TECHNISCHE UNIVERSITÄT
IN DER KULTURHAUPTSTADT EUROPAS
CHEMNITZ

Fakultät Informatik
Professur Neurorobotik

Research Internship

Topic: A Spiking Neural Network Framework for Multimodal Human Action Recognition via Event Camera and Skeleton Data Fusion in a Simulated Operating Room

Description:

The aim of this work is to investigate whether Spiking Neural Networks can serve as an energy-efficient, low-latency backbone for multimodal human action recognition in clinical settings, using event cameras and skeleton data as complementary sensing modalities. Rather than adapting an existing architecture, the thesis approaches this as a design-space exploration: you will first establish a simulated clinical environment in which virtual humans perform labelled actions, and from which synchronized event streams and skeletal sequences can be generated under controlled conditions. The core contribution is then the formulation and empirical validation of an SNN-based multimodal fusion principle—one that exploits the sparse, asynchronous nature of event data alongside the geometric compactness of skeleton data, without incurring the computational cost of conventional two-stream ANN pipelines. You will study training strategies suitable for spiking dynamics and evaluate the resulting system not only on recognition accuracy, but on the trade-off between energy consumption, memory footprint, and robustness to conditions that are difficult to reproduce in real recordings—such as rapid motion, illumination change, and occlusion.

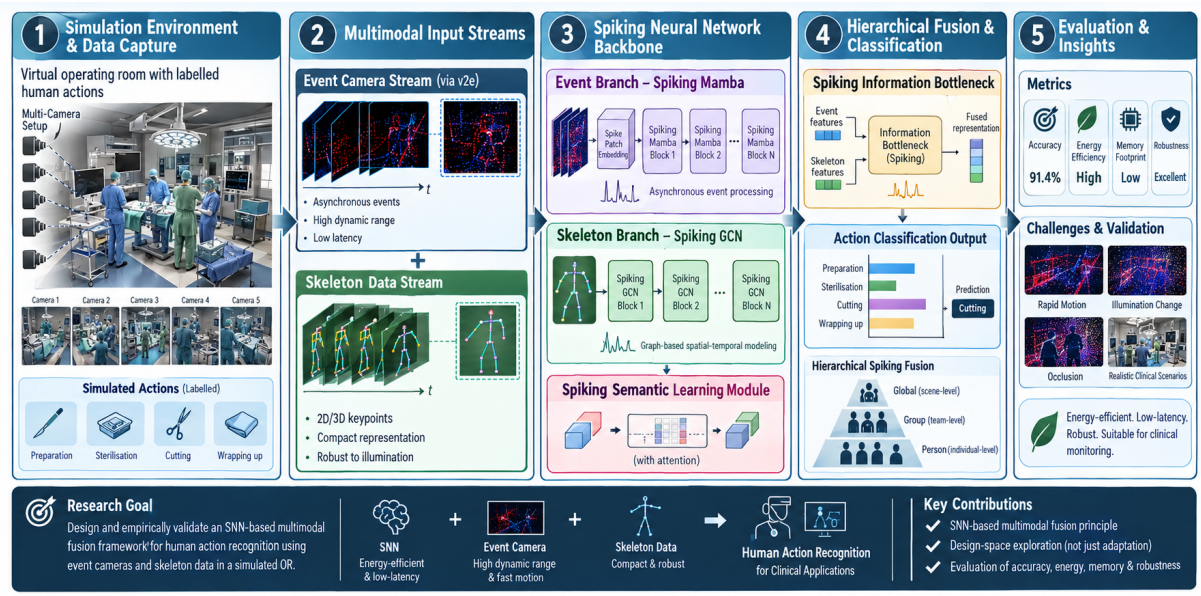


Figure 1: Overview of the proposed multimodal SNN framework. Simulated surgical actions generate event streams and skeletal sequences, processed by spiking Mamba and GCN branches. Their representations are refined and fused through a spiking information bottleneck for action classification and evaluation of accuracy, energy efficiency, memory footprint, and robustness.

Motivation:

Multimodal human action recognition based on the fusion of RGB and skeleton data is effective, but it is constrained by high computational complexity, excessive memory consumption and substantial energy demands, particularly when implemented with conventional Artificial Neural Networks (ANNs). These limitations restrict its applicability in resource-constrained scenarios, such as continuous monitoring of clinical staff in the operating room. Skeleton data are

robust to illumination and background, yet removing appearance information leads to a loss of detail when describing complex actions and subtle limb movements. RGB data retain this detail but are sensitive to lighting and background and require large amounts of storage and computation. Event cameras, with their high dynamic range, low latency and low power consumption, capture only changes in the scene and are therefore well suited to fast motion and high-contrast scenes, although they can introduce noise under frequently changing illumination. Existing multimodal approaches mainly fuse RGB with skeleton data, or events with RGB and language, typically using two-stream ANN architectures, while SNNs have so far been applied mostly to single-modality action recognition, where they reduce energy consumption but often lose accuracy. The fusion of event and skeleton modalities, which combines high temporal resolution and dynamic range with compact geometric information, has received little attention and has only recently been addressed with SNNs by Zheng et al. Since their dataset is derived from conventional 30 fps recordings and real event cameras are not yet available in clinical environments, a simulation-based study offers a controlled and reproducible setting to generate high-quality event-skeleton data and to systematically evaluate this paradigm for operating room applications.

Reference Papers

- N. Zheng, H. Xia, Z. Liang, and Y. Du: SNN-Driven Multimodal Human Action Recognition via Sparse Spatial-Temporal Data Fusion (arXiv:2502.13385, 2025).
- N. Zheng, D. Cheng, H. Xia, and D. Liu: Signal-SGN: A Spiking Graph Convolutional Network for Skeletal Action Recognition via Learning Temporal-Frequency Dynamics (arXiv:2408.01701, 2024).
- Y. Hu, S.-C. Liu, and T. Delbruck: v2e: From Video Frames to Realistic DVS Events (IEEE/CVF CVPR Workshops, 2021, pp. 1312–1321).

Requirements

- **Programming**: Strong proficiency in Python and experience with deep learning frameworks such as PyTorch.
- **Spiking Neural Networks**: Basic understanding of SNN architectures and neuron models.
- **Computer Vision**: Familiarity with human action recognition, graph neural networks and sequence models.
- **Simulation**: Experience with 3D simulation or rendering tools (e.g. Blender, Isaac Sim) is advantageous.

Advisor: Majid Khan
Email: majid.khan@informatik.tu-chemnitz.de