Pattern Recognition 62 (2017) 45–55

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Trajectory aligned features for first person action recognition

Suriya Singh [a,*], Chetan Arora [b], C.V. Jawahar [a]

[a] CVIT, IIIT Hyderabad, India
[b] IIIT Delhi, India

ARTICLE INFO

Article history:
Received 21 September 2015
Received in revised form
17 April 2016
Accepted 23 July 2016
Available online 26 August 2016

Keywords:
Action and activity recognition
Egocentric vision
Video indexing and analysis
Video segmentation

ABSTRACT

Egocentric videos are characterized by their ability to have the first person view. With the popularity of Google Glass and GoPro, use of egocentric videos is on the rise. With the substantial increase in the number of egocentric videos, the value and utility of recognizing actions of the wearer in such videos has also thus increased. Unstructured movement of the camera due to natural head motion of the wearer causes sharp changes in the visual field of the egocentric camera causing many standard third person action recognition techniques to perform poorly on such videos. Objects present in the scene and hand gestures of the wearer are the most important cues for first person action recognition but are difficult to segment and recognize in an egocentric video. We propose a novel representation of the first person actions derived from feature trajectories. The features are simple to compute using standard point tracking and do not assume segmentation of hand/objects or recognizing object or hand pose unlike in many previous approaches. We train a bag of words classifier with the proposed features and report a performance improvement of more than 11% on publicly available datasets. Although not designed for the particular case, we show that our technique can also recognize wearer's actions when hands or objects are not visible.

1. Introduction

Advances in camera sensors and other related technologies have led to the rise of wearable cameras which are comfortable to use. In the past few years, the use of Google glass [1] and GoPro [2] has become increasingly popular. Such cameras are typically worn on the head or along with the eyeglasses and have the advantage of capturing from a similar point of view as that of the person wearing the camera. We refer to such cameras with first person view as egocentric cameras.

Excitement of sharing one's actions with friends and the community have made egocentric cameras like GoPro a de facto standard in extreme sports. Egocentric cameras can be used to capture visual logs for law enforcement officers leading to a significant decrease in complaints against the officers [3]. Daily logs from egocentric cameras are also useful in a video sharing application or simply as a memory aid for the wearer. For the visually challenged, researchers are trying to augment egocentric videos with meta data such as facial identity, place, text, etc. [4]. Even for people with regular vision, the promise of giving context aware suggestions is compelling. In spite of their popularity, egocentric videos can be difficult to watch from start to end because of the constant and extreme shake present due to natural motion of wearer's head.

Our focus in this paper is on recognizing wearer's actions from an egocentric video. Owing to their shakiness, egocentric videos are significantly more challenging to analyze than third person videos. Action recognition gives structure to such 'wild' videos which can then be used to search, index or browse. Action recognition is also usually a first step in many other egocentric applications, for example, video summarization, augmented reality, real time suggestions, etc. We follow the popular notation in the field to differentiate between 'activity' and 'actions'. *Activity* is a high level description of what a person is doing at a particular point of time. An activity is usually composed of many short *actions*, which are perceptually closer to the gestures performed by the person. For example, while making tea is an activity, picking the jar, opening the lid and taking sugar are the actions. Other types of actions popular in computer vision are sitting, standing, jumping, etc.

Egocentric videos are different from their third person counterparts, not only because of the change in camera perspective but also because of change in camera motion profile. Many of the accepted techniques for third person video analysis do not work as is for egocentric videos, and the community has been trying to adapt or develop from scratch solutions to these problems in the new context. Works done in the last few years have ranged from

* Corresponding author.
E-mail addresses: suriya.singh@research.iiit.ac.in (S. Singh), chetan@iiitd.ac.in (C. Arora), jawahar@iiit.ac.in (C.V. Jawahar).

http://dx.doi.org/10.1016/j.patcog.2016.07.031
0031-3203/

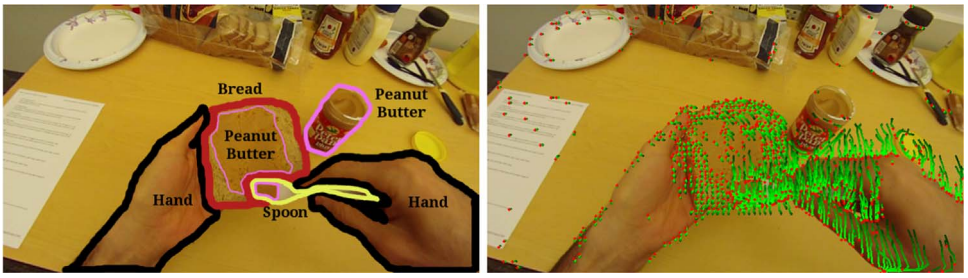

Fig. 1. The focus of this paper is on recognizing wearer's actions from egocentric videos. Earlier work in this area has suggested complicated image segmentation followed by hand or object recognition (left image). We observe that salient objects (hands or handled objects) in such actions are also the objects moving dominantly with respect to the background and can be captured easily using trajectory aligned features (right image) without any prior image segmentation or hand or object recognition. The example images shown here are from GTEA database [5].

tackling simpler problems like object recognition [5–7] and activity recognition [8–15], to more complex problems like video summarization [16–18], and understanding social interactions [19]. Interesting ideas which exploit special properties of egocentric videos have also been proposed for problems like temporal segmentation [20,21], frame sampling [22,23], hyperlapse [24], gaze detection [25] and camera wearer identification [26,27].

Wearer's action recognition from egocentric video is harder compared to regular third person action recognition due to associated unstructured and wild motion of the camera caused by wearer's natural head movement. Different speeds of performing actions and widely varying operating environment also cause difficulties. Fig. 2 gives some examples of the actions we are interested in recognizing.

Given the unique perspective of the egocentric camera, which makes unavailable, the view of the actor or his/her pose, standard action recognition techniques from third person actions are not applicable as is. Also quickly changing view field in typical egocentric videos makes it hard to develop models from foreground or background objects. Therefore, the techniques developed for wearer's action recognition have so far remained independent of work done in third person actions. The earliest work in wearer's action recognition used global features (GIST) for the task [11]. Later works focussed on objects present in the scene for recognition [12,28]. Position and pose of hand are important cues for action recognition involving object handling and have been explored by the researchers as well [5]. In action categories which do not involve any handled object, researchers have typically exploited the optical flow observed in the video, which for an egocentric video is indicative of head motion and is highly correlated with the kind of action being performed by the wearer [21,20]. Eye-motion and ego-motion have also been used to recognize indoor desktop actions [14].

Object or hand pose is an important cue for wearer's action but detecting them in an egocentric video is a difficult task and the dependence of the action recognition on such explicit detection/recognition affects the overall action recognition accuracy, besides making the system more complex and inefficient. We show in this paper that such prior information is not necessary. We observe that in any egocentric action scenario involving handled objects, the dominantly moving objects in the scene are typically hands and handled objects only (Fig. 1). Optical flow observed for the background is due to motion of the wearer's head. Such motion causes three dimensional rotation of the camera and can be easily compensated by cancelling frame to frame homography. This leads to a simple algorithm for extraction of hands and objects. We further show that complicated models of hand pose or object recognition are not necessary for the action recognition task, and instead, simple trajectory based features, combining motion profile and the visual features around these trajectories alone are

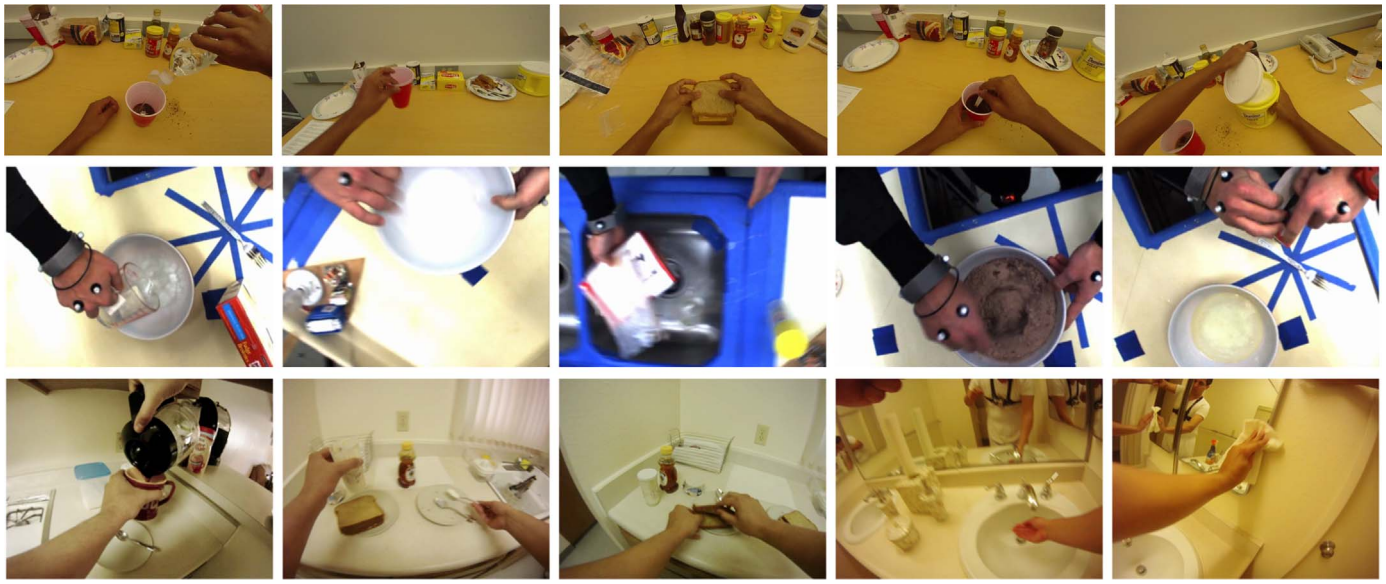

Fig. 2. Examples of wearer's action categories we propose to recognize in this paper from different datasets: GTEA [5] (top row), Kitchen [11] (middle row) and ADL [12] (bottom row). First, second and third columns across all rows are 'pour', 'take' and 'put' actions respectively. Fourth and fifth columns are 'stir' and 'open' actions for top and middle rows, and 'wash' and 'wipe' actions for bottom row. The actions vary widely across datasets in terms of appearance and speed of action. Features and technique we suggest in this paper is able to successfully recognize wearer's actions across different presented scenarios, showing robustness of our method.

sufficient to reach state of the art accuracy. This significantly simplifies the whole processing pipeline for first person action recognition. The simplification also allows to easily generalize the proposed technique to various kinds of actions, not possible with current state of the art, as we show later in the paper.

The focus of the paper is on first person actions in egocentric videos. The state of the art so far has overly stressed on object detection and hand segmentation. Our thesis and an important contribution is the observation that while objects and hands are important, explicitly segmenting or recognizing them is not necessary. This follows from the fact that the region of dominant motion in an egocentric video (after cancelling camera motion due to head) implicitly captures these salient objects and can be used directly for action recognition. We acknowledge that similar observations have been made in third person action recognition as well, where the state of the art uses features based on trajectories instead of complex segmentation/recognition. So far the important intuition has somehow failed to be applied in the egocentric domain. This may be partially attributed to the fact that unless looked at carefully, first person and third person videos look very different because of extreme camera shake due to head motion of the wearer. Therefore, trajectory based features from third person cannot be applied directly. Here, our second observation comes handy that camera motion due to head can be approximately but effectively cancelled by a simple homography. Rest of the contribution thereafter is application of various features that we found to be useful in our context. We understand that other types of features which keep the basic observations (1: Features along dominant motion regions after camera motion cancellation, 2: Motion cancellation by simple homography) intact can also be potentially used as well.

We believe that the simplicity is the strength of the our approach. Dependence on trajectory aligned features only allows our approach to generalize to datasets which are significantly different from each other. In contrast, none of the earlier state of the art has been shown to apply on all the datasets at the same time.

Most of the features used in this paper have been suggested earlier in principle. The choice of these features for the suggested approach is deliberate to some extent to show that once the basic two suggestions of the paper are followed, many commonly used ideas from third person action recognition start to become useful for first person as well. The paper, therefore, also serves to establish a bridge between first person and third person features which might be developed in future.

Contributions: We propose a novel representation of egocentric actions based on simple feature trajectories. Importantly, the proposed features can be computed using tracking alone. The features implicitly capture the visual and motion cues of hands and handled objects. This novel observation along with camera motion cancellation allows us to bypass the complicated steps (object detection, hand detection or image segmentation). Our paper is the first to propose the use of trajectory aligned features for egocentric action recognition. We use a bag of words model to learn action representation from trajectory based features. Our experiments on publicly available datasets show that the proposed technique improves the state of the art by more than 11%. We have explored the generalization of our features for action recognition when the wearer's hands or handled objects are not visible. We release an annotated database of 60 videos for 18 such action classes performed by different subjects.[1] Interestingly, our technique, not designed for such actions, gives an accuracy of 51.20% on the dataset. Even with a significantly simplified computing pipeline, we achieve state of the art results on all the publicly available egocentric datasets. This implies that the proposed features can be used for a variety of datasets with significant difference in appearance and actions. We note that none of the earlier proposals have been shown to apply on all the datasets at the same time.

[1] http://cvit.iiit.ac.in/projects/FirstPersonActions/

2. Related work

Action recognition has been a popular problem in computer vision. However, this is typically done from a third person view, for example, from a static or a handheld camera. A standard line of work is to encode the actions using keypoints and descriptors. This is done by extending spatial domain descriptors to space–time descriptors. These descriptors are then matched using Euclidean distance or other similar measures. Some techniques also rely on supervised learning with these descriptor vectors. Some notable contributions in this area include STIP [29], 3D-SIFT [30], HOG3D [31], extended SURF [32], and Local Trinary Patterns [33]. Methods that follow the pipeline of keypoint detection followed by an action descriptor usually work on a cuboidal video volume. They tend to merge the optical flow and the appearance information from the foreground and objects present in the scene. There have been proposals to demerge these two. Such attempts track feature points in a video and use these trajectories as cues for the action recognition. Some recent methods [34–37] show promising results for action recognition by leveraging the motion information of trajectories.

Camera motion is very common in real-world videos and poses a significant challenge to any action recognition technique. Wang et al. [38] propose a descriptor based on motion boundaries to reduce the interference from camera motion. They compute motion boundaries by a derivative operation on the optical flow field. Thus, motion due to locally translational camera movement is canceled out and relative motion is captured. There have been various improvisations on the technique [39–41] decomposing visual motion into dominant and residual motions both for extracting trajectories and computing descriptors.

Egocentric cameras have certain distinct advantages as well as constraints for action recognition. While having much lesser occlusion for objects is extremely useful, natural head motion of the wearer brings in large camera motion. Spriggs et al. [11] proposed to recognize first person actions using a mixture of GIST [42] features and IMU data. Their results confirm the importance of head motion in first person action recognition. Pirsiavash and Ramanan [12] attempt to recognize the activity of daily living (ADL). Their thesis is that first person action recognition is "all about the objects," and in particular, "all about the objects being interacted with." To recognize the objects from a first person view, they develop representations including (1) temporal pyramids, which generalize the well-known spatial pyramids to approximate temporal correspondence when scoring a model; and (2) composite object models that exploit the fact that objects look different when being interacted with. McCandless and Graumann [28] extend the work by using spatio-temporal pyramid histograms of objects appearing in the action. They devise a boosting approach that automatically selects a small set of useful spatio-temporal pyramid histograms among a randomized pool of candidates. In order to efficiently focus on the candidates, they propose an "object-centric" scheme that prefers candidates involving objects prominently involved in the actions. Fathi et al. [5] recognize the importance of hands in first person action recognition. They propose a representation for egocentric actions based on hand–object interactions and include cues such as optical flow, pose, size and location of hands in their feature vector. There is an assumption on the availability of hand, object and background labels in the video.

Objects are not always the most important cue in first person action recognition. In a sports video, when there are no prominent handled object, Kitani et al. [21] use motion based histograms recovered from the optical flow of the scene (background) to recognize the actions of the wearer. Ogaki et al. [14] use eye-motion and ego-motion to recognize indoor desktop actions. Recently, Ryoo et al. have suggested pooled motion features tracking how descriptor values are changing over time and summarizing them to represent an action in the video [43]. In a parallel independent work, Li et al. have also proposed a feature descriptor based upon dense trajectories [44]. However they also use complex patterns like gaze and hand pose, which we show are not necessary to reach state of the art accuracy. Convolutional neural networks (CNNS) have emerged as a useful tool for many computer vision tasks. Castro et al. [45] have tried to predict the daily life activities from egocentric images using deep neural networks. Recently, Singh et al. [46] have proposed to use the descriptors learned from multiple stream neural networks for first person action recognition.

3. Descriptor for first person actions

Motion of handled objects and hands is an important cue in first person action recognition. However, unlike previous approaches, we believe that segmentation and object recognition are not necessary for first person action recognition. We propose a novel idea to use simple point tracking based on the observation that in an egocentric video, the trajectory aligned features implicitly capture the visual and motion cues of hands and handled objects (Fig. 3). Existing trajectory based features cannot be used directly in an egocentric setting due to severe camera shake. To overcome this, we propose using simple camera motion cancellation as a preprocessing step. Our experiments corroborate the efficacy of this technique. Dependence on trajectory aligned features only allows our approach to generalize to datasets which are significantly different from each other and achieve state of the art results on all the datasets at the same time.

We propose an action descriptor based on the feature tracks obtained from egocentric video. The descriptor is an ensemble of different feature vectors obtained from feature tracks as well as from visual cues. We construct a bag of words representation separately for each such feature vector. To motivate the importance of each feature vector independently, we explain them sequentially below along with improvement in the accuracy by adding that feature vector in the descriptor.

We present the detailed run time analysis for each step in Table 5. Unlike the previous methods which involve hand and object segmentation, our method, though involves computation of many features, is fast and highly parallelizable. This is mainly due to the fact that we rely only on low level feature descriptors.

3.1. Baseline: dense trajectories

In third person action recognition, the constraints of feature representations derived from regularly shaped video volumes are well recognized. Therefore, the newer approaches rely on features computed along the trajectories. Typical keypoint detectors produce sparse feature trajectories affecting the quality of results. Use of feature points sampled on a regular grid has been proposed as a remedial measure. This leads to dense trajectories and improves stability and performance of the algorithms.

We use dense trajectory based feature [38] as a baseline for our work. As suggested by Wang et al. [38], we extract dense trajectories for multiple spatial scales. Feature points are sampled on a grid spaced by W pixels and tracked in each scale separately. Each point $P_t = (x_t, y_t)$ at frame t is tracked to the next frame $t+1$ by

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (\mathbb{M} * \omega)|_{(\bar{x}_t, \bar{y}_t)}$$

where $\mathbb{M}$ is the median filtering kernel, $\omega = (u_t, v_t)$ is a dense optical flow field, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of (x_t, y_t). Tracked points in subsequent frames are concatenated to form a trajectory: $(P_t, P_{t+1}, P_{t+2}, \ldots)$.

The shape of a trajectory encodes local motion patterns. Given a trajectory of length L, we describe its shape by a sequence $S = (\Delta P_t, \ldots, \Delta P_{t+L-1})$ of displacement vectors $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$. The resulting vector is normalized by the sum of the magnitudes of the displacement vectors as

$$S' = \frac{(\Delta P_t, \ldots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|}.$$

The vector is referred to as *trajectory descriptor*.

To leverage additional motion and appearance information in dense trajectories, we compute HOG and HOF descriptors within a space–time volume around the trajectory. The size of the volume is $N \times N$ pixels and L frames. The volume is subdivided into a spatio-temporal grid of size $n_\sigma \times n_\sigma \times n_\tau$. We use the default sampling step size of $W=5$ and 8 spatial scales spaced by a factor of $1/\sqrt{2}$, and parameters $N=32$, $n_\sigma = 2$, $n_\tau = 3$. Length of a trajectory is set to $L=15$ frames. Both HOG and HOF orientations are quantized into 8 bins using full orientations, with an additional zero bin for HOF. Both descriptors are normalized with their L_2 norm.

In order to classify the action at frame m, we take a sliding window of size $M+1$ frames and extract dense trajectories within this window. A sliding window centered at frame m consists of $(m - M/2, \ldots, m, \ldots, m + M/2)$ frames. In all our experiments, each sliding window consists of 31 frames ($M=30$). Frames at the border of the video are appropriately padded by reflection. We use a bag of words (BOW) model to represent the video segment. Vocabulary for each feature is built separately. For vocabulary construction, we randomly select 10% of training data and then use

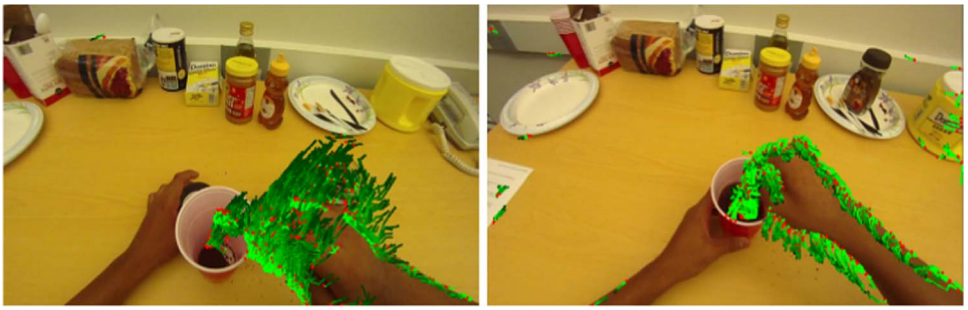

Fig. 3. First and second columns show the object and camera trajectories for 'pour' and 'stir' actions. There is enough information in the cues to classify first person actions. Similar works in egocentric vision use complex image segmentation algorithms to arrive at the labeling of hands and handled objects.

KMeans clustering and vector quantization for hard vocabulary assignment. Wang et al. [38] use vocabulary of size 4000 for each feature and concatenated histograms are used for classification using a one-*vs.*-rest SVM classifier with χ^2 kernel. In a similar way, we use vocabulary size of 2000 for all features for all experiments on GTEA dataset [5]. Later, the histograms corresponding to each feature are concatenated for classification. For classification, we train a one-*vs.*-rest SVM classifier using χ^2 kernel. The classifier parameters are estimated using 4-fold cross validation.

The experiments conducted using dense trajectories with HOG and HOF descriptors resulted in an accuracy of 50.17% and 30.16% respectively on GTEA dataset [5]. We give more details in Section 5.

3.2. Motion cues: motion boundary histogram

Dalal et al. [47] proposed the motion boundary histogram (MBH) descriptor for human detection from a moving camera, where derivatives of flow instead of raw optical flow itself are used. In egocentric videos, the use of the gradient of the optical flow counters the effect of head motion by suppressing the flow of the background. Therefore, we also use MBH in the proposed scheme. For computing the MBH descriptor, we compute the spatial derivative of the optical flow field $I_\omega = (I_x, I_y)$, and orientation information is quantized into histograms, similar to the HOG descriptor. We then obtain an 8-bin histogram for each component (MBHx and MBHy) and normalize them separately with the L_2 norm.

The experiments conducted using MBH descriptor resulted in 48.69% accuracy on GTEA dataset and using HOG+HOF+MBH improves the accuracy to 50.83%. For $L=15$, the lengths of descriptors are 30, 96, 108 and 192 dimensions for trajectory descriptor, HOG, HOF and MBH respectively.

Our comparison with PoT features [43] while using same underlying feature descriptors (HOG+HOF+MBH) confirms the effectiveness of our simple trajectory aligned features. The experiments show that using PoT features on raw video segment gives 45.60% accuracy on GTEA dataset while our method achieves 54.61%.

3.3. Action in reverse: bi-directional trajectories

We observe that human beings can recognize an action even if it is played in reverse. Flow fields, and hence, HOF as well as MBH histograms are different but meaningful when features are computed in the reverse direction. By adding extra information from reverse playback into the feature allows us to detect the action by using information from both playback directions in one go. We use features from forward and reverse trajectory as if they are obtained from independent trajectories, and hence the name 'bi-directional' trajectories. The lengths of descriptors obtained from a bi-directional trajectory are same as traditional dense trajectory descriptors described in the earlier section. The trajectories obtained from both playback directions are used to build BOW representation together (instead of separate histogram for each) and hence do not affect the BOW histogram size.

Using bi-directional trajectories improves the frame level first person action recognition accuracy from 50.83% to 54.61% on the GTEA dataset.

3.4. Handling wild motion: head motion cancellation

The motion of the camera due to head motion of the wearer pollutes the observed trajectories in an egocentric video. By applying head motion cancellation on the flow (see Fig. 4), the observed trajectories tend to be smooth and enhance object and hand motion. We model the observed motion due to head movement as 2*D* affine and cancel such motion from trajectory descriptor computed earlier. We observe an improvement in accuracy from 54.61% to 56.87% after cancelling head motion. Interestingly, we observe that camera stabilization as pre-processing also leads to similar gains.

3.5. Fast and slow actions: temporal pyramids

The bag of words representation of trajectory aligned features that we have presented so far ignores the temporal structure of activities. To overcome the limitation, we represent features in a temporal pyramid, where the top level is a histogram over full temporal extent of the video segment, the next level is the concatenation of two histograms obtained by temporally segmenting the video into two halves (while quantization) and so on. The frame where the trajectory first appears is used to decide the histogram to which it is assigned. All levels of pyramid have the same BOW histogram size that we have discussed earlier. We obtain a coarse-to-fine representation by concatenating all such histograms together. We use a three-level pyramid for HOG and HOF in our experiments. This makes feature dimension size 14,000 or $2000 \times (1 + 2 + 4)$ for HOG as well as HOF and 4000 for MBH. Using temporal pyramid further improves the frame level action recognition to 58.50% on GTEA dataset.

3.6. Kinematic and statistical features

As kinematic features, we use local first-order differential scalar quantities computed on the flow field around the trajectories. We consider the divergence, the curl and the hyperbolic terms similar to [39]. They encode the physical pattern of the flow which is useful for action recognition. We also use statistics related to trajectories from entire video segment as features. These features are number of the trajectories, and the average and standard deviation for x and y coordinates of the trajectory. Trajectory length as well as net displacement of tracked points in horizontal and vertical directions is added as features. A number of trajectories heading towards each quadrant normalized by total number of trajectories are also appended to it. Kinematic and statistical features improve frame level action recognition on GTEA to 60.11%.

Fig. 4. Motion of the egocentric camera is due to 3*D* rotation of wearer's head and can be easily compensated by a 2*D* homography transformation of the image. Left: optical flow overlayed on the frame. Right: compensated optical flow followed by thresholding. Almost all camera motion has been compensated by this simple technique.

3.7. Egocentric cues: camera activity

Camera motion in an egocentric video is due to motion of the wearer's head and is an important cue for action recognition. We represent the camera motion as a global frame to frame 2D translation, denoted as $\Delta c_M = (\Delta x, \Delta y)$. For a video consisting of $M+1$ frames, a camera activity descriptor C is described by a sequence $C = (\Delta c_1, \ldots, \Delta c_M)$ of displacement vectors. The vector C is normalized by the sum of the magnitudes of the displacement vectors as

$$C' = \frac{(\Delta c_1, \ldots, \Delta c_{M-1})}{\sum_{j=1}^{M-1} \| \Delta c_j \|}.$$

We concatenate C', total displacement, displacement average and standard deviation to represent camera movement and refer to it as *camera activity feature*. Using camera activity feature improves the frame level, first person action recognition accuracy from 60.11% to 61.23% on GTEA dataset.

3.8. Semantically meaningful temporal segmentation using proposed features

We have described our features within the context of a classification problem so far. However, the features can also be used for temporal segmentation of egocentric videos. For such semantic segmentation, we pose our problem as a probabilistic graphical model (MRF) where likelihood is derived from classifier score and smoothness prior is used as the regularizer. Modelling the problem in this way helps to overcome the difficulties in recognizing an action boundary by only likelihood based formulation without prior. We formulate the segmentation problem as follows. Consider a weighted graph where each vertex is a frame which can be labelled with an action label. Each pair of neighboring frames with the same action label will be connected by a low weight edge whereas a pair of neighboring frames with different action (action boundary) will be connected with a higher weight edge. Neighborhood of each frame is defined as 5 temporally adjacent frames on both sides (past and future). We assign edge weight using Euclidean distance between global HOF histograms of two neighboring vertices.

The intuition here is that the change in action between frames should cause a significant change in flow magnitudes and directions in neighboring frames. We proceed to estimate the minimum energy cut using the α-expansion algorithm. We report segmentation accuracy of 62.50% using the proposed formulation on GTEA dataset. Fig. 5 illustrates the segmentation result and errors using the proposed approach for GTEA dataset.

4. Datasets and evaluation protocol

In our work, we use four different publicly available datasets of egocentric videos: GTEA [5], Kitchen [11], ADL [12] and UTE [16]. Out of these, only GTEA and Kitchen datasets have frame level annotations for first person actions. For ADL and UTE datasets, where similar action level labelling was not available, we selected a subset of the original dataset and manually annotated the short term actions in the parts where a wearer is manipulating some object. Other kind of actions such as walking, watching television, etc. is labelled as 'background'. Statistics related to datasets are shown in Table 1.

GTEA dataset: This dataset consists of 28 videos, captured using head mounted cameras. There are 4 subjects, each performing 7 long term activities in a kitchen. Each activity is approximately 1 min long. We follow the experimental setup of [8] and use videos of subject 'S2' for testing and others for training. There are 11 action classes, viz., 'close', 'pour', 'open', 'spread', 'scoop', 'take', 'fold', 'shake', 'put', 'stir', and 'background'.

Kitchen dataset: The original dataset consists of videos of 43 subjects performing 3 activities, captured using head mounted camera and IMUS. Camera point of view is from top, and severe camera motion is quite common. Similar to [11], we select 7 subjects from 'Brownie' activity, train using videos of 6 subjects and test on the video of remaining subject. There are 29 classes of actions in this dataset. The action classes are 'Open cupboard (bowls)', 'Get fork', 'Open cupboard (brownie)', 'Walk to fridge', 'Open fridge', 'Get eggs', 'Close fridge', 'Walk to counter', 'Break one egg', 'Beating egg(s)', 'Pour in water in bowl', 'Get oil from cupboard', 'Pour oil in cup', 'Put oil away', 'Open brownie box', 'Pour in brownie mix', 'Pour oil in bowl', 'Stir brownie mix', 'Get baking pan', 'Spray with Pam', 'Put Pam away', 'Set stove settings', 'Pour mix in baking pan', 'Put pan in oven', 'Pour tap water in cup', 'Put cap on', 'Get Pam from cupboard', 'Remove cap', and 'Read recipe'.

ADL videos dataset: The original dataset consists of videos of 20 subjects performing 18 daily life activities, captured using chest mounted camera with 170° viewing angle. We selected 5 subjects and manually annotated the short term actions with 21 action labels. Similar to [8], we use videos of one subject for testing and the rest for training. The action classes are 'stir', 'cut', 'shake', 'switch on/off', 'take', 'open', 'close', 'fold', 'put', 'flip', 'pour', 'wash', 'write', 'scoop', 'wipe', 'wear', 'tear', 'dip', 'spray', 'type' and 'background'.

UTE dataset: Original UTE dataset [16] contains 4 videos captured from head-mounted cameras. Each video is about 3–5 h long, captured in a natural, uncontrolled setting. We select three parts where hand motion is dominant from two subjects and manually annotate the short term actions. The action labels are 'stir', 'cut', 'shake', 'switch on/off', 'take', 'open', 'close', 'fold', 'put', 'flip', 'pour', 'wash', 'wipe', 'tear', 'tap', 'mix', 'peel', 'scrub', 'rub', 'move' and 'background'.

IIIT Extreme Sports: Most of the egocentric action databases we have come across contain actions where wearer's hands or objects are visible. We are also interested in the performance of our features when such cues are not available. Kitani et al. [21] have suggested unsupervised clustering of such actions for sports videos but the dataset provided by them is quite small (6 categories each with only one video). We are introducing a new bigger dataset of similar actions with this paper. We refer to the dataset as 'IIIT Extreme Sports'. The dataset contains 60 videos, amounting to

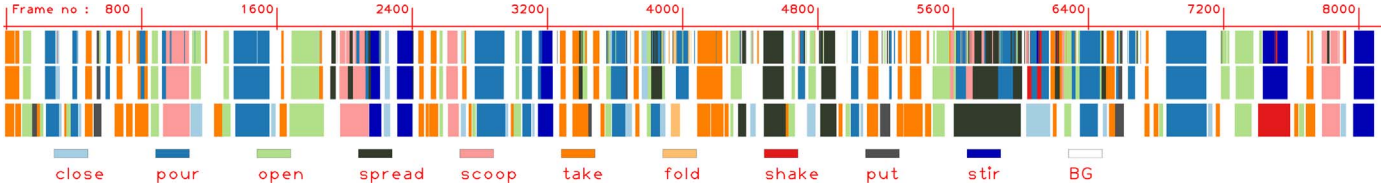

Fig. 5. Semantically meaningful temporal segmentation using proposed features: error visualization on all test frames (7 videos) of GTEA dataset. Each action label has been color coded. We use MRF based method for refining predicted label. We assign penalty according to difference in global HOF histogram of a frame when compared with that of its neighbors. Predicted action labels using classifier score are shown in the top row, action labels after MRF based temporal segmentation in the middle row and ground truth action labels in the bottom row. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 1
Statistics of egocentric videos datasets used for experimentation. The baseline accuracy is as achieved using dense trajectory method of Wang et al. [38]. The proposed approach uses various trajectory aligned features and improves the baseline as well as the state of the art result on all the datasets tested. The datasets vary widely in appearance, subjects and actions being performed, and the improvement on these datasets validates the generality of the suggested descriptor for egocentric action recognition task. Note that originally, ADL dataset has been used for activity recognition and UTE for video summarization and not for action recognition as in this paper. Therefore, comparative results are not available for these datasets.

Dataset	Subjects	Videos	Frames	Classes	Baseline accuracy [38]	State of the art accuracy	Our accuracy	Temporal segmentation
GTEA [5]	4	28	31,253	11	45.15%	47.70% [8]	61.23%	62.50%
Kitchen [11]	7	7	48,117	29	44.80%	48.64% [11]	59.74%	61.42%
ADL [12]	5	5	93,293	21	20.10%	–	31.40%	35.16%
UTE [16]	2	3	208,230	21	31.78%	–	52.62%	55.20%
IIIT Extreme Sports	60	60	412,250	18	43.81%	–	51.20%	53.30%

Fig. 6. Sample frames from the 'IIIT Extreme Sports' dataset introduced by us. The figure shows examples for 'jump' action in different sports categories: ski, jetski, mountain biking and parkour. Note the variations among the samples which makes the dataset extremely challenging for action recognition task.

nearly 8 h, from 5 extreme sports categories (mountain biking, jetski, skiing, speedflying and parkour). We have annotated the videos manually into 18 short term ego-actions similar to Kitani et al. [21]: 'forward', 'bumpy forward', 'curve-left', 'curve-right', 'turn-left', 'turn-right', 'left-right', 'jump', 'slide-stop', 'run', 'walk', 'roll', 'flip', 'climb', 'vault', 'lift', 'fly' and 'spin'. Fig. 6 shows some samples from the database.

These sports categories are very different from each other in terms of terrain, nature and types of actions. Due to fast movement nature of extreme sports, severe camera shake and motion blur are very common. We selected first person action classes similar to ego-actions used in [21]. Each video is captured using head-mounted cameras in diverse terrain (mountain, snow, river, sea, and air), weather and lighting.

4.1. Evaluation protocol

We consider short term actions performed by different subjects while performing different activities. Speed and nature of actions vary across subjects and activities (e.g., consider the action 'open' in two scenarios, 'open' water bottle and 'open' cheese packet). Formally, classification accuracy for first person action recognition task is defined as the number of frames (or video segment) classified correctly divided by the total number of frames (or number of video segments) in the videos used for testing. Frame level action recognition is important for continuous video understanding. This is also crucial for many other applications (e.g., step-by-step guidance based on wearer's current actions). We also evaluate our method for action recognition at the video segment level. In this case, there is only one action in each video segment. However, the length of the segment is not fixed. In this setting, we have an approximate knowledge of action boundaries which naturally improves action recognition results. Segment level action recognition is different from temporal segmentation as each segment is independent of each other. For temporal, segmentation we perform labelling of each frame without explicit knowledge about action boundaries.

5. Experiments and results

We first present our experiments and analysis of the proposed action descriptor on GTEA dataset to bring out the salient aspects of the suggested approach. Experiments with other datasets are described later. Note that these datasets are quite different from each other, and performance improvement on all these datasets compared to the current state of the art show the generality of our features.

In these datasets the duration of action varies from a few frames to a few hundred frames. The size of the sliding window plays a crucial role in correctly classifying an action. There can be more than one action within a sliding window at the action boundaries, leading to noisy training data. Due to this reason, we do not use features extracted from frames at the action boundaries for vocabulary construction and SVM training. However, all the frames are used for testing.

The annotated dataset and the source code for the paper are available at the project page: http://cvit.iiit.ac.in/projects/FirstPersonActions/

5.1. Results on different datasets

We follow the experimental setup of Fathi et al. [8] for GTEA dataset. They perform joint modelling of actions, activities and objects, on activities of three subjects and predict actions on activities of one subject. They have reported an accuracy of 47.70% using their method. Table 2 summarizes our analysis of the effect of different parameters on the performance of our descriptor on the dataset.

We have done a comparison (Table 3) with Pooled Time Series feature [43] using their released code. Note that, when using same underlying local feature descriptors (HOG, HOF, and MBH) and same temporal pooling extent (3 level temporal pyramid), PoT gives an accuracy of 49.14% on GTEA dataset. In contrast, our result is 58.50%, which clearly outperforms the pooled features by a significant margin. When combined with kinematic, statistical and camera activity features, the result further improves to 61.23%. The

Table 2
Effect of different parameters on the performance of our algorithm. The experiments are done on GTEA dataset. We use trajectory, HOG, HOF and MBH features using 2K vocabulary size for each feature for the experiment. Accuracy reported is computed per frame.

Method	Feature	Accuracy (%)
Uni-directional trajectory	Trajectory	25.36
	HOG	50.17
	HOF	30.16
	MBH	48.69
	HOG + HOF + MBH	50.83
Bi-directional trajectory	Trajectory	27.09
	HOG	51.25
	HOF	35.41
	MBH	48.87
	HOG + HOF + MBH	54.61
Affine-flow compensation	HOG + HOF + MBH	56.87
Camera stabilization	HOG + HOF + MBH	57.10
With 3 level pyramid	$HOG^{Pyr} + HOF^{Pyr} + MBH$	58.50
Combined	$HOG^{Pyr} + HOF^{Pyr} + MBH$ and Kinematic + Statistical + Camera activity	61.23

Table 3
Comparisons with pooled features (PoT) [43] using three-level temporal pyramid (HOG + HOF + MBH) on (a) GTEA dataset and (b) IIIT Extreme Sports dataset.

Input	PoT	Ours
(a) Comparison on GTEA dataset		
Raw video segment (%)	45.60	54.61
Stabilized video segment (%)	49.14	61.23
(b) Comparison on IIIT Extreme Sports dataset		
Raw video segment (%)	47.51	50.08
Stabilized video segment (%)	48.28	51.20

primary reason might be that pooled features do not seem to consider salient regions specially. We expect that there might be some merits in considering trajectory aligned pooled features. We have also done a similar comparison on our IIIT Extreme Sports dataset as well (see Table 3(b)). Note that the action classes in our IIIT Extreme Sports dataset are similar to UEC dataset [21], however, the dataset itself is much larger in size. In this case, the hands or objects are not visible and our method is not specifically designed for such actions, but we still observe the superior performance of our method.

We extend our experiments to other publicly available egocentric video datasets. Results on these datasets are shown in Table 4. We follow the same experimental setup as [11] and perform frame level action recognition for 'Brownie' activity for 7 subjects. Spriggs et al. [11] report an accuracy of 48.64% accuracy when using first person data alone and 57.80% when combined with IMU data. We achieve 59.74% accuracy using our method on egocentric video alone.

Table 4
Our results for first person action recognition on different egocentric videos datasets. Sliding window based approach for classification used in our algorithm performs poorly at action boundaries. Therefore, the accuracy for segment level classification, when the action boundaries are clearly defined, comes out higher.

Dataset	Accuracy (%)		
	Frame level	Segment level	Chance level
GTEA [5]	61.23	77.40	9
Kitchen [11]	59.74	60.00	3.4
ADL [12]	31.40	31.82	4.7
UTE [16]	52.62	55.12	4.7
IIIT Extreme Sports	51.20	55.74	5.5

The ADL dataset has been used for long term activity recognition by [12] in the past. We annotated the dataset with the short term actions and tested our method on it. Similar to our experiment on GTEA, we test our model on one subject while using the other for training. We achieve 31.40% accuracy at frame level and 31.82% at the video segment level using our method. Note that, ADL dataset is much larger and challenging dataset when compared to others. ADL contains actions from a diverse set of 18 activities while GTEA contains 7 activities and Kitchen dataset contains only one activity.

The UTE dataset has been used for video summarization by [16] in the past. Motion blur and low image quality is fairly common in this dataset. For action recognition, we achieve 52.62% accuracy at frame level and 55.12% at the video segment level using our method.

On our IIIT Extreme Sports dataset, where objects and hands are not visible, our method achieves similar performance, 51.20% at frame level and 55.74% at segment level. Short trajectories prove to be useful for short term actions even in severe camera or head motion, which is fairly common in first person videos of extreme sports.

The proposed action descriptor improves the baseline as well as the state of the art on all the five datasets tested upon (see Table 1 for the details about dataset and comparison details with baseline). Fig. 2 shows some of the actions from different datasets correctly classified by our approach. Note the difference in appearance.

5.2. Failure analysis

We rely on motion and appearance based cues for action recognition. While statistical and trajectory aligned features are useful for all the action classes, camera activity feature is particularly helpful with actions that have specific camera motion such as 'pour', 'stir' and 'shake'. Though highly discriminatory, we do see the instance when such features fail to classify correctly because of either dominant visual similarity or motion similarity or both. Yet some other errors arise due to limited capability of the proposed action descriptors to describe the action complexity and various ways in which the same action could have been performed. Fig. 7 shows some failure cases and possible reason for the failures.

Fig. 8 gives the confusion matrix of the proposed approach for the GTEA dataset. A large portion of observed errors occur on the action boundaries where the features from two actions merge. Part of it may be attributed to inherent ambiguity in the problem itself. One cannot say at which instant the action has started or has ended. For example, consider the action 'open' with the object 'water bottle'. One may consider the instant the hand starts interaction with water bottle is the start of action, while other may agree that the moment the hand starts twisting the cap of water bottle as the start of action (Fig. 7(c)). Also, most action occurs before or after 'BG' (see Fig. 5), hence the most common confusion with almost all the actions. Also note that, a high percentage error for some classes (e.g., fold, shake, put, etc.) is due to very few samples of those actions in the dataset. For example, 'fold' action accounts for less than 0.5% of all the actions in the dataset and has only 82 frames for training and 54 frames for testing. Handling multiple complex actions and the action boundaries are the weak points of the proposed framework and directions for our future research. The presence of multiple actions poses another

Fig. 7. Some failure cases of our method. (a) 'shake' classified as 'stir' due to high visual and motion similarity. On the right, a similar frame with 'stir' action classified correctly. (b) 'pour' classified as 'spread' due to hand movement. Notice the high similarity between 'pouring' mayonnaise and 'spreading' jam. On the right, a frame classified correctly as 'spread'. A large portion of observed errors occur on the action boundaries where the features from two actions merge. (c) shows two frames which are at action boundary 'open' (left, predicted correctly) and 'BG' (right, predicted as 'open'), and (d) on the left, 'fold' classified as 'pour' due to very few samples for 'fold' available in the dataset. 'fold' action accounts for less than 0.5% of all the actions in the dataset and has only 82 frames for training and 54 frames for testing. On the right, a frame classified correctly as 'pour'. Same objects present in left and right images might have led to the confusion. We believe, our method requires more examples of such scarce actions to distinguish between these cases.

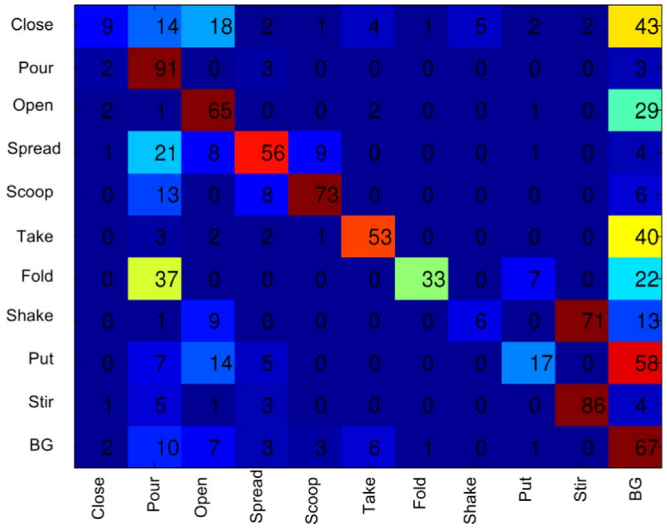

Fig. 8. Confusion matrix for our method on GTEA dataset. We observe that many errors occur because action boundary is not clearly defined. 'close' is commonly confused with 'open' due to similarity in the nature of the action. Also, most action occurs before or after 'background', hence the common confusion. High percentage error for some classes (e.g., fold, shake, put, etc.) is because very few sample of those actions in the dataset.

challenge. Enhancing the proposed action descriptor when the two actions are being performed jointly is another area of future research.

5.3. Implementation details and runtime analysis

We start by extracting frame to frame dense optical flow using the algorithm by Färneback [48] as implemented in the OpenCV library. We found this algorithm to be a good compromise between accuracy and speed. We further apply 5×5 median filter to smoothen the optical flow which is then used for multiscale point tracking as mentioned in [38]. As discussed earlier, we use the default value for parameters $N=32$, $n_\sigma = 2$, $n_\tau = 3$. Length of a trajectory is set to $L=15$ frames and length of video segment is set to 31 frames. Around these trajectories we extract

- trajectory descriptor (30 dimensional);
- texture and appearance descriptor: HOG (96 dimensional);
- motion descriptors: HOF (108 dimensional), MBH (96 dimensional each for MBHx and MBHy, or 192 dimensional);
- kinematic features (288 dimensional);
- statistical features (20 dimensional);
- camera activity descriptor (60 dimensional).

Table 5
Detailed runtime analysis of our method. For a video segment at 15 fps and with 31 frames, feature extraction from our bi-directional trajectories takes 0.71372 s while the complete pipeline takes 2.2047 s on an average. The runtime for each component is averaged over 100 iterations. We use serial single thread CPU implementation for all the steps in our pipeline. We limit all trajectories to length of 15 frames. The overall runtime is for feature extraction from a video segment of 31 frames at a spatial resolution of 360×240 pixels. We use vocabulary of size 2K for each feature.

Component	Type	Time (s)
Optical flow (Färneback algorithm)	Frame to frame	0.02658
Affine-flow compensation	Frame to frame	0.00483
Camera stabilization	Frame to frame	0.00531
Feature extraction (uni-directional)	Per video segment	0.35285
Feature extraction (bi-directional)	Per video segment	0.71372
Vocabulary assignment	Per video segment	0.54194
Vocabulary assignment (3 level pyramid)	Per video segment	1.47098
Overall runtime (video at 15 fps)	Per video segment	2.2047

Feature extraction from a video segment of 31 frames around bi-directional trajectories takes 0.71372 s on an average on an Intel Core i7-4790K CPU at 4.0 GHz (see Table 5).

We use statistical and camera activity feature as it is. For other features, we build a BoW representation using a uniform vocabulary size of 2K. For vocabulary construction, we randomly select 10% of all trajectories from training videos for clustering as the

total number of trajectory is too large. We use hierarchical KMeans for clustering for its speed and efficiency. For vocabulary assignment, we use Fast Approximate Nearest Neighbor Search (FLANN) [49] with four randomized kd-trees, which we found to have good accuracy versus speedup trade-off when compared to nearest neighbor search. The complete pipeline (optical flow, video stabilization, feature extraction and computing BoW representation) takes 2.2047 s on an average from a video segment of 31 frames on Intel Core i7-4790K CPU at 4.0 GHz (see Table 5).

Total feature dimension when using temporal pyramid as discussed earlier is 24,080 or 14, 000 + 4000 + 6000 + 20 + 60. Such BOW histograms feature is very sparse. With the mentioned features, we train a multiclass Support Vector Machine (SVM) using Liblinear library [50]. We use VLFeat's [51] homogeneous kernel map, which is a finite dimensional linear approximation of homogeneous kernels, including the intersection and χ^2 kernels. In all experiments, we use homogeneous kernel map for χ^2 kernel of order 3. Using homogeneous kernel mapping helps us reduce the SVM training time by a significant amount.

6. Conclusions

We propose a new action descriptor for first person action recognition from egocentric videos. In the absence of wearer's pose, the important cues for such action recognition tasks are objects present in the scene, how they are being handled and the motion of the wearer. The proposed descriptor accumulates all such cues by a novel combination of features from trajectories, HOG, HOF, MBH, as well as kinematic and statistical features. We also explore the importance of head motion and capture it using camera activity features. The proposed feature and bag of words model is able to adequately learn the representation and improves the state of the art in terms of accuracy by more than 11%. We validate the proposed descriptor by testing on widely varying egocentric action dataset. The performance improvement on all the datasets validates the generalizability of the proposed descriptor. Our method gives similar performance for action recognition even when handled objects or wearer's hands are not visible.

The thesis of our work and an important conceptual contribution is the observation that while objects and hands are important in first person actions, explicitly segmenting or recognizing them is not necessary. It may be noted that trajectory based features cannot be applied as is to egocentric actions, as shown in our baseline in Table 1. This is due to the extreme shake present in egocentric videos because of motion of wearer's head. Our second thesis is that for the purpose of egocentric actions, such motion can be adequately compensated using homography alone.

Another crucial contribution is to create a bridge between first person and third person action recognition techniques. Many of the proposed features have been used in problems from areas other than egocentric. Their use for egocentric actions now looks obvious after our experiments and findings. However, none of the prior art for egocentric actions cited in the paper have used such features.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.patcog.2016.07.031.

References

[1] Google glass, ⟨https://www.google.com/glass/start/⟩.
[2] Gopro, ⟨http://gopro.com/⟩.
[3] Bbc news, ⟨http://www.bbc.com/news/world-us-canada-30281735⟩.
[4] Orcam, ⟨http://www.orcam.com/⟩.
[5] A. Fathi, X. Ren, J.M. Rehg, Learning to recognize objects in egocentric activities, in: CVPR, 2011.
[6] X. Ren, M. Philipose, Egocentric recognition of handled objects: Benchmark and analysis, in: CVPRW, 2009.
[7] X. Ren, C. Gu, Figure-ground segmentation improves handled object recognition in egocentric video, in: CVPR, 2010.
[8] A. Fathi, A. Farhadi, J. M. Rehg, Understanding egocentric activities, in: ICCV, 2011.
[9] A. Fathi, Y. Li, J.M. Rehg, Learning to recognize daily actions using gaze, in: ECCV, 2012.
[10] M.S. Ryoo, L. Matthies, First-person activity recognition: what are they doing to me?, in: CVPR, 2013.
[11] E.H. Spriggs, F. De La Torre, M. Hebert, Temporal segmentation and activity classification from first-person sensing, in: CVPRW, 2009.
[12] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: CVPR, 2012.
[13] S. Sundaram, W.W.M. Cuevas, High level activity recognition using low resolution wearable vision, in: CVPRW, 2009.
[14] K. Ogaki, K.M. Kitani, Y. Sugano, Y. Sato, Coupling eye-motion and ego-motion features for first-person activity recognition, in: CVPRW, 2012.
[15] K. Matsuo, K. Yamada, S. Ueno, S. Naito, An attention-based activity recognition for egocentric video, in: CVPRW, 2014.
[16] Y.J. Lee, J. Ghosh, K. Grauman, Discovering important people and objects for egocentric video summarization, in: CVPR, 2012.
[17] Z. Lu, K. Grauman, Story-driven summarization for egocentric video, in: CVPR, 2013.
[18] O. Aghazadeh, J. Sullivan, S. Carlsson, Novelty detection from an ego-centric perspective, in: CVPR, 2011.
[19] A. Fathi, J.K. Hodgins, J.M. Rehg, Social interactions: a first-person perspective, in: CVPR, 2012.
[20] Y. Poleg, C. Arora, S. Peleg, Temporal segmentation of egocentric videos, in: CVPR, 2014.
[21] K.M. Kitani, T. Okabe, Y. Sato, A. Sugimoto, Fast unsupervised ego-action learning for first-person sports videos, in: CVPR, 2011.
[22] B. Xiong, K. Grauman, Detecting snap points in egocentric video with a web photo prior, in: ECCV, 2014.
[23] Y. Poleg, T. Halperin, C. Arora, S. Peleg, Egosampling: fast-forward and stereo for egocentric videos, in: CVPR, 2015.
[24] J. Kopf, M. Cohen, R. Szeliski, First-person hyperlapse videos, in: TOG, 2014.
[25] Y. Li, A. Fathi, J.M. Rehg, Learning to predict gaze in egocentric video, in: ICCV, 2013.
[26] Y. Hoshen, S. Peleg, Egocentric video biometrics, CoRR.abs/1411.7591.
[27] Y. Poleg, C. Arora, S. Peleg, Head motion signatures from egocentric videos, in: ACCV, 2014.
[28] T. McCandless, K. Grauman, Object-centric spatio-temporal pyramids for egocentric activity recognition, in: BMVC, 2013.
[29] I. Laptev, On space–time interest points, in: IJCV, 2005.
[30] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: ACMMM, 2007.
[31] A. Klaser, M. Marszalek, A spatio-temporal descriptor based on 3D-gradients, in: BMVC, 2008.
[32] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: ECCV, 2008.
[33] L. Yeffet, L. Wolf, Local trinary patterns for human action recognition, in: ICCV, 2009.
[34] P. Matikainen, M. Hebert, R. Sukthankar, Trajectons: Action recognition through the motion analysis of tracked features, in: ICCVW, 2009.
[35] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: ICCV, 2009.
[36] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, J. Li, Hierarchical spatio-temporal context modeling for action recognition, in: ICCV, 2009.
[37] J. Sun, Y. Mu, S. Yan, L.-F. Cheong, Activity recognition using dense long-duration trajectories, in: ICME, 2010.
[38] H. Wang, A. Klaser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in: CVPR, 2011.
[39] M. Jain, H. Jégou, P. Bouthemy, Better exploiting motion for better action recognition, in: CVPR, 2013.
[40] H. Wang, C. Schmid, Action recognition with improved trajectories, in: ICCV, 2013.
[41] E. Kraft, T. Brox, Motion based foreground detection and poselet motion features for action recognition, in: ACCV, 2014.
[42] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, in: IJCV, 2001.
[43] M. S. Ryoo, B. Rothrock, L. Matthies, Pooled motion features for first-person videos, in: CVPR, 2015.
[44] Y. Li, Z. Ye, J.M. Rehg, Delving into egocentric actions, in: CVPR, 2015.
[45] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, I. Essa, Predicting daily activities from egocentric images using deep learning, in: ISWC, 2015.
[46] S. Singh, C. Arora, C. V. Jawahar, First person action recognition using deep learned descriptors, in: CVPR, 2016.
[47] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of

flow and appearance, in: ECCV, 2006.
[48] G. Färneback, Two-frame motion estimation based on polynomial expansion, in: Scandinavian Conference on Image Analysis, 2013.
[49] M. Muja, D. G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration, in: VISAPP, 20009.
[50] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification, in: IJMLR, 2008.
[51] A. Vedaldi, B. Fulkerson, VLFeat: An Open and Portable Library of Computer Vision Algorithms, in: ACMMM, 2010.

Suriya Singh is currently pursuing M.S. in Computer Science at IIIT Hyderabad. His areas of interest are computer vision and machine learning.

Chetan Arora received his Bachelor's degree in Electrical Engineering, in 1999, and the Ph.D. degree in Computer Science, in 2012, both from IIT Delhi. He is currently an Assistant Professor at IIT Delhi. His broad areas of research include computer vision and discrete optimization.

C.V. Jawahar is a Professor at IIIT Hyderabad, India. His areas of research include robotic and computer vision, machine learning and document image analysis.