MVSA-Net: Multi-View State-Action Recognition for Robust and Deployable Trajectory Generation

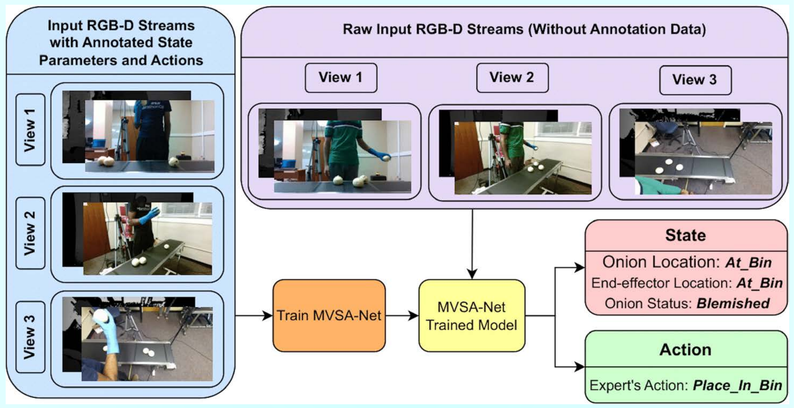
Ehsan Asali, Prashant Doshi, Jin Sun
School of Computing, The University of Georgia
ehsanasali@uga.edu

➢ Abstract

We introduce MVSA-Net, an approach that harnesses multiple viewpoints and gating networks for state-action recognition. Tailored for the 'learn-from-observation' (LfO) paradigm, it surpasses single-view system limitations employing a mixture-of-experts methodology. This method ensures superior accuracy, paving the way for the next era of robust and deployable robotic trajectory systems within the LfO context.

➢ MVSA-Net Overview

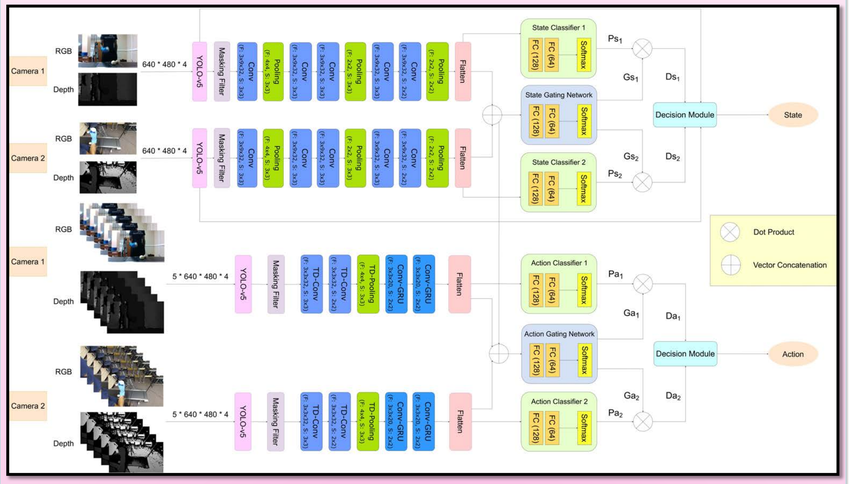
An overview of how to employ MVSA-Net for trajectory generation in a custom domain (onion-sorting) having three input RGB-D streams:



➢ MVSA-Net Methodology

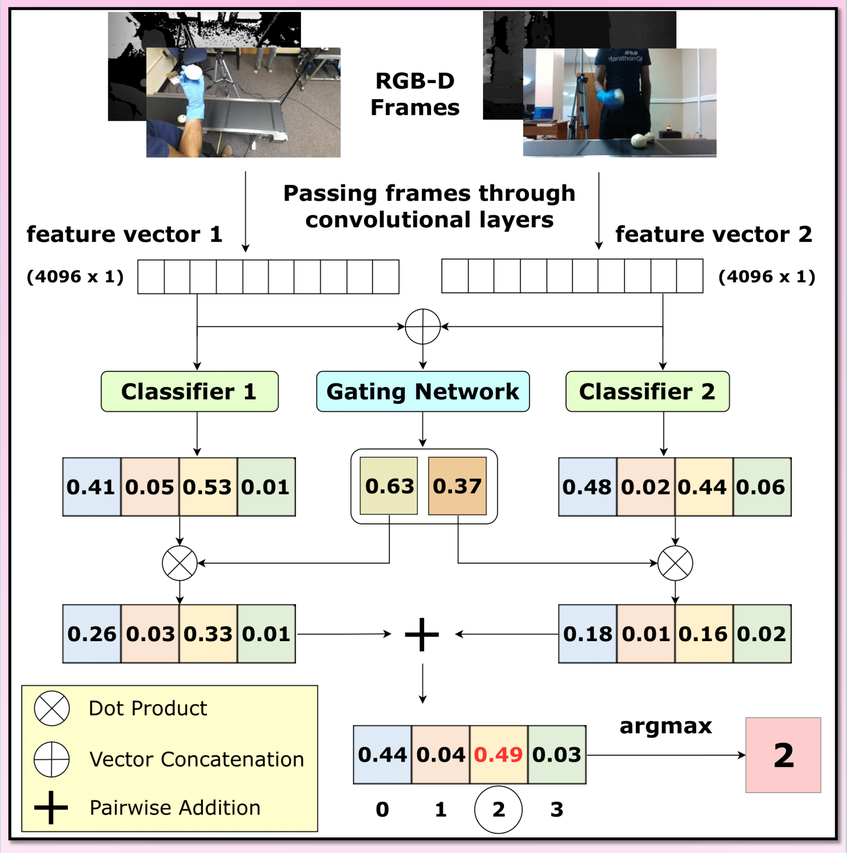
MVSA-Net processes synchronized RGB-D frames using deep convolutional and recurrent networks, enhanced by YOLO-v5 for having precise state-action predictions. These streams are first processed through convolutional layers, extracting vital spatial details. Importantly, the spatial features extracted are shared between the state and action recognition networks, ensuring a consistent and comprehensive understanding of the scene. For state recognition, the system analyzes the latest frames, flattening the data and directing it to a dedicated state classifier. Concurrently, for action recognition, MVSA-Net captures a series of consecutive frames, leveraging Time Distributed convolutions and GRU layers to discern temporal dynamics [1]. A pivotal component, the gating network, integrates insights from all available views [2]. The last component of MVSA-Net is a decision module that takes the weighted predictions for all classifiers and, considering specific conditions, provides the predicted state and action.

➢ MVSA-Net Architecture



➢ Gating Network

The gating network dynamically weighs multiple viewpoints for optimal state-action recognition in MVSA-Net. Here is an example:



➢ Experimental Results

We evaluate the prediction accuracy of MVSA-Net and the baselines under varied conditions using 5-fold cross-validation. The following table shows the results on onion-sorting domain:

Condition	Method	Onion Location	End-Effector Location	Onion Status	Action
Normal	SA-Net (front view)	85.7 ± 0.1	91.02 ± 0.1	40.06 ± 0.2	93.7 ± 0.1
	SA-Net (top-down view)	72.09 ± 0.3	72.76 ± 0.3	59.27 ± 0.7	80.07 ± 0.2
	Two-Stream VGG-Net (front view)	————	————	————	80.92 ± 0.1
	Two-Stream VGG-Net (top-down view)	————	————	————	69.51 ± 0.2
	Two-Stream VGG-Net (multi-view)	————	————	————	80.92 ± 0.1
	MVSA-Net without gating network	87.15 ± 0.1	92.11 ± 0.1	64.24 ± 0.2	95.66 ± 0.1
	MVSA-Net	**89.84 ± 0.1**	**95.68 ± 0.1**	**68.21 ± 0.2**	**97.67 ± 0.1**
Noisy Sensors	SA-Net w/ noise (front view)	62.79 ± 1.5	68.44 ± 1.2	0.0 ± 0.0	79.23 ± 1.1
	MVSA-Net w/o gating network w/ noise	86.05 ± 0.3	87.71 ± 0.3	62.58 ± 0.4	90.87 ± 0.2
	MVSA-Net w/ noise	**88.70 ± 0.2**	**89.69 ± 0.3**	**65.23 ± 0.3**	**92.97 ± 0.2**
Bad Lighting	SA-Net w/ bad lighting (front view)	84.17 ± 0.1	89.20 ± 0.2	31.79 ± 1.5	90.51 ± 0.2
	MVSA-Net w/o gating network w/ bad lighting	86.36 ± 0.1	90.02 ± 0.2	68.02 ± 0.2	92.88 ± 0.2
	MVSA-Net w/ bad lighting	**89.04 ± 0.1**	**90.70 ± 0.1**	**68.02 ± 0.2**	**92.88 ± 0.2**

Using MVSA-Net's predicted trajectories as input for an inverse RL algorithm (MAP-BIRL) resulted in a learned behavior accuracy (LBA) of 97.9%, which is significantly higher than the 83.3% achieved using a single-view SA-Net. MVSA-Net's multi-view fusion enhances LfO performance, adeptly handling occlusions and sensor noise for real-world applications.

➢ Contributions

- Achieves a higher prediction accuracy than baselines on two diverse LfO tasks
- Demonstrates robustness in varied lighting conditions and malfunctioning sensors
- Significantly improves the LfO trajectory performance

➢ Acknowledgements

This work was enabled in part by NSF grant #IIS-1830421 and a Phase 1 grant from the GA Research Alliance to PD. We also thank Prasanth Suresh for assistance with the experimentations.

➢ References

1. A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik. Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access, 6:1155–1166, 2017.
2. J. Chung, S. Ahn, and Y. Bengio. Gating networks. In Advances in Neural Information Processing Systems, pages 3473–3481, 2015.