Automatic, Real-Time Coding of Looking-While-Listening Children Videos Using Neural Networks

Yotam Erel*, Christine Potter**, Sagi Jaffe-Dax**, Casey Lew-Williams**, Amit H. Bermano*

Blavatnik School of Computer Science, TAU University* Department of Psychology, Princeton University**

Abstract

Infants' looking behavior, typically recorded in low-resolution videos, is a common tool for assessing infants' attention and learning. Despite its ubiquity in developmental science, estimating infants' gaze still typically involves laborious coding, or expensive setups that require calibration and involve data loss (e.g., Vencker et al., 2020). As a solution, we propose employing state-of-the-art computer-vision methods through automatic gaze estimation from low-resolution videos. We demonstrate our method on data collected from the common Looking-While-Listening procedure, where infants are expected to look at one of two locations on a screen. At the core of our method lies an artificial neural network that classifies gaze directions of infants in real-time. Using a large dataset of manually-annotated videos, we demonstrate performance that is on-par with human annotators and even replicates published results that used manual coding.

Methods

- 266 video sessions of infants and children (10-72m) participating in different studies using the Looking-While-Listening procedure**.**
- Trained research assistants had labeled each frame of each video for whether the infant was looking at one image ("left"), the other one ("right"), or neither ("away"). Yielding a complete dataset of ~500k frames.
- We trained a convolutional neural network on 186 sessions of manually-coded videos. We then tested on a sample of 8 held-out sessions and found 98% agreement with the manual coding of "left" or "right". Adding the "away" category reduced the network's performance, but still yielded ~90% agreement with manual results (compared to the ~96% agreement between different human annotators).
- We re-analyzed the automatically-labeled data and compared with published results (Potter and Lew-Williams, ICIS 2020).
- In another setup, we also adjusted & finetuned the state-of-the-art RT-GENE(Fischer et al., 2018) neural network to our problem set, and compared results.

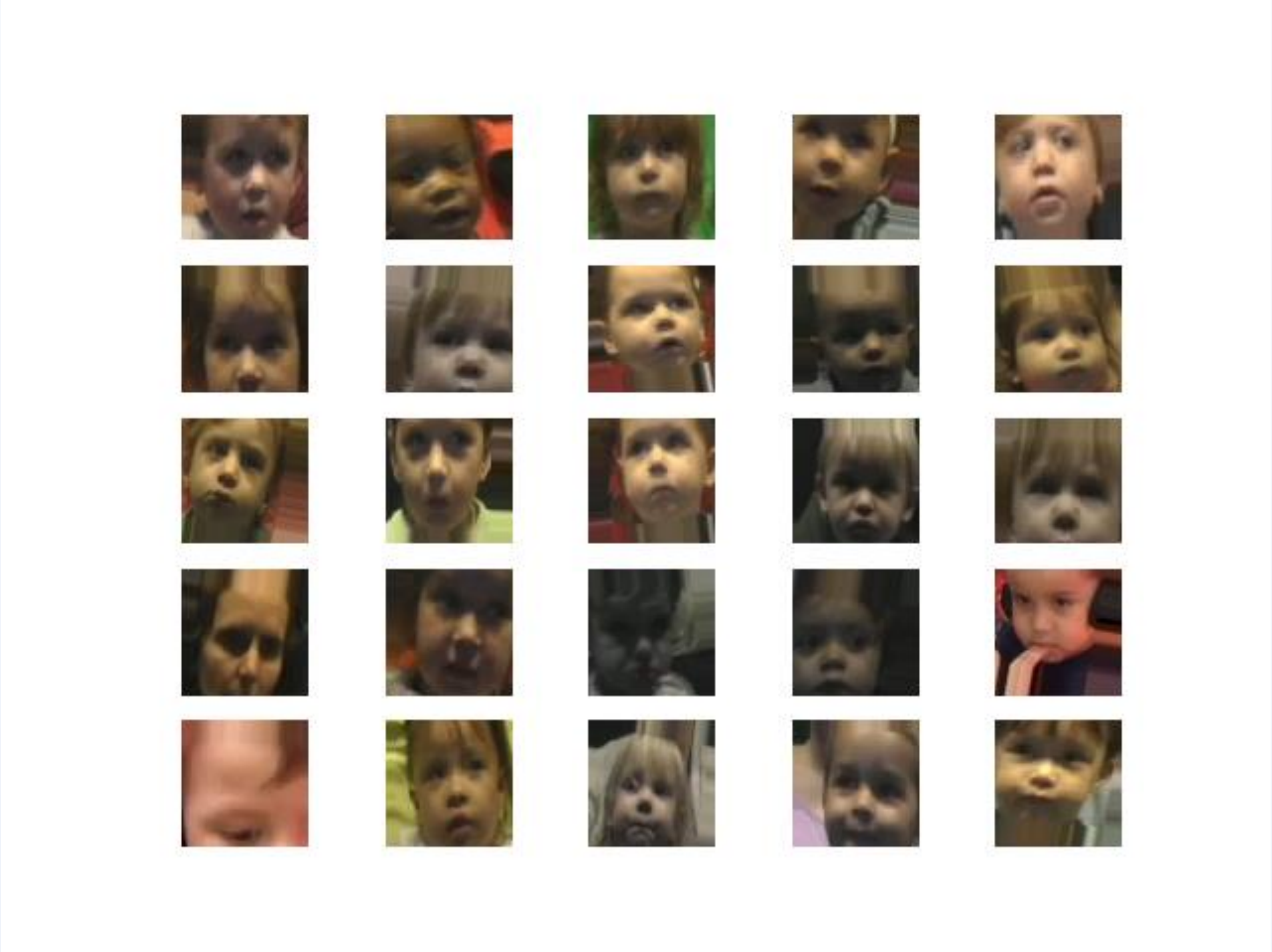

Figure 1 – Example of frames from our data set.

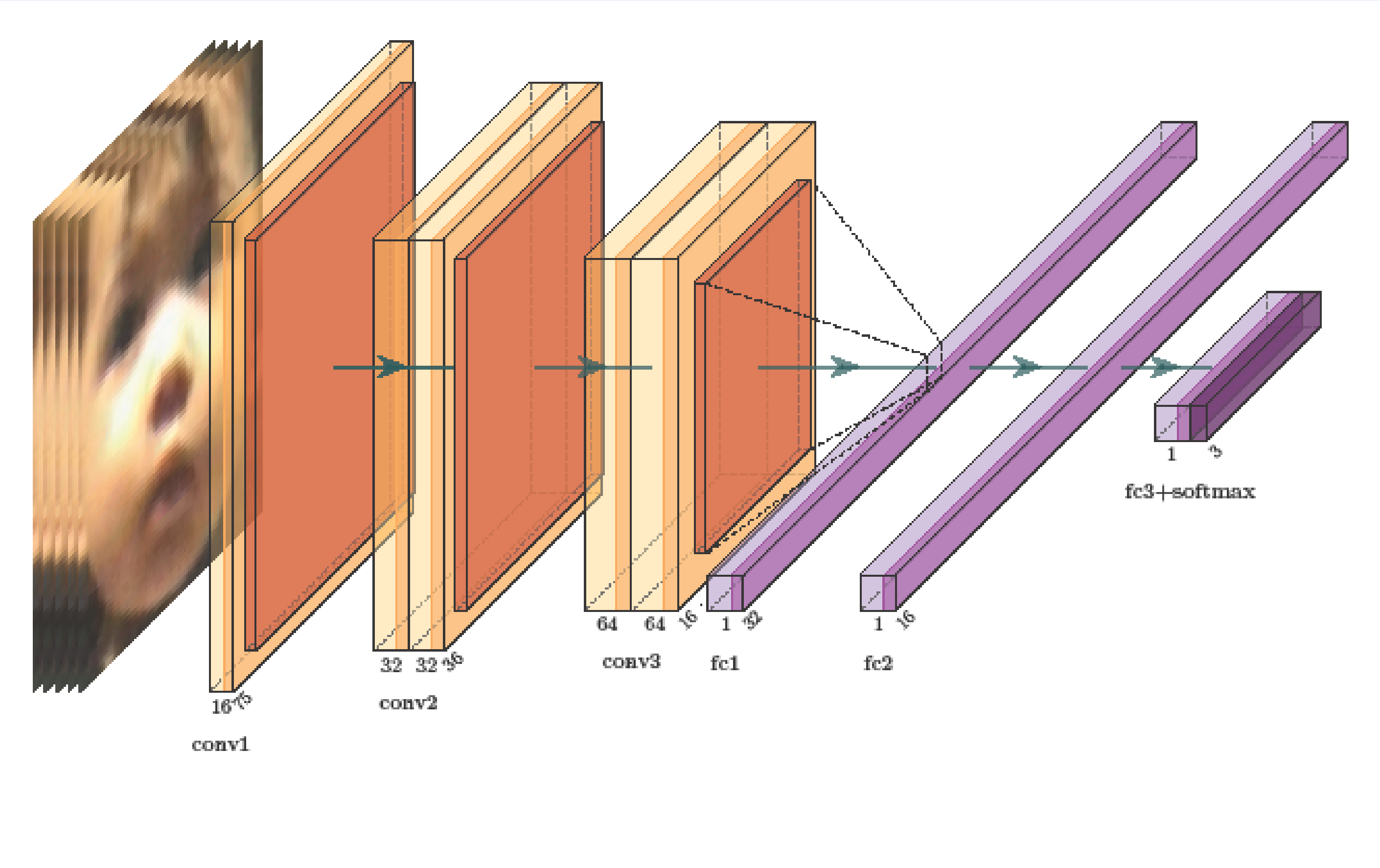

Figure 2 – CNN architecture used in our solution.

Results

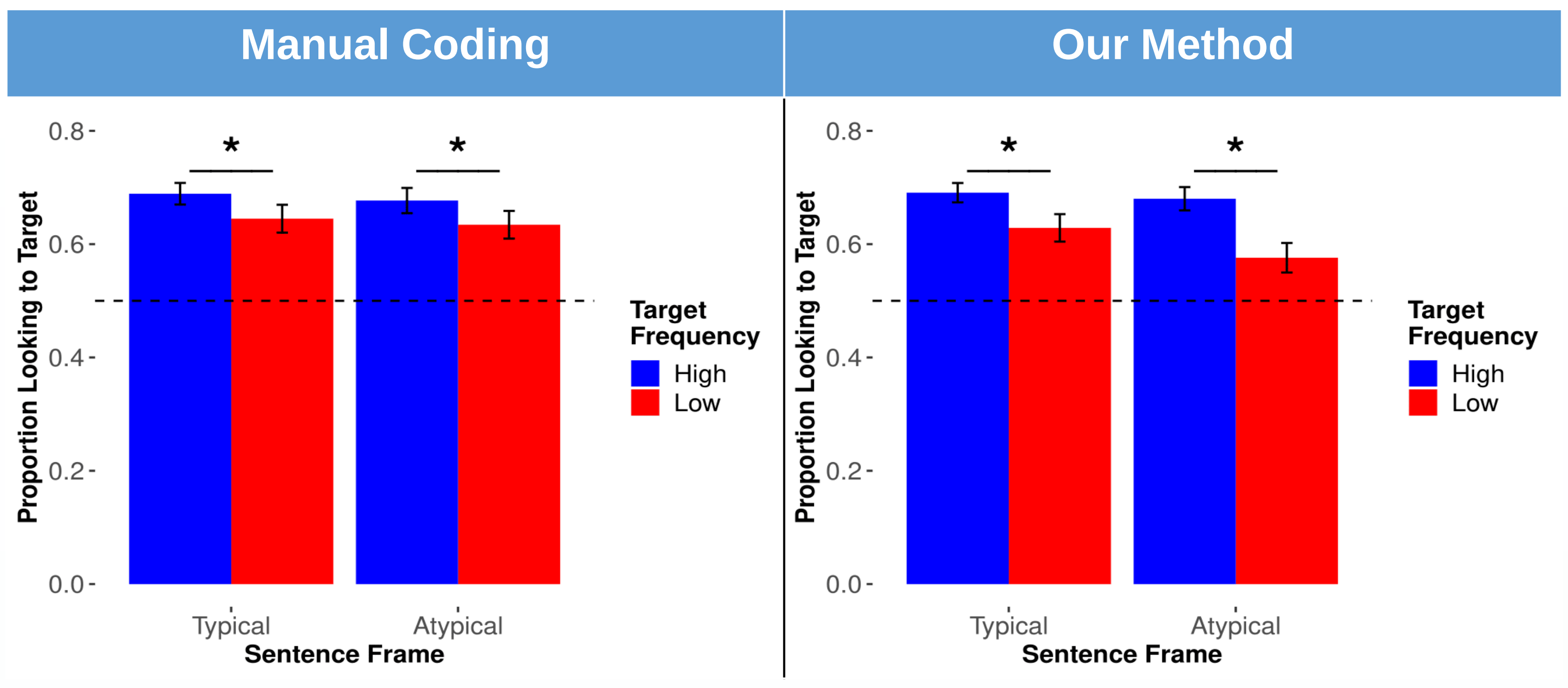

Figure 3 - We obtained similar patterns of data and replicated all significant effects (Potter and Lew-Williams, ICIS 2020).

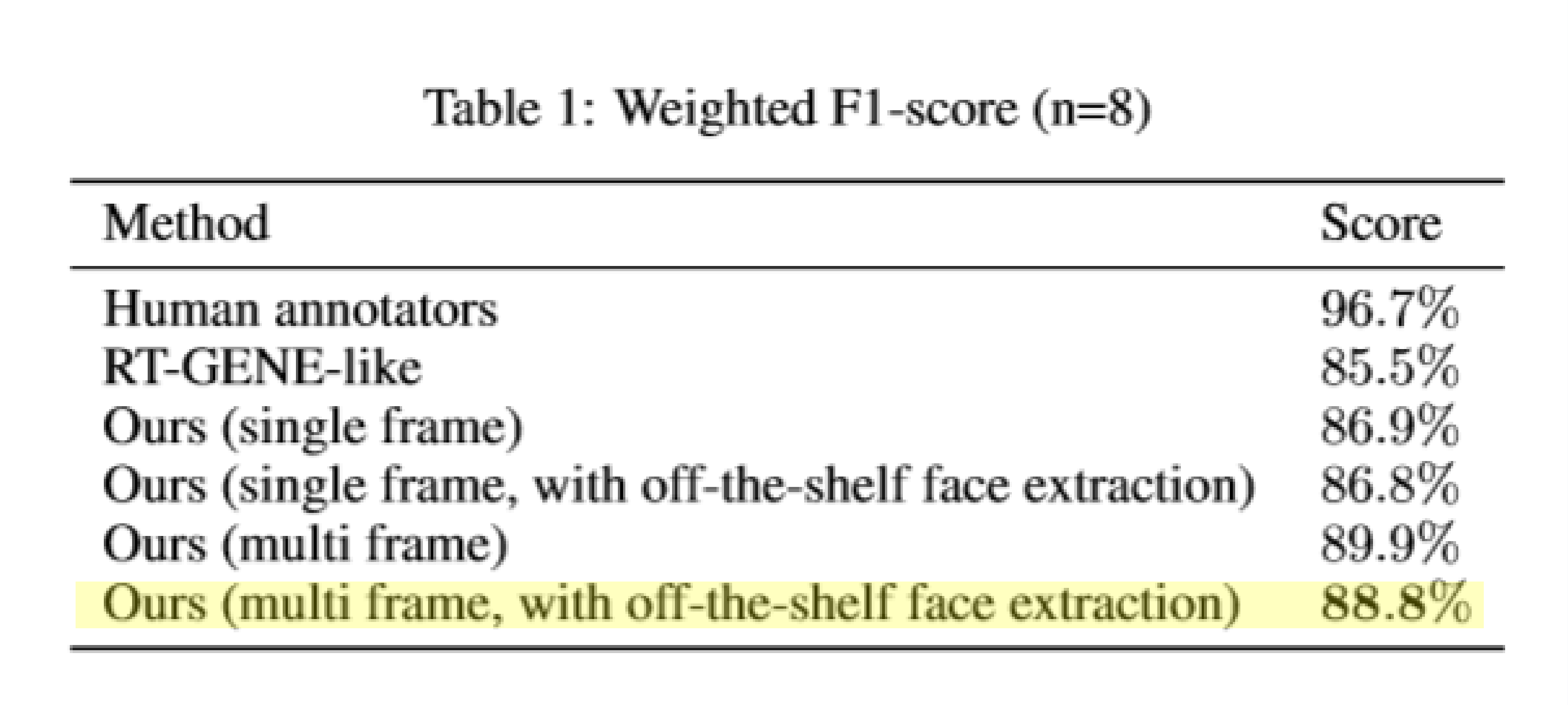

Table 1: Weighted F1-score (n=8)

Method	Score
Human annotators	96.7%
RT-GENE-like	85.5%
Ours (single frame)	86.9%
Ours (single frame, with off-the-shelf face extraction)	86.8%
Ours (multi frame)	89.9%
Ours (multi frame, with off-the-shelf face extraction)	88.8%

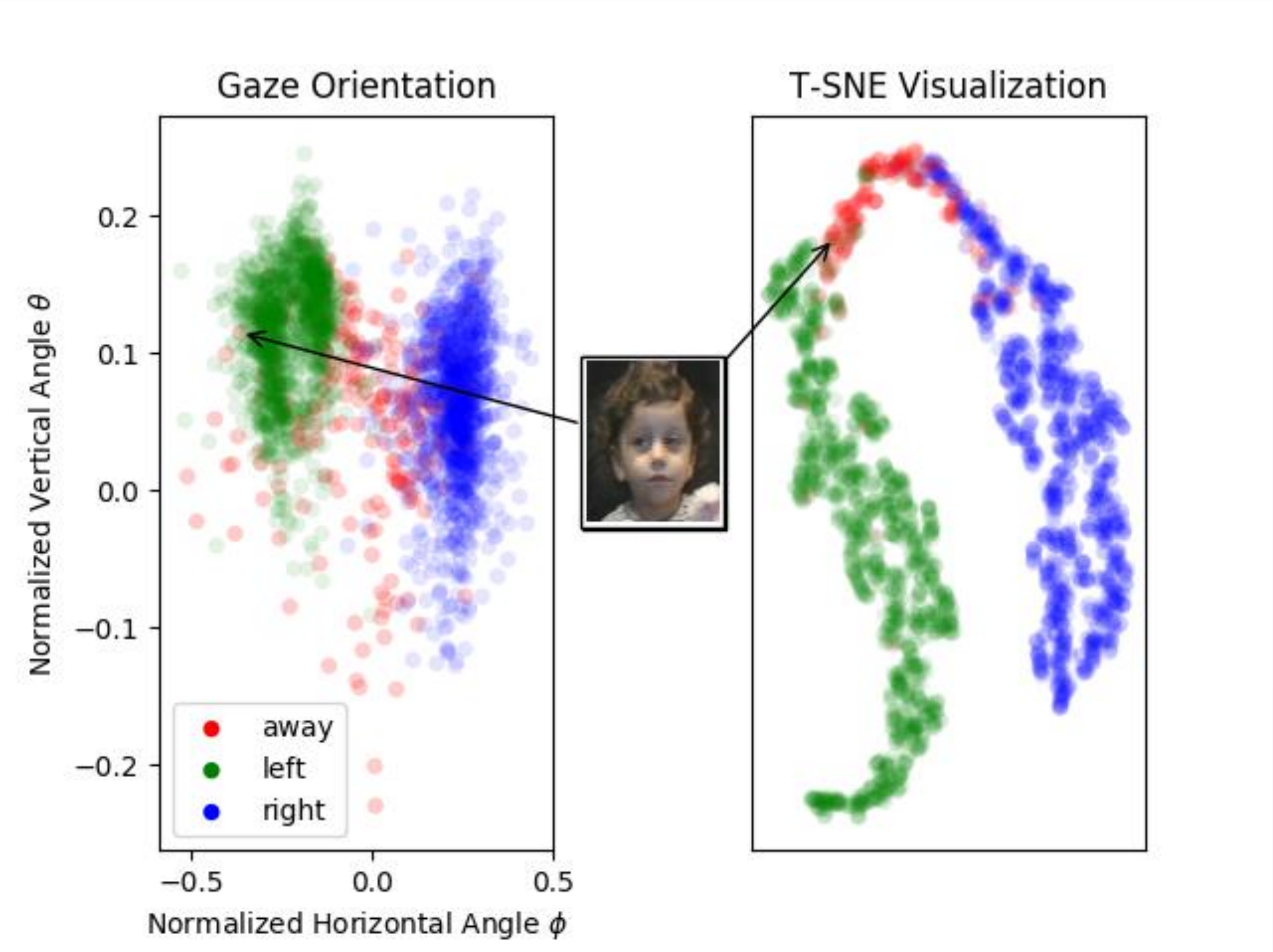

Figure 4 – Left: viewing angles as extracted by a "vanilla" version of RT-GENE for held-out video frames. Each frame is colored according to its manual coding. Right: the same frames, obtained at the output of the first fully connected layer within our version of the network, projected to 2D using T-SNE (van der Maaten et al., 2008). Notice how our network maps the frame features to more distinguishable clusters, as demonstrated by the pointed-out "away" frame in both graphs.

The RT-GENE-like solution under performed compared to our own model on the discrete 3-class problem but allowed continuous eye gaze to be inferred in real-time.

Discussion & Future Research

- Gaze-contingent paradigms can be developed using this method, for example, trials in which the infant was inattentive could be repeated, or infants could be presented with harder or easier trials depending on their performance on earlier trials.
- Our technique can be adjusted to produce continuous gaze directions (i.e. horizontal and vertical angles). Exploiting this feature could potentially induce more elaborate studies, such as those displaying evolving graphics or animations.
- We believe that incorporating specific experiment geometric properties into the data (such as screen size, distance from screen, room geometry) will improve results dramatically.

Contact

Yotam Erel
erelyotam@gmail.com