First-Person Activity Recognition from Micro-Action Representations using Convolutional Neural Networks and Object Flow Histograms

Panagiotis Giannakeris · Panagiotis C. Petrantonakis · Konstantinos Avgerinakis · Stefanos Vrochidis · Ioannis Kompatsiaris

Received: 20 Mar 2019 / Accepted: 14 Sept 2020

Abstract A novel first-person human activity recognition framework is proposed in this work. Our proposed methodology is inspired by the central role moving objects have in egocentric activity videos. Using a Deep Convolutional Neural Network we detect objects and develop discriminant object flow histograms in order to represent fine-grained micro-actions during short temporal windows. Our framework is based on the assumption that large scale activities are synthesized by fine-grained micro-actions. We gather all the micro-actions and perform Gaussian Mixture Model clusterization, so as to build a micro-action vocabulary that is later used in a Fisher encoding schema. Results show that our method can reach 60% recognition rate on the benchmark ADL

P. Giannakeris *
ITI-CERTH
Tel.: +30-2311-257-742
E-mail: giannakeris@iti.gr

P.C. Petrantonakis
ITI-CERTH
Tel.: +30-2311-257-804
E-mail: ppetrant@iti.gr

K. Avgerinakis
ITI-CERTH
Tel.: +30-2311-257-720
E-mail: koafgeri@iti.gr

S. Vrochidis
ITI-CERTH
Tel.: +30-2311-257-754
E-mail: stefanos@iti.gr

I. Kompatsiaris
ITI-CERTH
Tel.: +30-2311-257-774
E-mail: ikom@iti.gr

dataset. The capabilities of the proposed framework are also showcased by profoundly evaluating for a great deal of hyper-parameters and comparing to other State-of-the-Art works.

Keywords Activity recognition · Object detection · Egocentric vision · Ambient assisted living

1 Introduction

The continuous rise of the video format as a medium for communication has brought a digital video revolution to the modern connected world. It is safe to say that is has now surpassed the popularity of image and text formats judging by the countless online multimedia platforms that support it and the amount of video clips the web pages are filled with daily. The use cases are endless: from do-it-yourself tutorials, to marketing and live event broadcasting that are uploaded online, many popular public video repositories contain massive amounts of video content. It is not only the attractive combination of auditory and visual content that is making the medium popular, but also the technology of modern wearables that push seemingly every single person to carry a tiny video camera at all times, plus the convenient ways that exist for the videos to end up posted online for immediate consumption on social media.

In most of the videos uploaded online, humans are the center of attention and the thematic content is in one way or another moving around the activities that they perform. Multimedia processing and computer vision researchers have shown much interest in the exploitation of those huge databases. The proposed solutions can address the needs of several real life applications, such as video surveillance and security applications, human behavior understanding, video indexing and retrieval, human-machine interaction, etc.

In this work, we process videos captured by wearable devices and more precisely we focus on the recognition of the human activities such videos contain. Besides the use of wearables as entertainment devices mostly in outdoor environments, this technology can also be used effectively in order to monitor indoor activities of patients. Those patients that may have a critical disease are often called to live inside their own homes, as nursing homes and hospitals can not accommodate them in their own premises for too long. However, for some of them it is essential that a doctor or a carer should continue monitor their health and keep a log file of their behaviors throughout time. Thus, this work is mostly motivated by the need of efficiently recognizing human activities of daily living that are captured by wearable cameras in indoor environments.

Huge attention has been drawn to generic human activity recognition that capture the human subjects from distant cameras, or third person viewpoints, but the first-person activity recognition field is relatively understudied. There are several challenges when dealing with the task of first-person activity recognition, the main one being the lack of human actors in the field of view. The severe distortions that may also appear in egocentric videos, like field of view

distortions, and ego-motion from the user's movements, may negatively impact the process of extracting meaningful representations of activities using the State-of-the-Art methods proposed for third-person activity recognition. First-person video datasets have recently emerged, [36], [31], [38], as well as first-person activity recognition challenges [9], calling for more interest on the subject.

With human movements out of the field of view, most of the related work focuses on either human hand movements that may be still in the frame, or hand-object interactions [57], [3]. Many of the recent works also employ Deep Convolutional Neural Networks (CNNs) to study those interactions, as well as multiple CNNs that are tailored to complete a certain task in the whole framework [57], [51]. Our contribution lies close to the object-centric approaches, but we also focus on efficiency in all the steps of our methodology in order to recognize activities while maintaining feasible computational times. To this aim, we build upon a previous work of ours [18] by exploring a more appropriate dimensionality reduction scheme that exploits the sparsity of our representations to reduce the computational complexity during training whilst achieving comparable results. Additionally, we explore in this work the impact of motion compensation to our low-level descriptors.

In contrast with most of the previously published works, we not only detect relevant objects but also extract their individual motion patterns using object flow histograms. Moreover, aggregating the motion features by class over short temporal windows allows us to build discriminative representations that relate directly to the object manipulation patterns. In addition, we encode those patterns in a binning framework to understand their usage in short-term actions, which are fundamental building blocks of long-term activities. Unlike other State-of-the-Art (SoA) works, our current implementation does not rely on any hand movement information at all or other modalities, like sensory equipment or gaze information.

The rest of this paper is structured as follows. In section 2, we present the state-of-the-art related work, while in section 3, we show our proposed methodology. Section 4 describes our experimental work and evaluation results are included. Finally, conclusions are drawn in section 5.

2 Related Work

In this section we examine the previous work on activity recognition, and briefly on object detection, as it is an integral part of our approach.

2.1 Activity Recognition

Previous works on activity recognition can be categorized based on the appearance of human actors that perform the activities, or lack thereof. For activity recognition of the first category, where human actors appear in the activity

clips, most of the pre-CNN era works dealt with motion analysis using optical flow and the analysis of dense trajectories [24], [50], [49], [25], [1], which involved the process of extracting classic low-level descriptors like Histograms of Oriented Gradients (HOG), Histograms of Optical Flow (HOF) and Motion Boundary Histograms (MBH), to represent the visual and motion features around keypoints. Towards more temporally-aware methods, others have chosen to model activities as sequences of sub-actions focusing on the temporal structure of visual patterns [17], [39]. Shortly after the CNN impact, the direction was steered naturally towards deep learning approaches. Amongst the most popular was the multi-stream CNNs [51], [15], [45], that work by feeding various modes of video frames, mainly an RGB channel and an optical flow channel, in order to extract deep CNN visual and motion features and represent activities based on fusion of the two. In a later work, information from the actors pose was more effectively captured using pose-based CNN features [4]. Other methods that appeared later rely on RNNs, to more accurately model the temporal dynamics of activities. More specifically, the modern technique of deep visual attention was combined with RNNs in [42] and [11]. More recent works focused on refining existing techniques like in [44] where optical flow derivation features (OFF) where plugged in existing CNN-based action recognition schemes. In [14] the two-stream CNN approach was modernized by injecting residual connections, while the more recent TSN framework [52] works in sampled video segments with the aim to model long-range temporal structures more efficiently.

For first-person activity recognition other approaches have been proposed where, in contrast to the previous methods, human actors cant be seen performing the activities. Human hands or lower body parts are naturally the only information we can get as far as the actors movements are concerned. Therefore, in the context of representing activities of daily living, object manipulation and hand movements are the main source of visual information. As such, many of the works in this category propose to describe activities by an object-centric manner following the information that derives from the existence of specific objects in the scene [13], [38], [34], [57]. Moreover, scene understanding is also used in [47]. In [54] a multi-task clustering framework tailored to first-person view (FPV) activity recognition is presented. Another more recent approach is to use deep CNN architectures [53] to learn deep appearance and motion clues. Deep CNNs are also used to learn hand segmentations in order to understand the activities that a user performs and his interaction with other users that might also appear in the video frame [57], [3], [2]. More recent works focus on multi-modal analysis of egocentric cameras and information from other wearable sensor equipment with the deployment of early or late fusion schemes [35], [6], [5].

2.2 Object Detection

Since our method is heavily dependent on object detection, we provide here a brief overview of the literature on this subject as well. Several works have been proposed to solve this task with outstanding results in challenging datasets [32], [10], [30], [12]. The breakthrough of deep CNNs that were thoroughly examined for this task include works like the seminal work of [19], which deployed a proposal generator [46] step to feed the network. Later, the bounding box proposal network was incorporated into an end-to-end deep architecture in Faster R-CNN [41], achieving better performance and faster prediction during testing. Others have focused on deep end-to-end single shot detectors that predict classes and box coordinates directly using the last convolutional feature maps, like the SSD [33] and YOLO [40] detectors. Since then more works have been proposed that focused on faster detection time like in [7] and [28], and were based on sharing convolutions to multiple layers. The most recent high performance detectors were based on minor tweaks to vanilla models, but have produced significant performance boost nevertheless [26], [43], [56], [55], [58].

3 Methodology

In this section we take a closer look at how our proposed framework processes the activity clips. The motivation behind the modeling of micro-actions is first discussed, and then a detailed description for each processing stage is given.

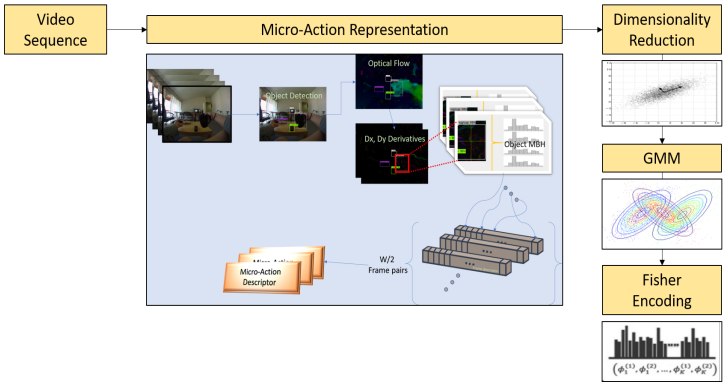


Fig. 1 Block diagram of the proposed methodology.

Activities of daily living such as “book reading”, “hand washing” or “preparing breakfast” usually take place in long segments inside an egocentric video, lasting on average a couple of minutes. Instead of trying to capture directly long-term dynamics, it is suitable to get a deep understanding of the lower level actions the actors are performing in order to accomplish the large scale activities. For example the activity “preparing breakfast” involves the fine-grained

actions "opening the fridge", "grabbing butter", "closing the fridge", "taking a knife", "spreading the butter", etc. This group of *micro-actions* as we call them, does not always need to form a complicated sequence for every activity. For example, the activity "reading a book", except for the actual "reading" activity, usually involves one micro-action performed repeatedly, i.e., "turning the page". For those reasons, we seek a way of extracting a representation of the full duration of an activity clip which will be informative towards the set of micro-actions that are involved and have a strong ability to uniquely describe the activity.

It is very well established in the literature [13] [38] [34] [57] that every activity is related closely to a group of active objects and a group of passive objects. The first group contains objects that are handled by the person during the activity, and the second group contains objects that are simply within the view of the camera when the activity is performed. Objects are good indicators of certain activities such as the TV in the "watching television" activity or the book in the "reading a book" activity. We further elaborate this notion by hypothesizing that not only the presence, but also the characteristic motions of the objects in the scene are powerful enough to discriminate between active and passive ones. For example, motion information from dishes that are being washed combined with the presence of a tap in the scene can uniquely describe the "washing dishes" activity.

The overall framework is shown in figure 1. The above assumptions are taken into account in our activity recognition method. First, we detect objects using a Deep CNN architecture that combines a deep feature extraction network and a bounding box coordinate regression network, that predicts object classes and locations in the video frames. We combine the powerful detector with a tracking algorithm, eliminating the need to utilize the deep architecture for every frame, in order to achieve near real time object detection. Then, every detected object's motion is processed using HOF or MBH [8] features, so as to form the lower level micro-action representations that appear in short time windows over the full activity sequence. The resulting micro-action descriptors go through a dimensionality reduction step, which keeps the representations compact with minimum loss of information. Finally, Gaussian Mixture Modeling (GMM) is used for clustering in order to extract prototype micro-actions, finding the most discriminative of the full set. Given a set of micro-action descriptors extracted for a single activity sequence and the GMM clustering centers, a Fisher encoding schema is used in order to yield the final descriptor of the full activity sequence in a Bag-of-Micro-Actions type of representation.

3.1 Object Detection

In order to detect the activity-related objects in the egocentric videos, we chose to extract deep image representations and predict pixel coordinates of bounding boxes using a deep CNN object detector. To this end, we adopt a modification of the Faster-RCNN which was originally proposed in [41].

A thorough evaluation of this model and comparisons with other SoA deep object detectors, presented in [23], reveal that the Faster-RCNN-resnet101 architecture achieves a good trade off between speed and accuracy. This model incorporates the resnet101 [21] deep feature extractor and a region proposal network, along with a bounding box classifier and coordinate regressors. In order to make the object detection procedure more efficient during inference time, we find useful to track the detected objects found in a frame into the next T frames of the video, instead of running the detector for each single frame. Since we do not expect dramatic cuts during an activity clip, we still manage to get good quality detections at far greater speeds. By assigning a detection rate of $T > 15$ our combined detector and tracker algorithm achieves near real time performance. We manually set the detection rate parameter to 15 following empirical evaluation after trials with other values ranging from 5 to 30. Intuitively, the detection rate defines the temporal resolution of the continuous object detection function. Lower detection rate means higher temporal resolution of the detector and vice versa. For a usual 30 fps video, by setting the detection rate to 15 the detector only operates between half-second intervals and the tracker works the rest of the time. This is expected to yield adequate temporal resolution, considering that it is very unlikely that an object will appear and disappear in less than half a second. The core functionality of our tracker is based on the KCF tracking algorithm that was proposed in [22].

3.2 Micro-Action Representation

Our method builds representations of short, fine-grained actions, of fixed temporal window W, from the motion patterns of the objects that are found in this window. More specifically, we first compute a dense optical flow field, to extract the full scene's motion between two consecutive frames. We use the OpenCV implementation of the dense inverse search algorithm proposed in [29]. In addition, doing the calculation every other frame inside the window, instead of every frame, leads to $W/2$ calculations which yields faster computation times. Having already detected the objects in a particular frame we take each bounding box as our region of interest and crop the dense optical flow map accordingly, taking only the portion that belongs to the object. Consequently, we can calculate HOF (histograms of optical flow) descriptors that represent an object's motion.

To calculate an object's HOF descriptor we apply a $2X2$ uniform grid on top of the bounding box region. For each one of the 4 cells, flow orientations are quantized into an 8 bin histogram weighted by their magnitude values. In addition, we chose to apply a soft binning method that distributes the votes between adjacent bins, based on the distances of the values from adjacent bins centers. This procedure results in a 32-dimensional motion descriptor that is extracted for every object in the scene. If multiple objects from the same class appear in one frame, we aggregate the vectors and divide by the number of objects, so as to get the average motion descriptor of that particular class. In

the case of absence of any objects from a particular class, the corresponding HOF descriptor is set to the zero vector. Let C be the number of classes the detector can predict and N_c the number of objects found for class c. The early object class motion descriptors are formed as follows:

$$D_c = \frac{1}{N_c} \sum_{j=1}^{N_c} HOF_{32}, \quad c = 1 \dots C \tag{1}$$

By concatenating L2 normalized motion descriptors for each class we get a complete description for a pair of consecutive frames in the window W:

$$R_f = \{D_1, D_2, \dots, D_c\} \tag{2}$$

Finally, we concatenate those descriptors throughout $W/2$ frame pairs to get a complete representation of a micro-action composed by the object's movement patterns that appeared in the window:

$$M = \{R_1, R_2, \dots, R_{W/2}\} \tag{3}$$

One problem with the accurate extraction of object motion from egocentric videos is that very frequently the wearable camera moves along with the person that is wearing it. As a result, ego-motion may overpower the delicate dynamics of the objects' motion that we are trying to capture. Therefore, we consider an alternative to the HOF descriptor, that is the MBH descriptor where the optical flow field is first separated into its x and y component and spatial derivatives are computed for each one of them. This time we obtain a 32-dimensional descriptor for each component (64-dimensional after concatenation) following the same procedure to obtain the final descriptor as in the previous case. Because MBH is the gradient of the optical flow, any motion that is happening constantly (global motion) is suppressed and only information about changes in the flow field (i.e., motion boundaries) is kept [48].

Given that there are N_c possible object classes, the dimensionality of a micro-action descriptor is given by $\frac{W}{2} \times N_c \times 32$ for HOF and $\frac{W}{2} \times N_c \times 64$ for MBH. It is expected that the dimensionality is increased dramatically for a high number of object classes or longer windows W. Moreover the descriptors can be very sparse because of total absence of certain objects classes, where the respective values are set to zeros. Therefore, we proceed with two alternative approaches to apply the dimensionality reduction stage, while at the same time we exploit the sparsity in a way that yields lower computational complexity.

3.3 Dimensionality Reduction

The high dimensionality of our micro-action descriptors severely affects the computational burden which we intended to alleviate through dimensionality reduction approaches. For dimensionality reduction two approaches were adopted, i.e., Principal Component Analysis (PCA) and random projections (RP).

PCA projects the data onto a lower-dimensional orthogonal subspace that captures as much of the variation of the data as possible. Due to the fact that PCA approach is quite expensive to compute for high-dimensional data sets we also investigate a computationally simpler method of dimensionality reduction that does not introduce a significant distortion in the dataset, i.e., the RP approach. In RP the original high-dimensional data, $X \in \mathbb{R}^{n \times D}$ is projected onto a lower-dimensional subspace using a random matrix whose rows have unit lengths. More formally, using matrix notation where $X \in \mathbb{R}^{n \times D}$ is the original set of n d-dimensional observations,

$$Y = XR \tag{4}$$

where $R \in \mathbb{R}^{D \times d}$ is the random matrix and $Y \in \mathbb{R}^{n \times d}$ is the projection of the data onto the lower d-dimensional space. The fundamental idea of random projection arises from the well-celebrated Johnson-Lindenstrauss lemma [27] which states that if point instances in a vector space are projected onto a randomly selected subspace of appropriate dimension, then distances are approximately preserved [16]. In this work we use random matrix whose elements are Gaussian distributed with zero mean and unit variance.

Due to its computational simplicity and our sparse feature vectors, RPs are ideal for the dimensionality reduction task in this work. In particular, the aforementioned random projection procedure is of order $O(Ddn)$ and taking account that X is in our case sparse (assuming l nonzero entries per row) the complexity is of order $O(ldn)$ [37].

3.4 Activity Recognition

For a given activity sequence, the extraction of micro-action descriptors that represents a small sequence of W frames takes place with a stride of S frames. We chose that value to be exactly 1 second in all our experiments. This simply means that for every micro-action descriptor M we skip 1 second into the video before we begin extracting the next micro-action descriptor. Contrary to using overlapping windows, the stride parameter was inserted to give our method a speed boost. The micro-action descriptors are extracted from fixed length temporal windows W. In contrast, the length of the activity clips are not expected to be constant. Therefore, the number of micro-action descriptors that are formed can vary, depending on an activity's duration and the length of W. Given that the micro-action window W is chosen sufficiently small, it can be guaranteed that the number of micro-actions that will be formed for an activity sequence will be enough for the activity to be adequately represented.

All micro-action descriptors extracted from all the training activity sequences are fed into a Fisher encoding schema. This way, a micro-action vocabulary based on the most discriminating ones is constructed. The computation of the most discriminating samples is performed by applying unsupervised clustering, using Gaussian Mixture Modeling, in the micro-action representation hyperspace.

Let $\{\mu_j, \Sigma_j, \pi_j; j \in R^L\}$ be the set of parameters for L Gaussian models, with μ_j, Σ_j and π_j standing respectively for the mean, the covariance and the prior probability weights of the j^{th} Gaussian. Assuming that the D-dimensional early descriptor is represented as $\overline{M}_i \in R^D; i = \{1, \ldots, N\}$, with N denoting the total number of descriptors, Fisher encoding is then built upon the first and second order statistics:

$$\begin{aligned} f_{1j} &= \frac{1}{N\sqrt{\pi_j}} \sum_{i=1}^{N} q_{ij} \sigma_j^{-1} (\overline{x}_i - \overline{\mu}_j) \\ f_{2j} &= \frac{1}{N\sqrt{2\pi_j}} \sum_{i=1}^{N} q_{ij} [\frac{(\overline{x}_i - \overline{\mu}_j)^2}{\sigma_j^2} - 1] \end{aligned} \tag{5}$$

where q_{ij} is the Gaussian soft assignment of descriptor M_i to the j^{th} Gaussian and is given by:

$$q_{ij} = \frac{exp[-\frac{1}{2}(M_i - \mu_j)^T \Sigma_j^{-1} (M_i - \mu_j)]}{\sum_{t=1}^{L} exp[-\frac{1}{2}(M_i - \mu_t)^T \Sigma_j^{-1} (M_i - \mu_t)]} \tag{6}$$

Distances as calculated by Eq. 5 are next concatenated to form the final $2LD$-dimensional Fisher vector, $F_X = [f_{11}, f_{21}, \ldots, f_{1L}, f_{2L}]$, that characterizes each activity sequence. The final Fisher encoding for a specific activity sequence can now be classified using an SVM or a Neural Network classifier.

4 Experimental Work

In this section, we first describe the experiments that we conducted, so as to select the best hyper-parameters for our activity recognition algorithm, while also comparing the performance of different descriptor options (HOF, MBH). The performance of our object detection and tracking algorithm is presented as well. Additionally, we extend our experimental work by studying alternative dimensionality reduction techniques in order to examine the validity of our assumptions. Furthermore, we applied camera ego-motion compensation, as in [20], to examine the improvement it may bestow upon our best models for both descriptors. We accumulated and present activity recognition results for each class, in the form of confusion matrices, and examine how each class performs depending on object detection performance. Finally, we present a comparison of our framework with other SoA works in terms performance in the ADL dataset in order to prove the applicability of our method.

4.1 Dataset

We performed our experiments on the ADL dataset [38]. It is composed of videos recorded with a wearable camera from 20 different persons. The videos contain realistic scenes of daily living and the benchmark is challenging due

to the existence of global camera motion. The objects are also in many cases occluded. From the 48 different classes of objects that are available, we select the 34 most frequently annotated to train our object detector. We also select a subset of 18 activity classes, as in [57], to train our activity recognition framework so as to present comparable results with previous works. The taxonomy of the activity classes is given in Figure 2. The activity classes can be divided in three major sub-categories.

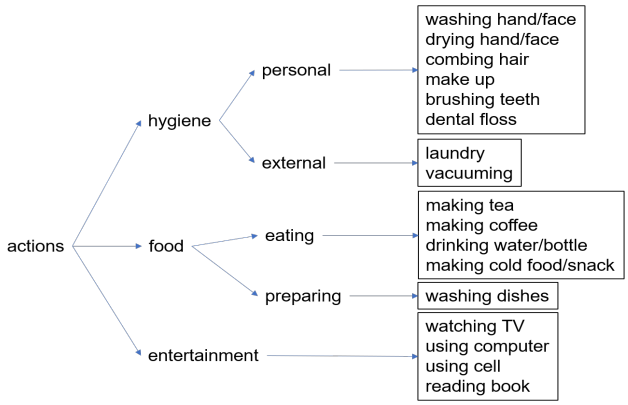


Fig. 2 Taxonomy of activities in the ADL dataset.

4.2 Object Detection

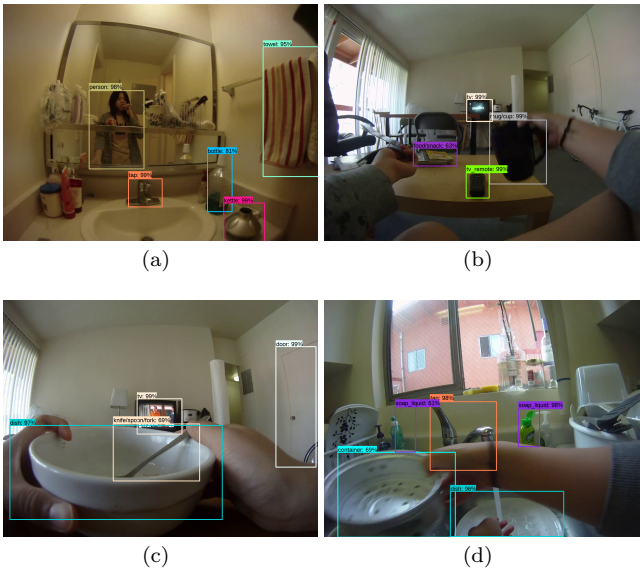
To train our object detector we used only the first 6 videos, since this is the typical way of splitting the dataset and is reported in previous works. During testing we set the detection rate to 15 frames and track the detected boxes, managing to achieve detection inference time at a rate of 12 fps on average. Our object detector achieves an overall 26.9% mAP on the 14 remaining test videos. A detailed performance evaluation per object category is shown in Table 1. The detector performs very well on several classes for which many annotated samples are provided in the training set (over 1000). However, it performs poorly on small objects like "towel" or "pills" and it even yields 0% mAP on three object classes. Small items that are handled by the actors are expected to be heavily occluded in comparison with big static objects such as a TV, an oven or a microwave. Figure 3 shows qualitative detection results in test frames of the ADL Dataset.

4.3 Hyper-parameter Selection

We experimented with two different durations for the temporal window: $W = 90$ and $W = 60$ frames. Those two values correspond to 3 seconds and 2

Table 1 Per-class performance evaluation of our object detector on the ADL dataset (mAP%).

tv	77.77	person	72.63	tv remote	69.03	tap	66.19
oven/stove	58.58	door	56.90	microwave	53.36	laptop	48.25
washer/dryer	41.97	container	37.53	fridge	37.47	soap liquid	36.12
dish	35.45	mug/cup	33.46	kettle	28.19	pan	24.09
tooth brush	18.13	book	17.69	bottle	16.75	knife/spoon/fork	16.06
tooth paste	11.90	cell phone	11.54	detergent	11.23	vacuum	8.30
food/snack	6.94	trash can	6.38	dent floss	3.65	pitcher	3.32
towel	2.69	pills	1.18	blanket	1.13	cell	0
tea bag	0	comb	0				



(a) (b) (c) (d)

Fig. 3 Object Detection qualitative results.

seconds respectively for the videos of the ADL Dataset which were recorded at 30fps. Considering that the activity average duration is in the order of minutes in this dataset, we manage to get enough micro-action descriptors assigned to each activity and simultaneously capture more complex object motions through time. Furthermore, we show that micro-actions of 3 or 2 seconds are long enough for our method to perform close to SoA levels. The two choices for our temporal window W proves to be convenient for algorithmic speed considerations as well.

The length of the micro-action descriptors before dimensionality reduction is $\frac{W}{2} \times 34 \times 32$ for the HOF descriptor and $\frac{W}{2} \times 34 \times 64$ for the MBH descriptor. In this stage, we use only two options for the dimensionality reduction stage, using PCA with 80 and 256 components, so as to focus on evaluating the other hyper-parameters. Later, we perform an extended study between various modes of dimensionality reduction on the most promising model configurations.

We also experiment here with two different vocabulary sizes, using 32 or 64 Gaussians. For the final stage, we deploy as our classifier, a fully connected neural network (NN1) with a depth of two layers, of width 512 and 256 accordingly, using RELU activations, 50% chance of dropout between layers and softmax activation in the output layer. Another similar architecture (NN2) was also deployed with half the amount of neurons for each layer (256 in the first layer and 128 in the second) and a linear SVM classifier as a third option for the sake of classifier comparison.

To evaluate the action recognition performance as in [57], we performed the leave-one-person-out cross-validation strategy for every hyper-parameter combination and we report the mean average precision (mAP) and standard deviation. Tables 2 and 3 present analytically our scores for every experiment.

Table 2 Activity recognition results for **HOF** descriptor

Model comparison (mAP%) for HOF descriptor			
	SVM	NN1	NN2
W 90 + PCA 80 + GMM 32	$43.19 \pm 15.1\%$	$52.40 \pm 16.3\%$	$47.07 \pm 14.4\%$
W 90 + PCA 80 + GMM 64	$46.22 \pm 12.4\%$	$51.04 \pm 13.1\%$	$51.56 \pm 17.1\%$
W 90 + PCA 256 + GMM 32	$45.21 \pm 17.4\%$	$\mathbf{52.86 \pm 15.4\%}$	$51.86 \pm 15.6\%$
W 90 + PCA 256 + GMM 64	$46.22 \pm 13.2\%$	$51.03 \pm 13.3\%$	$51.56 \pm 14.8\%$
W 60 + PCA 80 + GMM 32	$43.51 \pm 14.9\%$	$50.98 \pm 16.2\%$	$48.66 \pm 17.3\%$
W 60 + PCA 80 + GMM 64	$43.24 \pm 16.9\%$	$47.34 \pm 13.2\%$	$44.69 \pm 16.8\%$
W 60 + PCA 256 + GMM 32	$46.30 \pm 14.4\%$	$48.07 \pm 14.5\%$	$47.66 \pm 15.1\%$
W 60 + PCA 256 + GMM 64	$45.73 \pm 16.1\%$	$46.81 \pm 14.6\%$	$45.53 \pm 16.7\%$

Table 3 Activity recognition results for **MBH** descriptor

Model comparison (mAP%) for MBH descriptor			
	SVM	NN1	NN2
W 90 + PCA 80 + GMM 32	$41.06 \pm 14.9\%$	$52.96 \pm 13.3\%$	$53.88 \pm 15.6\%$
W 90 + PCA 80 + GMM 64	$39.89 \pm 15.2\%$	$49.16 \pm 16.3\%$	$51.23 \pm 14.3\%$
W 90 + PCA 256 + GMM 32	$41.34 \pm 14.1\%$	$53.12 \pm 14.7\%$	$50.02 \pm 15.9\%$
W 90 + PCA 256 + GMM 64	$43.61 \pm 15.6\%$	$54.88 \pm 12.5\%$	$50.25 \pm 16.1\%$
W 60 + PCA 80 + GMM 32	$49.37 \pm 15.4\%$	$57.09 \pm 13.8\%$	$54.57 \pm 14.8\%$
W 60 + PCA 80 + GMM 64	$47.17 \pm 15.1\%$	$52.60 \pm 14.2\%$	$50.93 \pm 15.6\%$
W 60 + PCA 256 + GMM 32	$45.62 \pm 16.6\%$	$\mathbf{57.14 \pm 15.7\%}$	$55.58 \pm 13.9\%$
W 60 + PCA 256 + GMM 64	$42.43 \pm 16.4\%$	$50.91 \pm 15.8\%$	$50.24 \pm 15.5\%$

As shown, choosing 256 components in PCA results in performance boost when combined with a larger temporal window. Choosing 80 components resulted in better performance in some cases of the shorter temporal window. However, as shown later, those two thresholds are limiting and more PCA or RP components lead to better performance overall. Increasing the size of

the GMM vocabulary from 32 to 64 failed to improve our results, especially when using the shorter temporal window. This proves that using a smaller vocabulary consisting of 32 words is enough to get good coverage of the most discriminant micro-actions of the entire dataset. Finally, we can see that the MBH descriptor almost entirely outperformed the HOF descriptor for every experiment with a temporal window of 60 frames and that the performance of the two was comparable for a window of 90 frames. This is an indication the MBH has to offer more when micro-action extraction is more refined in time. Overall, the best models came from the combination of 256 PCA components coupled with a GMM vocabulary of size 32 and the neural network architecture with the most learnable parameters (NN1).

4.4 Dimensionality Reduction

In this section, we keep the best models' parameters fixed, e.g. GMM vocabulary of size 32 and the NN1 classifier, and experiment upon using different dimensionality reduction techniques, for all the different descriptors and window sizes. The same evaluation scheme is also applied in this section as well, i.e. leave-one-person-out cross validation.

First we investigate selecting 1000 PCA components instead of 80 and 256, in order to explore the capabilities of our method for higher dimensional micro-action vectors. When reducing from some thousand components to only 256 it is possible that only a small portion of the dataset variance can be explained, thus the reduction step can become a bottleneck to the overall performance. In Table 4 the results indeed show significant improvement in performance for all settings, ranging from 3% to 5% mAP.

However, as discussed earlier PCA is rather expensive to compute mainly during training time for high dimensional data. Especially in our case, the micro-action descriptors, depending on the setting, are at least 30720-dimensional and up to 92160-dimensional vectors before the reduction stage. For those reasons, we chose to experiment with RP using 4 different settings: $d = 1000$, $d = 2500$, $d = 3500$ and $d = 5000$. Table 5 shows the results. We can see that with Random Projections close to 3500 components, the method scores either comparably or surpasses PCA's lower component settings (80, 256). Considering, all RP experiments took lesser time to produce, there exists a performance/speed trade-off when choosing one or the other during training. For full performance gain, it is evident that the 1000 PCA component setting is the ideal one.

4.5 Motion Compensation

Heavy ego-motion may appear between video frames of the ADL Dataset, as a result of the person moving around while performing the actions. As such, the real movement of objects may be overpowered by camera motion. We have

Table 4 Performance comparison with increased number of PCA components.

PCA components	Performance (mAP%) MBH W90	MBH W60	HOF W90	HOF W60
80	$52.9 \pm 13.3\%$	$57.1 \pm 13.8\%$	$52.4 \pm 16.3\%$	$50.9 \pm 16.2\%$
256	$53.1 \pm 14.7\%$	$57.1 \pm 15.7\%$	$52.8 \pm 15.4\%$	$48.1 \pm 14.5\%$
1000	$58.9 \pm 15.1\%$	$\mathbf{60.1 \pm 14.6\%}$	$\mathbf{56.8 \pm 15.6\%}$	$54.9 \pm 13.2\%$

Table 5 Performance comparison for Random Projection.

RP components	Performance (mAP%) MBH W90	MBH W60	HOF W90	HOF W60
1000	$51.1 \pm 16.7\%$	$54.5 \pm 17.9\%$	$50.3 \pm 15.2\%$	$49.6 \pm 15.9\%$
2500	$54.2 \pm 17.1\%$	$54.9 \pm 16.4\%$	$52.1 \pm 18.2\%$	$50.6 \pm 15.8\%$
3500	$54.3 \pm 13.5\%$	$55.9 \pm 15.3\%$	$52.3 \pm 16.6\%$	$51.8 \pm 16.2\%$
5000	$55.4 \pm 15.4\%$	$56.7 \pm 18.1\%$	$53.7 \pm 15.7\%$	$53.1 \pm 16.4\%$

already established the MBH descriptor as the best choice over HOF for activity recognition on the ADL dataset and the hyper-parameters that lead to the best performance. In this section, we apply an ego-motion compensation technique before the calculation of the descriptors, so as to determine its impact on the overall performance.

Having already computed the dense optical flow field, we select randomly 1000 points in the image and their displacement vectors and feed them to a RANSAC estimator of a projective transformation (3×3 homography) between consecutive frames. Then, the set of inlier samples can determine the camera displacement. We take the mean displacement of the inlier samples in each direction (x and y) and subtract it from the optical flow field. Then, the compensated optical flow field is used to calculate compensated versions of HOF and MBH descriptors. The performance comparison of the ego-motion compensated versions of our best performing HOF and MBH descriptors is shown in Table6. The impact of motion compensation upon the HOF descriptor is positive, and yields an additional performance improvement of 1.5% mAP. However, it still cannot surpass the best performing MBH descriptor. Contrariwise, a slight drop of performance on the compensated MBH indicates that it may not benefit from the use of motion compensation.

Table 6 Performance comparison using motion compensation.

	Performance (mAP%) no compensation	with compensation
HOF (W 90 + PCA 1000)	$56.8 \pm 15.6\%$	$\mathbf{58.3 \pm 15.7\%}$
MBH (W 60 + PCA 1000)	$\mathbf{60.1 \pm 14.6\%}$	$58.5 \pm 15.1\%$

4.6 Comparison with State-of-the-Art

In Table 7, we compared the accuracy rates of our best models, namely HOF and MBH variants with 32 GMM words and 1000 PCA components, to the ones that are mentioned in the literature. As already mentioned, we followed the evaluation procedure in [57], in order to present comparable results. As we can see, the MBH version of our method outperformed every other. The motion compensated HOF descriptor is also highly ranked.

Table 7 Comparison with SoA on the ADL Dataset

Method	Performance (mAP%)
Boost-RSTP [34]	33.7%
Boost-RSTP + OCC [34]	38.7%
Bag-of-objects [38]	32.7%
Bag-of-objects + Active model [38]	36.9%
Cascaded Interactional Network [57]	55.2%
Ours - Bag-of-Micro-Actions with HOF (best)	58.3%
Ours - Bag-of-Micro-Actions with MBH (best)	**60.1%**

4.7 Per-Class Evaluation of our Activity Recognition Framework

Next, we select our top two models (one for each descriptor) and train them for the first 6 videos of the dataset. We present the test set confusion matrices in Figures 4 and 5. As we can see, MBH performed better than HOF in most of the classes that heavy camera motion is expected, like the "washing dishes" or "drinking water" activities, because it simulates a compensated motion and it proves to be more appropriate when wearable cameras are used. Our framework is highly dependent on the performance of the object detector, as expected. The performance drops in instances that involve interactions with smaller objects, usually in hygiene activities. In addition, activities that involve the better performing object classes, like "watching tv" or "using computer", have higher recognition rates. Confusion seems to exist between the classes "making tea" and "making coffee" because they almost always involve person interactions with the same object classes. Another similar example is the confusion created between the "combing hair", "brushing teeth", and "dental floss" classes that are all taking place inside a bathroom with the same objects being visible from the camera. Hence, the need to directly deal with active vs passive object recognition is indicated here.

5 Conclusions

In this paper, we introduced a new approach for activity recognition from wearable cameras by detecting objects and then incorporating their motion

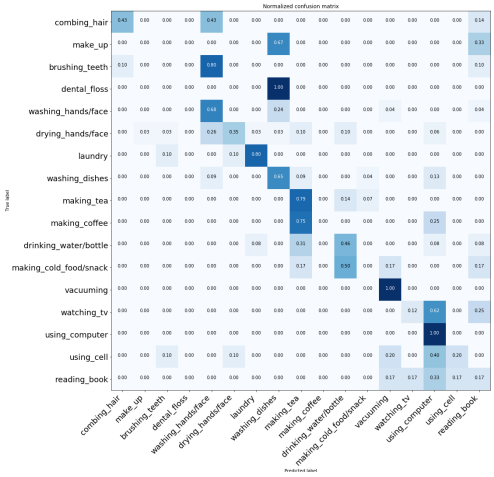


Fig. 4 Confusion matrix of our activity recognition method with HOF descriptors.

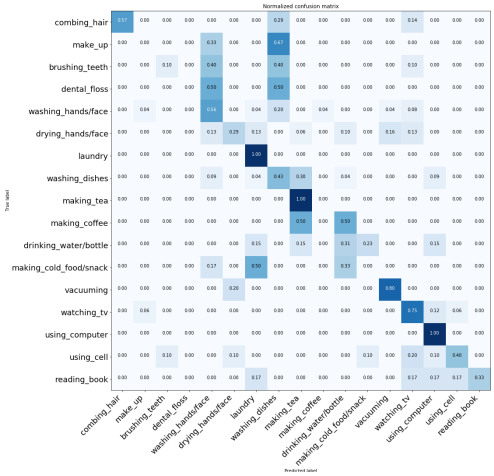


Fig. 5 Confusion matrix of our activity recognition method with MBH descriptors.

patterns into low level micro-action descriptors. We represented activities using a Bag-of-Micro-Actions schema, using GMM clustering and Fisher vector encoding. Comparison with SoA techniques on the ADL dataset validate the competitiveness of our approach.

Our future steps will be to develop an object detection algorithm that discriminates between active and passive objects so as to weight them differently and to leverage hand movements and include gesture patterns into the overall framework. Additionally, we plan to incorporate Deep Neural Networks at another stage in our framework, so as to replace the Fisher encoding schema and model temporal dependencies of active object movements using LSTMs.

Finally, evaluation on newer activity recognition datasets is also a future target.

Acknowledgements This work is supported by the project SUITCEYES that has received funding from the European Unions Horizon 2020 research and innovation program under grant agreement No 780814.

References

1. Avgerinakis, K., Briassouli, A., Kompatsiaris, Y.: Activity detection using sequential statistical boundary detection (ssbd). Computer Vision and Image Understanding **144**, 46–61 (2016)
2. Bambach, S., Crandall, D.J., Yu, C.: Viewpoint integration for hand-based recognition of social interactions from a first-person view. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 351–354. ACM (2015)
3. Bambach, S., Lee, S., Crandall, D.J., Yu, C.: Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In: Computer Vision (ICCV), 2015 IEEE International Conference on, pp. 1949–1957. IEEE (2015)
4. Chéron, G., Laptev, I., Schmid, C.: P-cnn: Pose-based cnn features for action recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 3218–3226 (2015)
5. Crispim-Junior, C.F., Buso, V., Avgerinakis, K., Meditskos, G., Briassouli, A., Benois-Pineau, J., Kompatsiaris, I.Y., Bremond, F.: Semantic event fusion of different visual modality concepts for activity recognition. IEEE transactions on pattern analysis and machine intelligence **38**(8), 1598–1611 (2016)
6. Crispim-Junior, C.F., Gómez Uría, A., Strumia, C., Koperski, M., König, A., Negin, F., Cosar, S., Nghiem, A.T., Chau, D.P., Charpiat, G., et al.: Online recognition of daily activities by color-depth sensing and knowledge models. Sensors **17**(7), 1528 (2017)
7. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp. 379–387 (2016)
8. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: European conference on computer vision, pp. 428–441. Springer (2006)
9. Damen, D., Doughty, H., Maria Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 720–736 (2018)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database (2009)
11. Du, W., Wang, Y., Qiao, Y.: Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3725–3734 (2017)
12. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
13. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 407–414. IEEE (2011)
14. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: Advances in neural information processing systems, pp. 3468–3476 (2016)
15. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1933–1941 (2016)
16. Frankl, P., Maehara, H.: The johnson-lindenstrauss lemma and the sphericity of some graphs. Journal of Combinatorial Theory, Series B **44**(3), 355–362 (1988)

17. Gaidon, A., Harchaoui, Z., Schmid, C.: Actom sequence models for efficient action detection. In: CVPR 2011, pp. 3201–3208. IEEE (2011)
18. Giannakeris, P., Avgerinakis, K., Vrochidis, S., Kompatsiaris, I.: Activity recognition from wearable cameras. In: 2018 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–6. IEEE (2018)
19. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448 (2015)
20. González-Díaz, I., Benois-Pineau, J., Domenger, J.P., Cattaert, D., de Rugy, A.: Perceptually-guided deep neural networks for ego-action prediction: Object grasping. Pattern Recognition **88**, 223–235 (2019)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
22. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 583–596 (2015)
23. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: IEEE CVPR (2017)
24. Jain, M., Jegou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2555–2562 (2013)
25. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 3192–3199 (2013)
26. Jin, P., Rathod, V., Zhu, X.: Pooling pyramid network for object detection. arXiv preprint arXiv:1807.03284 (2018)
27. Johnson, W.B., Lindenstrauss, J.: Extensions of lipschitz mappings into a hilbert space. Contemporary mathematics **26**(189-206), 1 (1984)
28. Kong, T., Sun, F., Yao, A., Liu, H., Lu, M., Chen, Y.: Ron: Reverse connection with objectness prior networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5936–5944 (2017)
29. Kroeger, T., Timofte, R., Dai, D., Van Gool, L.: Fast optical flow using dense inverse search. In: European Conference on Computer Vision, pp. 471–488. Springer (2016)
30. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 (2018)
31. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 619–635 (2018)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, pp. 740–755. Springer (2014)
33. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision, pp. 21–37. Springer (2016)
34. McCandless, T., Grauman, K.: Object-centric spatio-temporal pyramids for egocentric activity recognition. In: BMVC, vol. 2, p. 3 (2013)
35. Meditskos, G., Plans, P.M., Stavropoulos, T.G., Benois-Pineau, J., Buso, V., Kompatsiaris, I.: Multi-modal activity recognition from egocentric vision, semantic enrichment and lifelogging applications for the care of dementia. Journal of Visual Communication and Image Representation **51**, 169–190 (2018)
36. Ohnishi, K., Kanehira, A., Kanezaki, A., Harada, T.: Recognizing activities of daily living with a wrist-mounted camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3103–3111 (2016)
37. Papadimitriou, C.H., Raghavan, P., Tamaki, H., Vempala, S.: Latent semantic indexing: A probabilistic analysis. Journal of Computer and System Sciences **61**(2), 217–235 (2000)

38. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 2847–2854. IEEE (2012)
39. Pirsiavash, H., Ramanan, D.: Parsing videos of actions with segmental grammars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 612–619 (2014)
40. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788 (2016)
41. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99 (2015)
42. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. arXiv preprint arXiv:1511.04119 (2015)
43. Singh, B., Najibi, M., Davis, L.S.: Sniper: Efficient multi-scale training. In: Advances in Neural Information Processing Systems, pp. 9333–9343 (2018)
44. Sun, S., Kuang, Z., Sheng, L., Ouyang, W., Zhang, W.: Optical flow guided feature: a fast and robust motion representation for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1390–1399 (2018)
45. Tran, A., Cheong, L.F.: Two-stream flow-guided convolutional attention networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3110–3119 (2017)
46. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2), 154–171 (2013)
47. Vaca-Castano, G., Das, S., Sousa, J.P., Lobo, N.D., Shah, M.: Improved scene identification and object detection on egocentric vision of daily activities. Computer Vision and Image Understanding **156**, 92–103 (2017)
48. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 3169–3176. IEEE (2011)
49. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision **103**(1), 60–79 (2013)
50. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp. 3551–3558 (2013)
51. Wang, L., Xiong, Y., Wang, Z., Qiao, Y.: Towards good practices for very deep two-stream convnets. arXiv preprint arXiv:1507.02159 (2015)
52. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence (2018)
53. Wang, X., Gao, L., Song, J., Zhen, X., Sebe, N., Shen, H.T.: Deep appearance and motion learning for egocentric activity recognition. Neurocomputing **275**, 438–447 (2018)
54. Yan, Y., Ricci, E., Liu, G., Sebe, N.: Egocentric daily activity recognition via multitask clustering. IEEE Transactions on Image Processing **24**(10), 2984–2995 (2015)
55. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4203–4212 (2018)
56. Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., Ling, H.: M2det: A single-shot object detector based on multi-level feature pyramid network. arXiv preprint arXiv:1811.04533 (2018)
57. Zhou, Y., Ni, B., Hong, R., Yang, X., Tian, Q.: Cascaded interactional targeting network for egocentric video analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1904–1913 (2016)
58. Zhu, Y., Zhao, C., Wang, J., Zhao, X., Wu, Y., Lu, H.: Couplenet: Coupling global structure with local parts for object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4126–4134 (2017)