Spatio-Temporal Attention Reasoning for Human Activity Recognition with Privacy Protection

Tianqi Wen[1], Zihang Lyu[1], Yifei Fan[1], Jichen Zhang[1], Hon-Chung Lee[1], Evan Cheng[1], Rui Zhao[1*]
Hayden Crain[2], Van Khai Do[2], Nirosh Rambukkana[2], Alexander Young[2], Xiangjian He[2*]
[1]*Department of Electronic and Information Engineering, The Hong Kong Polytechnic University*
[2]*School of Electrical and Data Engineering, University of Technology Sydney, Australia*
rick10.zhao@connect.polyu.hk, xiangjian.he@uts.edu.au

***Abstract*—Egocentric video human activity recognition is a classic but energetic topic in computer vision applications. Current egocentric video activity recognition methods based on deep learning aim to recognize an activity based on the activity-related objects and the camera movement. In this paper, we go one step deeper to employ the spatial attention reasoning, instead of the object-level reasoning, for enhancing the spatial features. Furthermore, we adopt the second-order attention pooling for enhancing the temporal features. Our proposed spatio-temporal attention network (STANet) combines the attention in both the spatial and the temporal domains, which contributes significantly to the recognition performance. Moreover, egocentric videos usually contain some privacy information. Therefore, we further apply an effective privacy preservation technique to the egocentric videos to mislead general classifiers, while maintaining the video content. We evaluate our method by comparing it with current state-of-the-art methods. Experiment results demonstrate that our proposed STANet can achieve promising performance in egocentric video human activity recognition tasks, making it a potential solution to many practical problems.**

***Index Terms*—Spatio-temporal attention, human activity recognition, first-person vision analysis**

I. Introduction

With the increasing popularity of using wearable cameras to lifelog, a plethora of egocentric or first-person videos (FPVs) have been recorded. Being able to recognize the actions from these FPVs contributes significantly to many practical applications, including assisted living, activity tracking, health monitoring and sport activity analysis, etc. Although a large amount of research has been focused on action recognition in the third-person views, they may not be directly applicable to egocentric videos. This is because egocentric video analysis is mainly based on the object-level reasoning and the movement of camera wearers, instead of the human body postures. Therefore, there is an urgency for the research of first-person vision analysis for human activity recognition.

Previous research on FPV human activity recognition mainly focuses on human-related regions [1], object-level reasoning [2], and spatial-temporal dynamic features [3], etc. In recent years, the attention mapping approaches have attracted more and more attention for use in FPV human activity recognition. Sudhakaran et. al [4] proposed to generate spatial attention maps for locating interesting objects, which are highly related to the egocentric activities. Poleg et. al [5] proposed to make temporal segmentation with average pooling and multi-scale temporal windows to achieve efficient adaptation for activity recognition. Moreover, Cherian et. al [6] proposed an effective second-order pooling operation for generating a temporal attention map which can recognize human actions more accurately. However, only a few of the current studies consider a combination of the attention maps in both the spatial and temporal domain. Therefore, in this VIP Cup competition, we establish a novel FPV human activity recognition system, namely spatio-temporal attention network (STANet), using both the sRGB frames and the optical flows of a video. Specifically, we employ spatial attention based on the class activation maps on the sRGB features, and adopt the second-order temporal pooling for both sRGB and the optical flow feature maps. To evaluate the effectiveness of our method, we compare the proposed STANet with other state-of-the-art features associated with advanced classifiers, such as support vector machine (SVM) and K nearest neighbours (KNN), on the validation set of the FPV-OA database [7].

Since egocentric videos are captured from body cameras, the videos may result in a large amount of privacy or sensitive information leakage, such as human faces and PC screens. To address this issue, we further adopt an effective mechanism to enhance the privacy protection ability of the videos. The mechanism works as a misleader to mislead general object classifiers with unnoticeable changes to the image quality and utility. The objective of a privacy-preserving technique is to make effective but unnoticeable degradation to each frame of a video sequence so that a conventional intelligent model, such as ResNet50 [8], cannot accurately recognize it [9]. In this paper, in order to conceal privacy sensitive information, we apply a similar approach to [10] to all the frames of the training videos in the FPV-OA database. To further evaluate the performance, we randomly sample and apply the mechanism to several frames from the validation set of FPV-OA, and feed them to our pre-trained STANet to evaluate the misleading performance.

The main contributions of this paper can be summarized as follows:

- We propose a new FPV human activity recognition system, namely STANet, to more accurately classify human

IEEE SPS membership– Tianqi Wen: 96069425, Zihang Lyu: 96076666, Rui Zhao: 95740859, Hayden Crain: 96069753, Alexander Young: 96069780, Sean He: 41273744

actions by using spatio-temporal attention.

- We further enhance the performance of STANet based on the feature fusion of sRGB frames and optical flows.
- We adopt an effective mechanism to mislead general classifiers, while maintaining the video content, for protecting privacy information.

II. Proposed STANet for Human Activity Recognition with Privacy Protection

A. Network architecture

As shown in Fig. 1, the whole STANet consists of three blocks, including a sRGB network (sRGBNet), an optical flow network (Optical Flow Net), and a classifier with second-order pooling. The spatial attention and temporal attention are established by the spatial attention modules and a second-order pooling layer, respectively. We employ a 3-stage learning strategy to train up the whole network. Details on the learning strategy will be introduced in Section III.

B. Spatial and temporal attention

1) Spatial attention: The unfolded spatial attention module is shown in Fig. 2. It can be observed that the attention map $\boldsymbol{M}$ of a frame is generated by the weighted sum of its feature maps, and the weights are assigned based on the softmax score of the self-correlation of its fully-connected features. Therefore, the attention map can be formulated as follows:

$$\boldsymbol{M} = \sum_{i=1}^{n} w_i f_i(p), \tag{1}$$

where w_i and f_i are the self-correlation weight and the feature map of the i-th channel, respectively. p refers to a position of the feature map, and n is the total number of channels. Details of this spatial attention module can be found at [4]. Examples of this spatial attention are shown in Fig. 3. It can be seen that our spatial attention map can adaptively locate interesting objects for activity recognition.

2) Temporal attention: The temporal attention is achieved by replacing the conventional mean pooling or max pooling with the second-order pooling in the connection of a feature extractor (convolutional layers) and a classifier (fully-connected layers). Since the original second-order pooling is computationally intensive, in our method we adopt a low-rank approximation of it as follows:

$$\text{score}_{\text{attention}}(\boldsymbol{X}) = \boldsymbol{X}\mathbf{a}^T\boldsymbol{X}\mathbf{b}, \tag{2}$$

where $\boldsymbol{X}$ is the feature map for pooling, and $\mathbf{a}$ and $\mathbf{b}$ are two trainable parameters for temporal attention [12].

C. Privacy preservation mechanism

Privacy protection aims to distort video data to mislead classifier with the degradation perceptually unnoticeable. Therefore, in this work, we apply the private-fast gradient sign method (P-FGSM) [10] to all the frames of the FPV-OA database.

In this VIP Cup competition, since we are dealing with the egocentric videos recorded in office, we consider human faces, computer monitors, and mobile phones, as the key objects, which contain the privacy information. Therefore, our privacy protection mechanism aims to mislead a general classifier, such as ResNet50 [8] pre-trained on ImageNet [11], which can originally recognize these objects. P-FGSM misleads a classifier by defining a transformation T. We can consider this transformation as an adversarial noise generator as follows:

$$I^* = T(I) = I + \delta^*(I), \tag{3}$$

where I is an original frame of a video and δ^* is the adversarial noise generator. We need to optimize δ^* for maximizing the classification error as follows:

$$\delta^* = \arg\max_{\delta} \mathcal{L}_{\text{classifier}}(\theta, I + \delta(I); y), \tag{4}$$

where θ represents the classifier parameters, and y is the ground-truth label. We consider δ as a generative adversarial network and solve the non-convex problem, i.e. Eq. (4), by computing the gradients of the pixels of I, while fixing the classifier. Because of the domain difference between ImageNet and FPV-OA, we further fine-tune the pre-trained ResNet50 based on the FPV-OA images. The final classifier can accurately recognize the video frames, containing human faces, computer monitors, and mobile phones, from the FPV-OA database.

III. Experiments

A. Dataset generation

In this competition, we conducted experiments on the FPV-OA database [7]. FPV-OA consists of the egocentric videos of 12 different subjects with 18 different activities. All videos were captured by a chest-mounted camera with resolution 1280×720 pixels and frame rate 30 fps. Each subject video involves more than 25 activities, including ‘chat’, ‘drink’, ‘microwave’, etc.

We first extract video segments containing a single activity from these 12 videos. We feed the video segments extracted from Subjects 4 to 12 to STANet, with their corresponding labels, for training, and the video segments of Subjects 1 to 3 are used for validation. We further uniformly sample 25 frames from each video segment as the final video sequences. Each frame of a video segment is resized to 456×256, and we further crop the center region of size 224×224 from the resized frames. All sRGB frames are normalized with the respective means equal to 0.485, 0.456, 0.406, and the corresponding standard deviations equal to 0.229, 0.224, 0.225 for the R, G, B channels, respectively. Simple data augmentation operations, including rotation and flipping-over, are applied to the training segments to avoid overfitting. Overall, there are more than 320 segments for training and more than 80 segments for validation.

In terms of optical flow, we randomly sample 5 contiguous frames from a video segment and compute the optical flow for both the vertical and horizontal directions. In testing, instead of randomly sampling, we select the 5 frames located in the centre of a video segment for computing the optical-flow

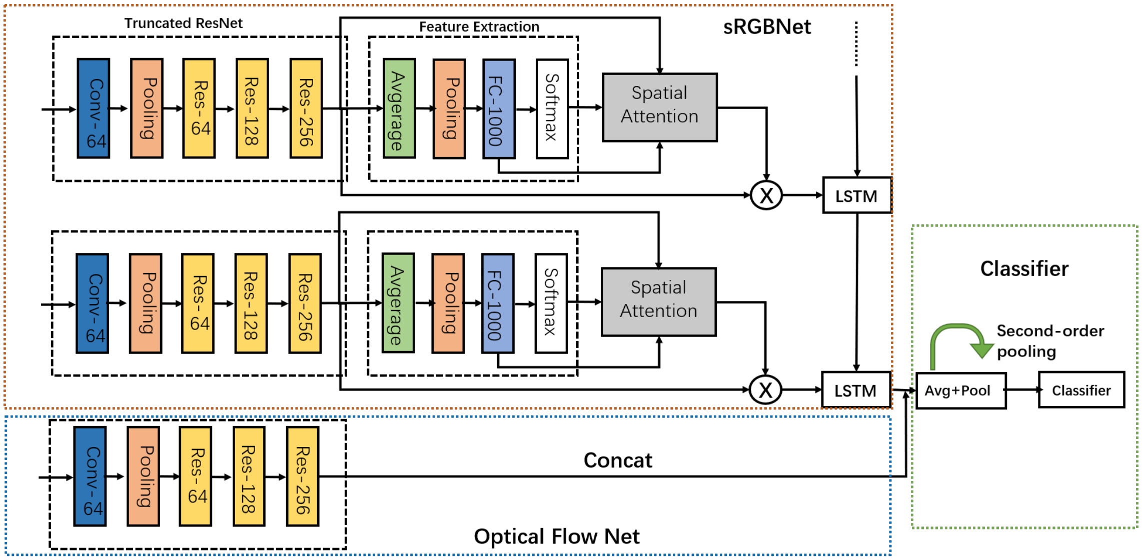


Fig. 1. The architecture of the proposed STANet. 'Res-$*$' represents a residual block with channel number $*$, and 'Conv-$*$' represents a convolutional layer with channel number $*$. LSTM is the convolutional long short-term memory for performing temporal feature encoding. We adopt ResNet34 pre-trained on ImageNet [11] as our backbone network for feature extraction. The generated spatial attention maps are multiplied with the original feature maps to achieve attention extraction. The 'Concat' represents the concatenating operation of the sRGB features and the optical-flow features.

TABLE I
THE PERFORMANCE OF STANET ON BOTH THE TRAINING AND VALIDATION SET OF FPV-OA, WITH AND WITHOUT PRIVACY PROTECTION.

Metric	chat	clean	drink	dryer	mach.	micr.	mobi.	paper	print	read	shake	stap.	take	type.	walk	wash	whit.	write	*Ave.*
Training set																			
Precision	75.0%	100%	92.3%	100%	92.3%	100%	100%	85.0%	100%	81.3%	100%	100%	100%	85.7%	98.5%	100%	88.6%	100%	94.37%
Recall	100%	78.5%	92.3%	100%	92.3%	100%	100%	100%	100%	81.3%	53.8%	100%	100%	80.0%	100%	93.8%	100%	72.7%	91.38%
F-score	85.7%	88.0%	92.3%	100%	92.3%	100%	100%	91.9%	100%	81.3%	70.0%	100%	100%	82.8%	99.3%	96.7%	93.9%	84.2%	92.13%
Validation set																			
Precision	62.5%	100%	25.0%	100%	50.0%	100%	100%	100%	100%	60.0%	0	100%	75.0%	100%	85.7%	100%	71.4%	50%	76.65%
Recall	62.5%	50.0%	33.3%	100%	33.3%	100%	100%	100%	100%	100%	0	100%	100%	66.7%	100%	100%	100%	33.3%	76.62%
F-score	62.5%	66.7%	28.6%	100%	40.0%	100%	100%	100%	100%	75.0%	0	100%	85.7%	80.0%	92.3%	100%	83.3%	40.0%	75.23%
Training set with privacy protection																			
Precision	73.9%	100%	91.7%	100%	85.7%	100%	100%	77.2%	100%	86.7%	87.5%	100%	100%	92.31%	98.5%	100%	88.5%	100%	93.45%
Recall	94.4%	78.6%	84.6%	100%	92.3%	100%	100%	100%	100%	81.2%	53.8%	100%	100%	80.0%	100%	93.7%	100%	72.7%	90.64%
F-score	82.9%	88.0%	88.0%	100%	88.9%	100%	100%	87.2%	100%	83.9%	66.7%	100%	100%	85.7%	99.3%	96.8%	93.9%	84.2%	91.41%
Validation set with privacy protection																			
Precision	62.5%	100%	33.3%	100%	50.0%	100%	100%	100%	100%	37.5%	0	100%	75.0%	100%	81.8%	100%	71.4%	0	72.87%
Recall	62.5%	50.0%	33.3%	100%	33.3%	100%	66.7%	100%	100%	100%	0	100%	100%	33.3	100%	100%	100%	0	71.06%
F-score	62.5%	66.7%	33.3%	100%	40.0%	100%	80.0%	100%	100%	54.55%	0	100%	85.7%	50.0	90.0%	100%	88.3%	0	69.22%

TABLE II
EVALUATION OF THE EFFICACY OF STANET ON THE VALIDATION SET OF FPV-OA. THE PRECISION, RECALL, AND F-SCORE ARE COMPUTED BASED ON THE AVERAGE ACROSS ALL THE CLASSES. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	Precision		Recall		F-score	
	SVM	KNN	SVM	KNN	SVM	KNN
PAF [13]	57%	54%	53%	55%	52%	50%
PAF-GP [13]	61%	51%	51%	53%	53%	50%
RMF [14] + PAF [13]	61%	55%	56%	57%	56%	52%
Ours	**76.65%**		**76.62%**		**75.23%**	

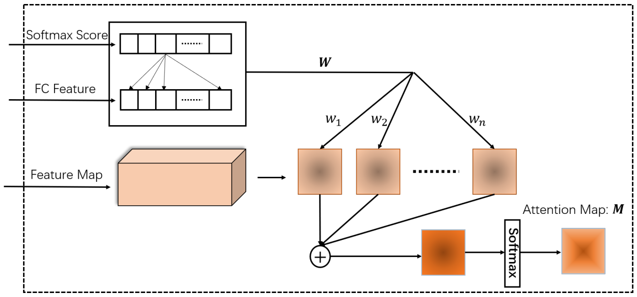
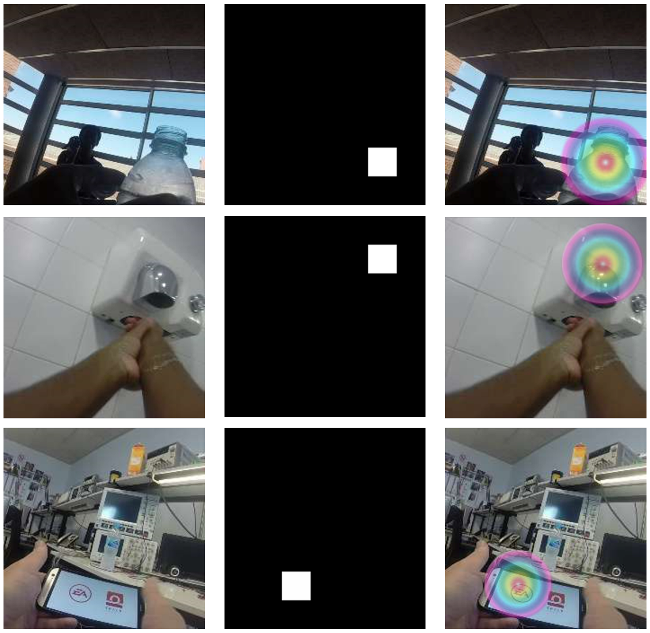


Fig. 2. Illustration of the spatial attention module. FC feature refers to the fully connected feature after average pooling.

features. Therefore, each video segment is associated with a 10-channel optical flow tensor.

B. Network training

To train up the whole STANet system, we divide our training process into 3 stages. In the first stage, we train the sRGB network using the sRGB video segments. In this stage, the second-order pooling is not applied in the classifier. We train the sRGB network using the optimizer in [15] with the weight decay 0.0005 and momentum 0.9 for minimizing the standard cross entropy loss in PyTorch. The learning rate is set to decrease from 10^{-3} to 10^{-7} in 300 epochs.



(a) Input video frames (b) Unnormalized spatial attention maps (c) Normalized spatial attention maps affect on the original frames

Fig. 3. Illustration for the impact of the spatial attention map on the original frames for human activity recognition. The labels of the first row, second row, and third row are ‘drink’,‘dryer’, and ‘mobile’, respectively.

In the second stage, we aim to train up the optical flow network. Similarly to Stage 1, the second-order pooling is not applied in the classifier. We train the network with the same optimizer as Stage 1 for 150 epochs.

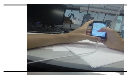
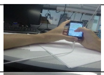
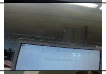
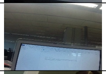
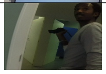
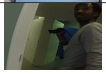
After training up the sRGB network and the optical flow network, we remove their classifiers, i.e. the fully connected layers, and concatenate their output features. In the third stage, we rebuild and train a classifier with the second-order pooling to classify the concatenated features. We train the rebuilt classifier using the same optimizer for another 100 epochs. The whole training process takes about 2 days, with two Nvidia GEFORCE GTX 1080 Ti GPUs.

C. Experimental results

We first investigate the performance of our model on both the training set and the validation set of FPV-OA. Results are listed in the first half of TABLE I. The proposed STANet performs well on basically all the activities, except the ‘shake’ activities. STANet tends to recognize the ‘shake’ activities as ‘chat’, which makes ‘chat’ have a high recall, but a relatively lower precision in training. The normalized confusion matrices of the training set and the validation set are shown in Fig. 4(a) and 4(b), respectively. It can be seen that the proposed STANet achieves promising performance on the FPV human activity recognition tasks as most of the classes are classified with over 70% accuracy.

Furthermore, we also evaluate the performance of the privacy enhancement approach we applied. The results are shown in TABLE III. Obviously, the mechanism effectively misleads the original classifier to enhance the privacy protection. More importantly, the adversarial noise added to the video frames is not perceptually noticeable, which ensures that the performance of STANet will not degrade too much.

TABLE III
THE EVALUATION OF THE PRIVACY PROTECTION MECHANISM FOR MISLEADING THE CLASSIFIER.

Frames	Class label	Probability	Frames	Class label	Probability
Before adding privacy protection			**After adding privacy protection**		
	mobile	0.9641		None	0.9941
	monitor	0.6332		None	0.9958
	human face	0.9470		None	0.9952

To investigate the generalization ability of STANet, we conducted experiments for recognizing the video segments with the privacy protection mechanism described in Section II. The recognition performance is tabulated in the second half of TABLE I, and the corresponding confusion matrices are shown in Fig. 4(c) and 4(d). Generally, the privacy protection mechanism results in a nearly 5% decrease of all the recognition metrics. However, STANet still achieves over 70% in terms of precision and recall, which indicates that STANet achieves its good generalization ability in defending the adversarial attacks.

We further evaluate STANet by comparing it with current state-of-the-art feature extractors and classifiers on the validation set of FPV-OA. Results are listed in TABLE II. It can be clearly observed that our proposed STANet outperforms these state-of-the-art handcrafted features associated with advanced classifiers by a large margin. Specifically, STANet achieves nearly 80% in terms of average precision and recall, and it achieves about 78% for F-score.

IV. CONCLUSION

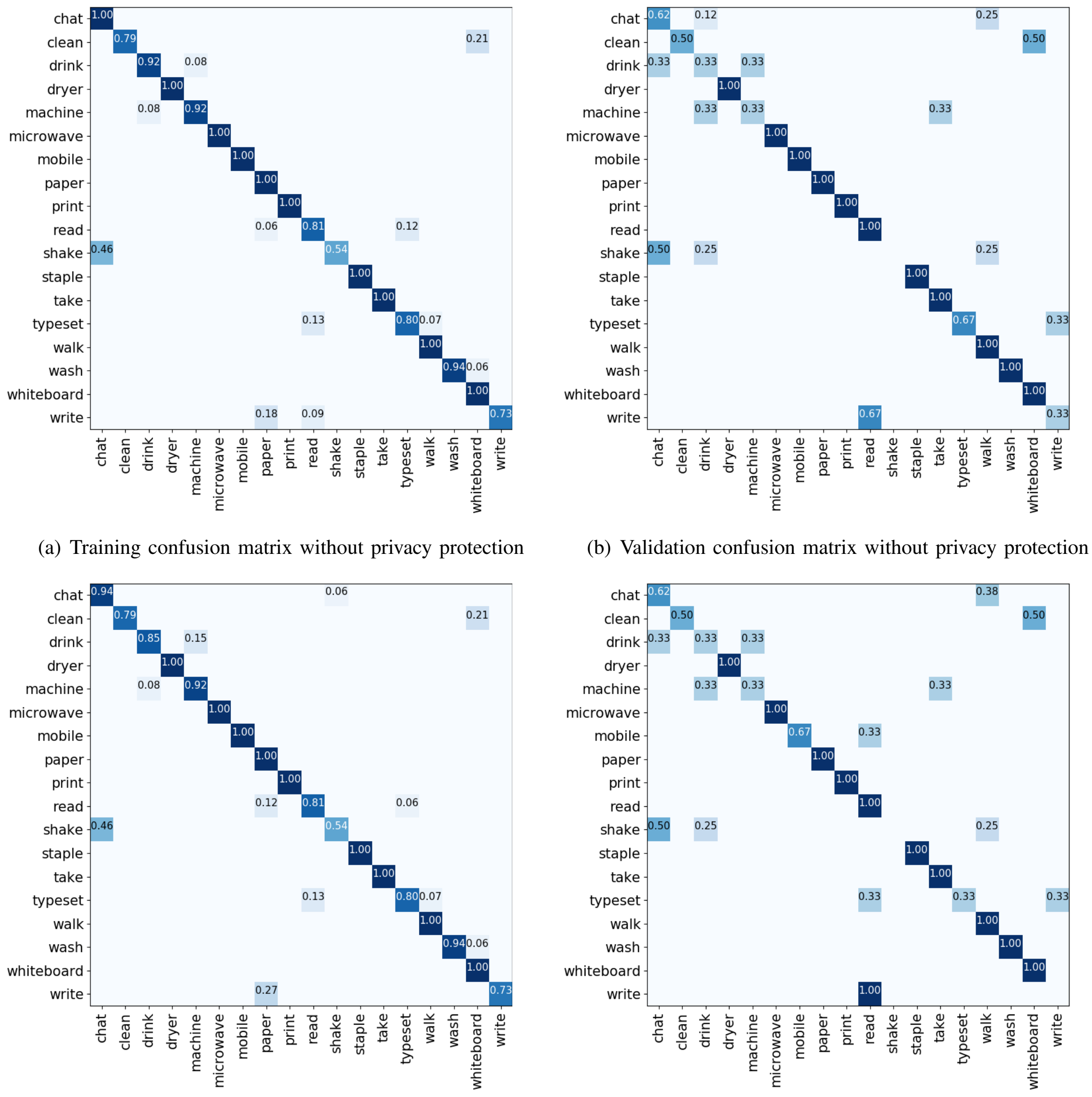
In this paper, we proposed a novel spatio-temporal attention network (STANet) for human activity recognition based on first-person videos. We perform the attention feature extraction in both the spatial and temporal domains to enhance the performance of our model. Furthermore, a privacy protection mechanism for egocentric videos is also established to mislead general classifiers with unnoticeable degradation on the video frames. Our extensive experiments have shown that STANet achieves appealing results on first-person view human activity recognition, which makes it an effective approach to various practical applications.

REFERENCES

[1] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, “Multi-stream cnn: Learning representations based on human-related regions for action recognition,” *Pattern Recognition*, vol. 79, pp. 32 – 43, 2018.

[2] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori, “Object level visual reasoning in videos,” in *ECCV*, 2018.

[3] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib, “Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition,” *Signal Processing: Image Communication*, vol. 71, pp. 76 – 87, 2019.



(a) Training confusion matrix without privacy protection

(b) Validation confusion matrix without privacy protection

(c) Training confusion matrix with privacy protection

(d) Validation confusion matrix with privacy protection

Fig. 4. The normalized confusion matrices of the training and the validation set of FPV-OA, with and without privacy protection. The vertical axis represents the true labels, and the horizontal axis represents the predicted labels.

[4] S. Sudhakaran and O. Lanz, "Attention is all we need: Nailing down object-centric attention for egocentric activity recognition," *CoRR*, vol. abs/1807.11794, 2018.

[5] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks for action recognition in videos," *CoRR*, vol. abs/1705.02953, 2017.

[6] A. Cherian and S. Gould, "Second-order temporal pooling for action recognition," *International Journal of Computer Vision*, vol. 127, pp. 340–362, Apr 2019.

[7] G. Abebe, A. Catala, and A. Cavallaro, "A first-person vision dataset of office activities," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction* (F. Schwenker and S. Scherer, eds.), (Cham), pp. 27–37, Springer International Publishing, 2019.

[8] [online], "Pixel privacy task mediaeval 2018," in *Available: http://www.multimediaeval.org/mediaeval2018/*, 2018.

[9] Yufeng Wang, W. Latif, C. C. Tan, and Yifan Zhang, "Security and privacy for body cameras used in law enforcement," in *2015 IEEE Conference on Communications and Network Security (CNS)*, pp. 173–181, Sep. 2015.

[10] C. Y. Li, A. Shahin Shamsabadi, R. Sanchez-Matilla, R. Mazzon, and A. Cavallaro, "Scene privacy protection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2502–2506, May 2019.

[11] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009.

[12] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," *CoRR*, vol. abs/1711.01467, 2017.

[13] M. S. Ryoo, B. Rothrock, and L. H. Matthies, "Pooled motion features for first-person videos," *CoRR*, vol. abs/1412.6505, 2014.

[14] G. Abebe, A. Cavallaro, and X. Parra, "Robust multi-dimensional motion features for first-person vision activity recognition," *Computer Vision and Image Understanding*, vol. 149, pp. 229 – 248, 2016.

[15] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *in COMPSTAT*, 2010.