First Person Action Recognition Using Deep Learned Descriptors

Suriya Singh [1] Chetan Arora [2] C. V. Jawahar [1]
[1] IIIT Hyderabad, India [2] IIIT Delhi, India

Abstract

We focus on the problem of wearer's action recognition in first person a.k.a. egocentric videos. This problem is more challenging than third person activity recognition due to unavailability of wearer's pose and sharp movements in the videos caused by the natural head motion of the wearer. Carefully crafted features based on hands and objects cues for the problem have been shown to be successful for limited targeted datasets. We propose convolutional neural networks (CNNs) for end to end learning and classification of wearer's actions. The proposed network makes use of egocentric cues by capturing hand pose, head motion and saliency map. It is compact. It can also be trained from relatively small number of labeled egocentric videos that are available. We show that the proposed network can generalize and give state of the art performance on various disparate egocentric action datasets.

1. Introduction

With availability of cameras from GoPro [2], Google Glass [1], Microsoft SenseCam [3] etc., the wearable cameras are becoming a commodity allowing people to generate more and more egocentric video content. By making first person point of view available, egocentric cameras have become popular in applications like extreme sports, law enforcement, life logging and home automation.

The egocentric community has been trying to develop or adapt solutions to a wide variety of computer vision problems in the new emerging context. Work done in last few years has ranged from problems like object recognition [10, 35, 36], activity recognition [7, 9, 27, 29, 31, 37, 41, 42] to more applied problems like summarization [4, 22, 26], and predicting social interactions [8]. Interesting ideas which exploit special properties of egocentric videos have been proposed for problems like temporal segmentation [16, 33], frame sampling [34, 49] and hyperlapse [18]. Newer areas specific to egocentric vision such as gaze detection [24] and camera wearer identification [11, 32] have also been explored.

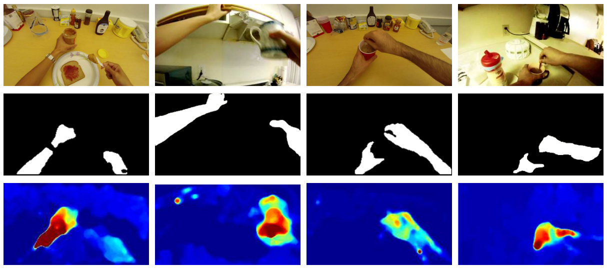

Figure 1: Hands and object motion pattern are important cues for first person action recognition. The first two columns show 'take' action whereas last two show 'stir'. Notice the wide difference in appearance. Second and third row shows hand mask and saliency map derived from dominant motion. In several cases hands are occluded by the handled object and hence the partial hand mask is obtained. We train a compact convolutional neural network using such egocentric cues. We achieve state of the art accuracy for first person action recognition. Our approach can be applied to datasets that differ widely in appearance and actions, without requiring any hand tuning. We further improve the performance by using pretrained networks for third person videos in a multi stream setting.

We focus on the recognition of wearer's actions (or first person actions) from egocentric videos in each frame. We consider short term actions that typically last few seconds, e.g., pour, take, open etc. We do not assume any prior temporal segmentation. First row in Figure 1 shows the frames corresponding to some example actions.

Recognition of wearer's actions is a natural first step in many egocentric video analysis problems. The problem is more challenging than third person action recognition because of unavailability of the actor's pose. Unlike third person actions where the camera is either static or smoothly moving, there are large shakes present in the egocentric videos due to head motion of the wearer. The sharp change in the viewpoint makes any kind of tracking impossible which makes it difficult to apply third person action recognition algorithms. Therefore, hands and handled objects become the most important cues for recognizing first person actions. Figure 1 helps in visualizing the same.

Researchers have understood the importance of egocentric cues for the first person action recognition problem. In last few years several features based on egocentric cues such as gaze, motion of hands and head, and hand pose have been suggested for first person action recognition [7, 9, 10, 25]. Object centric approaches which try to capture changing appearance of objects in the egocentric video have also been proposed [28, 31]. However, the features have been hand tuned in all these instances and shown to be performing reasonably well for limited targeted datasets.

Convolutional neural networks (CNNs) have emerged as a useful tool for many computer vision tasks. However, training such deep networks require huge amount of labeled samples. Unavailability of large amounts of data in egocentric context makes their direct use non-trivial for first person action recognition.

This paper proposes a framework for general first-person action recognition with following specific contributions:

1. We propose and demonstrate the utility of deep learned egocentric features for the first person action recognition. We show that these features alone can surpass the state of the art. They are also complementary to the popular image and flow features.

2. We provide an extensive evaluation of our deep learned features on various datasets widely different in appearance and actions. Our method performs well on four different egocentric video datasets with no change in the parameters.

3. The specific focus of the earlier works have restricted the standardization across various datasets and evaluation methodologies. To overcome this, we annotated a large number of publicly available egocentric first person action videos. We make the annotated datasets, network models and the source code available[1].

2. Related Work

Action recognition has traditionally been from a third person view, for example, from a static or a handheld camera. A standard pipeline is to encode the actions using hand crafted features based on keypoints and descriptors. Some notable contributions in this area includes STIP [21], 3D-SIFT [38], HOG3D [17], extended SURF [48], and Local Trinary Patterns [50]. Recent methods [12, 19, 45, 46] have shown promising results which leverage the appearance and motion information around densely sampled point trajectories instead of cuboidal video volume.

Deep learned features for action recognition have also been explored. Works like Convolutional RBMs [43], 3D ConvNets [13], C3D ConvNet [44], Deep ConvNets [15] take video frame as input. Recently, Two-stream ConvNets [39] introduced spatial and flow streams for action recognition.

Hand-crafted descriptors along trajectories lack discriminative property while deep learned features fail to capture salient motion. The TDD features proposed by [47] try to establish a balance by using deep convolutional descriptors along the trajectories.

We show later that both trajectory based techniques [45, 46] as well as deep neural network approaches [39, 47] do not capture the egocentric features. This restricts their performance for first person action recognition. Our egocentric features alone perform better than the state of the art. The proposed features can be complemented by the features suggested for the third person action recognition to get a further boost in the performance.

Appearance models are hard to develop from foreground or background objects due to quickly changing view field in typical egocentric videos. Spriggs *et al.* [41] have proposed to use a mix of GIST [30] features and IMU data to recognise first person actions. Their results confirm the importance of head motion in first person action recognition. Pirsiavash and Ramanan [31] attempt to recognise the activity of daily living (ADL). Their thesis is that the first person action recognition is “all about the objects” being interacted with. Lee *et al.* [22] present a video summarization approach for egocentric videos and use region cues indicative of high-level saliency such as the nearness to hands, gaze, and frequency of occurrence. Fathi *et al.* [10] recognize the importance of hands in the 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. In sports videos, where there are no prominent handled objects, Kitani *et al.* [16] use motion based histograms recovered from the optical flow of the scene (background) to learn the action categories performed by the wearer. Singh *et al.* [40] proposed a generic framework which uses trajectory aligned features along with simple egocentric cues for first person action recognition. They have released a challenging ‘Extreme Sports’ dataset of egocentric videos and showed that their method works even when there is no hand or object present in the video. Other works [6, 33] have focussed on recognising long term activities of the wearer lasting several minutes such as walking, running, working etc. Several of these methods are shown to be effective for limited targeted datasets of interest. In this paper we propose convolutional neural networks for first person action recognition trained using egocentric cues which can generalize to variety of datasets at the same time.

3. Ego ConvNet

Hand-eye coordination is a must to accomplish any object handling task. Whenever a wearer interacts with objects (grasps or reaches for the object), hands first reach out for the object, the pose of the hand is determined by the grasp

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

type for the object. Assuming that the wearer is looking straight, the head motion pattern is same as the gaze pattern, which follows the hands or the handled object. Hands often get occluded behind the handled objects. In this case, dominantly moving parts of the scene may give useful hints about the focus of the action. This coordination pattern is common for a vast variety of hand-object interactions. Capturing this information enables us to learn how the wearer interacts with his surrounding in order to carry out various actions.

We train an 'Ego ConvNet' to learn the coordination patterns between hands, head and eye movement. There are various challenges in this approach:

- Pixel level annotation of hand and object is very costly.
- Gaze or saliency map is usually captured with the help of separate eye tracking sensors.To make our approach generic we do not want to use extra sensors. We use egocentric video as the only input.
- Egocentric video analysis, being a relatively new field in computer vision, lacks a large enough annotated data for training a neural network.

To overcome these issues, we have used computer vision approaches to generate egocentric cues, which can be used for training a compact neural network with relatively small amount of labeled training samples.

3.1. Network Input

Hand Mask Hand pose in egocentric videos provides useful information for understanding hand-object manipulation and analyzing hand-eye coordination. However, egocentric videos present new challenges such as rapid changes in illuminations, significant camera motion and complex hand-object manipulations. Without relying on manually segmented hand mask as input for the network, we automatically segment out hand regions by modelling local appearance of hand pixels and global illumination. We learn the models from a small set of manually segmented hand regions over a diverse set of imaging conditions.

To account for different illumination conditions, Li and Kitani [23] have suggested to use a collection of regressors indexed by a global color histogram. We follow their approach and use the response of a bank of 48 Gabor filters (8 orientations, 3 scales, both real and imaginary components) to capture local textures. We create appearance descriptors using RGB, HSV and LAB color spaces from 900 super pixels.

The posterior distribution of a pixel x given a local appearance feature l and a global appearance feature g, is computed by marginalizing over different scenes c,

$$p\left(x \mid l, g\right) = \sum_{c} p\left(x \mid l, c\right) p\left(x \mid c, g\right)$$

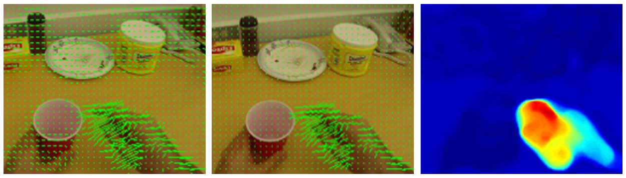

Figure 2: If the camera was static, salient regions in the image for an action recognition task can be computed as parts with moving objects. In egocentric setting, where the camera is also moving, we first cancel the component of flow due to camera motion. We note that the motion of the egocentric camera is 3D rotation for the considered actions. Such motion can easily be compensated by cancelling a 2D homography. Left and center images shows the original and compensated flow. We use the dominant motion direction from the compensated flow and use the component of flow in the direction of dominant motion direction to generate a saliency map (right image). See the text for the details.

where $p\left(x \mid l, c\right)$ is the output of a discriminative global appearance-specific regressor and $p\left(c \mid g\right)$ is a conditional distribution of a scene c given a global appearance feature g.

We perform k-means clustering on the HSV histogram of each training image. For each cluster we generate different global appearance models by learning separate random tree regressor. Histogram allows for encoding both the appearance as well as illumination of the scene. We train $n = 11$, models in our experiments. We assume that the hands viewed under similar global appearance share a similar distribution in the feature space. The conditional distribution $p\left(c \mid g\right)$ is approximated using a uniform distribution over the n nearest models.

Second row in Figure 1 shows the hand masks obtained for some example actions using the described method.

Head Motion Egocentric camera is often mounted on the wearer's head and mimics its motion. Due to the natural pivot of the head on the neck, the induced motion is a 3D rotation which can be easily captured using a 2D homography transformation of the image. We use optical flow for correspondence (avoiding hand regions) and estimate frame to frame homography using RANSAC. Use of optical flow instead of feature matching using local image descriptor such as SIFT or SURF avoids extra computation overhead. It also ensures robustness against moving objects that may be present in such videos. We set the bound on head motion between two consecutive frames to be between -5 pixels and $+5$ pixels.Head motion in x and y direction is then normalized to the range $[0, 255]$ and encoded as grayscale image separately.

Saliency Map The background in a first person action is often cluttered and poses serious challenges for an action classifier. There are a variety of objects present in the scene

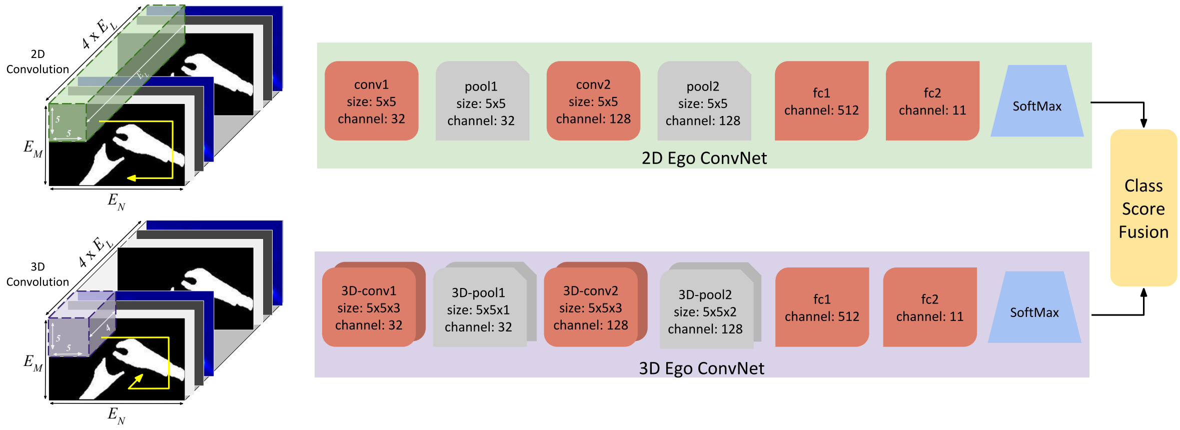
2D Convolution
4 x E_L
E_M
E_N
5
conv1 size: 5x5 channel: 32
pool1 size: 5x5 channel: 32
conv2 size: 5x5 channel: 128
pool2 size: 5x5 channel: 128
fc1 channel: 512
fc2 channel: 11
SoftMax
2D Ego ConvNet
Class Score Fusion
3D Convolution
4 x E_L
E_M
E_N
3D-conv1 size: 5x5x3 channel: 32
3D-pool1 size: 5x5x1 channel: 32
3D-conv2 size: 5x5x3 channel: 128
3D-pool2 size: 5x5x2 channel: 128
fc1 channel: 512
fc2 channel: 11
SoftMax
3D Ego ConvNet

Figure 3: Being a relatively new field, availability of labeled dataset for first person action recognition is limited. We propose a new compact CNN architecture for recognizing the wearer's actions, which takes as input egocentric cues and can be trained from limited training samples. The first and second rows show the architecture of the proposed Ego ConvNet with 2D and 3D convolutions respectively. We collect the information from a video segment of E_L frames and generate a 3D input stack of size $E_N \times E_M \times (4 \times E_L)$. The features learnt from 2D and 3D Ego ConvNet seem to be complementary and we use the two networks together achieving state of the art accuracy of 58.94% on GTEA dataset [10].

during the action, but usually only the objects being handled are important. Such salient objects can easily be determined if the gaze information is available. But this requires extra sensors. We note that there is hand-eye coordination in any first person action, and therefore saliency map obtained from the gaze resembles the dominantly moving parts in the scene. If the camera was static, such objects could be easily distinguished from others on the basis of observed optical flow only. In egocentric videos, motion of the camera has to be taken care of before using this approach. As described in the last section, this motion is easily approximated using a 2D homography. We use the homography to cancel the component due to camera motion in the observed flow. After camera motion compensation, the dominant motion in the scene comes from handled objects or hands. Figure 2 shows the same for an example action.

We use dominant flow regions to generate a saliency map per frame. We take orientation/direction of compensated optical flow and quantize it into 9 bins over an interval of $0 - 360°$. Each flow vector votes proportional to its magnitude. The bin with the highest number of votes is declared the dominant motion direction. Saliency value for a pixel is evaluated as the magnitude of flow in the direction of dominant motion direction. The saliency values are normalized to the range $[0, 255]$ and encoded as grayscale image for input to the network.

3.2. Architecture

We use a convolutional neural network to learn the coordination patterns between hands, head motion and saliency map while the wearer performs an action. We encode hand mask as a binary image. Camera motion (x and y directions separately) and saliency map are encoded as grayscale images. These are used as input to the network. We scale the input images to $E_N \times E_M$ pixels. We collect the information from a video segment of E_L frames generating a 3D input stack of size $E_N \times E_M \times (4 \times E_L)$. For each frame we use 4 different egocentric features. Hence, for E_L frames the input dimension depth becomes $4 \times E_L$. We preserve the original video aspect ratio while setting E_N and E_M.

Our Ego ConvNet architecture (Figure 3) consists of 2 convolution layers each followed by MAX pooling, RELU non-linearity, and local response normalization (LRN) layers and 2 fully connected layers. We use infogain multinomial logistic loss instead of popular multinomial logistic loss during training to handle class imbalance.

$$E = \frac{-1}{N}\sum_{n=1}^{N}\mathbf{H_{l_n}}log(\mathbf{p_n}) = \frac{-1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}H_{l_n,k}log(p_{n,k})$$

$$H_{m,p} = \begin{cases} 1 - \dfrac{|\mathcal{L}_m|}{\sum_{k=1}^{K}|\mathcal{L}_k|}, & \text{if } p = m \\ 0, & \text{otherwise} \end{cases}$$

where E is infogain multinomial logistic loss, N is training batch size, K is number of classes, l_n is ground truth label for sample n , $p_{n,k}$ is probability of sample n being classified as class k and $|\mathcal{L}_m|$ is number of training samples belonging to class m. We also use dropout ratio of 0.5 with fully connected layer to avoid overfitting during training phase. During test phase, we estimate the class label by applying SOFTMAX on $fc2$ layer output. Without any appearance and motion cues, i.e., using egocentric cues alone,

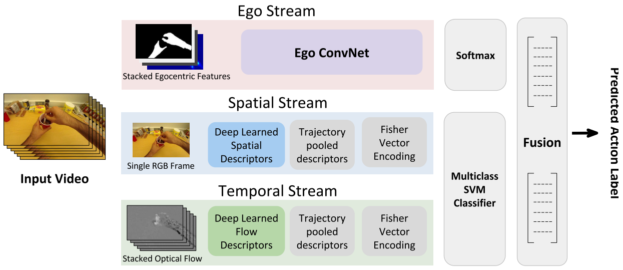

Figure 4: We extend our Ego ConvNet by adding two more streams corresponding to spatial and temporal streams in two-stream architecture [39]. Intuitively spatial and temporal streams capture the generic appearance and flow based features whereas the egocentric stream captures the coordination between wearer's hands, head and eyes, which is not captured by appearance and flow features. Using multiple stream architecture improves the accuracy of proposed method from 58.94% to 68.5% on GTEA dataset [10].

our Ego ConvNet is able to outperform state of the art by a large margin. This outlines the importance of egocentric cues for first person action recognition task.

3D Convolution Tran et al. [44] extended 2D convolution to allow 3D convolution and 3D pooling, which has been shown to capture the temporal structure of an action. C3D [44] performs 3D convolutions and 3D pooling, propagating temporal information across all the layers in the network. We have also experimented with 3D convolutions in the proposed network. The architecture for the 3D Ego ConvNet is slightly different from that of 2D Ego ConvNet. We use filters of size $5 \times 5 \times 4$, temporal pooling of size $5 \times 5 \times 1$ for first convolution layer and $5 \times 5 \times 2$ for second convolution layer where last dimension represents *depth* of the filter. The first pooling layer has kernel of depth 1 with the intention of avoiding the merge of the temporal signal too early. Figure 3 shows the architecture.

The performance of 2D and 3D Ego ConvNet are similar. However, the features learnt seem to be complementary. Therefore, we combine the two networks by using two-streams architecture (similar to the one proposed by Simonyan and Zisserman [39]) by adding a new SVM classifier at the end. While the accuracy for Ego ConvNet using 2D and 3D convolutions is 57.61% and 55.79% respectively, the accuracy when the two networks are used together is 58.94%.

4. Three-Stream Architecture

Dense Trajectory [45] and its variants [12, 19, 46] constitute a popular approach for third person action recognition. These techniques track interest points across time and compute hand designed flow and appearance based descriptors (HOG, HOF, and MBH) along the trajectories. Recently Wang *et al*. [47] have proposed to replace hand designed features with 'Deep Convolution' descriptors. They use two-stream architecture proposed by Simoyan and Zisserman [39] for computing deep convolution descriptors. The first 'Spatial Stream' uses individual video frame as input, effectively performing action recognition from still images using ImageNet [20] architecture. The spatial stream takes a single RGB video frame as input. The second 'Temporal Stream' uses stacked optical flow as input to capture motion information. The temporal stream takes the stacked dense optical flow displacement fields between L consecutive frames as input. The x and y components of flow fields are encoded as grayscale image separately after normalising to $[0, 255]$. The temporal convnet has the same architecture as the spatial stream.

We extend our Ego ConvNet by adding two more streams corresponding to spatial and temporal streams in the TDD or two-stream architecture. Intuitively, spatial and temporal streams capture the generic appearance and flow based features whereas the egocentric stream captures the coordination between the wearer's hands, head and eyes, which is not captured by appearance and flow features. Fusion of these cues results in a more meaningful feature to describe first person actions.

We perform fusion of the three streams: spatial, temporal and egocentric, by combining weighted classifier scores. The weights are learnt using cross validation. For egocentric features we use softmax score as the classifier score. For the spatial and temporal streams we use SVM classification score. We learn a multiclass SVM classifier using improved Fisher Vector representation of trajectory pooled features from appearance and flow features.

The architecture of our convnet can be seen in Figure 4. Using multiple stream architecture with pre-trained spatial and temporal streams improves the accuracy of proposed method from 58.94% to 68.5%.

5. Experiments and Results

5.1. Datasets and Evaluation Protocol

We consider short term actions performed by different subjects while performing different activities. In our work, we use four different publicly available datasets of egocentric videos: GTEA [10], Kitchen [41], ADL [31] and UTE [22]. Out of these, only GTEA and Kitchen datasets have frame level annotations for the 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 kinds of actions, such as walking, watching television etc. are labelled as 'background'. The speed and nature of actions vary across subjects and activities (e.g., consider the ac-

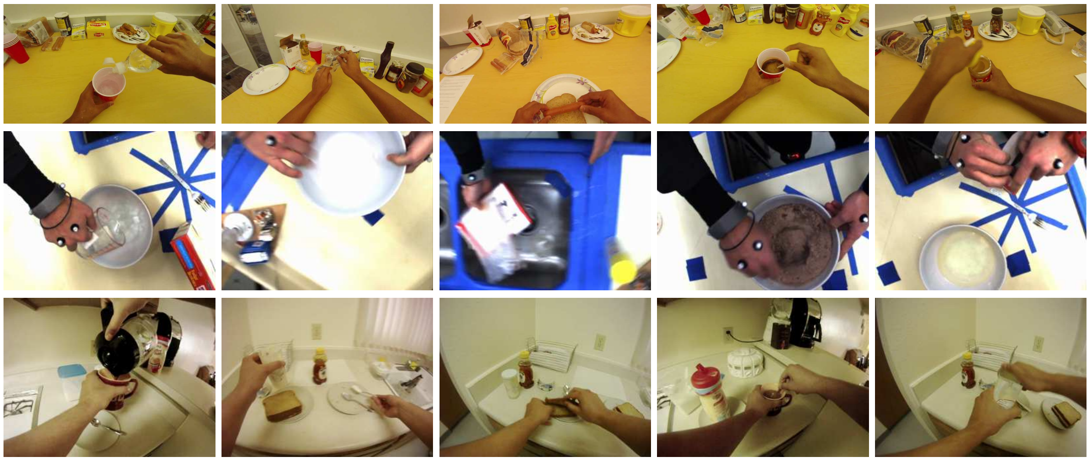

Figure 5: Examples of wearer's action categories we propose to recognize in this paper from different datasets: GTEA[10] (top row), Kitchen[41] (middle row) and ADL[31] (bottom row). The columns represent the actions 'pour', 'take', 'put', 'stir' and 'open'. The actions vary widely across datasets in terms of appearance and speed of action. Features and technique we suggest in this paper is able to successfully recognize the wearer's actions across different presented scenarios, showing the robustness of our method.

Dataset	Subjects	Frames	Classes	State of the art Accuracy	Ours	Ours (cross validated)
GTEA [10]	4	31,253	11	47.70 [7]	68.50	64.41
Kitchen [41]	7	48,117	29	48.64 [41]	66.23	66.23
ADL [31]	5	93,293	21	N.A.	37.58	31.62
UTE [22]	2	208,230	21	N.A.	60.17	55.97

Table 1: Statistics of egocentric videos datasets used for experimentation. The proposed approach uses deep learned appearance, motion and egocentric features and improves the state of the art on all the datasets we tested. Results are reported in terms of percentage of accuracy. The datasets vary widely in appearance, subjects and actions being performed, and the improvement on these datasets validates the generality of 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.

tion 'open' in two scenarios, 'open' water bottle and 'open' cheese packet). Statistics related to the datasets are shown in Table 1.

The GTEA dataset consists of 7 long term activities captured using head mounted cameras. Each activity is approximately 1 minute long. We follow 'leave–one–subject–out' experimental setup of [7] for all datasets.

Kitchen dataset [41] is captured using head mounted camera and IMUs. Camera point of view is from top, and severe camera motion is quite common. Similar to [41], we select 7 subjects from 'Brownie' activity. We use videos of 6 subjects for training and test on the video of the remaining subject. ADL dataset consists of videos of subjects performing daily life activities, captured using chest mounted cameras with 170 degrees of viewing angle. UTE dataset [22] contains 3 to 5 hours long videos captured from head-mounted cameras in a natural and uncontrolled setting. The annotated datasets and the source code along with the pre-trained CNN models for the paper are available at the project page: http://cvit.iiit.ac.in/projects/FirstPersonActions/.

Evaluation Protocol For systematic evaluation, we use leave–one–subject–out policy for training and validation and report classification accuracy on the unseen test subject. Formally, classification accuracy for first person action recognition task is defined as the number of frames (or video segments) classified correctly divided by total number of frames (or 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 at the video segment level. In this case, there is only one action in each video segment. However, length of the segment is not fixed. In this setting, we have an approximate knowledge of action boundaries which naturally improves action recognition results.

Unlike [25], we are interested in classification of first person action in different settings and not the specific object which is involved in the action. For example, the action where the wearer is 'pouring' 'water' from the 'bottle' into the 'cup' and where the wearer is 'pouring' 'mayonnaise' on to the 'bread'. In our experiments we consider both actions as 'pouring' despite different objects being involved. We believe that such evaluation is more challenging and removes the bias of object instance specific appearance while learning action models.

Features	Frame level Accuracy			
	GTEA [10]	Kitchen [41]	ADL [31]	UTE [22]
S+T	59.47	58.92	34.20	55.67
H+C+M (2D)	57.61	51.33	30.78	53.32
H+C+M (3D)	55.79	50.06	28.85	53.10
H+C+M (2D+3D)	58.94	54.35	32.12	53.42
Combined	68.50	66.23	37.58	60.17

Table 2: Detailed analysis of spatial, temporal and egocentric features used in our work. H: **H**and masks, C: **C**amera/Head motion, M: Saliency **M**ap, S: Deep learned **S**patial descriptors, T: Deep learned **T**emporal descriptors. Results are reported in terms of percentage of accuracy.

5.2. Implementation Details

We use Caffe's CNN implementation [14] due to its speed and efficiency. In all experiments, we normalise data to the range $[0, 1]$ and use infogain multinomial logistic loss to counter imbalanced class population in the datasets.

Pre-training Ego ConvNet We use 343 pre-segmented hand masks and implementation for hand segmentation provided by [23]. Since GTEA dataset is too small to properly train a convnet, we add more data by using videos from the Interactive Museum dataset [5], consisting of 700 videos at 800×450 resolution and 25 frames per second, all shot with a wearable camera mounted on the head. We manually segment hands from randomly selected 60 frames in order to train the hand models.

We chose Interactive Museum dataset for pre-training due to its similarity to GTEA dataset in which same actions are performed by various subjects. Further, the hand gesture videos emphasize on actions using hands under similar camera or head movement.

Prior to training on GTEA dataset, we pre-train the network on [5]. This takes care of the small dataset size available for training. Pre-training is done in leave-one-subject-out manner. Videos from subjects 1-4 are used for training and validation is done on videos from subject 5. We select non-overlapping windows of E_L frames (depending on architecture) as input to the network. Pre-training is done for 20 epochs in all experiments. For Kitchen, ADL and UTE datasets, we fine-tune Ego Convnet which has been pre-trained on Interactive Museum and GTEA datasets, to avoid training from scratch.

Training Similar to pre-training, training is done in leave-one-subject-out manner as well. We use video segments from subject 1 and 3 for training, subject 4 for validation and subject 2 for testing. A video segment is an overlapping sliding window of E_L frames (depending on architecture) as input to the network. Since the input is overlapping frames, the training set is shuffled at random to avoid all samples belonging to the same class for batch training. Frames are resized while maintaining the original aspect ratio, as we found warping to square size input reduces the performance. Training is done by SGD with minibatch size of 64 examples. Initial learning rate is 0.001, and is divided by 10 every 10K iterations. The optimization is stopped at 25K iterations (about 60 epochs).

Algorithm	Features	Accuracy
DT [45]	trajectory+HOG+HOF+MBH	45.15
iDT [46]	trajectory+HOG+HOF+MBH	52.37
TDD [47]	Spatial	58.61
TDD [47]	Temporal	57.12
TDD [47]	Spatial + Temporal	59.47

Table 3: Performance of trajectory based methods when used with various features for GTEA dataset. Accuracy in terms of percentage is reported for frame level action recognition.

Varying Network Architecture We trained a compact network with 2 convolution layers and 2 fully connected layers. Such a compact network has been shown to work fairly well despite having limited training data. To search for a good Ego ConvNet architecture, we varied both spatial, $E_N \times E_M$, (64×36, 32×18 and 16×9 pixels) and temporal, E_L, (28, 16, 10 and 5 frames) input dimensions and found 32 × 18 × 5 to be the best. This size is small enough to keep number of parameters low without losing relevant information. We also tried skipping frames (by selecting every third or sixth frames) from deeper input volume (15 frames or 30 frames) keeping input depth to 5. However, this approach did not improve recognition performance. We also investigated with different filter sizes (3 × 3, 5 × 5 and 7 × 7) and number of filters (32, 64, 128, 256 for convolution layers and 256, 512, 1024 and 2048 for fully connected layer) and found 5 × 5 filter with 32 filters in first convolution layer, 128 filters in second convolution layer and 512 channel output for $fc1$ layer to be the best performer. For 3D Ego ConvNet, keeping all other parameters same, we use filters having *depth-4*.

5.3. Results and Discussion

We first present our experiments and analysis of the proposed action descriptor on GTEA dataset to bring out salient aspects of the suggested approach. Experiments with other datasets have been described later.

We follow experimental setup of Fathi *et. al.* [7] 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 report an accuracy of 47.70%. Table 2 shows performance of our Ego ConvNet and Table 3 for trajectory based approaches on

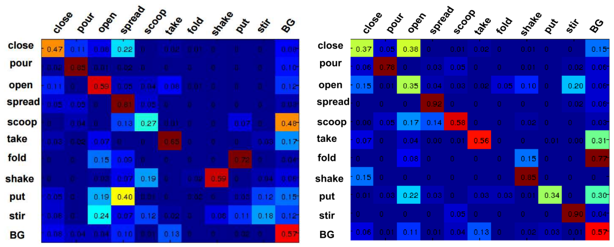

Figure 6: Confusion matrix for different deep learned descriptors on GTEA dataset. Classification using Ego ConvNet descriptors (left) and TDD descriptors (right). The Ego ConvNet and TDD features are clearly complementary and improve overall action recognition accuracy when used together.

Method	Features	Accuracy
Ego ConvNet (2D)	H	51.76
Ego ConvNet (2D)	H+C	54.13
Ego ConvNet (2D)	H+C+M	57.61
Ego ConvNet (3D)	H	50.82
Ego ConvNet (3D)	H+C	53.15
Ego ConvNet (3D)	H+C+M	55.79
Ego ConvNet (2D) + TDD	H+C+M+S+T	65.29
Ego ConvNet (3D) + TDD	H+C+M+S+T	66.96
Combined	H+C+M+S+T	68.50

Table 4: Effect of various CNN features on first person action recognition. The experiments are done on GTEA dataset. Accuracy reported is for frame level action recognition. H: **H**and masks, C: **C**amera/Head motion, M: Saliency **M**ap, S: Deep learned **S**patial descriptors, T: Deep learned **T**emporal descriptors.

GTEA dataset. Although, the standalone performance for two types of features is similar, we found the learnt features to be complementary. Figure 6 shows confusion matrices for the two sets of features to support our claim. Therefore, we fuse the trajectory pooled and egocentric features in a three-stream architecture. Table 4 gives the effect of different features on the performance of our algorithm on the dataset. Experimental protocol leaving subject 2 is what is done by Fathi *et. al.* in [7] and is followed by us to do a fair comparison. We also show cross-validated results in leave–one–subject–out manner (see Table 1).

We extend our experiments to other publicly available egocentric video datasets. Results on these datasets are shown in Table 2 and 5. We follow the same experimental setup as [41] and perform frame level action recognition for ‘Brownie’ activity for 7 subjects. Spriggs *et al.* [41] reports an accuracy of 48.64% accuracy when using first person data alone and 57.80% when combined with IMU data. We achieve 54.35% accuracy using our method with egocentric stream alone and 66.23% with our three-streams approach.

The ADL dataset has been used for long term activity recognition by [31] 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 other for training. We achieve 37.58% accuracy at frame level and 39.02% at video segment level using the proposed method.

Dataset	Accuracy		
	Frame level	Segment level	Chance level
GTEA [10]	68.50	82.40	11%
Kitchen [41]	66.23	71.88	3.4%
ADL [31]	37.58	39.02	4.7%
UTE [22]	60.17	65.30	4.7%

Table 5: 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.

The UTE dataset has been used for video summarization by [22] in the past. Motion blur and low image quality is fairly common in this dataset. For action recognition we achieve 60.17% accuracy at frame level and 65.30% at video segment level using the proposed method.

The proposed action descriptor improves upon the state of the art on all four datasets (see Table 1 for the details about dataset and comparison). Figure 5 shows some sample actions correctly classified by our approach. Note the difference in appearance between the datasets. The experiments show that the proposed approach consistently outperforms state of the art accuracy for each of the datasets.

6. Conclusion

Previous approaches for first person action recognition have explored various hand tuned features based on egocentric cues such as hand pose, head motion and objects present in the scene. These approaches do not leverage existing work in third person video analysis and do not generalize beyond the dataset considered.

We have proposed a new convolutional neural network based framework for first person action recognition. We propose a three-stream architecture which uses egocentric cues in the first stream and complements it with pre trained spatial and temporal streams from third person video analysis. We show that with the egocentric stream alone, we can achieve state of the art accuracy. The performance improves further after using complementary features from spatial and temporal streams. The generality of the approach is validated by achieving state of the art accuracy on all available public datasets at the same time.

Acknowledgement The authors would like to thank Kohli Center on Intelligent Systems (KCIS) for the partial financial support.

References

[1] Google glass. https://www.google.com/glass/start/. 1

[2] Gopro. http://gopro.com/. 1

[3] Microsoft sensecam. http://research.microsoft.com/en-us/um/cambridge/projects/sensecam/. 1

[4] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *CVPR*, 2011. 1

[5] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *CVPRW*, 2014. 7

[6] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa. Predicting daily activities from egocentric images using deep learning. In *ISWC*, 2015. 2

[7] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011. 1, 2, 6, 7, 8

[8] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012. 1

[9] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 1, 2

[10] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 1, 2, 4, 5, 6, 7, 8

[11] Y. Hoshen and S. Peleg. Egocentric video biometrics. *CoRR, abs/1411.7591*, 2014. 1

[12] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013. 2, 5

[13] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. In *TPAMI*, 2013. 2

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, , and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *arXiv preprint arXiv:1408.5093*, 2014. 7

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2

[16] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 1, 2

[17] A. Klaser and M. Marszalek. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 2

[18] J. Kopf, M. Cohen, and R. Szeliski. First-person hyperlapse videos. *ACM Transactions on Graphics*, 2014. 1

[19] E. Kraft and T. Brox. Motion based foreground detection and poselet motion features for action recognition. In *ACCV*, 2014. 2, 5

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 5

[21] I. Laptev. On space-time interest points. *IJCV*, 2005. 2

[22] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 1, 2, 5, 6, 7, 8

[23] C. Li and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *CVPR*, 2013. 3, 7

[24] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 1

[25] Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *CVPR*, 2015. 2, 6

[26] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 1

[27] K. Matsuo, K. Yamada, S. Ueno, and S. Naito. An attention-based activity recognition for egocentric video. In *CVPRW*, 2014. 1

[28] T. McCandless and K. Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. In *BMVC*, 2013. 2

[29] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato. Coupling eye-motion and ego-motion features for first-person activity recognition. In *CVPRW*, 2012. 1

[30] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001. 2

[31] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 1, 2, 5, 6, 7, 8

[32] Y. Poleg, C. Arora, and S. Peleg. Head motion signatures from egocentric videos. In *ACCV*, 2014. 1

[33] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *CVPR*, 2014. 1, 2

[34] Y. Poleg, T. Halperin, C. Arora, and S. Peleg. Egosampling: Fast-forward and stereo for egocentric videos. *CoRR, abs/1412.3596*, 2014. 1

[35] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010. 1

[36] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *CVPRW*, 2009. 1

[37] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013. 1

[38] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACMMM*, 2007. 2

[39] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 2, 5

[40] S. Singh, C. Arora, and C. V. Jawahar. Trajectory aligned features for first person action recognition. *CoRR, abs/1604.02115*, 2016. 2

[41] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPRW*, 2009. 1, 2, 5, 6, 7, 8

[42] S. Sundaram and W. W. M. Cuevas. High level activity recognition using low resolution wearable vision. In *CVPRW*, 2009. 1

[43] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *ECCV*, 2010. 2

[44] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2, 5

[45] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 2, 5, 7

[46] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 2, 5, 7

[47] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, 2015. 2, 5, 7

[48] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008. 2

[49] B. Xiong and K. Grauman. Detecting snap points in egocentric video with a web photo prior. In *ECCV*, 2014. 1

[50] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009. 2