Generic Action Recognition from Egocentric Videos

Suriya Singh [1] Chetan Arora [2] C. V. Jawahar [1]

[1] CVIT, IIIT Hyderabad, India [2] IIIT Delhi, New Delhi, India

***Abstract*—Egocentric cameras are wearable cameras mounted on a person's head or shoulder. With their ability to have first person view, such cameras are spawning new set of exciting applications in computer vision. Recognising activity of the wearer from an egocentric video is an important but challenging problem. The task is made especially difficult because of unavailability of wearer's pose as well as extreme camera shake due to motion of wearer's head. Solutions suggested so far for the problem, have either focussed on short term actions such as pour, stir etc. or long term activities such as walking, driving etc. The features used in both the styles are very different and the technique developed for one style often fail miserably on other kind. In this paper we propose a technique to identify if a long term or a short term action is present in an egocentric video segment. This allows us to have a generic first-person action recognition system where we can recognise both short term as well as long term actions of the wearer. We report an accuracy of 90.15% for our classifier on publicly available egocentric video dataset comprising 18 hours of video amounting to 1.9 million tested samples.**

I. Introduction

Advances in sensor technology has made wearing the camera on one's head practical and affordable. Such wearable cameras, popularized by Google Glass [1] and GoPro [2], are typically worn on the head or along with the eyeglasses and have the advantage of having a similar view as that of the person wearing the camera. We refer to such cameras as first person cameras or egocentric cameras. Egocentric cameras allow to capture from wearer's perspective and are able to capture wearer's social interactions as well interactions with surrounding environment and objects, giving useful insights into wearer's daily activities.

Availability of first person view, therefore, has rightly spawned a new set of exciting applications and challenges in egocentric vision. Understanding wearer's activities from egocentric videos can help in a range of applications such as smart homes, automation, remote assistance as well as medical emergencies. The egocentric videos can be especially useful for people with disabilities for daily visual logs or as a memory aid for the wearer. There are also attempts to augment egocentric videos with meta data such as face, place, text etc. for people with limited vision [3]. The extracted meta data is compelling even for people with regular vision, for its utility in giving context aware suggestions.

Clearly, a large number of applications rely on accurate activity and action recognition. Within egocentric community, first-person actions can be broadly divided into two broad categories: *short term* and *long term* actions (see Figure 1). As the name suggests short term actions occur over a period of few seconds and are more similar to gestures performed by the wearer, e.g., picking an object, opening a cap, shake etc.). Whereas, long term actions are similar to activities and typically occur over several minutes. e.g., walk, run, drive, stand etc. There are large differences in two class of videos. While short term actions typically involve handling some object using wearer's hands, long term activities do not have any such dominant handled object. Researchers have, therefore, developed separate techniques and features for two action classes. The researchers in short term actions have relied on hand-object interactions using cues such as optical flow, pose, size and location of hands in their feature vector. In contrast, research in long term actions largely rely on motion features from optical flow. This has led to a situation where the method and features are specific to one kind of action and does not generalize well to the other kind. For example, for short term action, state of the art techniques use trajectory based features. For long term actions, spanning several minutes, there are large variations in viewing directions due to motion or wearer's head making the tracking for long term virtually impossible.

978-1-4673-8564-0/15/$31.00 ©2015 IEEE

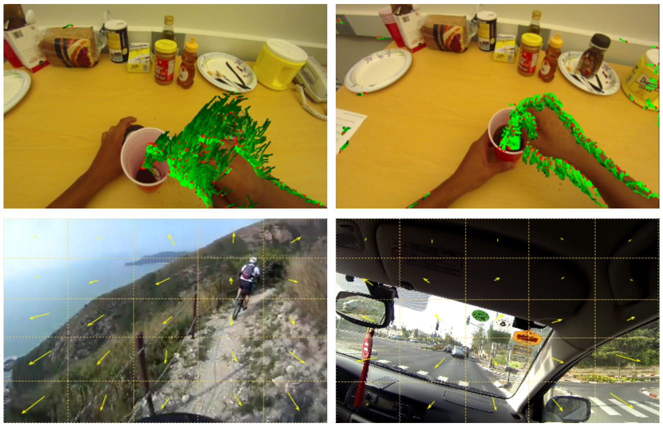

Fig. 1: First-person actions can be broadly divided into two categories *short term* and *long term* actions. Top row shows short term actions (left: 'scoop' and right: 'stir') from GTEA[7] dataset and trajectory aligned features used by [22] for action recognition. Bottom row shows long term actions (left: 'riding' and right: 'driving') from Egoseg [17] dataset and motion feature used by [17] for action recognition. In recent works, method and features are specific to one kind of action and does not work well for the other kind. The focus of this paper is on recognising kind of action in an egocentric video and identify appropriate features as well as method for further processing.

Our focus in this paper is on a generic action recognition system from egocentric videos. The objective is to leverage the existing research in the egocentric action recognition by adding a pre-processing classifier which can recognise whether a short term or a long term action is present in a video segment. We propose a novel *Dominant Motion* feature derived from optical flow for this task and report an accuracy of 90.15% on 18 hours (1.9 million frames) of egocentric videos. After this step, appropriate feature and method can be used for detailed action recognition. For example, one can use [19], [7] or [22] for parts of video classified as short term actions and [17] or [9] for long term actions.

II. 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 includes STIP [11], 3D-SIFT [18], HOG3D [10], extended SURF [23], and Local Trinary Patterns [24]. Some recent methods [13, 20, 22] show promising results for action recognition by leveraging the motion information of trajectories.

Egocentric cameras have certain distinct advantages as well as constraints for action recognition. While having much lesser occlusions for the objects in an egocentric video is extremely useful, natural head motion of the wearer brings in additional large camera motion, posing a challenge to any first-person action recognition algorithm. While all the methods focus on a specific kind of action and shows good results on public dataset, no technique generalizes to both short term and long term actions. Spriggs et al. [19] proposed to recognise first-person actions using a mix of GIST [15] features and IMU data. Pirsiavash and Ramanan [16] attempt to recognise the activity of daily living (ADL). Their thesis is that the first-person action recognition is "all about the objects", and in particular, "all about the objects being interacted with". McCandless and Graumann [12] extend the work by using spatio-temporal pyramid histograms of objects appearing in the action. They propose an "object-centric" scheme that prefers candidates involving objects prominently involved in the actions. Fathi et al. [7] 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. [7, 12, 16, 19] focus was only short term actions.

For first-person actions when there are no prominent handled object, Kitani et al. [9] use motion based histograms recovered from the optical flow of the scene (background) to recognise the short term actions of the wearer. Ogaki et al. [14] use eye-motion and ego-motion to recognise indoor desktop actions. For long term actions, Poleg et al. [17] use motion cues of the camera wearer for egocentric video segmentation into meaningful chapters. Notice that most of the technique for short term actions features rely on hand's and object's appearance and motion occurring due to interaction as cues. For long term actions, there are no dominant object in the view and the features are primarily derived from optical flow in the scene representing motion of wearer's head.

III. Dominant Motion Feature

We propose a novel feature based on optical flow for identifying types of first-person action in an egocentric videos. We note that motion of the egocentric camera is due to $3D$ rotation of wearer's head and can be easily compensated by a $2D$ homography transformation of the image. The remaining dominant motion in the scene always come from handled objects or hands and are used to describe a video segment in the proposed descriptor. We refer to our feature as *Dominant Motion feature*.

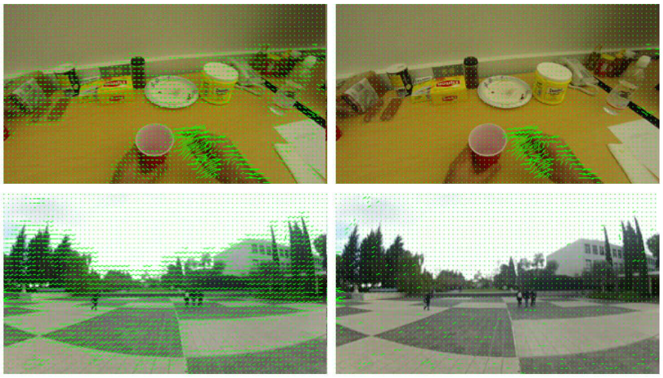

Fig. 2: Motion of the egocentric camera is due to $3D$ rotation of wearer's head and can be easily compensated by a $2D$ homography transformation of the image. Left: Optical flow overlayed on the frame. Right: Compensated optical flow followed by thresholding. Almost all camera motion has been compensated by this simple technique. It is the compensated flow that prove to be more useful for identifying type of action present in the video. Top row shows short term action 'take', bottom row shows long term action 'walking'.

We start by extracting frame to frame dense optical flow using the algorithm by Färneback [5] 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 as mentioned in [22]. In homogeneous image areas without any structure, it is impossible to compute accurate optical flow, we discard such flow in all further steps.

A. Camera Motion Compensation

Using optical flow as correspondences, we estimate frame to frame homography using RANSAC. We use optical flow as correspondences instead of feature matching using local image descriptor such as SIFT or SURF to avoid extra computation overhead as well as robustness against moving objects that may be present in such videos. Next, we subtract $2D$ homography transformation of the image from the optical flow which usually compensate much of the camera motion due to head movement. In case of sharp changes in the visual field of the egocentric camera, some residual flow is left and can be discarded by putting a threshold on flow magnitude (we use threshold of 2 pixels). Figure 2 shows optical flow and compensated flow for short term and long term actions.

B. Dominant Motion Feature from compensated flow

For computing dominant motion between a pair of frames, flow orientations (or directions) are quantized into 9 bins using full orientations ($360°$) and each flow vector votes for a corresponding bin (decided by flow orientation) proportional to its magnitude. The bin with highest vote is set to 1 and remaining bins are set to 0. This feature vector for each consecutive frame pair in a video is added up, resulting into a 9 dimensional feature vector. The feature vector is then normalized using L_1 normalization. The resulting video descriptor is called Dominant Motion (DM) feature. Clearly, DM feature is an extension of Histogram of Optical Flow (HOF). We compare DM and HOF for our task, later in the experimental section, and show that DM outperforms HOF by a significant margin.

Dataset	Subjects	Videos	Frames	Action
GTEA [7]	4	28	31,253	short term
GTEA Gaze(+) [6]	5	30	454,830	short term
ADL [16]	5	5	93,293	short term
First Person Hyperlapse [8]	3	3	96,271	long term
Egoseg [17]	7	29	857,942	long term
Youtube videos	60	60	412,250	long term
Example videos	3	3	43,423	short and long term

TABLE I: Statistics of egocentric videos datasets used for experimentation. The datasets vary widely in appearance, subjects and actions being performed.

C. Dominant Motion feature from trajectories

Inspired from the merits of trajectory aligned features in third person action recognition [22], we explore it in our case as well. As suggested by Wang *et al.* [22], 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: $(P_t, P_{t+1}, P_{t+2}, \ldots)$. To leverage motion information in dense trajectories, we compute HOF descriptors using compensated optical flow 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. Dominant motion is computed from each trajectory aligned HOF descriptors in similar way as discussed previously. For example, if a HOF descriptor is $[12, 345, 0, 0, 988, 12, 32, 90, 0]$ then its dominant motion is simply $[0, 0, 0, 0, 1, 0, 0, 0, 0]$. DM feature is then obtained by adding up all such dominant motion histogram followed by L_1 normalization.

D. Generic Action Recognition for First Person Actions

Given an egocentric video, our task is to classify each video segment into the type of first-person action, short term or long term, present in the video. This would enable us to use appropriate feature and method for the particular segment. We train a binary SVM classifier on publicly available egocentric videos using DM feature. We define each video segment as a set of 60 consecutive frames in our experiments. We extract 9 dimensional DM feature and use pre-trained SVM to predict the type of action present in the segment.

IV. Experiments and Results

A. Datasets and Evaluation Protocol

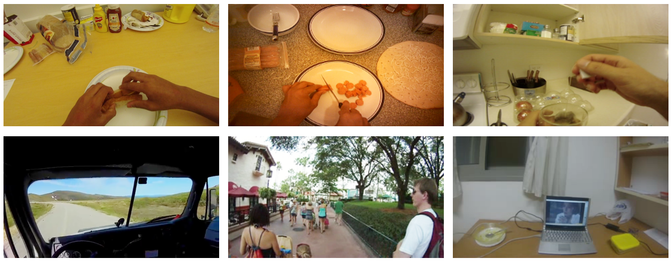

Fig. 3: Example frames from various datasets used for training our classifier using DM feature. Top row: short term actions. Bottom row: long term actions.

Method	# Samples		Accuracy	
	Class 1	Class 2	HOF	DM
Optical Flow	550K	1350K	57.21%	61.05%
Compensated Optical Flow	550K	1350K	78.73%	84.70%
Trajectories with compensation	550K	1350K	83.56%	90.15%

TABLE II: Binary classification results. DM feature is able to capture dominant motion information and outperform HOF in distinguishing type of action.

In our work, we make use of 5 publicly available datasets of egocentric videos, GTEA [7], GTEA Gaze(+) [6], ADL videos [16], Egoseg [17] datasets, Bike and Walking sequences from [8] and 60 other egocentric videos from youtube. Statistics related to datasets are shown in Table I and example frames are shown in Figure 3.

We consider two classes of first-person actions: short term and long term action. To evaluate the DM feature, we use the proposed feature for the binary classification task. Classification accuracy for identifying first-person action is defined as number of frames classified correctly divided by total number of frames. For classifying each frame we consider a window of 60 frames around it (30 frames on each side). 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).

All available egocentric video datasets focusses on only one type of action. As per our knowledge, there is no such egocentric video dataset available that contains both types of action. Therefore for our task, we capture 3 long egocentric videos. We use GoPro mounted on the head of wearer for recording. Our videos are of 1280×720 resolution at 30 frames per second rate. Each video has a mix of short term and long term actions, in some cases both occurring at the same time (e.g, 'typing' on mobile phone while 'walking'). These videos are challenging for both long term as well as short term action recognition systems.

B. Classification Results

Using DM feature we train a Two-Class Support Vector Machine (binary SVM) using Liblinear library [4]. We use VLFeat's [21] 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. We keep choosing a sequence at random until we get $20,000$ training samples/frames for each class. The remaining sequences are used for testing. SVM parameters were estimated using 4 fold cross validation.

Table II summarizes the binary classification result. DM

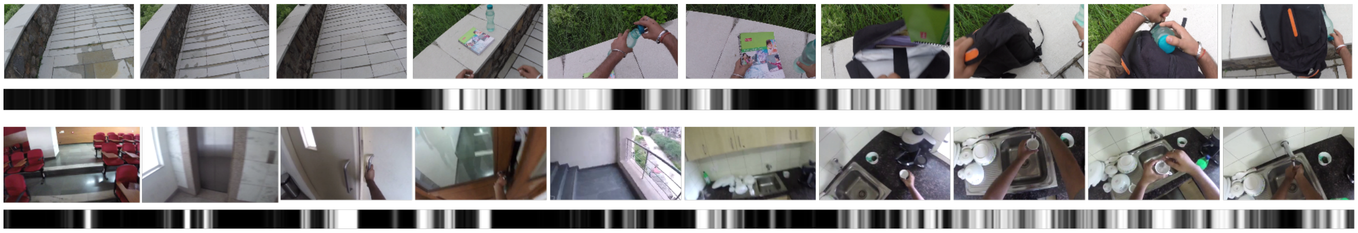

Fig. 4: Temporal Segmentation result visualisation on example videos consisting both types of action. Predicted label is shown below the frames. White: short term action, Black: long term action. Notice that short term action has been predicted with higher confidence when there is hand-object interaction.

feature can capture dominant motion information better than HOF and hence the better classification result. Further, notice that compensation for camera motion improves the result. This is because, without camera motion rest of the flow mainly comes from hand-object interactions. As the pixels belonging to hand and object move together in the same direction, we have been able to capture this information with our DM feature. Notice that generally there is no such dominant motion in case of long term action.

We also recognise that the trajectory based features work better than the optical flow as shown by various works in third person action recognition. For short term actions, region of dominant motion in an egocentric video (after cancelling camera motion due to head) implicitly capture hands and objects motion while for long term actions this is not the case. We believe this is the reason behind accuracy improvement when using trajectories.

We show qualitative result of temporal segmentation of an egocentric video using our technique in Figure 4. Here we show every 90^{th} frame from the video we captured along with predicted labels using our technique. It also serves to show that our method generalizes well to unseen samples (when none of the samples from the dataset was used for training). Notice that, at the boundaries of two actions (long term action immediately followed by short term action or vice-versa) prediction confidence for either class is low but increases after some time as expected. It may also be noted that short term actions are more sensitive to action boundaries due to short action duration. It is common approach to avoid such action boundaries for training action recognition model as it tends to add noise to the model.

V. Conclusion

We propose a system to classify a video segment based upon what kind of action, short term or long term, is present in it. The task is important as there is no single action recognition technique available for both kinds of first-person actions. We show that simple camera compensation based on homography enhance salient motion present in the video while suppressing the background motion. We propose a novel Dominant Motion feature based on optical flow which can be used for the classification task. The proposed system can be used as a pre-processing to select a suitable technique for action recognition for further processing each segment of the video.

References

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

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

[3] Orcam. http://www.orcam.com/. 1

[4] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. In *IJMLR*, 2008. 3

[5] G. Färneback. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*, 2013. 2

[6] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaz. In *ECCV*, 2012. 3

[7] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 1, 2, 3

[8] M. C. Johannes Kopf and R. Szeliski. First-person hyperlapse videos. In *SIGGRAPH*, 2014. 3

[9] 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

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

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

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

[13] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009. 2

[14] 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. 2

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

[16] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 2, 3

[17] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *CVPR*, 2014. 1, 2, 3

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

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

[20] J. Sun, Y. Mu, S. Yan, and L.-F. Cheong. Activity recognition using dense long-duration trajectories. In *ICME*, 2010. 2

[21] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. 3

[22] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 1, 2, 3

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

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