

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, Hong Kong
Email: tianshan.liu@connect.polyu.hk, enkmlam@polyu.edu.hk

Abstract—The popularity of wearable cameras has opened up a new dimension for egocentric activity recognition. While some methods introduce attention mechanisms into deep learning networks to capture fine-grained hand-object interactions, they often neglect exploring the spatio-temporal relationships. Generating spatial attention, without adequately exploiting temporal consistency, will result in potentially sub-optimal performance in the video-based task. In this paper, we propose a flow-guided spatial attention tracking (F-SAT) module, which is based on enhancing motion patterns and inter-frame information, to highlight the discriminative features from regions of interest across a video sequence. A new form of input, namely the optical-flow volume, is presented to provide informative cues from moving parts for spatial attention tracking. The proposed F-SAT module is deployed to a two-branch-based deep architecture, which fuses complementary information for egocentric activity recognition. Experimental results on three egocentric activity benchmarks show that the proposed method achieves state-of-the-art performance.

I. INTRODUCTION

Egocentric, or first-person, activity recognition has attracted increasing attention, due to the popularization of wearable cameras (e.g., GoPro) and the widespread use of some real-world applications [1] [2], such as human-robot interaction, video retrieval, autonomous driving vehicle, etc. Most of the current research on human activity recognition focuses on the third-person perspective [3] [4]. In these kinds of videos, the camera is usually far away from the subjects, without involving the activities. Different from the third-person activity recognition, the objective of egocentric activity recognition is to recognize the human activities targeting the camera wearer (observer) [5]. Thus, the invisibility of the camera wearer and the presence of ego-motion make the recognition task much more challenging.

In this paper, we focus on addressing the problem of fine-grained egocentric activity recognition. Since the activities usually involve complex human-object interactions in the egocentric scenarios, it is crucial to simultaneously identify hand motion patterns and the manipulated objects. One potential way is to locate the regions of relevant objects by leveraging large-scale fine-grained annotations, such as gaze information [6], hand segmentation, and object localization [7]. However, this approach is computationally intensive and unfeasible in practice.

Recent advances in attention mechanisms, combined with deep learning frameworks, have highly benefited various visual tasks [8] [9] [10], especially in the detection of regions of interest. It can also avoid extracting features from irrelevant regions. Some attempts have been made to integrate attention mechanisms into egocentric activity recognition [11]. Since videos consist of sequences of image frames, it is unreasonable to generate attention independently based on each frame, without considering temporal consistency [12]. In other words, simply applying the attention mechanism to still images may degrade the recognition performance of egocentric activities. Although some methods proposed utilizing a sequential fashion to generate attention in each sequential frame, it is not robust to track the spatial attention across the frames by only maintaining the historical information of the RGB modality. As the inaccurate spatial attention occurred in a frame will gradually degrade the accuracy of the whole model, additional guidance is necessary for accurate attention tracking. Motivated by the works in [13] [14], optical flow information can be employed as guidance, because it can capture the information of short-term motion patterns and facilitate spatial alignment between feature maps.

In order to highlight the discriminative features from relevant regions in each frame, we propose a flow-guided spatial attention tracking (F-SAT) module for egocentric activity recognition. The F-SAT module is designed based on a two-stage strategy. In the first stage, we employ a top-down attention mechanism to generate a coarse attention map based on the current RGB frame. In the 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. To provide relevant motion cues from moving parts for spatial attention tracking, we propose a new form of input, namely the optical-flow volume, by stacking the multiple, consecutive optical-flow images around each current frame. The whole network architecture consists of two branches, the appearance branch and motion branch, with the input of RGB frames and optical-flow volume, respectively. Each branch is built based on a hybrid network, by combining Convolutional Neural Networks (ConvNets) with LSTM cells. The overall network architecture is illustrated in Fig. 1.

The main contributions of this paper can be summarized as follows. First, we propose a flow-guided spatial attention

The following publication T. Liu and K. -M. Lam, "Flow-guided Spatial Attention Tracking for Egocentric Activity Recognition," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 4303-4308 is available at <https://doi.org/10.1109/ICPR48806.2021.9412512>.

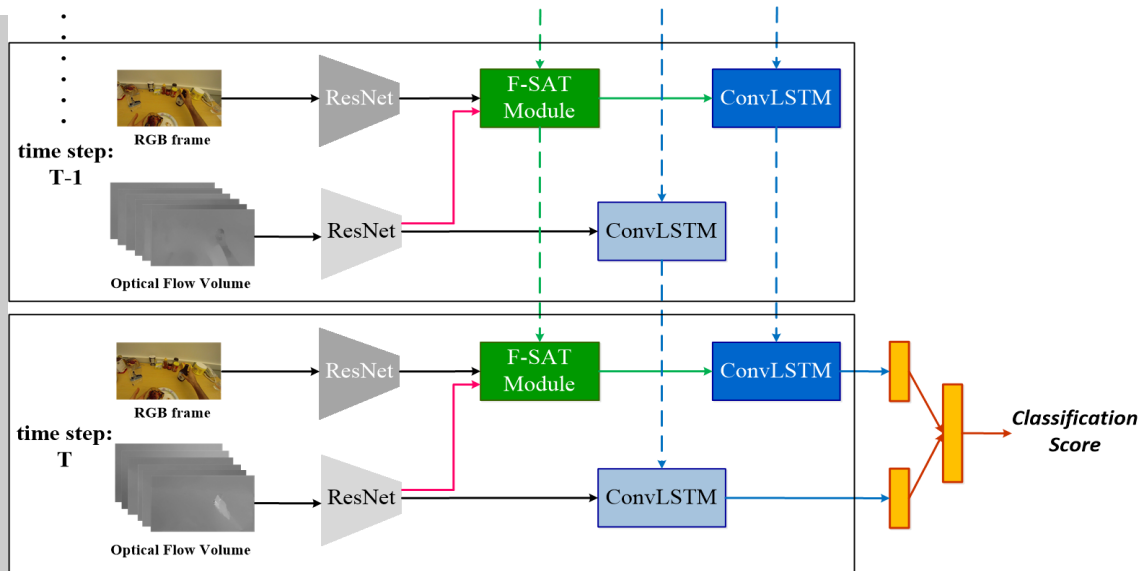


Fig. 1. The overall architecture of the proposed egocentric activity-recognition method.

tracking (F-SAT) module, which accurately localizes discriminative features of regions of interest across frames. Second, we insert the proposed F-SAT module into a two-branch-based architecture, which facilitates the fusion of different modalities and provides complementary information for egocentric activity recognition. Third, experimental results on three public egocentric activity data sets validate that the proposed method outperforms the current state-of-the-art methods.

II. RELATED WORK

Since egocentric activities involve a lot of human-object interactions, an intuitive way is to explore semantic cues, such as gaze information, hand segmentation, and object texture. Fathi et al. [15] presented a mid-level motion feature, based on the extraction of low-level optical flow, and further integrated gaze information [16] for egocentric activity recognition. With the success of deep-learning-based methods in visual tasks, a variety of attempts have been made to employ Convolutional Neural Networks (CNNs) for activity representation and analysis from the egocentric perspective. In order to localize the regions related to actions, Ma et al. [7] pre-trained two specialized CNNs for hand segmentation and object detection, which require a large amount of annotated data, in addition to the video-level labels, for training. Zaki et al. [17] used a temporal feature pooling function to align the features extracted from CNN, and then employed the Fourier Temporal Pyramid (FTP) algorithm to encode long-short-term characteristics. Nevertheless, this method ignores the temporal order of the frames. RNN-based techniques, such as LSTM [18] or ConvLSTM [19], have been widely used to encode temporal information from a sequence of frames. Shen et al. [20] presented a synchronous LSTM module to extract discriminative feature from the informative temporal segments for egocentric activity prediction, which requires gaze direction in the training stage.

Most of these state-of-the-art methods require additional annotated data to locate the regions of interest, which are related to the actions or interacted objects. However, it is impractical and time consuming to manually label all the frames in a video. It is therefore significant for the egocentric activity-recognition methods to be based on weak supervision or self-supervision.

Attention mechanism has been widely used, and is embedded into deep-learning-based frameworks, so as to focus on the features, which are most discriminative or representative for various visual tasks. Li et al. [1] proposed a relation-modeling method for describing human-human interactions. To encode the individual action representations, an attention module was proposed to localize the interactor based on human-segmentation guidance. Sudhakaran et al. [11] proposed the Ego-RNN model, by embedding the class-specific activation-based top-down attention mechanism into a ConvLSTM network for egocentric activity recognition. However, the Ego-RNN model ignores the temporal consistency in a sequence and only generates spatial attention via processing each frame independently. Zhang et al. [21] proposed an element-wise-attention-gate (EleAttG) to assign different levels of importance to each input frame. EleAttG employs a bottom-up approach, so that the weights are fixed after training, thus leading to generating attention without relating to input in the inference phase. Wang et al. presented a symbiotic attention mechanism using privileged information (SAP) [22] and deployed it into a three-stream-based architecture. This method leverages Faster R-CNN to detect position-aware object features via per frame processing. Long short-term attention (LSTA) [12] was proposed by simultaneously employing top-down and sequential methods to overcome the limitations of Ego-RNN and EleAttG. However, LSTA only considers the history of the relevant regions based on RGB

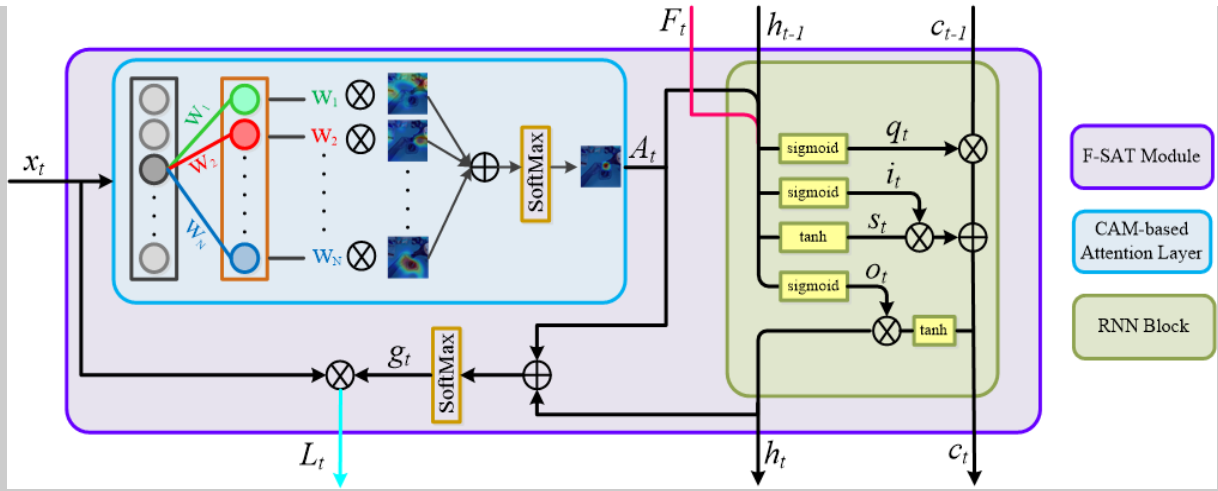


Fig. 2. Schematic diagram of the proposed flow-guided spatial attention tracking (F-SAT) module.

frames to enhance the temporal relationship, which is not robust to tracking spatial attention in complex scenarios. Since optical flow images focus more on moving parts in successive frames and are insensitive to cluttered background noise, our proposed method integrates optical flow as a guidance signal to generate more accurate attention across frames. This flow-guided spatial attention tracking module can be inserted into the two-stream-based architecture for egocentric activity recognition.

III. METHODOLOGY

In this section, the overall network architecture is first introduced to give an overview of our proposed method. Then, we present the concept of optical-flow volume in details. In order to accurately localize the informative spatial regions in sequential frames, a flow-guided spatial attention tracking module is designed and described.

A. Network Architecture

As illustrated in Fig. 1, the overall network is a two-branch-based framework. For the appearance branch, the input is a series of RGB frames from a video clip. At each time step, an individual RGB frame is fed into the ResNet to extract appearance information. To further highlight the features from those relevant spatial regions, a spatial attention tracking module, i.e. the F-SAT module, is applied to generate filtered discriminative features. Then, a convLSTM block is cascaded for temporal encoding. For the motion branch, we stack multiple consecutive optical-flow images around the current frame to generate an optical-flow volume as the input. After extracting the short-term motion information by the ResNet, the optical-flow signal is injected into the spatial attention tracking module in the appearance branch to provide guidance information. Meanwhile, the optical-flow signal is also fed into a convLSTM block for long-term temporal structure encoding. Finally, at the last time step, the features from both two branches are concatenated, followed by a fully-connected (FC) layer, for activity classification.

B. Optical-Flow Volume

To provide information about the relevant motion patterns for spatial attention tracking, we present the construction of an optical-flow volume, which is formed by stacking multiple consecutive optical-flow images around each time step and fed to the appearance branch. Specifically, suppose that a video clip $V = \{I_1^R, I_2^R, \dots, I_K^R\}$, consisting of K RGB frames, is given, and the corresponding optical-flow image sequence is denoted as $O = \{I_1^f, I_2^f, \dots, I_K^f\}$. For the current time step, suppose that the current RGB frame and optical-flow frame are denoted as I_t^R and I_t^f , respectively. The current RGB frame I_t^R is sampled as the input of the appearance branch. To form an optical-flow volume, $2L + 1$ consecutive optical flow images are stacked, with I_t^f at the center. The optical-flow volume is denoted as $OV_t = \{I_{t-L}^f, \dots, I_t^f, \dots, I_{t+L}^f\}$. Note that each optical-flow image consists of 2 channels, one for the x channel (horizontal component) and the other for the y channel (vertical component). The optical-flow volume is designed to provide short-term motion cues around each time step and functions as a guidance signal to facilitate more accurate spatial attention updates in the appearance branch. In addition, the optical-flow volume can be more naturally combined with RNN blocks for long-term motion-pattern modeling.

C. Flow-guided Spatial Attention Tracking Module

To accurately track the discriminative features from relevant regions, a flow-guided spatial attention tracking module is proposed, which mainly consists of a class activation map (CAM)-based attention layer and a newly designed recurrent block. The schematic diagram of this module is illustrated in Fig. 2. Specifically, given a convolutional feature tensor for the appearance branch $\mathbf{x}_t \in \mathbb{R}^{N \times W \times H}$, where t is the time index, N denotes the number of channels, W and H are the width and height, respectively. We first employ a top-down attention mechanism, called the class activation map (CAM) [23], to generate a coarse attention map based on the input.

Spatial average pooling is applied for each feature map of \mathbf{x}_t to obtain N units. Then, the CAM $\mathbf{A}_t^c(i)$ for class c can be calculated as:

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

where $\mathbf{x}_t^n(i)$ represents the activation of the n -th feature map at spatial location i , and w_n^c is the weight for the unit n of class c . Then, the class category with the highest score will be returned, and the corresponding CAM can be selected as the initial spatial attention map. Different from the work in [23] that leverages strong supervision to train the attention module, our method only utilizes video-level label to generate an attention map, without requiring other annotation information.

Considering the fact that video sequences have temporal consistency, the history of the attention maps can be used to achieve smooth attention tracking across the consecutive frames of a video sequence. In addition, optical flow naturally provides informative cues from moving parts in videos, which can be regarded as a guidance signal to align the feature maps of these two modalities. Therefore, a recurrent block is proposed to explore temporal context and integrate optical-flow signals, leading to generation of more accurate spatial attention. Specifically, at each time step, the coarse attention map \mathbf{A}_t and the flow signal \mathbf{F}_t are fed into a RNN unit with memory \mathbf{c}_t and hidden state \mathbf{h}_t . In order to preserve the spatial structure, the convolutional version of the RNN cell is adopted. The recurrent block works as follows:

$$(\mathbf{i}_t, \mathbf{o}_t, \mathbf{q}_t, \mathbf{s}_t) = (\sigma, \sigma, \sigma, \eta) (\mathbf{W} * \mathbf{A}_t + \mathbf{U} * \mathbf{F}_t + \mathbf{V} * \mathbf{h}_{t-1} + \mathbf{b}), \quad (2)$$

$$\mathbf{c}_t = \mathbf{i}_t \odot \mathbf{s}_t + \mathbf{q}_t \odot \mathbf{c}_{t-1}, \quad (3)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \eta(\mathbf{c}_t), \quad (4)$$

where \mathbf{i}_t , \mathbf{o}_t , \mathbf{q}_t and \mathbf{s}_t represent the input gate, output gate, forget gate and update candidate, respectively. σ and η represent the sigmoid and tanh activation functions, respectively. $*$ denotes the convolution operator, and \odot is the Hadamard product. $\{\mathbf{W}, \mathbf{U}, \mathbf{V}, \mathbf{b}\}$ is the parameter set of each RNN block. Then the hidden state is utilized as residual to adjust the coarse attention map \mathbf{A}_t to obtain a more accurate spatial attention. The final attention map \mathbf{g}_t is calculated based on a softmax operation as follows:

$$\mathbf{g}_t = \text{softmax}(\mathbf{A}_t + \mathbf{h}_t). \quad (5)$$

The filtered feature tensor \mathbf{L}_t is obtained by applying the attention map \mathbf{g}_t to \mathbf{x}_t using point-wise multiplication, as follows:

$$\mathbf{L}_t = \mathbf{g}_t \odot \mathbf{x}_t. \quad (6)$$

Then, the filtered feature tensor \mathbf{L}_t is fed to the conventional convLSTM block for further temporal encoding. For egocentric activity recognition, the features highlighted by spatial attention can be from the regions involving hand-object interactions.

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

IV. EXPERIMENTS

A. Data Sets

We evaluate the proposed method on three public egocentric activity-recognition data sets, including GTEA 61, GTEA 71 and EGTEA Gaze+ [24]. GTEA 61 and GTEA 71 contain 61 and 71 activity classes, respectively. EGTEA Gaze+ is a recently collected large-scale data set with 10,325 samples and 106 activity classes. For GTEA 61 and GTEA 71 data sets, we follow the same experimental setup in [12], by using the leave-one-subject-out cross-validation. For EGTEA Gaze+ data set, the averaged accuracy on three splits is reported.

B. Implementation Details

ResNet-34 is chosen as the backbone ConvNets for both the appearance and motion branches. We use a standard convLSTM block with 512 hidden units for temporal encoding. We uniformly sample 20 RGB frames from each video sequence to form the inputs to the appearance branch. We stack 3 consecutive optical flow images (i.e., 6 channels) for each optical-flow volume. The networks are trained by minimizing the cross-entropy loss. The appearance and motion branches are first trained separately. Then we employ a two-stage training strategy to jointly train the combined two-branch-based network. The ResNet-34, used in the appearance branch, is pre-trained on ImageNet. We follow the cross-modality initialization strategy [25] to initialize the motion branch. In the first stage, the F-SAT modules, convLSTM blocks and classifier are trained for 600 epochs. The learning rate is set at 0.001, and is reduced by half after 150, 300 and 450 epochs. In the second training stage, the layers to be trained include the final convolutional layer of ResNet-34 in the appearance branch, all the convolutional layers of ResNet-34 in the motion branch, and the layers trained in the first stage. The learning rate is initialized at 0.01, and decayed by 0.1 after 100 and 200 epochs, with a total of 300 epochs. The ADAM algorithm is adopted for optimizing the parameters of the network. The batch size is set to 32. During the training phase, random horizontal flipping and multi-scale corner cropping approaches are adopted for augmentation, so as to avoid overfitting. The center crop of the frame is used for classification in the inference stage.

C. Ablation Study

To evaluate the effectiveness of the proposed F-SAT module, we conduct ablation experiments using various configurations

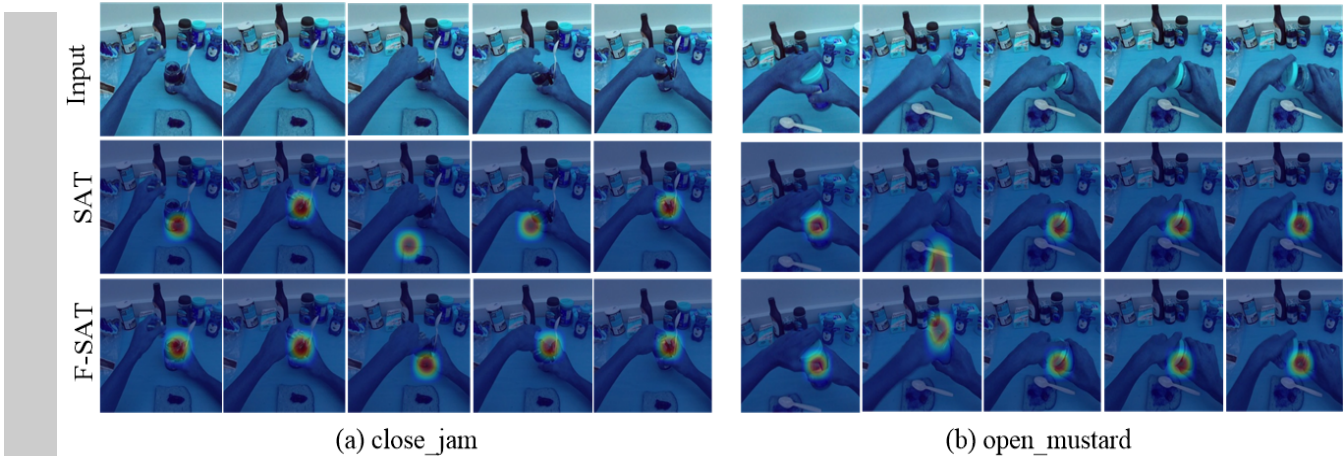


Fig. 3. Visualization of the attention maps generated by SAT and F-SAT on two video sequences.

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

on the GTEA 61 data set. The experimental results are presented in Table I. The motion branch and the appearance branch denote the baseline model, which simply combines ResNet and ConvLSTM. The input of these two networks are the RGB frame and optical-flow volume, respectively. The appearance branch achieves an accuracy of 51.68%. By adding the spatial attention tracking module (without flow interaction) to the appearance branch, a performance gain of 22.24% is obtained. The spatial attention tracking module not only enables the network to localize the informative regions based on the current input frame, but also considers the attention maps generated from the past frames. This leverages the temporal consistency in a video sequence, thus leading to a smooth tracking of the relevant regions. Furthermore, by introducing flow features into the spatial attention tracking module as guidance signal, the appearance branch (F-SAT) achieves a promising accuracy of 78.16%. The features extracted from the optical-flow volume can provide short-term discriminative cues from moving parts, which guides the module to generate more accurate spatial attention. To further intuitively demonstrate the effectiveness of the flow guidance, we visualize the attention maps generated by SAT and F-SAT, as illustrated in Fig. 3. In the “close_jam” sequence, SAT does not track the manipulated object (e.g., the third column in Fig. 3(a)), thereby result in mis-classification. Moreover, we also find that

SAT fails to locate the relevant regions in the “open_mustard” sequence (e.g., the second column in Fig. 3(b)) when severe occlusion occurs in the region of hand-object interactions. F-SAT achieves smooth and accurate attention tracking in these two sequences, which can be largely attributed to the enhancement of the motion patterns provided by optical flow.

Incorporating both the appearance branch (F-SAT) and the motion branch results in a satisfactory accuracy of 81.29%. We can find that, although a single motion branch obtains an accuracy of 46.72% only, combining it with the appearance branch (F-SAT) to form the two-branch-based model can still significantly improve the overall performance. This reveals the fact that fusing the RGB and optical-flow modalities can provide complementary information for egocentric activity recognition.

D. Comparison with State-of-the-Art Methods

The proposed method is compared with other egocentric activity recognition methods, and the results are presented in Table II. We can find that our model outperforms eight state-of-the-art methods on the three egocentric activity data sets. The first two methods listed in Table II utilize additional annotations, such as gaze information [24], object location and hand segmentation [7]. The two-stream model [26] and TSN [25] are classical frameworks, originally proposed for

third-person activity recognition. EleAttG [21], Ego-RNN [11] and LSTA-two stream [12] integrate attention mechanism with RNN cell from different aspects. SAP [22] leverages object detection model to extract position-aware attention without exploiting spatio-temporal relationships. Although EleAttG models the temporal context of a sequence, it simply computes a weight matrix, based on the bottom-up attention mechanism, without being related to the input. The proposed method outperforms EleAttG, because the F-SAT module employs a top-down attention method to generate attention maps based on the input. Ego-RNN considers each frame independently to generate spatial attention, which leads to the loss of temporal consistency. LSTA further improves Ego-RNN by introducing a recurrent unit to maintain the past information of spatial attention. The proposed F-SAT module not only explores temporal context from the RGB modality, but also integrates optical-flow signal to provide informative cues from moving parts. This enables the network to achieve smooth and accurate spatial attention tracking. Therefore, the performance of the proposed method outperforms both Ego-RNN and LSTA.

V. CONCLUSION

In this paper, we propose a flow-guided spatial attention tracking (F-SAT) module for fine-grained egocentric activity recognition. By exploring temporal context and integrating optical flow as a guidance signal, the proposed F-SAT module is capable of localizing the discriminative features from relevant regions across the frames in a video. In addition, optical-flow volume is introduced as a new form of input to provide effective short-term motion patterns for spatial attention tracking for each time step. Furthermore, 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] H. Li, Y. Cai, and W. Zheng, "Deep Dual Relation Modeling for Egocentric Interaction Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7924–7933.
- [2] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Actor and Observer: Joint Modeling of First and Third-Person Videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7396–7404.
- [3] W. Zhang, J. Cen, and H. Zheng, "Temporal Inception Architecture for Action Recognition with Convolutional Neural Networks," in *24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3216–3221.
- [4] J. Zhu, W. Zou, and Z. Zhu, "Two-Stream Gated Fusion ConvNets for Action Recognition," in *24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 597–602.
- [5] M. S. Ryoo and L. Matthies, "First-Person Activity Recognition: Feature, Temporal Structure, and Prediction," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 307–328, 2016.
- [6] 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.
- [7] 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.
- [8] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective Sparse Sampling for Fine-Grained Image Recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6598–6607.
- [9] Z. He, J. Li, D. Liu, H. He, and D. Barber, "Tracking by Animation: Unsupervised Learning of Multi-Object Attentive Trackers," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1318–1327.
- [10] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6077–6086.
- [11] 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.
- [12] 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.
- [13] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal Multiplier Networks for Video Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7445–7454.
- [14] Z. Zhu, W. Wu, W. Zou, and J. Yan, "End-to-End Flow Correlation Tracking with Spatial-Temporal Attention," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 548–557.
- [15] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
- [16] A. Fathi, Y. Li, and J. M. Rehg, "Learning to Recognize Daily Actions Using Gaze," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012, pp. 314–327.
- [17] H. F. M. Zaki, F. Shafait, and A. Mian, "Modeling Sub-Event Dynamics in First-Person Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1619–1628.
- [18] C. Cao, Y. Zhang, Y. Wu, H. Lu, and J. Cheng, "Egocentric Gesture Recognition Using Recurrent 3D Convolutional Neural Networks with Spatiotemporal Transformer Modules," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3783–3791.
- [19] S. Sudhakaran and O. Lanz, "Convolutional Long Short-Term Memory Networks for Recognizing First Person Interactions," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2339–2346.
- [20] Y. Shen, B. Ni, Z. Li, and N. Zhuang, "Egocentric Activity Prediction via Event Modulated Attention," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 202–217.
- [21] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "EleAtt-RNN: Adding Attentiveness to Neurons in Recurrent Neural Networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1061–1073, 2020.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] L. Wang et al., "Temporal Segment Networks for Action Recognition in Videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2740–2755, 2019.
- [26] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems (NIPS)*, 2014, pp. 568–576.