Joint Attention, Computer Vision, and Foundation Models: How State-of-the-art AI approaches can analyze instances of early childhood development

Priscilla Zhao, Raymond Zhang

Faculty Advisors: Phil Fisher, Nick Haber, Silvio Savarese, Jeannette Bohg

Introduction

The Filming Interactions to Nurture Development (FIND) is a video coaching program aimed at promoting positive interactions between caregiver and child. This translational neuroscience research initiative has been shown to improve the language skills of families of low socio-economic backgrounds[5]. Our project aims to help scale up this intervention by exploring ways in which computer vision can augment the film editing process. We believe this project will be a good demonstration of the 2D to 3D reconstruction methodologies and contribute to the field of translational neuroscience.

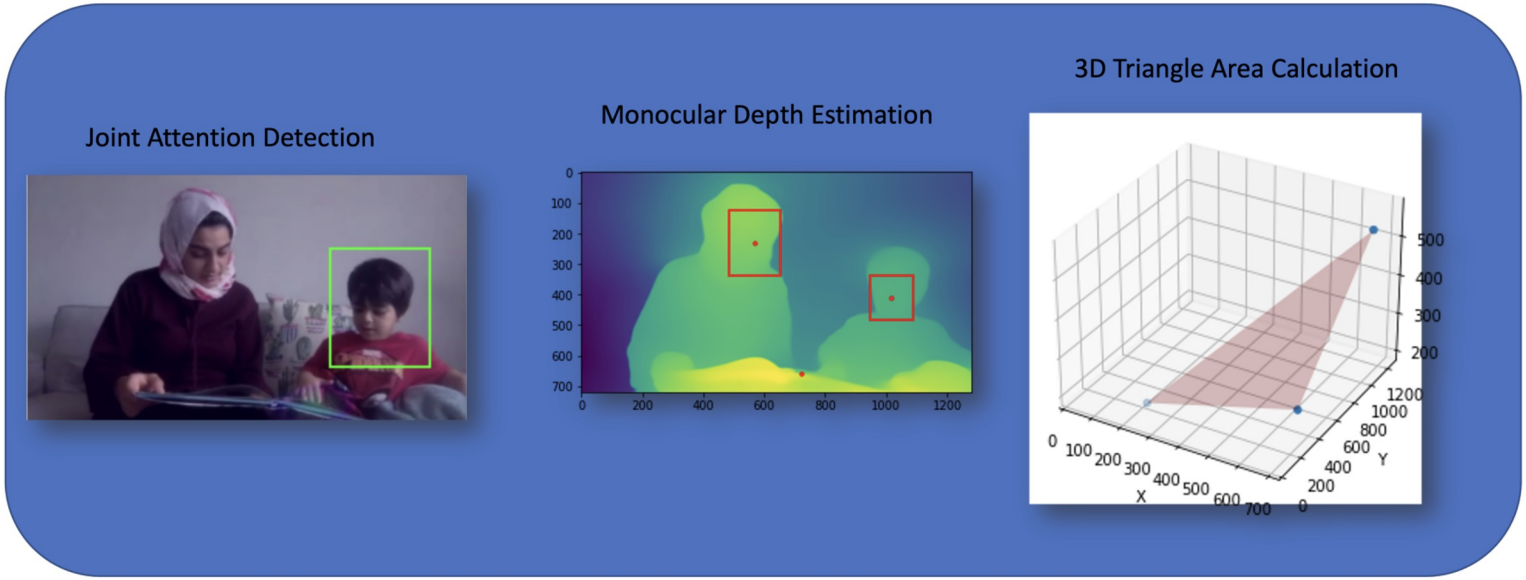


Figure 1. Output to joint attention pipeline

The overall goal of our research group is to create a human-in-the-loop AI video editing system that would speed up the existing manual editing process. By focusing on instances of joint attention this project could potentially speed up the video editing process of the FIND program resulting in an increased scale of video coaching sessions.

Background: Joint Attention

Joint attention is a behavior in which two people focus on an object or event for the purpose of interacting with each other. Prior work has shown that joint attention has been correlated with increases in the utterance object of shared attention, and could be beneficial to children's early language acquisition [13].

Problem Statement: Building a Joint Attention 3D Reconstruction Pipeline

This project aims to build a pipeline or a prototype to detect the 3D coordinates of the caregiver, child, and object that both agents are sharing joint attention of.

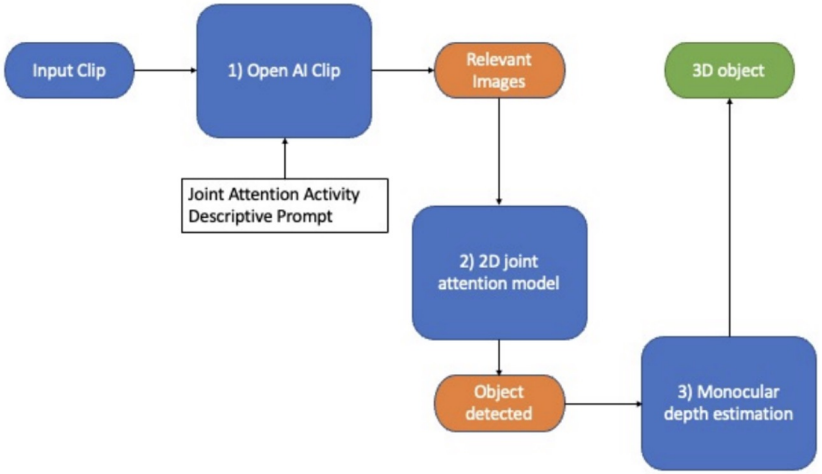


Figure 2. Overall Pipeline for Joint Attention 3D Reconstruction

Step 1: Joint Attention Detection Via CLIP

The first step of the project is to determine if a frame contains joint attention interaction or not. After experimenting with a few prompts the following prompts were used **"a photograph of an adult and a child both LOOKING at the SAME object"** and **"a photograph of an adult and children LOOKING at each other"**.

CLIPS Similarity Threshold

There are 260 image instances in total, out of which 140 are hand-labeled as True joint attention. The similarity scores of these cases are ranging from 0.185 to 0.276.

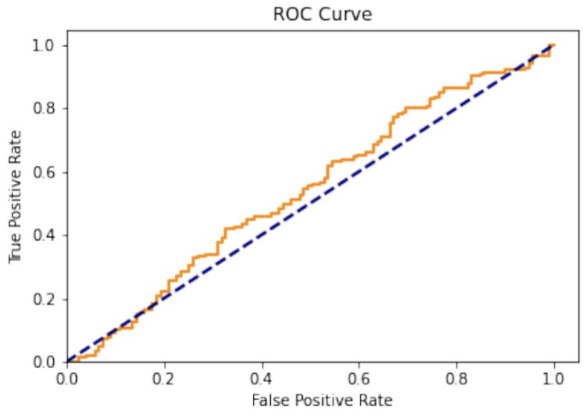


Figure 3. ROC Curve, Threshold for TPR = 0.8: 0.1941

Calculating TPR of 0.8 the resulting similarity threshold is 0.1941.

Step 2: 2D Joint Attention Model

Attention Flow model outputs an image with the headbox detected and a heat map around the area of joint attention.

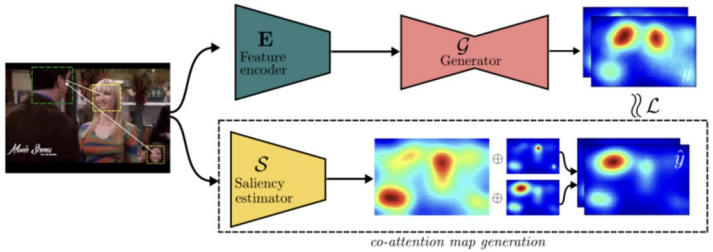


Figure 4. Attention Flow method is composed of three modules, (i) feature encoder, (ii) attention flow generator, and (iii) saliency-based ground truth generation.

Step 3: Monocular Depth Estimation

The model for monocular depth estimation is the Vision Transformers for Dense Prediction, which proposes a new architecture by using vision

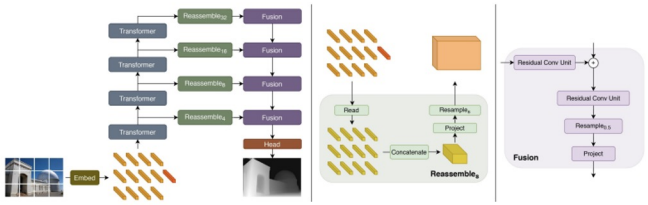


Figure 5. DPT-Large Model Architecture

Using the depth estimation image it is possible to estimate the z coordinate which gives the resulting point in three dimensions. Given the three points of the child, caretaker, and object it is possible to construct a triangle that will represent the relationship between the three entities.

Analysis and Results

Pipeline Performance

Step	Accuracy	Sample Size
Step 1: CLIP	80%	112
Step 2: Joint Attention	88%	9

Above is the accuracy percentage of the first 2 steps of the joint attention 3D reconstruction pipeline.

Computational Analysis of Joint Attention

Activity	Triangle Area	N
Reading	0.074	4
Playing	0.087	5

Table 1. Traingle Area of Different Interactions

When there is joint attention and shared reading activity the area of the triangle is smaller compared to the playing behavior, this would indicate that during shared reading caregiver and child are closer together physically.

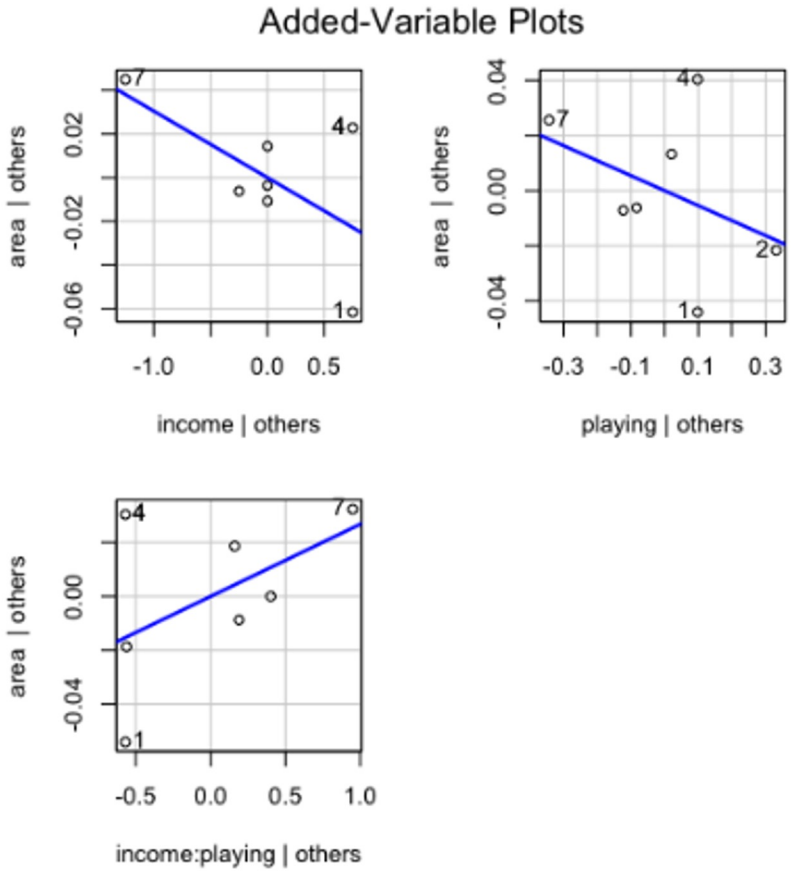


Figure 6. Multivariate Regression of 3D Triangle Area and Income and Play Behavior

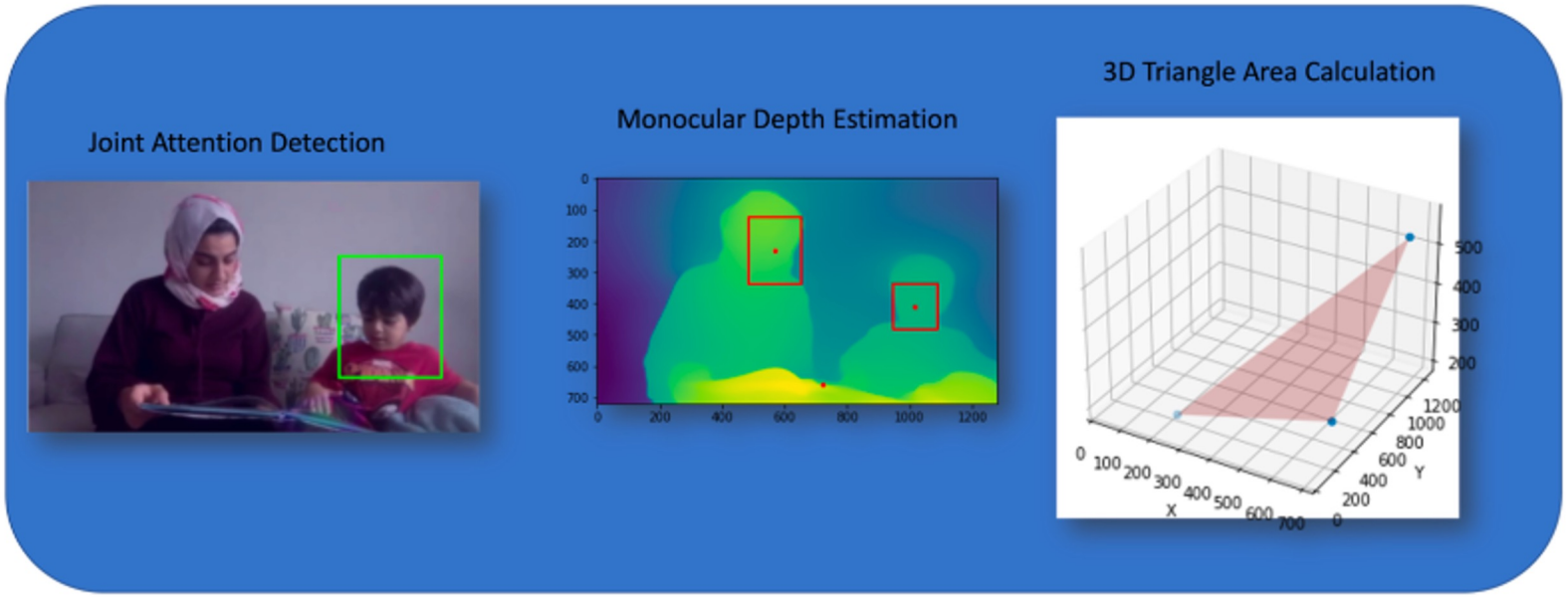
Regression shows that as income increases the area of the triangle decreases, similarly if shared reading is present the area of the triangle decreases.

Limitations

Limitations:

1. Evaluation data is mainly positive cases of joint attention
2. No ground truth of depth estimation in Step 3 of the pipeline
3. 3D Triangle area normalization is not accurate

The key limitation of this project is the unbalanced evaluation data set. For scene understanding tasks such as identifying joint attention, there needs to be expertly created train, dev, and test data sets that are balanced and without error.



References

[1] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[2] Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. Temporal localization and spatial segmentation of joint attention in multiple first-person videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 2

[3] Andrea Imhof, Sihong Liu, Lisa Schlueter, Tiffany Phu, Sarah Watamura, and Philip Fisher. Improving children's expressive language and auditory comprehension through responsive caregiving: Evidence from a randomized controlled trial of a strength-based video-coaching intervention. *Prevention Science*, 24(1):84–93, 2022. 1

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[5] Joseph Nelson. Prompt engineering: The magic words to using openai's clip, May 2021. 2

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2

[7] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *CoRR*, abs/2103.13413, 2021. 5

[8] Omer Sumer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[9] Ömer Sümer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. *CoRR*, abs/2001.03960, 2020. 4

[10] Michael Tomasello and Michael Jeffrey Farrar. Joint attention and early language. *Child Development*, 57(6):1454, 1986. 1

[11] Pamela J. White, Mark O'Reilly, William Streusand, Ann Levine, Jeff Sigafoos, Giulio Lancioni, Christina Fragale, Nigel Pierce, and Jeannie Aguilar. Best practices for teaching joint attention: A systematic review of the intervention literature. *Research in Autism Spectrum Disorders*, 5(4):1283–1295, 2011. 2

[12] Chen Yu and Linda B. Smith. Hand–eye coordination predicts joint attention. *Child Development*, 88(6):2060–2078, 2017. 2

Thank you!