# Automatic analysis of infant engagement during play: An end-to-end learning and Explainable AI pilot experiment

Marc Fraile
Department of Information Technology, Uppsala University
Uppsala, Sweden

Joakim Lindblad
Department of Information Technology, Uppsala University
Uppsala, Sweden

Christine Fawcett
Department of Psychology, Uppsala University
Uppsala, Sweden

Nataša Sladoje
Department of Information Technology, Uppsala University
Uppsala, Sweden

Ginevra Castellano
Department of Information Technology, Uppsala University
Uppsala, Sweden

## ABSTRACT

Infant engagement during play is an active area of research, related to the development of cognition. Automatic detection of engagement could benefit the research process, but existing techniques used for automatic affect detection are unsuitable for this scenario, since they rely on the automatic extraction of facial and postural features trained on clear video capture of adults. This study shows that end-to-end Deep Learning methods can successfully detect engagement of infants, without the need of clear facial video, when trained for a specific interaction task. It further shows that attention mapping techniques can provide explainability, thereby enabling trust and insight into a model's reasoning process.

## CCS CONCEPTS

• **Applied computing** → *Psychology*; • **Computing methodologies** → *Interest point and salient region detections.*

## KEYWORDS

infant engagement, video analysis, end-to-end learning, deep learning, explainable artificial intelligence, class activation mapping

**ACM Reference Format:**
Marc Fraile, Joakim Lindblad, Christine Fawcett, Nataša Sladoje, and Ginevra Castellano. 2021. Automatic analysis of infant engagement during play: An end-to-end learning and Explainable AI pilot experiment. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3461615.3485443

## 1 INTRODUCTION

Task engagement in infants during play has been identified as a relevant concept in child developmental studies [6]. However, the analysis of the relevant experimental data is a laborious process, requiring manual annotation of session data by one or more domain experts. This marks infant engagement detection as a prime target for automation. The automatic detection of engagement and other affect-related states in adults has been the subject of a growing body of research [16, 20], but most proposed methods rely on the automatic extraction of engineered facial, postural and audio features [7]. The available extraction tools [3, 4] are trained on clear video capture of adults, and do not generalize well to the facial and bodily characteristics of young children [5].

An alternative, under-explored path for classification of affective states is end-to-end training of Deep Learning models. While this approach has shown to have great predictive power, its black-box nature limits our trust in the obtained predictions. To combat this, a variety of *explainability* techniques have been developed [1]. In the context of convolutional networks for computer vision, *attention maps* are an important family of such techniques [2, 17, 18].

In this study, we display how an end-to-end Deep Learning approach can successfully predict infant engagement during guided play. This is done from a single video source capturing both the infant and the researcher from a lateral view, without dedicated facial capture or a dedicated feature extraction phase. We achieve this with a very limited amount of data, and use attention maps to validate the network's reasoning. We fine-tune a pre-trained video classification network on three different guided play tasks, with only 40-60 samples in each training set. Once satisfactory results are obtained, we showcase how a selection of attention mapping techniques can be used to increase trust, reveal possible modes of failure, and learn what behaviors are correlated with the labels.

## 2 RELATED WORK

### 2.1 Automatic Infant Engagement Recognition

The automatic recognition of affect in human-human interactions has been most often studied in adults and using video recordings of the participant's face. Sariyanidi et al. [16] provide a 2015 survey, predating the modern explosion of Deep Learning methods. They highlight the dominance of scales originating in the psychology literature, such as Facial Action Units (FAU). These measurements can be estimated using computer vision methods, and used as inputs for further classification methods. Popular later tools like OpenFace [3] and OpenPose [4] are based on the same principle: provide reliable estimates for facial and postural features, and let the end-user apply statistical models. This contrasts with the modern trend to perform


*ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada*

ACM ISBN 978-1-4503-8471-1/21/10…$15.00
https://doi.org/10.1145/3461615.3485443

*end-to-end* training: remove all assumptions about what features are best, and instead feed the raw input to a deep learning model. This often results in higher predictive power, and may reveal previously unknown domain-specific knowledge. While some works in this direction exist [20], the subject remains underexplored.

An area of particular interest is *engagement*. It has seen heavy interest in the context of education, due to its ties to student performance and institutional success [9]. Automatic detection of student engagement has seen similar developments as affect recognition: facial, postural and auditory hand-crafted features are extracted from frontal video, and detection is performed on these. A recent example is the EmotiW Challenge [7] , which has included student engagement detection in several editions. Analysing the accepted papers, we see a repeated use of OpenFace and OpenPose as feature extractors, often paired with engineered audio features.

Task engagement has also been studied in-depth in the context of infant development. Infant engagement has been shown to correlate with cognitive performance [6], and is an active area of research. Studies on infant engagement typically rely on manual annotation of videos, where domain experts make subjective judgements based on features such as positive emotion, gaze direction, and goal-directed movements. This is a time-consuming process, which can lead to coarse labelling and relatively small amounts of available data, making it difficult to obtain statistical significance. For these reasons, reliable automatic engagement detection would be a very useful tool for research, but it has remained virtually unexplored.

One possible reason for this lapse on research is the lack of tools: face and pose estimation algorithms trained on adults do not necessarily generalize to young children, and might require re-training. For example, Chambers et al. [5] re-train OpenPose with a dedicated infant dataset. Relevant studies might opt to use a human expert for classification, even in real-time applications [14]. If an infant is too young to speak, we cannot rely on audio features for affect recognition. If they are too young to sit still, it can be hard to obtain clear facial video capture. When automatic analysis is explored, machine learning approaches are avoided. For example, Egmose et al. [8] use kinetic energy estimates to estimate joint attention between mother and child.

## 2.2 Explainability

End-to-end Deep Learning models often outperform methods based on hand-crafted features, at a lower development complexity. However, it's typically hard to explain why an input is mapped to its output, leading us to treat the system as a *black box* whose contents are unknown. If the collected data contains unexpected or unwanted correlations, the model can obtain high statistical scores through faulty reasoning.

The set of techniques used to alleviate the black-box problem is known as Explainable Artificial Intelligence (XAI), and has seen a surge of interest in recent years [1]. Explainability has further been identified by the European Union as a necessary principle to attain trustworthy AI [10]. Some prior work has applied XAI techniques in the context of automatic affect recognition. Lin et al. [12] use end-to-end learning to predict affective labels in adults from physiological signals. They feed four separate types of sensory data into independent models, and fuse the results using a Random Forest classifier. They then calculate a relevance score for each stream as an explainability step. Pandit et al. [15] use convolutional networks to predict arousal and valence from extracted FAU. They then simplify the model to its shallowest as an explainability step. To the best of our knowledge, neither end-to-end learning has been applied to infant engagement, nor XAI techniques have been used.

*Attention maps* are an important family of XAI methods in image processing. They attach an importance score to each input pixel, given an output decision. Since defining "importance" is open-ended, this has led to a variety of methods with different execution speeds and interpretations. Notable examples include guided backpropagation [18], Grad-CAM [17], and LRP [2]. A selection of attention mapping methods was applied in this study.

# 3 METHOD

## 3.1 Dataset

We collected a dataset including videos of 22 14-month-old infants (11 girls; mean age = 14 months, 6 days) participating in three different interaction tasks with an adult experimenter. Infants were recruited from a local list of families who were interested in participating in research with their child. Before the tasks, parents received information about the study and signed a consent form. The procedure was approved by the local ethical committee.

During the tasks, the infant was seated in a high chair at a table with the parent seated behind them and the experimenter seated across from them. A Sony Handycam HDR-CX260 camera (1440 × 1080px @ 25fps) was used to record the interaction. It produced a profile view of the infant and experimenter.

In the *"people"* task, four round boxes were attached to the table. The yellow boxes directly in front of the infant and the experimenter each contained 10 wooden dolls. The boxes to the left (red) and the right (blue) of the child were empty. The experimenter began by naming the boxes "sun house" (red) and "moon house" (blue), and placing a doll in one of the boxes. She then asked the infant if they would like to try and removed the cover from the infant's doll box. The experimenter placed half of her dolls into one of the boxes one at a time, and then switched to placing them in the other box.

In the *"eggs"* task, the experimenter showed the infant an egg-shaped shaker and began to shake it at either 150 or 170bpm for 10 seconds. Then she gave the infant an egg shaker. Infants could play with the shaker for 30 seconds. The experimenter then pretended to drop her egg on the ground and when she picked it up, she began shaking it at the other rhythm for another 30 seconds.

In the *"drums"* task, the experimenter showed the infant a drum and tapped on it with a drumstick at one of the predetermined rhythms, as in the previous task. She moved the drum to the middle of the table and gave the infant their own drumstick and encouraged them to join in drumming. After 30 seconds, she flipped the drum over and began drumming at the other rhythm.

A coder watched each video, and rated the child for their level of engagement with the task. Each task was divided into 30-second segments, and each segment was labelled either *playfully engaged object*, if the child was playing with the relevant object, or *not engaged*, otherwise. Between two and five labelled clips were obtained per session and task. This totalled 77 samples for "people", 54 samples for "eggs", and 50 samples for "drums".

## 3.2 Classification Algorithm

While the performance of video classification networks has improved drastically over the last few years, successful models tend to be large and complex, and progress has been driven by the growing availability of ever-larger datasets [21]. Thus, pre-training is a crucial step when working with smaller sample sizes. Despite this, not all successful networks are equally big, and a smaller modern network runs a lower risk of over-fitting the data. These factors led us to choose the Mixed Convolution Network provided by the `torchvision` package [19]. It is a smaller model in the ResNet family, and consists of an embedding unit followed by a logistic regression classifier head. The embedding unit consists of convolutional layer blocks with skip connections, using 3D convolutions at the earlier blocks and 2D convolutions at the latter blocks. It comes pre-trained on the Kinetics-400 dataset [11], a collection of videos containing 400 categories, with over 400 videos per category.

Training was done independently for each task, resulting in 3 separate binary classifiers. As a pre-processing step, 80% of the recorded sessions were split into a training fold, with the remaining 20% reserved for testing. Each relevant 30-second annotated clip was then extracted from the session recording, and downsampled to a spatial resolution of 208x160 pixels. This resulted in 40-60 clips in each training set, and 10-20 clips in each test set. The test data framerate was uniformly reduced to 1/8th (3.125fps down from 25fps), while the training data was reduced to between 1/7th and 1/9th as a data augmentation step. At training time, a video data augmentation pipeline was run: cutting the videos to 60 frames; applying a small amount of rotation, scaling, and stretching; extracting a random 112x112 pixel crop; possibly applying a horizontal flip, color biasing, Gaussian blur, and/or normal white noise.

Training consisted of two steps: a *head training* phase, where the single-layer 400-class classification head for Kinetics was substituted for a single-output binary classification head and trained independently; and a *fine-tuning* phase, where the whole network was trained at a lower learning rate. For head training, a hyperparameter grid search was employed; best results were obtained using L-BFGS [13] with class weights and no parameter decay. For fine-tuning, ADAM optimizer with weight decay gave good results.

## 3.3 Attention Maps

For this study, we evaluated a selection of attention mapping methods, starting with *gradient visualization*. It consists in running the backpropagation algorithm, and displaying the gradient at the input layer. In multi-class classification, only the relevant class is fed into the algorithm. This simple method can already reveal useful information on the network's attention, but tends not to be class-discriminative, and susceptible to high-frequency noise.

One approach to address these issues is *guided backpropagation* [18]. It modifies the backpropagation algorithm so that negative gradients are discarded when propagating through a ReLU layer. This change tends to produce sparser attention maps, with greater focus on what areas of the image provide positive evidence for a decision. It also suffers from a lack of class discrimination, but better captures the relevant fine details.

Another approach is *Grad-CAM* [17]. It is a class-aware low-frequency solution for convolutional networks: take the activation maps on the last convolutional layer, weigh each map according to the mean of its gradients, sum them, and discard negative values. The resulting grayscale map will typically be low-resolution (10x10 pixels in our case), and can then, if needed, be upsampled to match the input resolution. In contrast, *guided Grad-CAM* [17] marries the fine detail of guided backpropagation with the class-awareness of Grad-CAM by multiplying both maps together. It produces relevant but sparse visualizations.

**Table 1: Accuracy and F1 score for infant engagement in each task (convolutional network vs. manual annotation).**

| task | people | eggs | drums |
|---|---|---|---|
| test accuracy | 75% | 83% | 92% |
| test F1 score | 82% | 83% | 89% |

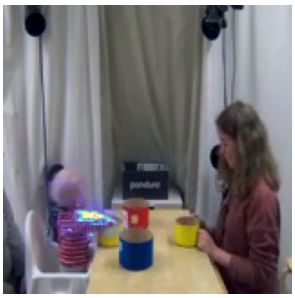


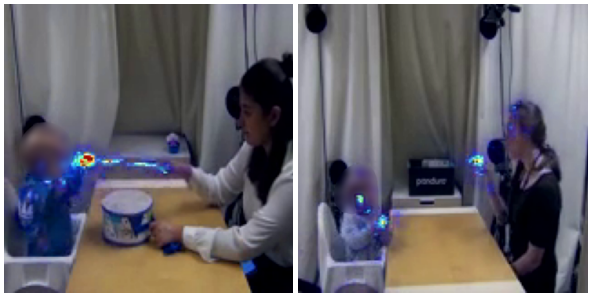
**Figure 1: Infant engaging in the "people" task, with *guided backpropagation* overlayed. We can see that the child's arm movement is relevant to the classification result.**

**Figure 2: Infant engaging in the "people" task, with *gradient visualization* overlayed. We can see the network focuses on the infant. The same frame is shown in Figure 1, displaying greater legibility.**

# 4 RESULTS AND ANALYSIS

The performance of the most successful network per task (judged by test F1 score) is listed in Table 1. While these numbers are promising, the network could be relying on accidental relationships in the data. To investigate if this is the case, we applied the techniques listed in Section 3.3 to correctly classified samples, and performed subjective analysis of the results.

In all three tasks, *gradient visualization* gave a noisy indication that the network was focusing on the infant and experimenter, and reacting to their motion. Figure 2 shows the gradient on a

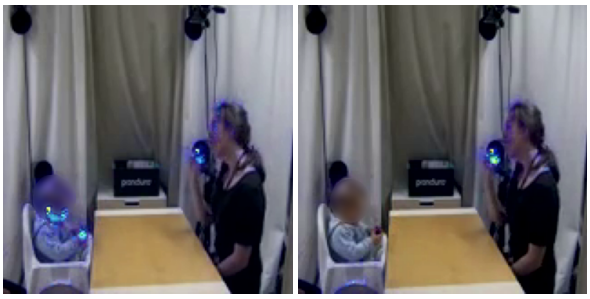


(a) Engaging in "drums". (b) Engaging in "eggs".

**Figure 3: Infant engaging in the "drums" and "eggs" tasks, with *guided backpropagation* overlayed. (a) the network reacts to the infant's hand, and both drumsticks. (b) the network reacts to moving oval shapes.**

(a) Correct class (not engaged). (b) Opposite class (engaged).

**Figure 4: Infant not engaged in the "eggs" task. *Grad-CAM* comparison for the correct class vs. the opposite class, showing which elements in the frame support each outcome.**



(a) Guided backpropagation. (b) Guided Grad-CAM.

**Figure 5: Infant engaged in the "eggs" task. Comparison between *guided backpropagation* and *guided Grad-CAM*. The network focuses on the experimenter's shaker.**

positive sample for the "people" task. We can see that in this case the network is solely focusing on the child, but there is too much noise to deduce further details.

*Guided backpropagation* displayed greater clarity, and showed that the network focuses on task-relevant forms of motion. Figure 1 shows the same sample as Figure 2, enabling comparison of the attention maps produced by different methods. We can see that guided backpropagation focuses more clearly on the infant's arm, which is extended into the toy box. Figure 3a shows a positive example for the "drums" task. It shows the network is paying attention to the infant's hand, and both drumsticks. This reveals a possible mode of failure: the network could fail to distinguish the infant from the experimenter. Figure 3b shows an "engaged" sample in the "eggs" task. Similar to drums, the network is focusing on the moving shakers. Despite the good classification performance, we see a strong reaction to the experimenter's motion and to other oval shapes in the image, again hinting at possible modes of failure.

*Grad-CAM* gave very coarse maps, but could clearly separate what constitutes evidence for engagement, and what constitutes evidence against it. Figure 4 shows a negative (not engaged) sample from the "eggs" task. Figure 4a shows evidence for the correct class. The network is (correctly) focusing on the child's general position to determine "child not engaged". Figure 4b shows the present evidence for the opposite class ("child engaged"), which is (erroneously) highlighting the experimenter's motion. The network might be using undesired environmental clues to detect engagement.

*Guided Grad-CAM* gave very sparse maps, making it problematic to derive conclusions in some cases. When enough information was visible, it provided both class discrimination and fine detail. Figure 5 shows a positive example in the "eggs" task, comparing guided backpropagation (5a) and guided Grad-CAM (5b). We can see that the network focuses on the experimenter's shaker to determine "child engaged", coinciding with our observation in Figure 4b.

## 5 CONCLUSIONS

In this study, we have shown that an end-to-end Deep Learning model can successfully predict infant engagement during play, when trained for a specific interaction task. By leveraging a standard video classification architecture pre-trained on a large dataset (Kinetics 400), we succeeded despite very limited amounts of data (40-60 training samples per task), and without the usual limitations imposed by feature extraction steps (clear view of the face for facial features, clear view of all limbs for postural features, etc.).

Finally, we have shown how attention maps can increase confidence in a classifier (demonstrating that the network focuses on relevant motion in all 3 tasks), or help identify modes of failure (network reacts to all moving oval shapes in "eggs" task, considers experimenter's engagement). Both types of examples illustrate what behaviors are correlated with the labelling in the current dataset. Since this information is found without intervention, it is also free from human assumptions, and can help us understand the data.

Future work could involve revisiting the annotations to obtain more context-relevant information (Is the child engaged with the task but not the experimenter? Is the child engaging the experimenter?), and study how the network attention changes. It could be interesting to compare the aspects considered important by a human and the aspects highlighted by the network, opening a possibility for new knowledge discovery.

## 6 ACKNOWLEDGEMENTS

This work was partly funded by the Centre for Interdisciplinary Mathematics, Uppsala University, and the Swedish Research Council (grant n. 2020-03167).

# REFERENCES


[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[5] Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel R Pierce, Daniel K Bogen, Laura Prosser, Michelle J Johnson, and Konrad P Kording. 2020. Computer vision to automatically assess infant neuromotor risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 11 (2020), 2431–2442.

[6] Kaili Clackson, Sam Wass, Stanimira Georgieva, Laura Brightman, Rebecca Nutbrown, Harriet Almond, Julia Bieluczyk, Giulia Carro, Brier Rigby Dames, and Victoria Leong. 2019. Do helpful mothers help? Effects of maternal scaffolding and infant engagement on cognitive performance. *Frontiers in psychology* 10 (2019), 2661.

[7] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. 2020. Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 784–789.

[8] Ida Egmose, Giovanna Varni, Katharina Cordes, Johanne Smith-Nielsen, Mette S. Væver, Simo Køppe, David Cohen, and Mohamed Chetouani. 2017. Relations between Automatically Extracted Motion Features and the Quality of Mother-Infant Interactions at 4 and 13 Months. *Frontiers in Psychology* 8 (2017), 2178. https://doi.org/10.3389/fpsyg.2017.02178

[9] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research* 74, 1 (2004), 59–109.

[10] AI HLEG. 2019. High-level expert group on artificial intelligence: Ethics guidelines for trustworthy AI. *European Commission, 09.04* (2019).

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[12] Jionghao Lin, Shirui Pan, Cheng Siong Lee, and Sharon Oviatt. 2019. An explainable deep fusion network for affect recognition using physiological signals. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2069–2072.

[13] Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming* 45, 1 (1989), 503–528.

[14] Setareh Nasihati Gilani, David Traum, Arcangelo Merla, Eugenia Hee, Zoey Walker, Barbara Manini, Grady Gallagher, and Laura-Ann Petitto. 2018. Multimodal dialogue management for multiparty interaction with infants. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 5–13.

[15] Vedhas Pandit, Maximilian Schmitt, Nicholas Cummins, and Björn Schuller. 2020. I see it in your eyes: Training the shallowest-possible CNN to recognise emotions and pain from muted web-assisted in-the-wild video-chats in real-time. *Information Processing & Management* 57, 6 (2020), 102347.

[16] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2014. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence* 37, 6 (2014), 1113–1133.

[17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[18] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

[19] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.

[20] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309.

[21] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. 2020. A Comprehensive Study of Deep Video Action Recognition. *arXiv preprint arXiv:2012.06567* (2020).