A first-person vision dataset of office activities

Girmaw Abebe[1], Andreu Catala[2], and Andrea Cavallaro[3]

[1] Institute of Biomedical Engineering, University of Oxford, UK
girmaw.abebe@eng.ox.ac.uk
[2] Universitat Politecnica de Catalunya, Barcelona, Spain
andreu.catala@upc.edu
[3] Centre for Intelligent Sensing, Queen Mary University of London, UK
a.cavallaro@qmul.ac.uk

Abstract. We present a multi-subject first-person vision dataset of office activities. The dataset contains the highest number of subjects and activities compared to existing office activity datasets. Office activities include person-to-person interactions, such as chatting and handshaking, person-to-object interactions, such as using a computer or a whiteboard, as well as generic activities such as walking. The videos in the dataset present a number of challenges that, in addition to intra-class differences and inter-class similarities, include frames with illumination changes, motion blur, and lack of texture. Moreover, we present and discuss state-of-the-art features extracted from the dataset and baseline activity recognition results with a number of existing methods. The dataset is provided along with its annotation and the extracted features.

Keywords: Wearable camera · First-person vision · Dataset

1 Introduction

First-person vision (FPV) uses wearable cameras to record a scene from the point of view of the wearer. FPV applications include lifelogging, video summarisation, and activity recognition. Datasets are important to support the development and testing of algorithms and classification pipelines for FPV applications. Publicly available FPV datasets are mainly focused on activities such as cooking [4, 6, 15], sports [2, 7], and ego-activities such as *going upstairs/downstairs* and *walking* [11]. Office-related activity datasets are instead limited in both number and range of activities [3, 9, 10, 14].

The 2-hour office dataset (UTokyo) by Ogaki et al. [10] contains only five activities (*reading a book, watching a video, copying text from screen to screen, writing sentences on paper* and *browsing internet*) performed by five subjects and recorded with a head-mounted camera. Other activities (e.g. *conversing, singing* and *random head motions*) are considered part of a *void* class. Each subject recorded each activity for about two minutes twice, thus resulting 60 videos. An eye-tracker was used in addition to a GoPro Hero camera to help estimate the attention (gaze) of the wearer. The 30-minute NUS first-person dataset by Narayan et al. [9] covers eight interaction activities (*handshake, waving, throwing*

an object, passing an object, open and go through a door, using a cellphone, typing on a keyboard, and *writing on a board/paper*) captured with a head-mounted GoPro camera and from a third-person perspective. The 13-minute life-logging egocentric activities (LENA) dataset by Song et al. [14] contains 13 activities performed by 10 subjects using Google Glass. Each subject recorded two 30-second clips for each activity. The activities are grouped as motion (*walking and running*), social interaction (e.g. *talking on a phone and to people*), office work (*writing, reading, watching videos, browsing internet*), food (*eating and drinking*) and house work. LENA includes different varieties of *walk* activity, which are *walk straight*, *walk back and forth* and *walk up/down*, as different activities, and challenges such as scene and illumination variations.

In this paper, we present FPV-O, a dataset of office activities in first-person vision available at `http://www.eecs.qmul.ac.uk/~andrea/fpvo.html`. FPV-O contains 20 activities performed by 12 subjects with a chest-mounted camera (see Fig. 1). The activities include three person-to-person interaction activities (*chat*, *shake* and *wave*), sixteen person-to-object interaction activities (*clean* and *write on a whiteboard*, *use a microwave*; *use a drink vending machine*, *take a drink from the vending machine*, *open* and *drink*; *use a mobile phone*, *read* and *typeset on a computer*, *take a printed paper*, *staple* and *skim over* and *handwrite on a paper*, *wash hands* and *dry*), and one proprioceptive activity (*walk*). The more stable chest-mount solution is often preferred [2, 8, 20, 21] to the head-mount solution, which is affected by head motion [7, 11]. The larger number of activities and subjects in FPV-O compared to other datasets create classification challenges due to intra-activity differences and inter-activity similarities (see Fig. 2). The dataset is distributed with its annotation and features extracted with both hand-crafted methods and deep learning architectures, as well as classification results with baseline classifiers. FPV-O includes around three hours of videos and contains the highest number of both subjects (12) and activities (20) compared to other office activity datasets. The dataset will be made available with the camera ready paper.

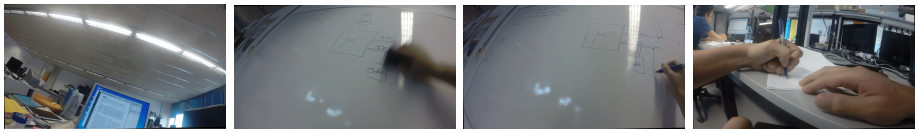
The paper is organized as follows. Section 2 provides the details of the FPV-O dataset, the definition and duration of the activities as well as the contribution of each subject to the dataset. Section 3 describes state-of-the-art features extracted for the recognition of activities in FPV. Section 4 presents the baseline results of the feature groups (and their concatenation) in classifying office activities. Finally, Section 5 concludes the paper.

2 The FPV-O dataset

We used a chest-mounted GoPro Hero3+ camera with 1280x720 resolution and 30 fps frame rate. Twelve subjects (nine male and three female) participated in the data collection. Each subject recorded a continuous video sequence of approximately 15 minutes on average, resulting in a total of 3 hours of videos. The FPV-O activities (see Table 1) extend the scope of existing office activity datasets [9, 10, 14] by including, for example, more object-interactive activities

such as *using a printer, microwave, drink vending machine, stapler, computer*, which are commonly performed in a typical office environment. The number of classes (20) is larger compared to existing office activity datasets [9, 10, 14].

(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

Fig. 1: Keyframes of activities from sample videos in the FPV-O dataset.

The detailed contribution of each subject in the FPV-O dataset for each class is given in Table 3. The annotation includes start and end times of each class segment for each video sequence. The ground-truth labels were generated using ELAN [16].

The challenges in FPV-O include intra-activity differences, i.e. an activity performed differently by subjects due to their pose differences (see Fig. 2). Chest-mounting may also result in different field-of-views for male and female subjects

Table 1: Definition of activities in the FPV-O dataset. P-P: person-to-person interactions; P-O: person-to-object interactions; P: proprioceptive activity.

	Category			
Label	P-O	P-P	P	Definition
Typeset	✓			Typeset using a computer keyboard
Print	✓			Take out a printed paper from a printer
Staple	✓			Staple printed papers using a stapler
Paper	✓			Read a printed paper
Read	✓			Read/navigate on a computer
Clean	✓			Clean a whiteboard using a duster
Whiteboard	✓			Write on a whiteboard using a marker
Write	✓			Handwrite on a paper using pen/pencil
Machine	✓			Place an order on a drink vending machine
Take	✓			Take a bottle/can out of a drink vending machine
Open	✓			Open a bottle/can to drink
Drink	✓			Drink, e.g. from a bottle/can
Mobile	✓			Navigate on smartphone apps
Microwave	✓			Use a microwave
Wash	✓			Wash hands
Dry	✓			Use a hand dryer
Wave		✓		Wave to a colleague
Shake		✓		Shake hands with a colleague
Chat		✓		Chat with a colleague
Walk			✓	Walk naturally

Table 2: Summary of the FPV-O dataset that describes the number of video segments (# segments) per subject, S_i, $i \in \{1, 12\}$, and overall duration (Dur.) in minutes (min). M: male; F: female.

	Subjects												
	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**

(see Fig. 2a). Additional challenges include inter-activity similarities, e.g. *chat* and *shake* (Fig. 2b); *read* and *typeset* (Fig. 2c). Some of inter-activity similarities could be avoided by merging them into a macro activity, e.g. *read* and *typeset* can be merged as *using a computer* activity. Some activities may occur only for a very short duration, e.g. *open* has only 146 frames, whereas *typeset* has $28,561$ frames (see Table 3). This exemplifies class imbalance in FPV-O as data-scarce activities, such as *open*, *wave* and *shake*, have limited information for training. There are also illumination changes as indoor lighting often mixes with daylight (Fig. 2d). Moreover, feature extraction is made challenging by motion blur and lack of texture (Fig. 2e).

Table 3: Details for the contribution of each subject, S_i, $i \in \{1, 12\}$, in FPV-O for each class in number of frames. Note that some activities are not performed by some subjects.

	Subjects												
	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**

3 The features

We selected three frequently employed FPV-based existing methods to extract discriminant features for office activity classification in FPV-O. These are *average pooling* (AP) [17–19], *robust motion features* (RMF) [1, 2] and *pooled appearance features* (PAF) [1, 12], which we describe below.

Let $\mathbf{V} = (V_1, \cdots, V_n, \cdots, V_N)$ be N temporally ordered activity samples from a subject. Each sample, V_n, contains a window of L frames, i.e. $V_n = (f_{n,1}, f_{n,2}, \cdots f_{n,i}, \cdots f_{n,L})$. Successive V_n pairs may overlap. Each of AP, RMF and PAF provides a feature representation for V_n. AP and RMF mainly exploit motion using optical flow, whereas PAF encodes appearance information. Grid optical flow of V_n is $G_n = (g_{n,1}, g_{n,2}, \cdots g_{n,i}, \cdots g_{n,L-1})$, where $g_{n,i} = g^x_{n,i} + jg^y_{n,i}$ represents a flow vector between successive frames, $f_{n,i}$ and $f_{n,i+1}$, $i \in [1, L-1]$. The superscripts x and y represent horizontal and vertical components, respectively. γ is the number of grids in each of horizontal and vertical components, hence results γ^2 grids per frame. AP [18] applies average pooling of each element across $L-1$ grid flow vectors in G_n, which helps discard noise. After smoothing, the final representation of AP is derived as a concatenation of the horizontal and vertical grid components.

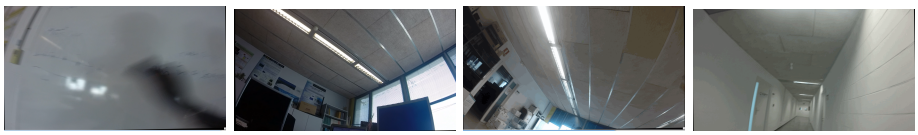
RMF [2] extracts more discriminative features by encoding the direction, magnitude and frequency characteristics of G_n. RMF contains two parts: grid optical flow-based features (GOFF) and centroid-based virtual inertial features (VIF).

(a) Intra-activity differences among four *staple* clips

(b) Inter-activity similarity, e.g. *chat* (top row) and *shake* (bottom row)

(c) Inter-activity similarity, e.g. *read* (top row) and *typeset* (bottom row)

(d) Illumination changes



(e) Lack of texture and motion blur.

Fig. 2: Sample frames that illustrate the challenges in the FPV-O dataset.

GOFF is extracted from the histogram and Fourier transform of motion direction, $G_n^\theta = arctan2(G_n^y/G_n^x)$, and motion magnitude, $|G_n| = \sqrt{|G_n^x|^2 + |G_n^y|^2}$. The histogram representations quantize G_n^θ and $|G_n|$ with β_d and β_m bins, respectively. The frequency representations are derived from grouping the fre-

quency response magnitude of G_n^θ and $|G_n|$ into N_d and N_m bands, respectively. VIF is a virtual-inertial feature extracted from the movement of intensity centroid, $C_n = C_n^x + jC_n^y$, across frames in V_n. The intensity centroid is computed from first-order image moments as $C_n^x = \mathcal{M}_{O1}/\mathcal{M}_{OO}$ and $C_n^y = \mathcal{M}_{1O}/\mathcal{M}_{OO}$. For a frame, $f_{n,i}$, which is H-pixels high and W-pixels wide, its first-order image moments are calculated from the weighted average of all the intensity values as $\mathcal{M}_{pq}^i = \sum_{r=1}^{H}\sum_{c=1}^{W} r^p c^q f_{n,i}(r,c)$, where $p, q \in \{0,1\}$. Once the centroid locations are computed across frames, C_n, then successive temporal derivatives are applied to obtain corresponding velocity, $\dot{C}_n$, and acceleration, $\ddot{C}_n$, components. VIF is extracted from $\dot{C}_n$ and $\ddot{C}_n$ as inertial features from accelerometer and gyroscope data, e.g. *minimum, maximum, energy, kurtosis, zero-crossing* and *low frequency coefficients*.

PAF [12] features are motivated by exploiting appearance features using different pooling operations. We test two types of appearance features: *Overfeat* and *HOG*. Overfeat [13] is a high-level appearance feature that is extracted from the last hidden layer of a deep convolutional neural network -Overfeat [13], which was pretrained with a large image dataset (ImageNet [5]). HOG (histogram of oriented gradients) is a commonly used frame-level appearance descriptor [1]. A simple averaging can be applied for each feature element in HOG and Overfeat across frames to obtain a representation for a video sample, V_n. Gradient pooling (GP) can also be applied to encode variation of the appearance features across frames. The gradient is computed by applying a first-order derivative on each feature element across time, and the pooling operations include sum and histogram of positive and negative gradients [12].

We follow the parameter setups for the corresponding methods as in their authors' choices. Hence, we used $L = 90$ frames (equivalent to three seconds duration) and $\gamma = 20$ for each of horizontal and vertical grid components. Thus AP becomes 800-D. For GOFF of RMF, $\beta_d = 36$, $\beta_m = 15$, $N_d = N_m = 25$, resulting 137-D feature vector. For VIF, we extracted 106-D feature vector that is composed of 60-D frequency features, i.e. 10-D low frequency coefficients from 6 inertial time-series components (2 velocity, 2 acceleration and their magnitude (2)). RMF concatenates GOFF and VIF resulting in 243-D feature vector. For PAF, HOG is 200-D extracted using 5-by-5-by-8 spatial and orientation bins. Overfeat [13] is extracted from the last convolutional layer resulting 4096-D feature.

4 Baseline classification results

In this section, we describe the setups employed to validate different state-of-the-art features on the FPV-O dataset using multiple classifiers. The baseline results, evaluated using various performance metrics, are thoroughly discussed.

We employed support vectors machines (SVM) and k-nearest neighbours (KNN), which are the most frequently employed, respectively, parametric and non-parametric classifiers for activity recognition in FPV [2]. We apply one-vs-all (OVA) strategy for the training of SVM. Since FPV-O consists of 12 subjects,

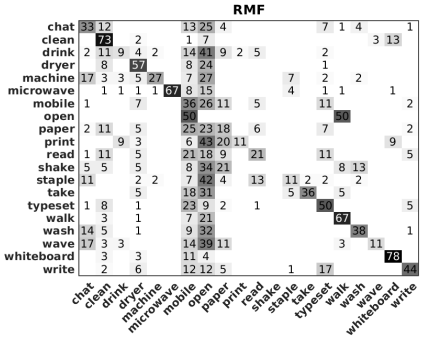
Table 4: Performance of different state-of-the-art features on the FPV-O dataset with an SVM and a KNN classifier. The performance metrics are $\mathcal{P}$: precision; $\mathcal{R}$: recall and $\mathcal{F}$: F-score. Key – AP: average pooling; VIF: virtual inertial features; GOFF: grid optical flow-based features; RMF: robust motion features; PAF: pooled appearance features; GP: gradient pooling applied on the appearance features. RMF [2] + PAF [12] represents a concatenation of existing motion- and appearance-based features.

	SVM			KNN		
Features	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
AP [18]	17	10	10	15	10	9
VIF [2]	24	16	15	21	20	19
GOFF [2]	53	42	44	43	44	41
RMF [2]	51	38	41	45	43	41
PAF [12]	57	53	52	54	55	50
PAF-GP [12]	**61**	51	53	51	53	50
RMF [2] + PAF [12]	**61**	**56**	**56**	**55**	**57**	**52**

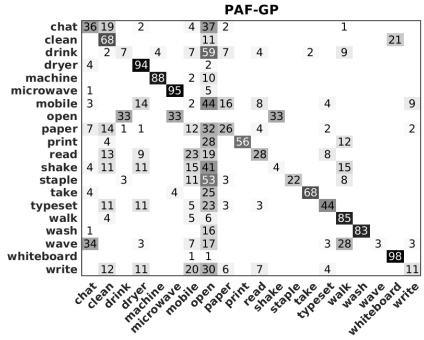
we experiment *one-subject-out* validation, which reserves one subject for testing and uses the remaining for training in each iteration.

We employ precision ($\mathcal{P}$), recall ($\mathcal{R}$) and F-score ($\mathcal{F}$) metrics to evaluate the classification performance. Other performance metrics such as accuracy and specificity are not used as they are less informative of the recognition performance in the OVA strategy [2]. Given true positives (TP), false positive (FP) and false negative (FN), the metrics are computed as $\mathcal{P} = \frac{TP}{TP+FP}$, $\mathcal{R} = \frac{TP}{TP+FN}$ and $\mathcal{F} = \frac{2*\mathcal{P}*\mathcal{R}}{\mathcal{P}+\mathcal{R}}$. For each one-subject-out iteration, $\mathcal{P}$, $\mathcal{R}$ and $\mathcal{F}$ are evaluated for each activity. The final recognition performance is computed by averaging, first, over all activities, and then over all subjects. The confusion matrix is also given to visualize the misclassification among activities. All experiments were conducted using Matlab2014b, i7-3770 CPU @ 3.40GHz, Ubuntu 14.04 OS and 16GB RAM.

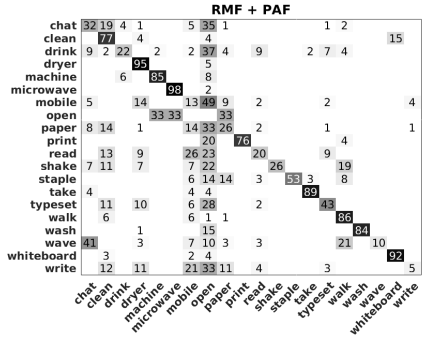
The performance of the selected methods on the FPV-O dataset is shown in Table 4. AP [17, 18] only concatenates smoothed horizontal and vertical grid components, and does not use magnitude and direction information, which is important in this context. As a result, the performance of AP are the lowest both with SVM and KNN. RMF outperforms AP as it encodes motion magnitude, direction and dynamics, using multiple feature groups (GOFF and VIF). The performance of VIF of RMF are inferior to GOFF as the intensity centroid of a frame hardly changes over time since the subjects remain stationary for different activities, e.g. *read*, *mobile* and *typeset*. While both AP and RMF are designed to encode motion information in FPV-O, PAF exploits appearance information that is more discriminative for interaction-based activities. As a result, PAF has the highest performance among the selected methods. PAF achieve equivalent performance with and without gradient pooling (GP) (see Table 4). This also confirms the superiority of appearance information for this dataset as variation encoding using GP did not provide significantly discriminative characteristics.



(a) RMF [2]



(b) PAF-GP [12]



(c) Concatenation of RMF [2] and PAF [12]

Fig. 3: Confusion matrices based on the SVM classifier using RMF [2], PAF-GP [12] and the concatenation of RMF and PAF.

The concatenation of both motion and appearance features outperform all the remaining feature groups.

The confusion matrices shown in Fig. 3 replicate the corresponding performance of motion-based (RMF [2]), appearance-based (PAF-GP [12]) and their concatenation (RMF [2] + PAF [12]). The concatenation of RMF [2] and PAF [12] improved the recognition performance of *drink* from 9% with RMF and 7% with PAF-GP to 22%. The same is true for *print* whose recognition performance was improved from 11% with RMF and 56% with PAF-GP to 76% with the concatenation. On the other hand, the combination of motion and appearance features worsened the recognition performance of *write*. Note also the frequent misclassification with *open* due to the class imbalance problem (see Table 3).

5 Conclusions

We collected, annotated and distributed a dataset of 20 office activities from a first-person vision perspective (FPV-O) at http://www.eecs.qmul.ac.uk/~andrea/fpvo.html. Moreover, we employed and discussed state-of-the-art features extracted using both handcrafted methods and deep neural architectures, and baseline results of different feature groups using SVM and KNN classifiers.

FPV-O covers about three hours of egocentric videos collected by 12 subjects and contains the highest number of office activities (20) compared to existing datasets with similar activities. FPV-O contains challenging intra-activity differences and inter-activity similarities in addition to motion blur and illumination changes. We hope that this dataset and associated baseline results will support and foster research progress in this area of growing interest.

References

1. Abebe, G., Cavallaro, A.: Hierarchical modeling for first-person vision activity recognition. Neurocomputing **267**, 362–377 (December 2017)
2. Abebe, G., Cavallaro, A., Parra, X.: Robust multi-dimensional motion features for first-person vision activity recognition. Computer Vision and Image Understanding (CVIU) **149**, 229 – 248 (2016)
3. Asnaoui, K.E., Hamid, A., Brahim, A., Mohammed, O.: A survey of activity recognition in egocentric lifelogging datasets. In: Proc. of IEEE Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS). pp. 1–8. Fez, Morocco (April 2017)
4. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. arXiv preprint arXiv:1804.02748 (2018)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. Miami, USA (June 2009)
6. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: Proc. of European Conference on Computer Vision (ECCV). pp. 314–327 (2012)
7. Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR). pp. 3241–3248. Colorado, USA (June 2011)
8. Nam, Y., Rho, S., Lee, C.: Physical activity recognition using multiple sensors embedded in a wearable device. ACM Transactions on Embedded Computing Systems **12**(2), 26:1–26:14 (February 2013)
9. Narayan, S., Kankanhalli, M.S., Ramakrishnan, K.R.: Action and interaction recognition in first-person videos. In: Proc. of IEEE Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 526 – 532. Columbus, USA (June 2014)
10. Ogaki, K., Kitani, K.M., Sugano, Y., Sato, Y.: Coupling eye-motion and ego-motion features for first-person activity recognition. In: Proc. of IEEE Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1 – 7. Providence, USA (June 2012)
11. Poleg, Y., Ephrat, A., Peleg, S., Arora, C.: Compact CNN for indexing egocentric videos. In: Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. New York, USA (March 2016)

12. Ryoo, M.S., Rothrock, B., Matthies, L.: Pooled motion features for first-person videos. In: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR). pp. 896–904. Boston, USA (March 2015)
13. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. In: Proc. of International Conference on Learning Representations (ICLR). Banff, Canada (April 2014)
14. Song, S., Chandrasekhar, V., Cheung, N.M., Narayan, S., Li, L., Lim, J.H.: Activity recognition in egocentric life-logging videos. In: Proc. of Asian Conference on Computer Vision (ACCV). pp. 445–458. Singapore (November 2014)
15. Spriggs, E.H., De La Torre, F., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: Proc. of IEEE Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 17–24. Miami, USA (June 2009)
16. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: Elan: a professional framework for multimodality research. In: Proc. International Conference on Language Resources and Evaluation (LREC). pp. 1556–1559. Genoa, Italy (May 2006)
17. Zhan, K., Faux, S., Ramos, F.: Multi-scale conditional random fields for first-person activity recognition. In: Proc. of IEEE International Conference on Pervasive Computing and Communications (PerCom). pp. 51–59. Budapest, Hungary (March 2014)
18. Zhan, K., Faux, S., Ramos, F.: Multi-scale conditional random fields for first-person activity recognition on elders and disabled patients. Pervasive and Mobile Computing **16, Part B**, 251–267 (January 2015)
19. Zhan, K., Ramos, F., Faux, S.: Activity recognition from a wearable camera. In: Proc. of IEEE International Conference on Control Automation Robotics & Vision (ICARCV). pp. 365 – 370. Guangzhou, China (December 2012)
20. Zhang, H., Li, L., Jia, W., Fernstrom, J.D., Sclabassi, R.J., Mao, Z.H., Sun, M.: Physical activity recognition based on motion in images acquired by a wearable camera. Neurocomputing **74**(12), 2184–2192 (June 2011)
21. Zhang, H., Li, L., Jia, W., Fernstrom, J.D., Sclabassi, R.J., Sun, M.: Recognizing physical activity from ego-motion of a camera. In: Proc. of IEEE International Conference on Engineering in Medicine and Biology Society (EMBC). pp. 5569–5572. Buenos Aires, Argentina (August 2010)