Sirish Gambhira

424-440-0682 | Personal Website | sirishgam001@gmail.com | linkedin.com/in/sirish-gambhira

Education

Georgia Institute of Technology Exp: May 2026
M.S in Computer Science, Specialization in ML, GPA - ***4.0 / 4.0*** *Atlanta, GA*

Indian Institute of Technology, Kharagpur Aug 2017 - May 2022
B.Tech + M.Tech, Electronics & Electrical Communication Engineering, Minor in CS - ***9.33 / 10.0*** *Kharagpur, India*

Technical Skills

Languages: Python, C/C++, JavaScript, TypeScript, SQL, HTML/CSS, C#

Frameworks/Libraries/Tools: Pytorch, Tensorflow, CUDA, Triton, CMake, React, Redux, ASP.NET, FastAPI, Node.js, Express, Phaser3, Azure, NSight Compute, Docker, Postman, Git, YAML pipelines

Experience

Research Engineer | Microsoft Research India July 2022 - July 2024

- Primary contributor to HyWay, an interactive browser-based audio-video platform for in-person and remote participation.
- Implemented face detection module to asynchronously detect faces from video stream; telemetry module to extract and generate user-related analytics using Kusto Query Language; integrated enterprise data using JWT access token.
- Built an image-editing tool enabling users to create event spaces. The tool allows users to upload a custom image and edit it using drag/drop, group/ungroup, resize/rotate and duplicate different shapes (e.g., PowerPoint). The user actions are built using event listeners and use vector transformations. Stored user actions as objects in a array enabling undo-redo.
- **Performance Optimisations**: Fixed system memory leakage by performing in-place mutation instead of creating new objects for user actions. Migrated global variables from react state to redux store and reduced the number of re-renders. Maintained client-side synchronization using periodic server polling, and limited system memory by freeing stale resources.
- Our system hosted multiple internal events such as poster sessions supporting more than **300 users per session**.

Indoor User Localization and Representation

- Developed a computer-vision system to detect (YOLO), track (OCSORT), localize and represent (RetinaFace) multiple users present in an indoor environment using multiple cameras with non-overlapping views.
- Proposed an inter-camera matching algorithm to associate identical people across different camera viewpoints.
- Optimized an open-source face detection lib (RetinaFace) and significantly reduced the end-to-end processing time of our system.
- Real world evaluations showed our mean inter-camera matching algorithm accuracy is around **90%**, mean indoor localization error is **1.037m** and the overall pipeline achieved **15 FPS**, supporting real-time operations.

Multi-Sensor User State Modeling for Distraction Detection

- Developed a windows-application for automatically determining whether a user is distracted or not during online meetings (e.g., Teams, Zoom etc..) in a privacy-preserving manner (user's data shall not leave their device). The application collects data from multiple sources such as camera, microphone, speaker, and screen activity (e.g., mouse, keyboard) during the call.
- Implemented an acoustic echo cancellation module to filter local user's speech using microphone and speaker streams.
- Proposed high-confidence rules to automatically generate supervised training data corresponding to distracted behavior.
- Our approach used temporal convolutional networks with resnet backbone for video-inputs and combined these predictions with rule-based outcomes for final predictions.
- This enabled different usecases such as: auto-mute A/V devices, recap missed-out content, meeting effectiveness.

Data and Applied Scientist Intern | Microsoft India May 2021 - July 2021

- AI graph is a user-centric graph with user's emails, meetings, documents as nodes and topics, associated people as edges.
- Developed a graph parsing algorithm to extract user's topics and used heuristic scores to quantify user-topic relevance.
- Used Holt-Winters models for time-series analysis and visualized interactive spider & radar charts using D3.js

Selected Projects

Caching policies using L2 persistence for DNN inference Aug 2024 - Dec 2024

- Observed that default caching policy leads to sub-optimal performance for weight-shared DNNs inference (metric: latency)
- Implemented dynamic prefetching: fetch next layer's weights during current layer's computation [CUDA stream concurrency]
- Observed improvement in latency for hand-written CUDA kernels for linear and convolutional networks on low-end GPUs
- Extended policies for custom PyTorch models (e.g., OFA MobileNet) and observed upto **8.27%** compared to default policy

Can DB Queries Exploit Tensor Cores? Oct 2024 - Nov 2024

- Implemented database hash join operation as matrix multiplication in Triton.
- Implemented alternative approach to hash join - hash lookup in CUDA and compared the performance against Triton.
- Results indicated that average run-time for CUDA kernel is 90.65us and for triton is 13.78ms (**150x slower**).
- Nsight Compute showed that Triton achieved occupancy of **18.67%** compared to **74.14%** for CUDA

Parallel bitonic sort using CUDA Aug 2024 - Dec 2024

- Implemented parallel bitonic sort algorithm over 10M elements on L40S GPU using shared memory.
- Identified excessive memory accesses using NSight Compute and improved memory throughput from **67%** to **93%**
- Explored host-device data transfer optimizations such as pinned-memory to reduce transfer time from **15ms** to **5ms**
- Proposed approach achieved **110x** speedup over CPU based sorting.

3D Object Reconstruction using multi-view 2D images Aug 2021 - May 2022

- Implemented an encoder-decoder architecture to generate independent 3D point clouds from each multi-view image and fused them into a unified 3D point cloud object.
- Extracted 2D key points from the generated 3D objects for supervision, eliminating the need for labelled 3D ground truth.
- Obtained a Chamfer distance of **5.12**, comparable to the state-of-the-art single-view reconstruction result of **3.48**.
- Recognised as one of the **Top 5** theses and nominated for the **Best Thesis** award in the department

Deep Clean Aug 2020 - May 2021

- Implemented a recurrent variational autoencoder network to denoise electroencephalogram data from artifacts.
- Observed that model trained with mean squared error (MSE) loss resulted our predictions to converge to zero.
- **Improved** the model accuracy over MSE using alternative loss functions such as Gaussian Negative LogLikelihood

Publications

Sirish Gambhira et al. *HyWay: Enabling Mingling in the Hybrid World*. **ACM UbiComp 2023. US Patent Filed.**