Flow-guided Spatial Attention Tracking for Egocentric Activity Recognition

Tianshan Liu and Kin-Man Lam
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University
Email: tianshan.liu@connect.polyu.hk, enkmlam@polyu.edu.hk

Introduction

❑ The Problem & Challenges

- ✓ The objective of egocentric activity recognition is to recognize the human activities targeting the camera wearer (observer).
- ✓ The invisibility of the camera wearer and the presence of ego-motion make the recognition task much more challenging.
- ✓ It is crucial to simultaneously identify hand motion patterns and the manipulated objects.

❑ Related Work & Motivations

- ✓ One potential way is to locate the regions of relevant objects by leveraging large-scale fine-grained annotations [1] [2]. This pipeline is computationally intensive and unfeasible in practice.
- ✓ Ego-RNN [3] generates attention independently based on each frame, without considering temporal consistency.
- ✓ It is not robust to track the spatial attention across the frames by only maintaining the historical information of the RGB modality.

❑ Contributions

- ✓ We propose a *flow-guided spatial attention tracking (F-SAT) module*, which accurately localizes discriminative features of regions of interest across frames.
- ✓ We insert the proposed F-SAT module into a *two-branch-based architecture*, which provides complementary information for egocentric activity recognition.

Methodology

❑ Overall Architecture

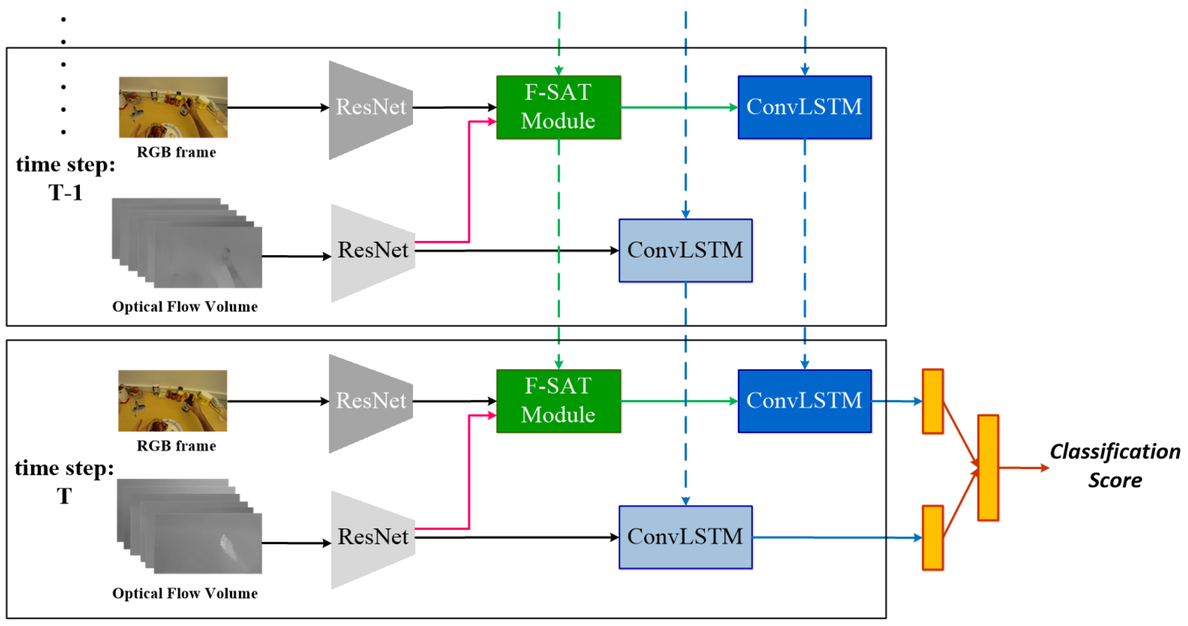


Fig. 1. The overall architecture of the proposed method.

❑ Flow-guided Spatial Attention Tracking Module

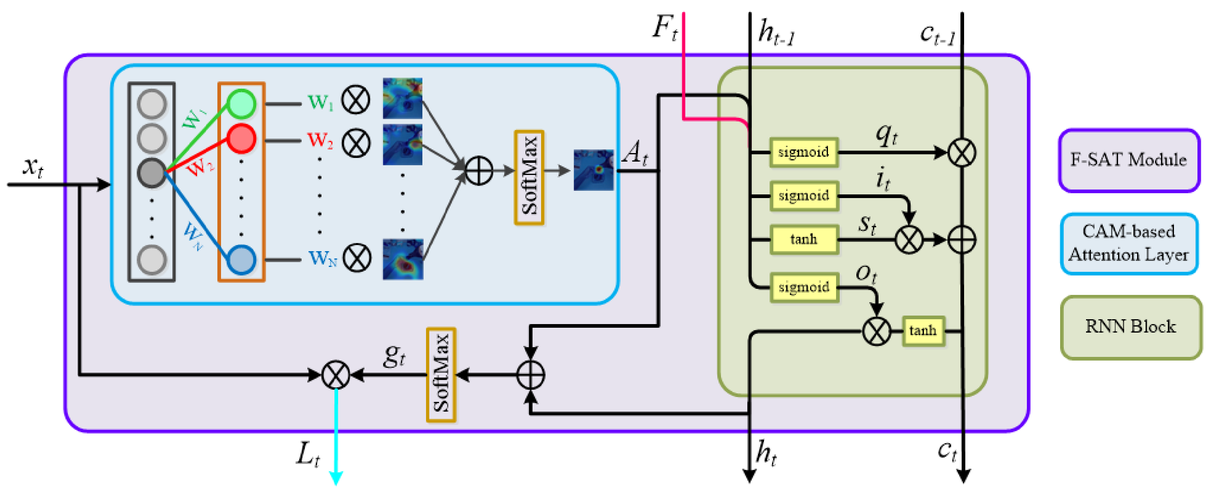


Fig. 2. Schematic diagram of the proposed flow-guided spatial attention tracking (F-SAT) module.

❑ Flow-guided Spatial Attention Tracking Module

- ✓ First stage: we employ a top-down attention mechanism, i.e., class activation map (CAM) [4], to generate a coarse attention map based on the input:

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

- ✓ Second stage: a novel recurrent block, which aims to fine-tune the coarse attention map, is proposed by exploiting contextual information and guidance based on optical flow:

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

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

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

- ✓ Residual connection-based recalibration & feature filtering:

$$\mathbf{g}_t = softmax(\mathbf{A}_t + \mathbf{h}_t)$$

$$\mathbf{L}_t = \mathbf{g}_t \odot \mathbf{x}_t$$

Experimental Results

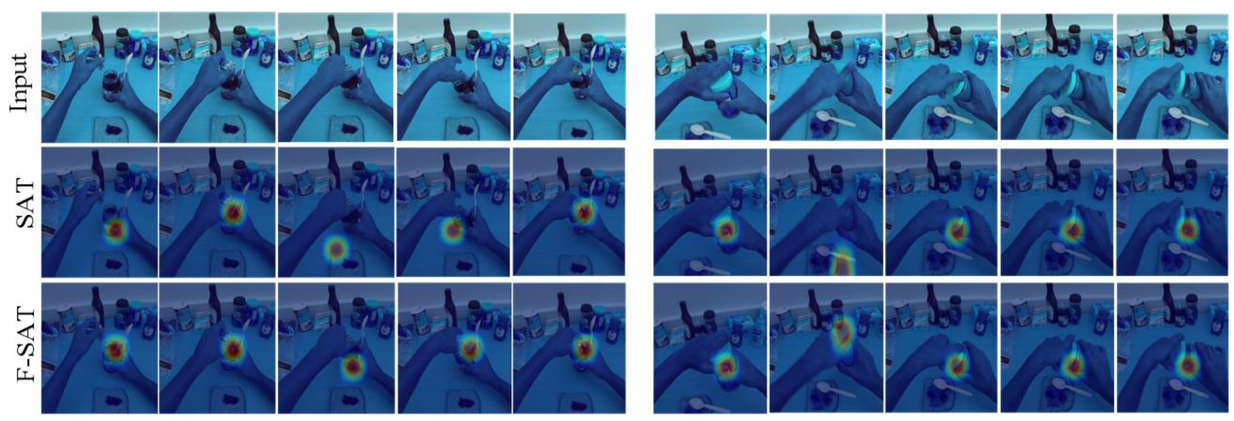
❑ Datasets: GTEA 61, GTEA 71 and EGTEA Gaze+

❑ 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**



(a) close_jam (b) open_mustard

Fig. 3. Visualization of the attention maps generated by SAT and F-SAT on two video sequences.

❑ 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 [5]	64.00	62.10	46.50
Action+object-Net [1]	73.02	73.24	-
Two-stream model [2]	51.58	49.65	41.84
TSN [6]	69.33	67.23	55.93
EleAttG [7]	66.67	60.83	57.01
Ego-RNN [3]	79.00	77.00	60.76
LSTA-two stream [8]	80.01	78.14	61.86
SAP [9]	-	-	62.70
F-SAT-two stream	**81.29**	**79.02**	**62.78**

Conclusion

- ✓ By exploring temporal context and integrating optical flow as a guidance signal, the proposed F-SAT module is capable of highlighting the discriminative features from relevant regions across the frames.
- ✓ We validate the practical effectiveness of the F-SAT module by inserting it into a two-branch-based CNN-LSTM network.
- ✓ Evaluation results on three egocentric activity data sets demonstrate that our method can achieve better performance, compared with state-of-the-art algorithms.

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," in Proceedings of the 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," in 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," in British Machine Vision Conference (BMVC), 2018, pp. 1–12.
[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.
[5] Y. Li, Y. Zhefan, and J. M. Rehg, "Delving into egocentric actions," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 287–295.
[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang and L. V. Gool, "Temporal Segment Networks for Action Recognition in Videos," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 41, no. 11, pp. 2740–2755, 2019.
[7] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "EleAttRNN: Adding Attentiveness to Neurons in Recurrent Neural Networks," IEEE Transactions on Image Processing, vol. 29, pp. 1061–1073, 2020.
[8] S. Sudhakaran, S. Escalera, and O. Lanz, "LSTA: Long Short-Term Attention for Egocentric Action Recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9946–9955.
[9] X. Wang, Y. Wu, L. Zhu, and Y. Yang, "Symbiotic Attention with Privileged Information for Egocentric Action Recognition," in The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI), 2020.