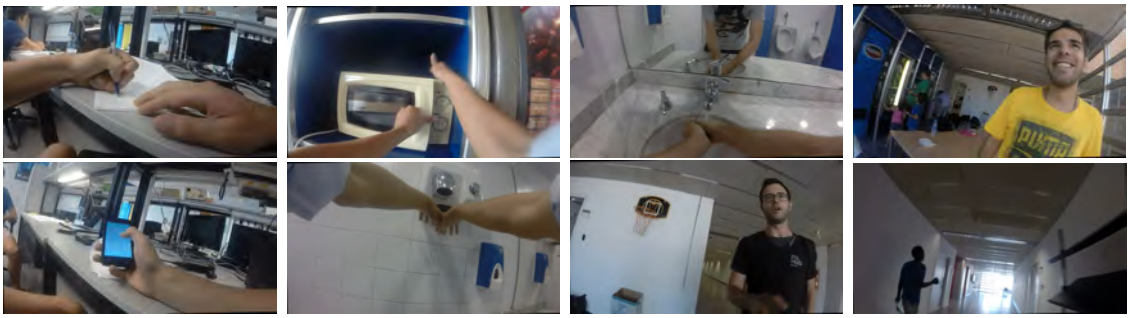
2019 IEEE Video and Image Processing Cup
(VIP-Cup)



Activity Recognition from Body Cameras

Supported by:

IEEE Signal Processing Society (SPS)
Computational Health Informatics Group, Oxford University
IBM Research Africa
Centre for Intelligent Sensing, Queen Mary University of London

2019 IEEE Video and Image Processing (VIP) Cup

Activity Recognition from Body Cameras

Organisers

Girmaw Abebe Tadesse, University of Oxford (girmaw.abebe@eng.ox.ac.uk)
Oliver Bent, University of Oxford
Kommy Weldemariam, IBM Research Africa
Andrea Cavallaro, Queen Mary University of London

1 Introduction

The increasing availability of wearable cameras enables the collection of first-person videos (FPV) for the recognition of activities at home, in the workplace and during sport activities. 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 due to other people captured by the camera), motion blur, illumination changes and self-occlusions.

The 2019 VIP-Cup challenge focuses on FPV from a chest-mounted camera and on the privacy-aware recognition of activities, which include generic activities, such as *walking*, person-to-person interactions, such as *chatting* and *handshaking*, and person-to-object interactions, such as *using a computer* or *a whiteboard*. As videos captured by body cameras may leak private or sensitive information about individuals, the evaluation of the IEEE VIP-Cup challenge entries will include *privacy enhancing solutions* jointly with the recognition performance.

A dataset of activities from several subjects is provided with the annotation for training and validation (`http://www.eecs.qmul.ac.uk/~andrea/fpvo`). The evaluation will be performed based on separate test datasets.

2 Schedule

- April, 30: Participation Guidelines and Initial Training Dataset released;
- May, 5: Additional Training Dataset released
- June, 15: Submission deadline
- July, 15: Finalists (best three teams) announced
- July, 30: Test Dataset 1 released
- September, 22: Competition on Test Dataset 2 at ICIP 2019

3 Eligibility

We invite teams satisfying the following eligibility criteria to participate in the VIP-Cup. Each team must be composed of:

- One faculty member[1] (the Supervisor);
- At most one graduate student[2] (the Tutor);
- At least three but no more than ten undergraduate students[3] (the Team Members).
  - All undergraduate students are eligible to participate (i.e. third and fourth year undergraduate students are eligible).
  - Students who are in their fourth year of a 5-year program that will culminate in a Master's degree (and do not hold a Bachelors degree) are eligible to participate as regular undergraduate Team Members.
- At least three undergraduate members of a finalist team must be either IEEE Signal Processing Society (SPS) members or SPS student members.

The VIP-Cup is a competition for undergraduate students and therefore Master's students, regardless of the duration of their Bachelor's degree, cannot participate as regular Team Members (one graduate student can however be the Tutor for a team).

[1] Postdocs and research associates are not considered as faculty members

[2] A graduate student is a student having earned at least a 4-year University degree at the time of submission

[3] An undergraduate student is a student without a 4-year degree

4 Tasks

The 2019 IEEE VIP-Cup focuses on the recognition of 18 *activities* in videos from a chest-mounted body camera and, through appropriate changes (transformations, obfuscations, redactions, etc.) in the video frames, on the preservation of the *privacy* of the wearer and of people, objects, or places captured by the body camera.

The list and definition of the activities to recognise is given below.

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 on 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

Methods for *activity* recognition from body cameras use deep learning approaches [1, 2, 3, 4], hand-crafted features [5, 6, 7, 8], or a combination of hand-crafted and deep features and/or transfer learning [9, 10, 11].

Examples of mechanisms for the preservation of the *privacy* in videos include detecting and masking sensitive regions such as faces (or bodies) of individuals, computer screens, keyboards and places such as restrooms [12, 13, 14, 15].

In summary the 2019 IEEE VIP-Cup includes three tasks:

- **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

The dataset, annotation and baseline features of the initial dataset can be downloaded at http://www.eecs.qmul.ac.uk/~andrea/fpvo. A reference source code is provided at: https://github.com/girmaw/VIPCUP, which contains Matlab scripts to implement feature extraction from the continuous videos and classification of activities using support vector machines (SVM) and k-nearest neighbours (KNN) classifiers. These scripts are listed as follows.

- *office_activities_classification_March_2019.m* loads two types of motion features from first-person videos of activities
- *GOF_computation_office.m* computes grid optical flow vectors from videos
- *goff_feature_extraction.m* extracts optical-flow based features, both in time and frequency domains
- *centroid_computation_office.m* computes the intensity centroid per each frame
- *image_moments.m* computes the first-order image moments that are necessary to find the intensity centroid for each frame
- *virtual_inertial_feature_extraction.m* extracts virtual-inertial features from the displacement of intensity centroid across frames
- *arrange_train_test_office.m* takes the available data, apply train-test split, train and test two classifiers (SVM and KNN), and return the results.

5 Datasets

5.1 Training and Validation dataset

The train and validation dataset was collected using a chest-mounted GoPro Hero3+ camera with a 1280x720 pixels resolution and a 30 fps frame rate. Nine male and three female subjects participated in the data collection. Each subject recorded a video of approximately 15 minutes on average, resulting in a total of 3 hours of videos. The contribution of each subject S_i for each class in terms of number of frames is shown below[4] (note that a few subjects did not perform some activities).

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	**Total**
Chat	550	1765	1722	1160	1420	2368	3419	1750	2045	3830	423	347	**20799**
Clean	499	243	236	671	597	1010	928	873	250	0	435	202	**5944**
Drink	338	216	145	101	207	288	145	555	88	179	52	35	**2349**
Dryer	646	1040	540	848	854	641	663	506	1060	1003	1453	0	**9254**
Machine	615	87	182	138	324	266	142	491	93	87	134	131	**2690**
Microwave	138	569	614	587	887	698	686	564	850	308	280	347	**6528**
Mobile	1818	2050	1061	0	2173	2697	2098	1972	3234	0	0	2781	**19884**
Open	104	0	0	0	0	0	0	0	42	0	0	0	**146**
Paper	470	1133	989	1510	2152	2415	2428	2590	3402	3405	1831	1921	**24246**
Print	149	108	153	98	110	155	108	110	121	119	109	0	**1340**
Read	938	1477	959	2446	1034	1792	2638	2590	1564	3159	4046	2242	**24885**
Shake	165	168	163	145	90	156	134	123	95	152	123	96	**1610**
Staple	33	249	63	249	271	454	105	99	194	129	0	0	**1846**
Take	191	138	93	99	111	147	169	108	109	163	147	85	**1560**
Typeset	1807	1319	1241	1079	1139	1888	3255	3090	3537	3114	3945	3147	**28561**
Walk	1116	1350	1292	2116	1872	2890	1707	1280	1200	2219	2023	2366	**21431**
Wash	474	464	213	234	683	567	427	363	700	333	525	0	**4983**
Wave	222	47	60	212	60	267	465	43	70	0	0	0	**1446**
Whiteboard	1043	2256	621	1905	1941	1828	2479	2525	3083	2821	2917	2448	**25867**
Write	1218	1648	1360	1412	0	1864	1785	2197	3294	2650	2634	2478	**22540**
Total	**12534**	**16327**	**11707**	**15010**	**15925**	**22391**	**23781**	**21829**	**25031**	**23671**	**21077**	**18626**	**227909**

Please place the videos downloaded from http://www.eecs.qmul.ac.uk/~andrea/fpvo in a single directory. Each video is labelled after the ID of the subject that participated in the data collection. The number of video segments (# segments) per subject, S_i (M: male; F: female), and the overall duration (Dur.) in minutes (min) is given below.

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	**Total**
Gender	M	M	M	F	M	M	M	F	M	F	M	M	**12**
# segments	32	27	28	29	32	29	27	36	25	26	27	26	**344**
Dur. (min)	11	12	11	12	12	16	17	17	17	20	19	17	**181**

Participants are required to use the first three subjects for validation and use the remaining nine subjects for training.

The annotation includes Start and End times of each class segment for each video sequence. The ground-truth labels, generated using ELAN [16], are available as

[4]Due to their significantly lower number of samples, *Open* and *Wave* are excluded from the official evaluation of the challenge. However, Teams are welcome to employ their strategy to alleviate the data imbalance problem in the dataset.

gt_office.mat, which is a structure containing arrays for start time (in milliseconds), final time (in milliseconds) and corresponding label (e.g. '*typeset*') for each segment of a subject. After loading *gt_office.mat* in Matlab workspace, the structure will be as follows:

```
gt_office =

  struct with fields:

    subject10: [1×1 struct]
    subject11: [1×1 struct]
    subject12: [1×1 struct]
     subject1: [1×1 struct]
     subject2: [1×1 struct]
     subject3: [1×1 struct]
     subject4: [1×1 struct]
     subject5: [1×1 struct]
     subject6: [1×1 struct]
     subject7: [1×1 struct]
     subject8: [1×1 struct]
     subject9: [1×1 struct]
```

```
>> gt_office.subject10

ans =

  struct with fields:

    start_time: [26×1 double]
    final_time: [26×1 double]
         label: {26×1 cell}
```

5.2 Testing datasets

Test dataset 1 and *Test dataset 2* will be used to evaluate Teams in two rounds. The best three teams (finalists) will be selected in the first round using *Test dataset 1*, whereas *Test dataset 2* will be used on the final VIP CUP day to rank the finalists and determine the winning Team.

These two datasets contain the same set of activities that appear in the Training and Validation dataset, but the videos are recorded in new scenes and with new subjects (five subject for each dataset). Each subject has a total of 20 minutes of video covering the 18 activities.

Additional Training dataset: Videos from three additional subjects will be provided to the Teams for training before the submission deadline.

6 Evaluation Criteria

6.1 Activity Recognition Score (Task 1 and Task 3)

The classification performance will be evaluated using precision, P, recall, R, and the F score, F. Given true positives, T_P, false positives, F_P, and false negatives,

F_N, for each class, the performance measures are

$$P = \frac{T_P}{T_P + F_P} \tag{1}$$

$$R = \frac{T_P}{T_P + F_N} \tag{2}$$

$$F = 2 \times \frac{P \times R}{P + R} \tag{3}$$

The measures that will be considered for the analysis of the methods are P, R and F for each activity and their average across all the activities (with their standard deviation). The activity recognition score will be generated on the original (raw, non-protected) video data and on the privacy-protected video data. The average F across all the activities will be used as activity recognition score to rank the Teams for Task 1 and Task 3.

6.2 Privacy assessment (Task 2)

Task 2 will be evaluated based on the effectiveness of the automated privacy-preserving technique(s) employed. A privacy-preserving technique is considered effective if it can conceal privacy sensitive information in the videos, such as for example the face of a person or the content of a computer screen. The goal is to mislead a classifier without unnecessarily distorting the content of a frame. What protection mechanism(s) to employ is up to each Team.

For privacy assessment evaluation (Task 2), a panel of observers will rank the effectiveness and distortion of the protected video content on a five-category scale. The five categories are: no protection when needed (score: 0); effective but disturbing protection (score: 0.25); effective but distracting protection (score: 0.5); effective, noticeable but not distracting protection (score: 0.75), effective and not noticeable protection (score: 1).

6.3 Overall assessment

The overall score to rank the Teams is the average of the F-scores for Task 1 and Task 3 and the privacy score of Task 2. Moreover, for the ranking of the three finalists, up to an additional 0.3 score bonus will be assigned based on the quality of the presentation and on the applicability of the method used by a Team.

7 Submission package

Each team shall submit a ".zip" package with all the files included in a single directory named VIPCup_"TeamName".zip, (for example, VIPCup_TeamA.zip), where the name of each file must be one string without spaces. The zip package shall include:

- Confirmation of eligibility in PDF (e.g. *TeamA_eligibility.pdf*). The Supervisor shall sign a letter with the university letterhead listing the participants and confirming the eligibility of each team member.
- Source code (.zip). Participants are free to use the programming language of their choice. All the necessary scripts or functions should be included in a single and zipped directory (e.g. *TeamA_source_code.zip*). Also the trained model is required to be saved in this directory. A *ReadMe* file, which describes how to run these scripts and reproduce the results, should also be included in the source code directory. In addition to activity classification, the source code directory should also include scripts used for privacy protection.
- Executable or script to infer the activity label (e.g. *TeamA_classification_executable.exe*), based on the following instructions:
  - A clear instruction to generate the activity prediction on a set of video segments in a test directory, checked to work on Windows and Unix operating systems. Necessary libraries are also required to be clearly listed.
  - The executable file must take the path to the "test-directory-folder" as a first argument. "test-directory-folder" contains the test data, which consists of multiple video segments from previously unseen subjects. Each segment contains single activity, and a subject may have multiple segments per activity.
  - The "test-directory-folder" could be "VIP_Cup19_Test_Data" or any other folder which has similar contents (with different folder names).
- Executable or script to generate protected videos (e.g. *TeamA_privacy_executable.exe*), based on the following instructions:
  - A clear instruction to generate the privacy-protected video version for each video segment in a test directory, with similar specifications (e.g. frame rate and resolution), checked to work on Windows and Unix operating systems. Common video playing softwares, e.g. VLC media player, should be able to open and play the generated protected videos. Necessary libraries are also required to be clearly listed.
  - The executable file must take the path to the "test-directory-folder" and "protected-directory-folder" as first and second arguments, respectively. The "test-directory-folder" contains the test video segments, and

"protected-directory-folder" is an output directory that contains the corresponding protected videos.

- Report in PDF. The report file should be formatted as double-column, single-spaced IEEE conference proceeding manuscript[5]; and be no longer than 6 (six) pages. The report should include a description of:
  - context and background information;
  - key ideas and principles of the selected method;
  - a description of the privacy enhancement approach, including what sensitive information the approach aims to protect;
  - any signal/image processing used;
  - the classifier(s) and any machine learning methods used;
  - details on other data sources, trained models, and strategies used for training the model(s);
  - a table with the results on the training and validation dataset and a confusion matrix for all the activities. First, the recognition metrics, P, R and F, should be shown for each activity followed by the average across all the activities. In the confusion matrix, the vertical axis represents the ground truth whereas the horizontal axis represents the predicted classes. This must be done for both the clean data (the original dataset) and on the protected data (the dataset modified with your method to protect privacy).
  - and a discussion on the results.
  - Appropriate references when necessary.

If a team submits more than once (up to a maximum of three times) before the deadline, only the last submission before the deadline will be considered for selecting the finalists.

8 Registration

All submissions will be managed by the IEEE Signal Processing Society, and please register at https://www2.securecms.com/VIPCup/VIPRegistration.asp.

For technical difficulties regarding registration and submission the above address, please contact spcup@cmsworkshops.com.

[5]IEEE - Manuscript Templates for Conference Proceedings can be downloaded form the following link: https://www.ieee.org/conferences/publishing/templates.html

9 Q&A platform

Please join the VIPCup2019 page created on the Piazza to interact with organizers and for your submission.

To join the VIPCup2019 Piazza page:

- Go to `https://piazza.com`.
- Click the "Sign Up" tab on the top left corner of the page, and select "Students Get Stared".
- In the "Search Schools:" type in "IEEE SPS" and select the pop-up option.
- Select Summer 2019 term.
- In Class 1 tab:
  - Write "VIPCUP2019: IEEE Video and Image Processing (VIP) Cup 2019".
  - Password is "sps009".
  - Select "Join as: Student".
  - Click on "Join the Class".
- Check your inbox for your confirmation email, and activate your piazza account for your enrollment in SPS 009.
- Enter the validation code from the confirmation email to access your classes.
- Proceed with setting up your piazza account.

Any questions about the 2019 VIP-CUP should be directed to Girmaw Abebe Tadesse girmaw.abebe@eng.ox.ac.uk

Bibliography

[1] 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.

[2] Y. Han, P. Zhang, T. Zhuo, W. Huang, and Y. Zhang, "Going deeper with two-stream convnets for action recognition in video surveillance," *Pattern Recognition Letters*, vol. 107, pp. 83–90, 2018.

[3] 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.

[4] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora, "Compact CNN for indexing egocentric videos," in *Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV)*, New York, USA, March 2016, pp. 1–9.

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

[6] H. Zhang, L. Li, W. Jia, J. D. Fernstrom, R. J. Sclabassi, Z.-H. Mao, and M. Sun, "Physical activity recognition based on motion in images acquired by a wearable camera," *Neurocomputing*, vol. 74, no. 12, pp. 2184–2192, June 2011.

[7] K. Zhan, S. Faux, and F. Ramos, "Multi-scale conditional random fields for first-person activity recognition on elders and disabled patients," *Pervasive and Mobile Computing*, vol. 16, Part B, pp. 251–267, January 2015.

[8] Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," in *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*, Ohio, USA, June 2014, pp. 2537–2544.

[9] G. Abebe and A. Cavallaro, "Hierarchical modeling for first-person vision activity recognition," *Neurocomputing*, vol. 267, pp. 362–377, December 2017.

[10] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled motion features for first-person videos," in *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, March 2015, pp. 896–904.

[11] G. Abebe and A. Cavallaro, "A long short-term memory convolutional neural network for first-person vision activity recognition," in *Proc. of International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017, pp. 1339–1346.

[12] E. T. Hassan; Rakibul Hasan; Patrick Shaffer; David Crandall; Apu Kapadia, "Cartooning for enhanced privacy in lifelogging and streaming videos," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[13] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia, "Enhancing lifelogging privacy by detecting screens," in *Proc. of Conference on Human Factors in Computing Systems (CHI)*, ser. CHI '16, 2016, pp. 4309–4314.

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

[15] C. Li, A. Shamsabadi, R. Sanchez-Matilla, R. Mazzon, and A. Cavallaro, "Scene privacy protection," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Procesing (ICASSP)*, Brighton, UK, May 2019.

[16] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *5th International Conference on Language Resources and Evaluation (LREC 2006)*, 2006, pp. 1556–1559.