 97% of the faces (Table 1).[4]

After extraction of AUs, we trained a Support Vector Machine (SVM) on either ChiCo (experiment 1) or on SMILES (Experiment 2),[5] and then classified images in ChiCo's test set into smile or no-smile using their AUs.

3.2.2 End-to-end model. The end-to-end approach classifies directly from the data (whereas the Openface-based model involves AUs extraction first and classification second). We used a CNN architecture (see Figure 1) that takes as input the pre-processed grayscaled images of 64-by-64 pixels and passes them to three convolutional layers followed, each, by a max pooling layer. Information is then passed through a fully connected layer of size 128 to the output layer, ending with a sigmoid activation function to classify the image as either containing a smile or not. The CNN was trained for 20 epochs on batches of 64 images each with a learning rate of 0.001. The Adam optimizer was used to optimize the binary cross-entropy loss.[6]

The CNN was either trained and tested on ChiCo (Experiment 1) or via transfer learning by first being trained on SMILE data and then tested on ChiCo data with or without fine-tuning (Experiments 2 and 3).

3.3 Gazing behavior detection

For this task, the goal is to determine whether or not a child (or parent) is looking at the screen (a proxy for looking at the interlocutor) vs. looking away (a proxy for gaze aversion). Unlike the case of smiles, we could not find an available dataset annotated for this specific looking behavior. Therefore, we used ChiCo both for

[2]Note that this is the best resolution we could do for the cropped faces, given that an entire frame from zoom videos (i.e., including the background) has a resolution of 1280 by 720.
[3]https://gitlab.com/Thom/fea_tool
[4]The reason OpenFace was able to recognize much fewer faces in SMILE than in ChiCo is perhaps since faces in SMILE were too tightly cropped, sometimes hiding some key features of the face.
[5]Fine-tuning, that is, Experiment 3, was only done with the CNN-based approach.
[6]We trained the models on an M2 chip with an 8-core GPU, 8-core CPU, and 16GB of memory. We used Python 3.8 and TensorFlow 2.6 to implement our models. Training and inference for all models are done within an eight hours limit.

	Original	Retinaface	OpenFace
smile	51909 (12.00%)	51418 (12.02%)	51324 (12.28%)
no-smile	380523 (88.00%)	376225 (87.98%)	366574 (87.72%)
gaze	314297 (72.68%)	312173 (73.00%)	304254 (72.81%)
no-gaze	118135 (27.32%)	115470 (27.00%)	113644 (27.19%)
total	432432	427643	417898

Table 1: Number of frames of all videos in the ChiCo corpus broken down by the smile condition or by the gaze condition (original). The table also shows the remaining frames which were successfully processed with Retinaface and the final distributional that was successfully processed/recognized with OpenFace.

training and testing (no transfer learning), especially since gaze data were less biased (compared to smiles, see Table 1).

3.3.1 Feature-based model. We first applied OpenFace to derive coordinates of gaze for both eyes in addition to head pose estimate, both can play a role in detecting gaze aversion (see Figure 2). Next, we trained an SVM classifier that takes in these features and predicts whether or not the person is looking at their screen.

3.3.2 End-to-end model. We used a similar CNN architecture to classify gaze as we did for smiles (Figure 1). We also trained the CNN for 20 epochs, using batches of size 64 with a learning rate of 0.001.

4 RESULTS

The results for both smiles and gaze are represented synthetically in Figure 3, using balanced accuracy,[7] which provides the most intuitive outcome given imbalanced data in ChiCo – unlike other measures like accuracy and F-score. The latter can sometimes be inflated with imbalanced data and can provide misleading scores when comparing across tasks with different data distributions (here, smile vs. gaze, see Table 1).

All scores we show in Figure 3 reflect the ability of the models, not only to learn but also to *generalize* to novel participants that were not seen by the models during training or fine-tuning.

Smile

We found that Experiments 1 and 3 underperformed. Only Experiment 2 (trained on SMILE and directly tested on ChiCo) led to noticeable scores for both feature-based and end-to-end methods. The results of this experiment are shown in Figure 3 (left). It may seem counter-intuitive that providing the models with zoom data input in training or fine-tuning hurts performance. However, this is because a diverse dataset like SMILE allows for better mapping of AUs to smiles across thousands of faces allowing a better generalization to unseen people. It is no wonder that adding zoom videos – made of a large number of frames, the majority of which is redundant – does not help, or even hurt performance when the model is tested on unseen videos.

[7] In our (binary) case, balanced accuracy is defined as the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate), or the area under the ROC curve with binary predictions.

The Feature-based model performs relatively well on all test videos. The End-to-End model, however, performs poorly. The low performance of the End-to-End model is interesting: When using the same CNN architecture to both train *and* test on the SMILE dataset (using an 80/20 training/test split), the CNN predicted smiles with high (balanced) accuracy: 91.68%. This high performance, however, did not transfer to the novel domain/context of Zoom videos for either children or adults.

Gazing behavior

The results for Gazing behavior (Figure 3, right) showed a reversed pattern compared to smiles: The Feature-based model performed exactly at chance, while the End-to-End model performed relatively well for both children and adults. Investigation of the Feature-based models showed that the SVM classified all frames as looking at screen. We tried to improve the OpenFace-based model, for example, by using another classifier (e.g., Multilayer perceptron or MLP) instead of the SVM or by selecting an equal number in each category for the model to see in the training phase. None of these methods led to a noticeable increase in the Feature-based model's performance.

Reproducing Results of Hand Annotation

Here we illustrate how our automatic coding translates into specific research questions. We investigate a) the proportion of time children/adults smile in video chats, and b) the proportion of time they avert their gaze (e.g., look away). We compared the answer to these questions using i) hand-coded video frames, and ii) automatically coded frames predicted by the best model.[8] The results in Figure 4, broken down by dyads (1 to 8), show variability in the quality of automatically-coded data, i.e., in their ability to mimic the hand-coded one. This variability correlated with the accuracy of the annotation: The videos for which there was the highest mismatch between hand and automatic coding are also the ones with the lowest model accuracy (and vice versa). That said, despite imperfect accuracy scores (Figure 3), the automatic coding allowed – when data is aggregated – (see 'ALL' in Figure 4) to derive similar overall conclusions about the relative frequency of gaze aversion and smiles in video chats (although the numbers were overestimated for the smiles).

5 DISCUSSION

This paper provided a preliminary investigation into the possibility of automatizing the detection of multimodal signals in children's video calls. We focused on two behaviors: smiling and gazing, not only because they are two of the most important signals in face-to-face conversations, but also because both can be approached with similar modeling techniques: They can be detected from static images (or isolated frames in a video), in contrast to, e.g., "head nods," which is also a crucial interactive signal but requires a dynamic representation of the data.

To address this question, we examined two modeling approaches: a) A feature-based approach that relies primarily on extracting features, followed by classification into the categories of interest, b) An end-to-end approach that consists in using Neural Network

[8] that is, the Feature-based model in the first question and the End-to-end mode in the second.

Figure 4: The proportion of frames in a video where children (top) or adults (bottom) were smiling (right) or/and averting gaze (left). We show the numbers obtained on each of the 8 dyads in ChiCo and the average ('ALL'), using both hand vs. automatically coded frames.

architecture, trained to optimize the categories of interest directly from the images or video frames.

The main findings of the study were as follows. Using the AUs generated by OpenFace provided a better solution than the end-to-end models we tested for smile detection. In the case of gazing behavior, using the OpenFace's features (i.e., eye-gaze coordinate and head pose estimate) appeared to be of little use, and using simple end-to-end architectures proved to be much more helpful in classifying videos frames into looking at screen vs. looking away.

More generally, there was a difference between, on the one hand, a communicative behavior like a smile whose detection depends only on the facial features of the communicator and, on the one hand, the behavior of gazing at someone or something, which depends not only on the gaze estimates of the person doing the gazing but also on the *relative* location of the target object of gaze (here, the computer screen). In the first case, using relevant features (i.e., AUs) instead of all the face (as in end-to-end) lowers the problem's dimensionality, which leads to better performance.

In the second case (i.e., gazing behavior), the fact that there is a large between-subject variability in how participants position themselves with respect to the webcam – in addition to hardware differences (e.g., the size and angle of the screen) – would lead to large variability in the values of gaze or head pose estimates that correspond to looking at the screen, causing the classifier to fail in finding one correct mapping. In this case, using raw data allows an end-to-end model to discover – by itself – cues that are most useful to the task.

When using the models to answer specific research questions (Figure 4), we found that automatic coding can diverge from the gold standard at the individual level, but it tends to converge on reasonable scores when the videos are aggregated (especially the one for gazing behavior). This suggests that, at least given their current abilities, our models can be used to draw broad conclusions from large-scale data, but cannot yet be reliably used to draw precise conclusions about individual videos.

In order to achieve higher accuracy scores, the models can be improved in several ways. We found the feature-based method with AUs to be the most promising for the smiling behavior, especially when combined with transfer learning. This can be further optimized by increasing and diversifying the training using other datasets for smiles such as GENKI-4K, or even by curating new datasets. As for gazing behavior, the end-to-end method is more promising, and it can be further improved by increasing the size of the annotated data of video calls. In both cases, as data size increases, we can start using more sophisticated CNNs (e.g., deep residual networks) to make the most out of the data. Finally, the fact that we obtained similar scores for adults and children suggests that using training data from adults' video calls (which are easier to collect) can help with the detection of children's communicative signals as well.

Finally, in addition to insights we gained from the above modeling experiments, one broader goal of this work is to provide the basis of a collaborative toolkit for the automatic detection of children's communicative signals in video chat, thus facilitating and speeding up research on children's online multimodal interactions. While the current work focused on smiles and gaze (the code of which will be made available to the community to use/optimize), in future work we will tackle other important signals like head nods/shakes [28], interactive alignment [26], prosody [10], Backchannel [5], laughter [24], dialog acts [27], and contingency [1].

ACKNOWLEDGMENTS

This research was supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), ANR-21-CE28-0005-01 (MACOMIC). The work of Leonor Becerra-Bonache has been performed during her teaching leave granted by the CNRS (French National Center for Scientific Research) at the Laboratoire Parole et Langage of Aix-Marseille University.

REFERENCES

[1] Kirsten Abbot-Smith, Julie Dockrell, Alexandra Sturrock, Danielle Matthews, and Charlotte Wilson. 2023. Topic maintenance in social conversation: What children need to learn and evidence this can be taught. *First Language* (2023), 01427237231172652.

[2] Abhishek Agrawal, Jing Liu, Kübra Bodur, Benoit Favre, and Abdellah Fourtassi. 2023. Development of Multimodal Turn Coordination in Conversations: Evidence for Adult-like behavior in Middle Childhood. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018).* 59–66. https://doi.org/10.1109/FG.2018.00019

[4] Kübra Bodur, Mitja Nikolaus, Fatima Kassim, Laurent Prévot, and Abdellah Fourtassi. 2021. ChiCo: A Multimodal Corpus for the Study of Child Conversation. In *Companion Publication of the 2021 International Conference on Multimodal Interaction* (Montreal, QC, Canada) *(ICMI '21 Companion).* Association for Computing Machinery, New York, NY, USA, 158–163. https://doi.org/10.1145/3461615.3485399

[5] Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. 2023. Using video calls to study children's conversational development: The case of backchannel signaling. *Frontiers in Computer Science* 5 (2023).

[6] Junkai Chen, Qihao Ou, Zheru Chi, and Hong Fu. 2017. Smile detection in the wild with deep convolutional neural networks. *Machine vision and applications* 28 (2017), 173–183.

[7] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. 2021. Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark. arXiv:2104.12668 [cs.CV]

[8] Jeffrey F Cohn and Edward Z Tronick. 1987. Mother–infant face-to-face interaction: The sequence of dyadic states at 3, 6, and 9 months. *Developmental psychology* 23, 1 (1987), 68.

[9] Zhoucong Cui, Shuo Zhang, Jiani Hu, and Weihong Deng. 2014. Evaluation of Smile Detection Methods with Images in Real-World Scenarios. In *ACCV Workshops*.

[10] Maureen de Seyssel, Marvin Lavechin, Hadrien Titeux, Arthur Thomas, Gwendal Virlet, Andrea Santos Revilla, Guillaume Wisniewski, Bogdan Ludusan, and Emmanuel Dupoux. 2023. ProsAudit, a prosodic benchmark for self-supervised speech models. *arXiv preprint arXiv:2302.12057* (2023).

[11] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. arXiv:1905.00641 [cs.CV]

[12] Starkey Duncan and Donald W Fiske. 2015. *Face-to-face interaction: Research, methods, and theory*. Routledge.

[13] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[14] Yotam Erel, Christine E. Potter, Sagi Jaffe-Dax, Casey Lew-Williams, and Amit H. Bermano. 2022. iCatcher: A neural network approach for automated coding of young children's eye movements. *Infancy* 27, 4 (2022), 765–779. https://doi.org/10.1111/infa.12468 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/infa.12468

[15] Alan Fogel, Sueko Toda, and Masatoshi Kawai. 1988. Mother-infant face-to-face interaction in Japan and the United States: A laboratory comparison using 3-month-old infants. *Developmental Psychology* 24, 3 (1988), 398.

[16] Shreya Ghosh, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, and Qiang Ji. 2022. Automatic Gaze Analysis: A Survey of Deep Learning based Approaches. arXiv:2108.05479 [cs.CV]

[17] Xin Guo, Luisa Polania, and Kenneth Barner. 2018. Smile Detection in the Wild Based on Transfer Learning. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 679–686. https://doi.org/10.1109/FG.2018.00107

[18] Antonia F de C. Hamilton and Judith Holler. 2023. Face2face: advancing the science of social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences* 378, 1875 (2023), 20210470. https://doi.org/10.1098/rstb.2021.0470 arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2021.0470

[19] Daniel D. Hromada, Charles Tijus, S. Poitrenaud, and Jacqueline Nadel. 2010. Zygomatic Smile Detection: The Semi-Supervised Haar Training of a Fast and Frugal System: A Gift to OpenCV Community. In *2010 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*. 1–5. https://doi.org/10.1109/RIVF.2010.5633176

[20] Hui-Chin Hsu, Alan Fogel, and Daniel S Messinger. 2001. Infant non-distress vocalization during mother-infant face-to-face interaction: Factors associated with quantitative and qualitative differences. *Infant Behavior and Development* 24, 1 (2001), 107–128.

[21] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica* 26 (1967), 22–63.

[22] Bin Li and Dimas Lima. 2021. Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering* 2 (2021), 57–64.

[23] Jing Liu, Mitja Nikolaus, Kübra Bodur, and Abdellah Fourtassi. 2022. Predicting backchannel signaling in child-caregiver multimodal conversations. In *Companion publication of the 2022 international conference on multimodal interaction*. 196–200.

[24] Chiara Mazzocconi, Benjamin O'Brien, Kevin El Haddad, Kübra Bodur, and Abdellah Fourtassi. 2023. Differences between mimicking and non-mimicking laughter in child-caregiver conversation: A distributional and acoustic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

[25] Scott A Miller. 2012. *Theory of Mind: Beyond the Preschool Years*. Psychology Press.

[26] Thomas Misiek and Abdellah Fourtassi. 2022. Caregivers exaggerate their lexical alignment to young children across several cultures. *Proceedings of SemDial* (2022).

[27] Mitja Nikolaus, Juliette Maes, Jeremy Auguste, Laurent Prevot, and Abdellah Fourtassi. 2021. Large-scale study of speech acts' development using automatic labelling. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Vienna, Austria. https://hal.science/hal-03234620

[28] Patrizia Paggio, Manex Agirrezabal, Bart Jongejan, and Costanza Navarretta. 2020. Automatic Detection and Classification of Head Movements in Face-to-Face Conversations. In *Proceedings of LREC2020 Workshop "People in language, vision and the mind" (ONION2020)*. European Language Resources Association (ELRA), Marseille, France, 15–21. https://aclanthology.org/2020.onion-1.3

[29] Dinh Viet Sang et al. 2017. Facial smile detection using convolutional neural networks. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 136–141.

[30] Bogdan Smolka and Karolina Nurzynska. 2015. Power LBP: A Novel Texture Operator for Smiling and Neutral Facial Display Classification. *Procedia Computer Science* 51 (2015), 1555–1564. https://doi.org/10.1016/j.procs.2015.05.350 International Conference On Computational Science, ICCS 2015.

[31] Saiyed Umer, Ranjeet Kumar Rout, Chiara Pero, and Michele Nappi. 2022. Facial expression recognition with trade-offs between data augmentation and deep learning features. *Journal of Ambient Intelligence and Humanized Computing* (2022), 1–15.

[32] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE international conference on computer vision*. 3756–3764.

[33] Yu Xia, Di Huang, and Yunhong Wang. 2017. Detecting Smiles of Young Children via Deep Transfer Learning. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 1673–1681. https://doi.org/10.1109/ICCVW.2017.196

[34] Marygrace E Yale, Daniel S Messinger, Alan B Cobo-Lewis, and Christine F Delgado. 2003. The temporal coordination of early infant communication. *Developmental psychology* 39, 5 (2003), 815.

[35] Kaihao Zhang, Yongzhen Huang, Hong Wu, and Liang Wang. 2015. Facial smile detection based on deep learning features. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* (2015), 534–538.