Multimodal Deep Neural Network for Behavior Recognition with FOS-R for Children with ASD

Zhenhao Zhao[1], Eunsun Chung[2*], Kyong-Mee Chung[2] and Chung-Hyuk Park[1]

[1]Department of Biomedical Engineering, The George Washington University, United States of America

[2]Department of Psychology, Yonsei University, Republic of Korea

Background

- The prevalence of Autism Spectrum Disorder (ASD) has been steadily increasing over the past decades, becoming a significant global concern (Chiarotti and Venerosi, 2020)
- Current statistics in the United States: 1 in 36 children is diagnosed with ASD (Maenner et al., 2023)
- Children with autism often experience symptoms such as somatosensory disturbances and atypical developmental patterns, which greatly impact daily social functioning (Mayes and Calboun, 1999)
- The **Family Observation Schedule (FOS)**: A comprehensive tool used to assess family interactions in different contexts.
  - FOS in autism research: Provide valuable insights for diagnosing, treating, and supporting children with autism by examining their social contexts and dynamics. (Lee and Chung, 2016)
- Presently, FOS data is encoded manually, a time-consuming and labor-intensive process. An automated FOS encoding algorithm could alleviate the burden on clinicians and researchers, ultimately benefiting numerous children with autism

Purpose

- This study proposed an efficient multi-modal deep learning model for FOS data encoding, using task-specific videos of children with autism as training data to optimize model efficacy.

Methods

❖Dataset Description

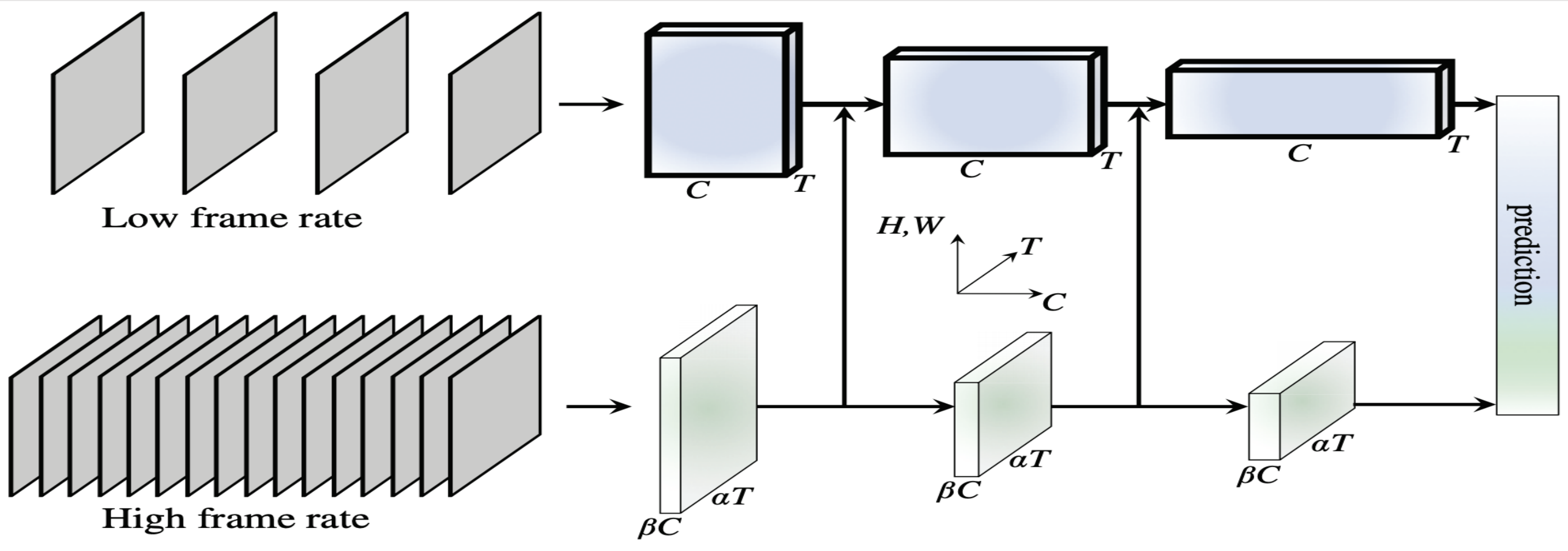
- 216 videos of children with autism or their parents/guardians
  - Duration: 5 – 10 mins
  - Subjects age: 1 – 12 years old
- The children in the videos performed three different task: **(1) playing with specific types of toys**, **(2) performing a series of specific instructions**, and **(3) free playing alone**
- Ground truth of the videos is the interaction styles (IS):
  - Some describing the IS of parents: Praise (**P**) and Affection (**AF**)
  - Some other describing IS of Child: on-compliance (**NC**) and Oppositional (**OP**)

❖Data preprocessing

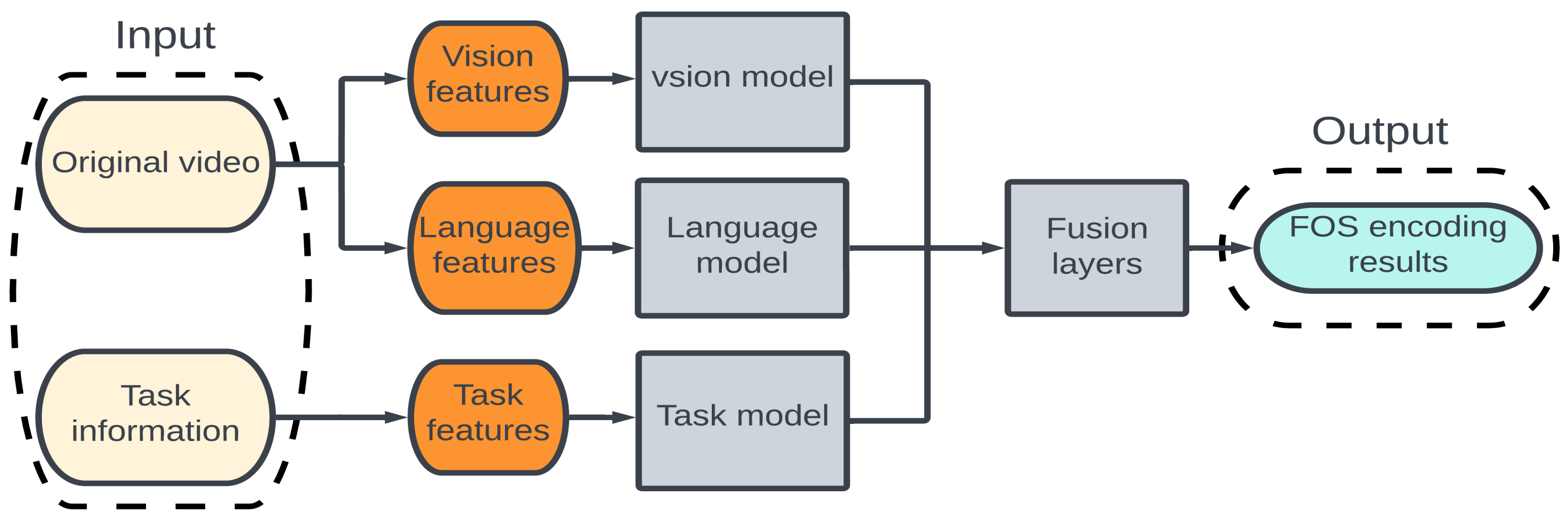
- Segmented the original video into 10-second intervals for each encoding
- Do the data argumentation by key frame random sampling
  - One video can be trained multiple time to study more features

❖Model training

- We divided the original dataset to three parts based on the ISs
  - Vision modality IS: primarily identified by action-based information, such as Physical Negative (**PN**)
  - Language modality IS: identified through language-based information, such as Praise (**P**)
  - Combined modality IS: Both language and vision information will used to make prediction, such as Non-compliance (**NC**)
- Currently, we designed vision-based model and continue working on the language-based mode and combined model
- Our vision-based model is modified from the **fast-slow Resnet** (Feichtenhofer et al., 2019), as shown in Fig. 1
- In the future research, we will also **integrate the task information** into our model to increase the recognition precision
- Fig. 2 shows the general structure of our research



<Figure 1> Basic structure of the fast-slow Resnet



<Figure 2> General structure of the multi-modality IS recognition.

Initial Results

- We finished training the **vision-based model**
- Evaluate dataset performance
  - General prediction accuracy: **82%**
  - Prediction F1 score: 0.59
  - Prediction precision: 0.51
  - Prediction recall: 0.68
- The language-based model's architecture is not decided yet
- We have decided to use a structure based on self-attention for multi-modal recognition, to fully leverage task information and language information, aiming to improve the predictive accuracy of the model

Conclusions

- **Deep learning-based behavior recognition of children with autism** : *Automate FOS data encoding, training* by task-specific videos of children with autism
- **Integrated task information and other multi-modal data**, such as video and audio, to enhance the model's efficacy, addressing a gap in previous studies
- The collected dataset includes 216 videos of children with autism or their guardians performing different tasks, which are then encoded to describe their interaction styles
- **A three-step sequential training** was carried out for the model based on vision, language, and combined vision-language modalities
- **Preliminary results showed promising accuracy of over 80%** for the initial visual information classification model, built on the fast-slow Resnet architecture
- Future work: commence the construction of a language model and a multi-modal approach and improve the overall model accuracy

References

- [Chiarotti and Venerosi, 2020] Chiarotti, F. and Venerosi, A. (2020). Epidemiology of autism spectrum disorders: A review of worldwide prevalence estimatessince 2014. Brain Sci, 10(5):274.
- [Maenner et al., 2023] Maenner, M. J., Warren, Z., Williams, A. R., and et al. (2023). Prevalence and characteristics of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2020. MMWR Surveill Summ, 72(No. SS-2):1–14.
- [Mayes and Calboun, 1999] Mayes, S. D. and Calboun, S. L. (1999). Symptoms of autism in young children and correspondence with the dsm. Infants & Young Children, 12(2):11–23.
- [Lee and Chung, 2016] Lee, M. and Chung, K. (2016). Development of parent child interaction-direct observation checklist (pci-d) for children with developmental disabilities. Journal of Rehabilitation Psychology, 23(2):367–395.
- [Feichtenhofer et al., 2019] Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In Proceedings of the IEEE/CVFInternational Conference on Computer Vision (ICCV).

More Information

Scan the QR for more information and a copy of the poster.

This study was supported by Institute of Ministry of Trade, Industry and Energy grant funded by the Government of South Korea (No. HI18C0458)

QR 코드 삽입 예정