Girmaw Abebe Tadesse, Oliver Bent, Lucio Marcenaro, Komminist Weldemariam, and Andrea Cavallaro

Privacy-Aware Human Activity Recognition From a Wearable Camera

Highlights From the IEEE Video and Image Processing Cup 2019 Student Competition

The Video and Image Processing (VIP) Cup is a student competition that takes place each year at the IEEE International Conference on Image Processing. Undergraduate students are encouraged to form teams to work on a specific challenge, which, for the 2019 IEEE VIP Cup, focused on video processing from a chest-mounted camera for the privacy-aware recognition of activities of the wearer.

The motivation behind this challenge is the increasing use of wearable cameras to collect first-person videos (FPVs) that can be used for the recognition of activities at home, in the workplace, and during sporting activities [1], [2]. FPV activity recognition has important applications, which include assisted living, activity tracking, and life logging. The main challenges of FPV activity recognition are the presence of outlier motions (for example, other people captured by the camera), motion blur, illumination changes, and self-occlusions [1]–[3]. Moreover, as videos captured by body cameras may leak sensitive/confidential information about individuals (e.g., bank details), the 2019 IEEE VIP Cup encouraged the design of privacy-enhancing solutions that protect privacy but are not detrimental to the activity recognition performance. An annotated data set of activities from several subjects (see http://www.eecs.qmul.ac.uk/~andrea/fpvo) and a so-called Getting Started code (see https://github.com/girmaw/VIPCUP) were provided [4].

In this article, we present an overview of the 2019 IEEE VIP Cup experience including the competition setup, the teams, and their technical approaches. Finally, we share the experience along with feedback obtained from the finalists.

Tasks, resources, and evaluation criteria

The 2019 IEEE VIP Cup focused on the recognition of 18 activities (see Table 1) in videos captured by a chest-mounted body camera and, through appropriate data transformations, on the preservation of the privacy of the wearer and of subjects captured by the camera. The tasks were:

- *Task 1*: Activity recognition from raw body camera videos
- *Task 2*: Privacy protection in body camera videos
- *Task 3*: Activity recognition from privacy-protected (transformed) body camera videos to evaluate the effectiveness of the privacy protecting methods in maintaining the discriminating characteristics for recognition of the activities.

The "Getting Started" code was provided, which implements optical flow-based [2] and centroid-based [3]

Table 1. A definition of activities considered in the IEEE VIP Cup 2019.

Activity	Label	Definition
Walking	Walk	Walking naturally
Chatting	Chat	Chatting with another person
Shaking hands	Shake	Shaking hands with another person
Reading from paper	Paper	Reading a printed document
Reading from screen	Read	Reading from a computer screen
Smartphone surfing	Mobile	Navigating smartphone apps
Typesetting	Typeset	Typesetting using a computer keyboard
Printing	Print	Taking out a printed document from a printer
Stapling	Staple	Stapling paper sheets using a stapler
Writing on paper	Write	Handwriting using a pen or a pencil
Writing on a board	Whiteboard	Writing on a whiteboard using a marker
Cleaning a board	Clean	Cleaning a whiteboard using a duster
Operating a machine	Machine	Placing an order at a vending machine
Taking	Take	Taking a bottle or can out of a vending machine
Drinking	Drink	Drinking from a bottle, a can, or a cup
Microwave heating	Microwave	Using a microwave oven
Washing	Wash	Washing hands in a sink
Drying	Dry	Using an electric hand dryer

Digital Object Identifier 10.1109/MSP.2020.2976158
Date of current version: 28 April 2020

FIGURE 1. The keyframes of activities from sample videos in the first training and testing data sets: (a) typeset, print, staple, and paper; (b) read, clean, whiteboard, and write; (c) machine, take, open, and drink; (d) mobile, microwave, wash and dry; (e) chat, shake, wave, and walk.

feature extraction from the videos and classification of activities using support vector machines (SVMs) and *k*-nearest neighbors classifiers.

The first training and validation data sets were collected with a chest-mounted GoPro Hero3+ camera with a 1,280 × 720 pixels resolution at 30 frames/s (Figure 1). Nine male and three female subjects participated in the data collection. Each subject recorded a video of approximately 15 min on average, resulting in a total of 3 h of videos. Teams were evaluated in two rounds on two data sets containing the same set of activities that appeared in the first training and validation data set, but with videos recorded in new scenes and with five new subjects for each data set (videos recorded in this different office environment were provided as an additional training data set). The three best-scoring teams (finalists) were selected in the first round using test data set 1, whereas test data set 2 was used to rank the finalists and determine the winner on the final day of the 2019 IEEE VIP Cup.

The classification performance was evaluated using precision (P), recall (R), and the F_1-score (F) for each activity and their average (and standard deviation) across all of the activities. The activity recognition F-score was calculated on the original video data (i.e., before privacy protection is applied), on the privacy-protected video data across all of the activities, and was also used to rank the teams for task 1 and task 3. Task 2 was instead evaluated based on the effectiveness of the automated privacy-preserving technique(s) employed. A privacy-preserving technique was considered effective if it could conceal (from a classifier or a human observer) privacy-sensitive information in the videos, such as, for example, the face of a person or the content of a computer screen which may expose login credentials. Two independent observers ranked the effectiveness and distortion of the protected video content on a five-category scale: 1) no

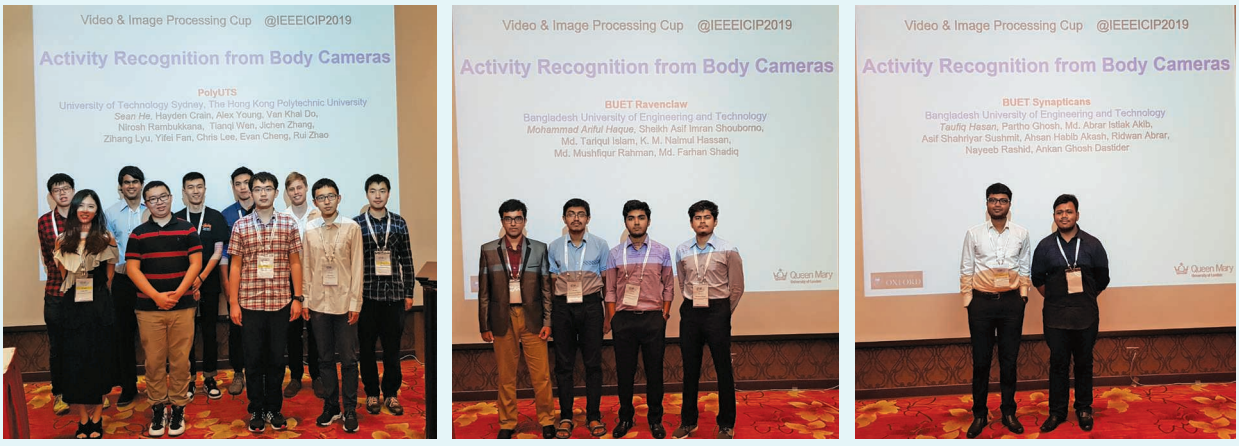

FIGURE 2. The members of the three finalist teams from the 2019 VIP Cup: (a) first place, the PolyUTS Team; (b) second place, the Ravenclaw Team; and (c) third place, the Synapticans Team.

protection when needed (score: 0) 2) effective but disturbing protection (score: 0.25) 3) effective but distracting protection (score: 0.5) 4) effective, noticeable but not distracting protection (score: 0.75) 5) effective and not noticeable protection (score: 1).

The overall score was the average of the *F*-scores for task 1 and task 3 and the privacy score of task 2. Moreover, up to an additional 0.3 score bonus was available for the ranking of the three finalists based on the quality of the presentation and on the applicability of the method used by a team.

The finalists

The finalist teams of the 2019 IEEE VIP Cup (see Figure 2) and their final ranking were:

Team PolyUTS (First place)

- *Affiliations*: University of Technology Sydney, The Hong Kong Polytechnic University
- *Students*: Hayden Crain, Alex Young, Van Khai Do, Nirosh Rambukkana, Tianqi Wen, Jichen Zhang, Zihang Lyu, Yifei Fan, Chris Lee, and Evan Cheng
- *Mentor*: Rui Zhao
- *Supervisor*: Sean He
- *Technical approach*: STANet [5] spatial temporal attention reasoning and classifier misleader via private-fast gradient sign method (P-FGSM) [6].

Team Ravenclaw (Second place)

- *Affiliation*: Bangladesh University of Engineering and Technology
- *Students*: Sheikh Asif Imran Shouborno, Md. Tariqul Islam, K.M. Naimul Hassan, Md. Mushfiqur Rahman, and Md. Farhan Shadiq
- *Supervisor*: Mohammad Ariful Haque
- *Technical approach*: Convolutional neural network (CNN)-based feature extraction [7], with multilayer perceptrons (MLPs), SVM classification, and object-detection (YOLOv3 [8]) with template matching and blurring.

Team Synapticans (Third place)

- *Affiliation*: Bangladesh University of Engineering and Technology
- *Students*: Partho Ghosh, Md. Abrar Istiak Akib, Asif Shahriyar Sushmit, Ahsan Habib Akash, Ridwan Abrar, Nayeeb Rashid, and Ankan Ghosh Dastider
- *Supervisor*: Taufiq Hasan
- *Technical approach*: Recurrent neural networks (RNNs) ensemble [9] (attention module) and Mask R-CNN [10] to identify and blur sensitive (recognizable) parts of a video.

Technical highlights

For the classification of the activities, both handcrafted features (mainly derived from the optical flow) and features extracted from hidden layers of existing deep neural networks were employed. Inception [11], DenseNet [12], and ResNet [13] architectures were considered to extract deep features. Different classifiers based on SVMs [14], MLPs [15], and RNNs [16] were used. To exploit discriminative information available in specific spatiotemporal instances, attention mechanisms were also applied [17], [18]. Ensembles of different models were also used to improve the performance of individual models. As for the training process, a variety of data augmentation techniques were employed, such as rotating, mirroring, and varying image resolution. Furthermore, Mixup [19] was employed, which takes different classes into account during the generation of an augmented sample, thus reducing the instability and improving generalizability of the model for practical problems.

Similarly, a variety of techniques were applied for the privacy-protection task. These methods can be broadly classified as blurring/masking [8], [10] and adversarial [6]. Blurring involves the detection of sensitive objects, such as a computer screen using YOLOv3 [8] and blurring or masking the detected region (see Figure 3). Postprocessing techniques, such as erosion, were applied to smooth the edge of the mask. Adversarial protection involves adding noise, such as using P-FGSM [6], to mislead a classifier.

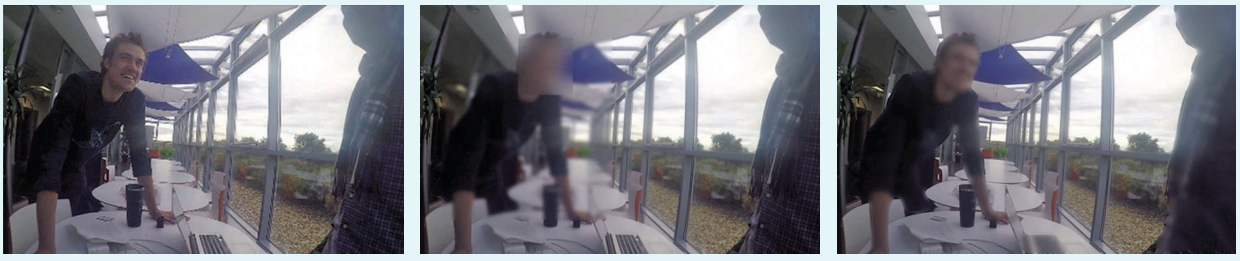

FIGURE 3. A sample privacy-protected frame: (a) Team PolyUTS using an imperceptible adversarial approach [6], and Team Ravenclaw (b) and Team Synapticans (c) using blurring techniques [10].

Finally, the methods were also evaluated in terms of execution time on a randomly selected 5-s video (see Table 2). The major difference in processing time is due to the privacy protection task for which the detection and blurring of sensitive objects is more time-consuming than the adversarial approach.

Summary of the 2019 VIP Cup experience

The teams gave very informative and well-organized presentations of their work on the final day. As for their experience with the VIP Cup, Fan Yifei found that "the competition gave a practical example to understand different network architectures for a specific application as well as the how computation improvements may need to be made for real world applications, in this case leading us to explore further the accelerations provided by Tensorflow." Abrar Istiak Akib observed that "the competition was the first signal processing problem we'd attempted with videos, as previous work in the team had been with static images. Using video was a whole new problem while we found task 2 to be especially novel in this case." Rui Zhao remarked that "Task 2 was certainly the most challenging; reading a paper from the organizers gave clarity for the team on how to mislead a classifier and served as a good starting point." Moreover, Naimul Hassan commented that "we found the competition pushed us outside of our lecture material, being an extracurricular activity to our studies of between 10–15 h a week, online resources helped us to accelerate development while we were able to use a dedicated PC as hardware for the team." As a final remark, Mushfiqur Rahman affirmed that "it was a great experience and platform to learn new things and apply these to real world problems, while I'm also inspired to continue to participate in these types of competitions."

Table 2. The time elapsed (in seconds) for each task for a sample video of 5-s.

Team	Task 1/3	Task 2
PolyUTS	30.31	32.67
Ravenclaw	47.25	649.16
Synapticans	94.42	1,868.34

Acknowledgments

The organizers of the 2019 VIP Cup would like to express their utmost gratitude to all who made this adventure a reality, including, but not limited to, the participating teams, the local organizers, and IEEE Signal Processing Society Membership Board. In addition, Andrea Cavallaro wishes to thank the Alan Turing Institute (EP/N510129/1), which is funded by the Engineering and Physical Sciences Research Council, for its support through the project PRIMULA.

Authors

Girmaw Abebe Tadesse (girmaw.abebe.tadesse@ibm.com) received his Ph.D. degree in electronic engineering from Queen Mary University of London under the Erasmus Mundus Double Doctorate Degree Program in Interactive and Cognitive Environments. He is currently a research scientist at IBM Research Africa. He was previously a postdoctoral researcher at the University of Oxford. He is a Member of the IEEE.

Oliver Bent (oetbent@robots.ox.ac.uk) received his M.Eng. degree in engineering science from the University of Oxford and is currently completing his Ph.D. degree with the Machine Learning Research Group under the supervision of Prof. Stephen Roberts.

Lucio Marcenaro (lucio.marcenaro@unige.it) is an assistant professor in the Department of Electrical, Electronics, and Telecommunication Engineering and Naval Architecture, University of Genoa, Italy. He chairs the Student Services Committee of the IEEE Signal Processing Society, supporting the Video and Image Processing Cup and the IEEE Signal Processing Cup. He is a Member of the IEEE.

Komminist Weldemariam (k.weldemariam@ke.ibm.com) received his B.S. degree from Addis Ababa University, Ethiopia, his M.Tech. degree from the Indian Institute of Technology, Bombay, and his Ph.D. degree from the University of Trento, Italy, all in computer science. He is a Master Inventor (holding more than 150 patents and applications) and is the chief scientist of IBM Research Africa with responsibilities for overseeing scientific and technical direction. He is a fellow of the Next Einstein Forum and was honored as an Emerging Young Scientist by the World Economic Forum.

Andrea Cavallaro (a.cavallaro@qmul.ac.uk) is a professor of multimedia signal processing at Queen Mary University of London (QMUL) and a Turing fellow with the Alan Turing Institute. He is the founding director of the Center for Intelligent Sensing at QMUL and fellow of the International Association for Pattern Recognition. He is a Member of the IEEE.

References

[1] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora, "Compact CNN for indexing egocentric videos," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, New York, Mar. 2016, pp. 1–9. doi: 10.1109/WACV.2016.7477708.

[2] G. Abebe, A. Cavallaro, and X. Parra, "Robust multi-dimensional motion features for first-person vision activity recognition," *Comput. Vis. Image Underst.*, vol. 149, no. C, pp. 229–248, Aug. 2016. doi: 10.1016/j.cviu.2015.10.015.

[3] G. Abebe and A. Cavallaro, "Hierarchical modeling for first-person vision activity recognition," *Neurocomputing*, vol. 267, pp. 362–377, Dec. 2017. doi: 10.1016/j.neucom.2017.06.015.

[4] G. A. Tadesse and A. Cavallaro, "Visual features for ego-centric activity recognition: A survey," in *Proc. 4th ACM Workshop Wearable Systems and Applications, se WearSys '18*, 2018, pp. 48–53. doi: 10.1145/3211960.3211978.

[5] L. Wen et al., Drone-based joint density map estimation, localization and tracking with space-time multi-scale attention network. 2019. [Online]. Available: https://arxiv.org/abs/1912.01811

[6] C. Li, A. Shamsabadi, R. Sanchez-Matilla, R. Mazzon, and A. Cavallaro, "Scene privacy protection," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Procesing (ICASSP)*, Brighton, U.K., May 2019, pp. 2502–2506. doi: 10.1109/ICASSP.2019.8682225.

[7] G. Abebe and A. Cavallaro, "A long short-term memory convolutional neural network for first-person vision activity recognition," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2017, pp. 1339–1346. doi: 10.1109/ICCVW.2017.159.

[8] J. Redmon and A. Farhadi, Yolov3: An incremental improvement. 2018. [Online]. Available: https://arxiv.org/abs/1912.01811

[9] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *Proc. 15th Annu. Conf. Int. Speech Communication Association*, 2014, pp. 1915–1919.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 2961–2969. doi: 10.1109/ICCV.2017.322.

[11] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Boston, June 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708. doi: 10.1109/CVPR.2017.243.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[14] W. S. Noble, "What is a support vector machine?" *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006. doi: 10.1038/nbt1206-1565.

[15] H. Taud and J. Mas, "Multilayer perceptron (MLP)," in *Geomatic Approaches for Modeling Land Change Scenarios*, M. Camacho Olmedo, M. Paegelow, J. F. Mas, and F. Escobar, Eds. New York: Springer-Verlag, 2018, pp. 451–455.

[16] Z. C. Lipton, J. Berkowitz, and C. Elkan, A critical review of recurrent neural networks for sequence learning. 2015. [Online]. Available: https://arxiv.org/abs/1506.00019

[17] Z. Lin et al., A structured self-attentive sentence embedding. 2017. [Online]. Available: https://arxiv.org/abs/1703.03130

[18] S. Sudhakaran and O. Lanz, Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. 2018. [Online]. Available: https://arxiv.org/abs/1807.11794

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, Mixup: Beyond empirical risk minimization. 2017. [Online]. Available: https://arxiv.org/abs/1710.09412

SP