THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Flow-guided Spatial Attention Tracking for Egocentric Activity Recognition

Tianshan Liu and Kin-Man Lam

International Conference on Pattern Recognition

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University

Outline

- Introduction
- Methodology
- Experiments
- Conclusion

Introduction

- Egocentric Activity Recognition

✓ Wide range of real-world applications

✓ Invisibility of camera wearer & Ego-motion

✓ Identify hand motion patterns & manipulated objects

Introduction

➢ Related Work

✓ Leveraging large-scale fine-grained annotations

- Gaze information [1]
- Hand segmentation and object localization [2]

✓ Attention mechanisms

- Ego-RNN (spatial attention) [3]
- LSTA (sequential attention) [4]

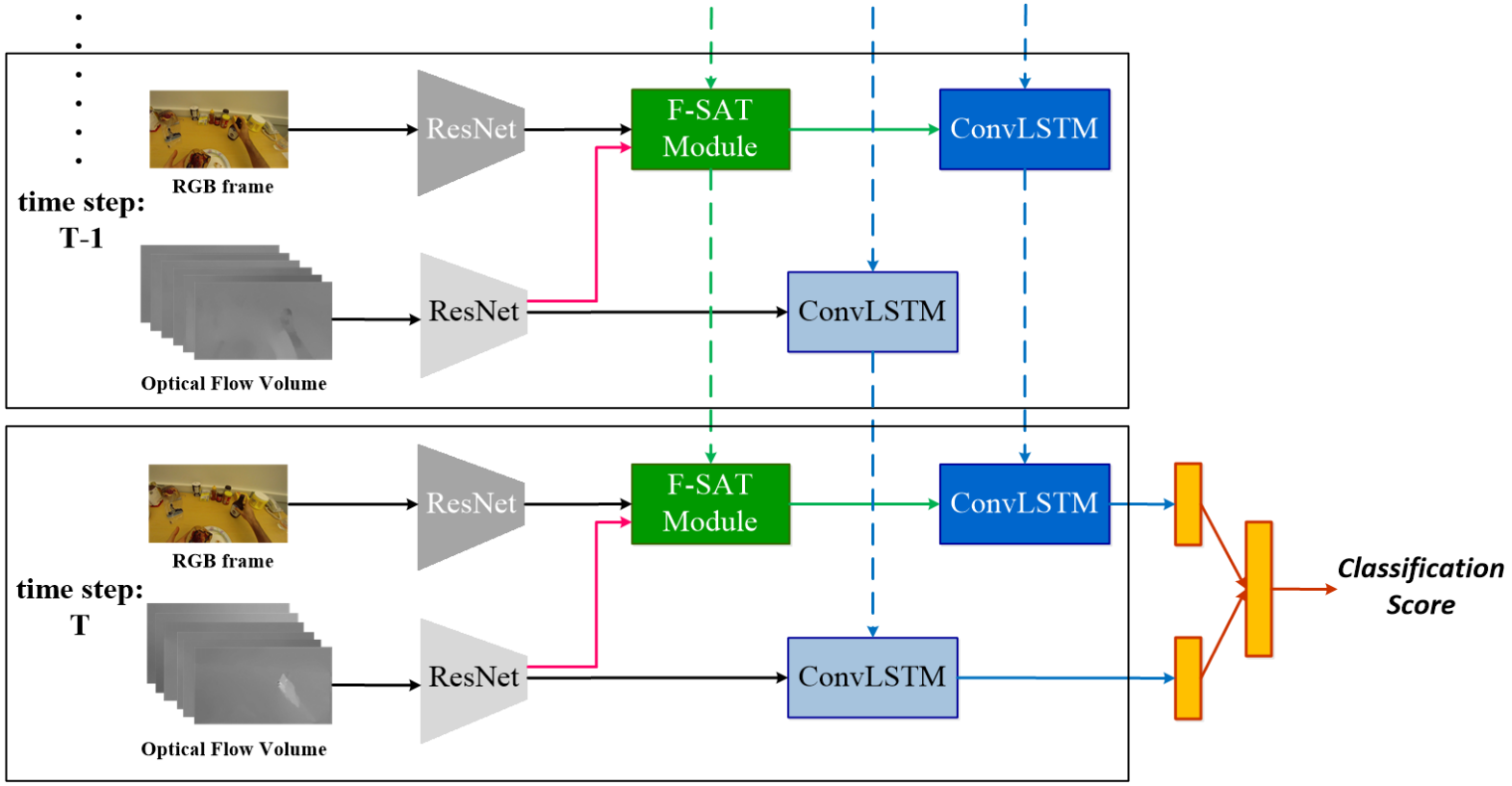
Introduction

➢ Contributions

- ✓ We propose a flow-guided spatial attention tracking (F-SAT) module, to highlight the discriminative features from relevant regions across frames.
- ✓ We insert the proposed F-SAT module into a two-branch-based architecture, to provide complementary information.
- ✓ Evaluation on three public egocentric activity data sets.

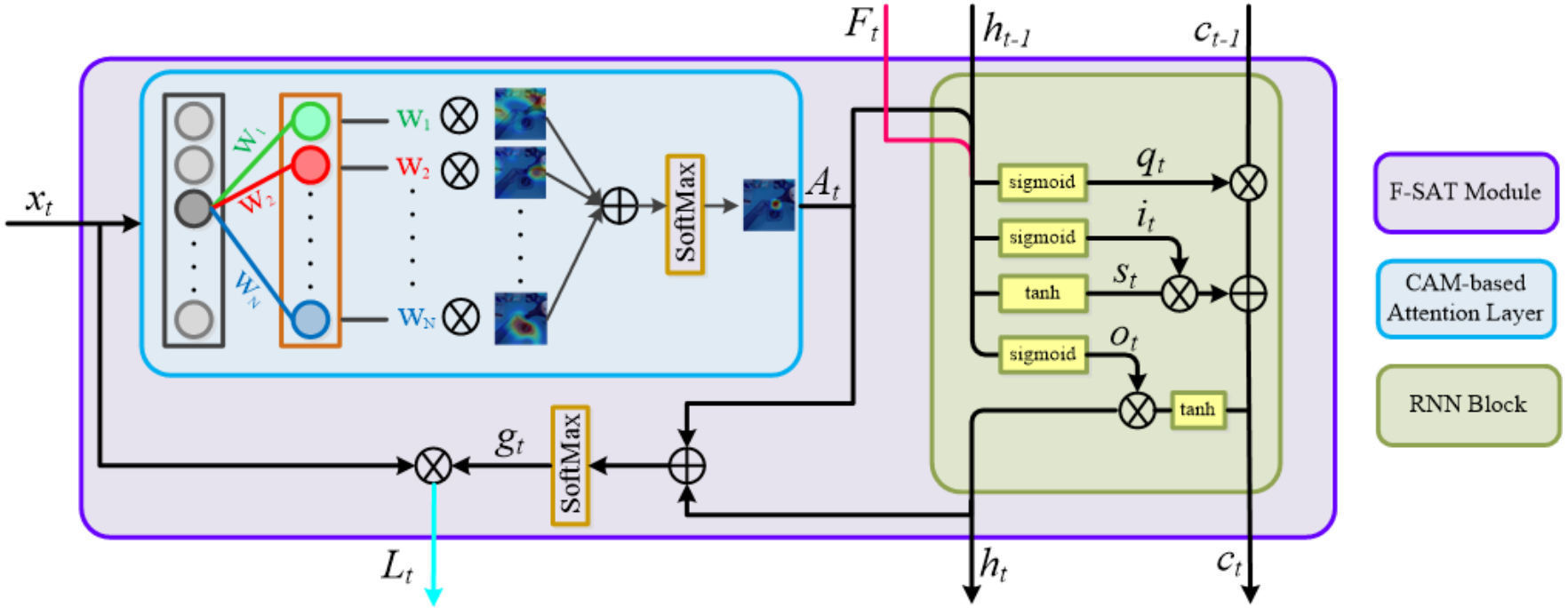
Methodology

- Overall Network Architecture



Methodology

- Flow-guided Spatial Attention Tracking Module



Methodology

➢ Flow-guided Spatial Attention Tracking Module

- Class activation map (CAM) [5]:

$$\mathbf{A}_t^c(i) = \sum_{n=1}^{N} w_n^c \, \mathbf{x}_t^n(i)$$

- Flow signal integrated RNN unit:

$$(\mathbf{i}_t, \mathbf{o}_t, \mathbf{q}_t, \mathbf{s}_t) = (\sigma, \sigma, \sigma, \eta)(\mathbf{W} * \mathbf{A}_t + \mathbf{U} * \mathbf{F}_t + \mathbf{V} * \mathbf{h}_{t-1} + \mathbf{b})$$

$$\mathbf{c}_t = \mathbf{i}_t \odot \mathbf{s}_t + \mathbf{q}_t \odot \mathbf{c}_{t-1}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \eta\,(\mathbf{c}_t)$$

Experiments

- Datasets
  - GTEA 61 : 61 activity classes
  - GTEA 71 : 71 activity classes
  - EGTEA Gaze+: 10,325 samples & 106 activity classes

Experiments

➢ Ablation Study

- ✓ Effectiveness of the F-SAT module
- ✓ Effectiveness of multi-branch fusion

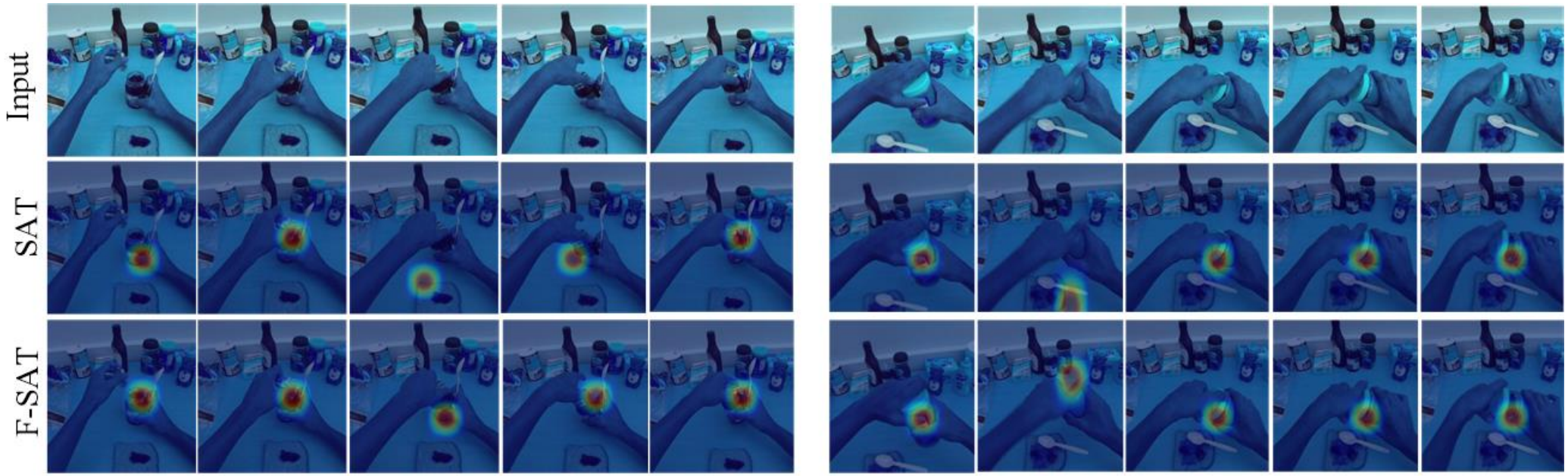
Table I Ablation experiment results on the GTEA 61 data set.

Ablation Setting	Accuracy (%)
Motion branch	46.72
Appearance branch	51.68
Appearance branch (SAT)	73.92
Appearance branch (F-SAT)	78.16
Two-branch (F-SAT)	**81.29**

Experiments

➢ Ablation Study

✓ Visualization of the attention maps generated by SAT and F-SAT



(a) close_jam (b) open_mustard

Experiments

- Comparison with State-of-the-Art Methods

Table II Comparison results on three egocentric activity data sets.

Methods	GTEA 61	GTEA 71	EGTEA Gaze+
DEA [24]	64.00	62.10	46.50
Action+object-Net [7]	73.02	73.24	-
Two-stream model [26]	51.58	49.65	41.84
TSN [25]	69.33	67.23	55.93
EleAttG [21]	66.67	60.83	57.01
Ego-RNN [11]	79.00	77.00	60.76
LSTA-two stream [12]	80.01	78.14	61.86
SAP [22]	-	-	62.70
F-SAT-two stream	**81.29**	**79.02**	**62.78**

Conclusion

- ✓ The proposed F-SAT module is capable of localizing the discriminative features from relevant regions across the frames, by exploring temporal context and integrating optical flow as a guidance signal.
- ✓ We validate the practical effectiveness of the F-SAT module by inserting it into a two-branch-based CNN-LSTM network.

References

[1] Y. Li, M. Liu, and J. M. Rehg, "In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video," *European Conference on Computer Vision (ECCV)*, 2018, pp. 639–655.

[2] M. Ma, H. Fan, and K. M. Kitani, "Going Deeper into First-Person Activity Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1894–1903.

[3] S. Sudhakaran and O. Lanz, "Attention is All We Need: Nailing Down Object-centric Attention for Egocentric Activity Recognition," *British Machine Vision Conference (BMVC)*, 2018.

[4] S. Sudhakaran, S. Escalera, and O. Lanz, "LSTA: Long Short-Term Attention for Egocentric Action Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9946–9955.

Thank You!