Action and Interaction Recognition in First-person videos

Sanath Narayan
Dept. of Electrical Engg.,
IISc, Bangalore
sanath@ee.iisc.ernet.in

Mohan S. Kankanhalli
School of Computing,
NUS, Singapore
mohan@comp.nus.edu.sg

Kalpathi R. Ramakrishnan
Dept. of Electrical Engg.,
IISc, Bangalore
krr@ee.iisc.ernet.in

Abstract

In this work, we evaluate the performance of the popular dense trajectories approach on first-person action recognition datasets. A person moving around with a wearable camera will actively interact with humans and objects and also passively observe others interacting. Hence, in order to represent real-world scenarios, the dataset must contain actions from first-person perspective as well as third-person perspective. For this purpose, we introduce a new dataset which contains actions from both the perspectives captured using a head-mounted camera. We employ a motion pyramidal structure for grouping the dense trajectory features. The relative strengths of motion along the trajectories are used to compute different bag-of-words descriptors and concatenated to form a single descriptor for the action. The motion pyramidal approach performs better than the baseline improved trajectory descriptors. The method achieves 96.7% *on the JPL interaction dataset and* 61.8% *on our NUS interaction dataset. The same is used to detect actions in long video sequences and achieves average precision of* 0.79 *on JPL interaction dataset.*

1. Introduction

With wearable cameras becoming popular, recognizing actions and interactions in videos captured from first-person perspective cameras is essential for many applications such as video-logging, behavior understanding, retrieval, *etc*. Action recognition is a well researched area in computer vision. However, most of the research till now is focused on third-person perspective/view. Egocentric vision is an emerging area and in this work we evaluate the performance of the present state-of-the art action recognition method, in third-person perspective (TPP), on actions in first-person perspective (FPP). The improved trajectories method by Wang *et al*. [16] is the current state-of-the art on many challenging action recognition datasets. The egocentric videos can be looked at as being captured in a different view when compared to the normal TPP. Hence action recognition methods which are able to handle view changes in TPP videos should be applicable to FPP videos as well. To this end, we analyze the performance of the "improved trajectories" [16] approach for actions in FPP.

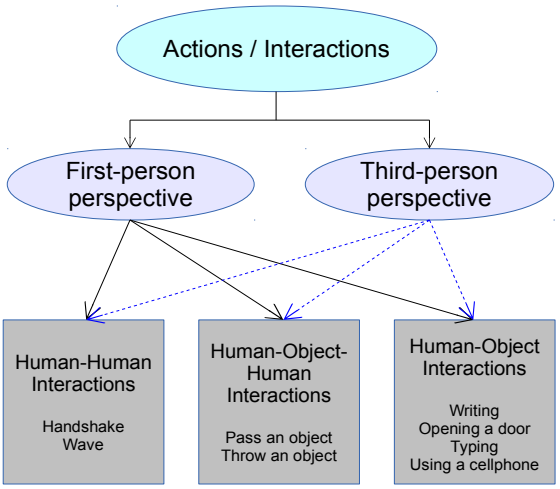

Figure 1. Illustration of the types of interactions that can be encountered by a first-person observer. The first-person observer may be actively involved in an interaction or passively observe an interaction from a third-person perspective. The interactions can be broadly classified into human-human interaction, human-object-human interaction and human-object interaction. Some examples for each group are given in the corresponding blocks.

To make the "improved trajectories" robust for wearable cameras, an adaptation based on motion pyramid is used. The video is segmented into different regions based on the magnitude and direction of motion. Direction of flow vectors is used to for segmenting into foreground and background while magnitude is used to assign a relative score to the pixels in foreground in every frame. Segmenting video into different regions with motion as criterion is useful in case of first person actions, since the main action will

generally be in the region associated with a higher motion score and any other unrelated background action or camera motion will be in the region associated with lower score. This is because the first-person actions are near to the camera. Identifying the dominant camera (head) motion helps in discarding the motion trajectories generated due to head motion.

In realistic scenarios, detecting actions in long and continuous video sequences is an important requirement. Continuous videos may contain multiple actions over time with periods of zero activity (no actions being performed) in between. We use the same approach along with sliding window technique to detect the actions by taking a decision based on the predicted class of the neighboring segments in such long video sequences.

In the real-world, the first-person observer will invariably observe passively (not being part of the interaction) from a TPP as well. Hence recognizing interactions in both perspectives is necessary. To this end, we introduce a new dataset which contains interactions from both perspectives, captured from a head-mounted camera. Interactions can be broadly grouped into three categories, *viz*., human-human interaction, human-object interaction, human-object-human interaction. The categorization is given in Figure 1.

2. Related Work

Egocentric vision research is receiving significant attention due to technological advancement of wearable cameras. Some of the problems researched in first-person vision are object recognition [13, 6], object-based activity recognition [12, 4], video summarization or "life-logging" [9], social interaction recognition [5], interaction-level human action recognition [14]. The research on activity detection is object-based and mainly involve activities in which objects are held or manipulated by first-person observers.

Action recognition has been an important area of research in the vision community for a long time. Majority of the previous works have focused on third-person videos based on space-time interest points (STIP) features using various detectors like Harris3D [7], separable Gabor filters [3], *etc*. Often local features for the interest points are based on gradient information, optical flow [3, 8]. The trajectory-based methods for action classification are presented in [1, 10, 17, 15]. Wang *et al*. [15] use local 3D volume descriptors based on motion boundary histograms (MBH), histogram of oriented gradients (HOG) and histogram of optical flow (HOF) around dense trajectories to encode action. The MBH descriptors are known to be robust to camera motion. Recently in [16], Wang *et al*. estimate the camera motion and compensate for it and thereby improving the trajectories and the associated descriptors. The improved trajectories method yields state-of-the-art results on challenging datasets. We use [16] method as baseline in this work.

Ryoo and Matthies [14] recognize actions in egocentric videos captured by a humanoid. Global and local motion information are used as features. For the local motion, cuboid feature detector [3] is used to obtain video patches containing salient motion. The features are clustered using k-means and the activity is represented as histogram of words. Finally, the actions are classified using SVM. The humanoid does not move on its own during the interactions. There is motion during interactions like handshake, petting the robot, punching the robot. To emulate ego-motion and mobility of a real robot, wheels are placed under it and is pushed around by a human.

While it is plausible for the humanoid observer to not move on its own during the interaction, it is not the case for a human observer whose varying head motion will result in undesired camera motion. The variations of head motion for different persons and actions also makes it challenging to recognize actions when the first-person observer is human.

Contributions of our work based on these considerations are the following. (i) A simple yet effective trajectory scoring technique for grouping via foreground identification and relative motion map is discussed. (ii) To the best of our knowledge, this is the first paper which focuses on recognition of human interactions viewed from both perspectives, *viz*., first-person and third-person perspectives. (iii) A challenging dataset for interaction recognition in the two perspectives is introduced.

The rest of the paper is organized as follows. In section 3, the interaction recognition approach is discussed. The experiment setup and results on the interaction datasets are given in section 4. The adaptation of the recognition approach to action detection in long video sequences is detailed in section 5 and we conclude the paper in section 6.

3. Interaction recognition

In this section, we discuss the interaction recognition approach using improved trajectories. The camera motion identification and grouping descriptors, based on motion pyramid, are used to improve the performance of the recognition. The overall approach is illustrated in figure 2.

3.1. Foreground motion map

In videos captured by head-mounted cameras, the camera motion is prominent and pronounced due to the objects being near to the camera. This results in varying and unintended camera motion which is not representative of the motion present in an action. For *e.g*., "typing on keyboard" action is represented by the motion of the hands on the keyboard. However the motion of the head can introduce motions not representing the typing action. The motion boundary histogram (MBH) trajectory descriptor is known to be robust to camera motion. While the camera motion may

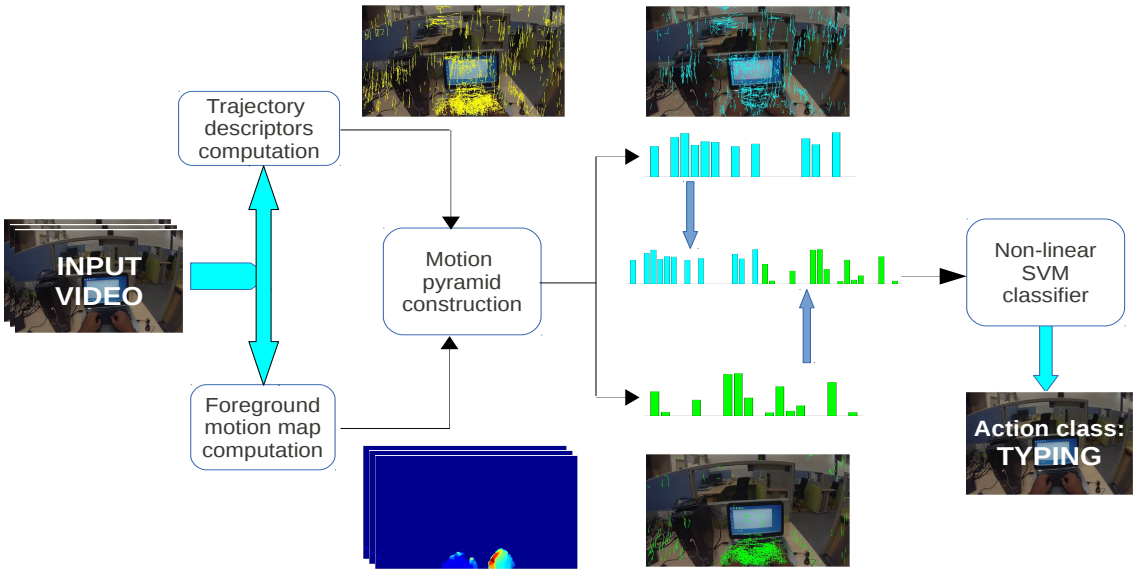

Figure 2. Illustration of the interaction recognition approach for first-person videos. The baseline improved trajectory features are computed. The foreground motion map for every frame is computed by multiplying the foreground mask and the motion magnitude map. The trajectory score is computed by adding the foreground motion map scores of the pixels the corresponding trajectory passes through. Based on the relative scores, different trajectories and corresponding features are grouped together and descriptors for each group (figure two groups) are computed. Concatenating the group descriptors results in a single descriptor for the video which is used for classification. (Best viewed in color).

be present in the third-person actions as well, since the actions take place relatively far from the camera when compared to first-person actions, the effect of camera motion is less in third-person actions. Hence identifying the camera/head motion in first-person videos helps in classifying actions better.

To estimate the motion due to camera, we assume that the camera motion is dominant compared to the foreground motion. Optical flow is computed between successive frames of the video. Each frame is divided into grids and histogram of flow in 9 directions (including zero degree bin) is computed for each grid. The flow directions are arranged in decreasing order of occurrence in the entire frame and the top 5 directions are checked for the following conditions. A binned flow direction is considered to be due to camera motion if it is present in multiple grids of a frame, or if at least 70% of pixels in a particular grid are associated with that flow direction. A foreground mask for each frame represents the pixels with flow vector directions which do not agree to the above conditions.

The motion map of a frame represents the magnitude of the flow vector at every pixel. The magnitude of flow vectors in the regions associated with camera motion may be varying across the frame and in turn may result in varied scores for trajectories associated with camera motion. To facilitate the grouping of trajectories, the trajectories belonging to the action must have similar scores. Hence, using this map directly to score and group the trajectories is not feasible. To this end, a foreground motion map for each frame is computed by multiplying the motion map with the corresponding foreground mask of the frame. This ensures that the scores for the regions associated with the foreground is higher than that of the background regions. This is illustrated in figure 3. The magnitude values in the foreground motion map are normalized by dividing them by the maximum magnitude of the foreground motion map in that frame.

3.2. Grouping trajectories

The motion trajectories are generated by tracking feature points in the video and is explained in section 4.1. The feature points which remain nearly static are discarded in the process. This ensures that only those feature points which move in the scene generate trajectories. But due to head motion, even the otherwise static background feature points move and generate trajectories which are not associated with the motion of the action being performed. This results in noisy descriptors for the action. In order to overcome this, we group the trajectories based on the magnitude scores in the foreground motion maps. The score of a trajectory is computed by adding the magnitude values of the pixels the trajectory passes through in the video. The magnitude values are from the foreground motion map.

The trajectories due to camera motion will pass through

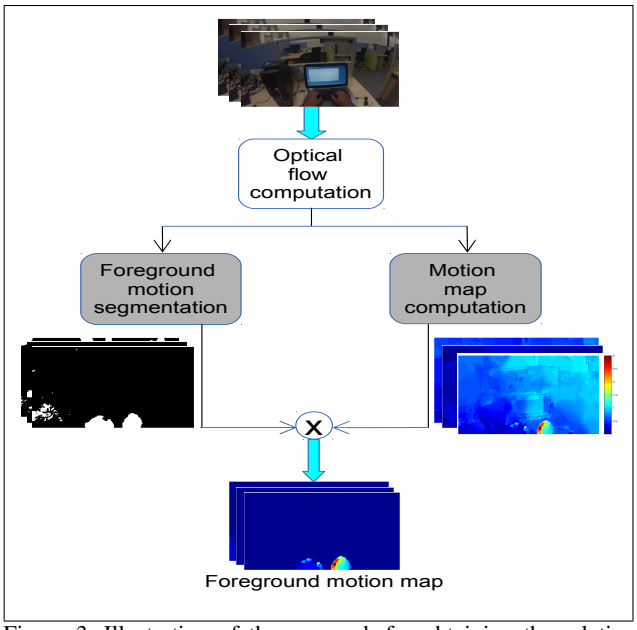

Figure 3. Illustration of the approach for obtaining the relative magnitude in the foreground region. Optical flow is computed for the video frames. The magnitude of flow in each frame is represented by the motion map. Foreground mask for each frame is generated based on the direction of flow vectors. The map and mask are multiplied to obtain the motion map for only the foreground.

regions with very low values in the foreground motion map in majority of the frames. Hence the trajectories which are generated due to camera motion receive lower scores than the trajectories associated with the interaction. After arranging the trajectories according to their scores in ascending order, they can be segmented into multiple groups. Majority of the trajectories due to head motion will have low scores and fall into the first few groups depending on the number of groups chosen. In figure 2, trajectories are separated into 2 groups. The group on top with trajectories plotted in cyan are mostly the camera motion trajectories and the bottom group with trajectories in green consists of action trajectories in majority. The histogram descriptors are computed for each group and concatenated into a single descriptor for the action. Figure 4 illustrates few examples of the grouping technique.

In this section, we have detailed the pre-processing stage of grouping trajectories to separate the trajectories generated due to camera motion and motion associated with the interaction. Finally, the improved trajectory descriptors are computed for each group of trajectories and concatenated to form a single descriptor and used as input to the classification stage.

Figure 4. Illustration of our trajectory segmenting technique on few examples. Each row corresponds to an interaction. The top 2 rows (FPP) are use cellphone and write on paper and bottom 2 rows (TPP) are open door and handshake. The left column contains trajectories (in yellow) with lowest motion scores. The rightmost column denotes trajectories (in cyan) with highest motion scores. The second and third columns denote trajectories (in green and blue) with in between motion scores. Majority of trajectories depicted in first 2 columns are from background while the 3rd and 4th columns denote most of the trajectories representing the action. (Best viewed when zoomed).

4. Experimental setup and Results

In this section, the details of the experimental setup with various parameter settings are provided. We use the improved trajectories as the baseline and modify it using trajectory grouping approach. The datasets used for the experiments are detailed in section 4.3 along with the performance of the approach in section 4.4.

4.1. Trajectory acquisition

The motions in the scene are represented quantitatively using dense optical flow. A pixel at $\mathbf{p}_t = (x_t, y_t)$ at frame $t \in [1, L-1]$ moves to

$$\mathbf{p}_{t+1} = (x_{t+1}, y_{t+1}) = \mathbf{p}_t + (u_t(\mathbf{p}_t), v_t(\mathbf{p}_t)) \quad (1)$$

at frame $t+1$ Here (u_t, v_t) represent the optical flow field at frame t. A trajectory is represented by $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_L)$. We use the code [1] provided by [16] with default settings to acquire the trajectories. The length of the trajectories is 15 samples. The trajectory features used in the experiments are TrajShape, HOF and MBH.

[1] http://lear.inrialpes.fr/people/wang/improved_trajectories

4.2. Feature encoding

We experiment with two types of feature encoding, *viz*., bag-of-words and Fisher vector encoding. For bag-of-words, the features are clustered using k-means and a codebook of 1000 is learnt for each descriptor. Each video is then represented by a histogram of words. We use an RBF-χ^2 kernel SVM to classify the actions. A *one-vs-all* SVM is used for multi-class classification. For combining different descriptors, the kernel matrices normalized by average distance are summed and the classifier is learnt.

Fisher vector [11] has shown an improved performance over bag of features for both image and action classification [2, 16]. It encodes encodes the first and second order statistics between the video descriptors and a Gaussian Mixture Model (GMM). The number of Gaussians is set to $K = 256$ and randomly sample a subset of $256,000$ features to estimate the GMM. The dimensions of the features are reduced by half using PCA. Each video is, then, represented by a $2DK$ dimensional Fisher vector for each descriptor type, where D is the reduced descriptor dimension after performing PCA. The fisher vectors are power and L2 normalized. Different descriptors are combined by concatenating their normalized Fisher vectors. For classification, a linear SVM with cost parameter $C = 100$ is used.

4.3. Datasets

The performance of the approach is evaluated on 2 first-person datasets. The first is JPL First-Person Interaction dataset and the second dataset is our new NUS First-person Interaction dataset.

JPL First-Person Interaction [14] dataset consists of 7 actions, including 4 friendly interactions with the observer, 1 neutral interaction, and 2 hostile interactions. Shaking hands with the observer, hugging the observer, petting the observer, and waving a hand to the observer are the four friendly interactions. The neutral interaction is the situation where two persons have a conversation about the observer while occasionally pointing it. Punching the observer and throwing objects to the observer are the two negative interactions. In total there are 84 videos with 12 instances or sets for each action. The number of groups for trajectory grouping for the dataset is 2. During classification, 6 sets are randomly chosen for training and the rest for testing. The experiment is run 100 times with random train-test splits and average classification accuracy is reported.

We introduce the **NUS First-person Interaction dataset** for interaction recognition with 8 interactions in 2 perspectives (first-person and third-person perspective) resulting in 16 classes in total. The dataset will be made publicly available at a later date. The dataset contains 2 human-human interactions, 2 human-object-human interactions and 4 human-object interaction classes. 'Handshake' and 'waving' are the human-human interaction classes. Human-object-human interactions involve 'throwing an object' and 'passing an object'. The 4 human-object interaction classes are 'open and go through door', 'using cellphone', 'typing on keyboard', and 'writing on board/paper'. Our dataset is significantly different from JPL dataset since the camera is worn by a human and the variations in head motion make it challenging for classification.

Our dataset contains 260 videos with at least 15 samples in each of the 16 classes. The dataset was collected using a GoPro camera head-mounted on the human observer. The videos were captured in 720p resolution at 60 frames per second to decrease the motion blur due to head-motion during capture and later down-sampled to $430 * 240$ frame resolution at 30 frames per second. To the best of our knowledge, this is the first dataset for interaction recognition which includes actions where the observer actively interacts and passively observes the interactions. The videos are captured by different people and with different actors in addition to view variations. Hence this a challenging dataset. The different classes of the dataset is illustrated in figure 5.

The trajectories are segmented into 4 groups (section 3.2) and since the head motion of human observers can be high in comparison to a humanoid, we discard the first group and concatenate the descriptors from the remaining three groups. For classification, we use random train-test splits with 50% each in training and testing sets and run the experiment 100 times and report the average classification accuracy.

4.4. Results

The results for the two datasets are reported in table 1. Improved trajectories are used with two feature encodings, *viz*., BoW and Fisher vector encoding. The performance for the Fisher vector encoding is better for both datasets when compared to BoW encoding performance. We use the motion pyramid and grouping technique on both the encodings and observe an increase in classification accuracy. The improvement for the BoW encoding is higher than that for Fisher vectors and the Fisher vector encoding with trajectory grouping performs the best for both the datasets.

For the JPL dataset, the average classification accuracy, over multiple runs, for improved trajectory features (ITF) with Fisher encoding is 96.1% with deviation of 0.3% and the proposed method performs marginally better at 96.7% with deviation of 0.2%. Since the camera motion during the action is relevant to the action itself and does not contain any unrelated motion, the increase in performance is only marginal.

The proposed method is evaluated on our dataset and the confusion matrix is shown in figure 7. The proposed method achieves an average accuracy of 61.8% while the baseline achieves 58.9%. The deviation from average was around

Figure 5. Our new interaction dataset is shown. The top and bottom rows correspond to first-person perspective and third-person perspective respectively. From left to right the eight interaction classes are Use cellphone, Open door and go through, Handshake, Pass an object, Throw an object, Type on keyboard, Wave, Write on board/paper.

Method	JPL Interaction	New dataset
Ryoo & Matthies [14]	89.6%	-
ITF (BoW)	93.2%	54.3%
ITF (Fisher)	96.1%	58.9%
Proposed (BoW)	95.4%	58.3%
Proposed (Fisher)	96.7%	61.8%

Table 1. Performance comparison on the two datasets using baseline method (ITF [16] - Improved Trajectory Features) and proposed approach. The baseline is evaluated using two feature encodings, *viz.*, bag-of-words and fisher vector. The proposed method is also evaluated under the two encodings.

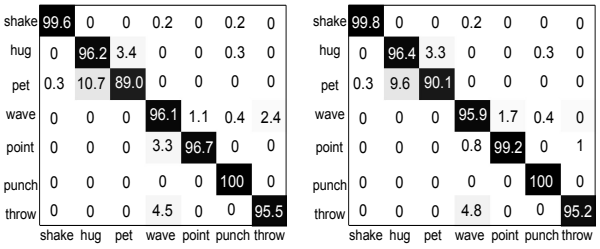

Figure 6. Confusion matrices for JPL dataset. Matrix on left is for baseline ITF and one on the right is by the proposed method.

0.7% in both cases. The proposed method of grouping trajectories helps in better classification of some actions like using cellphone, typing, waving in FPP. There is misclassification between actions from different perspectives. The misclassification of interaction in TPP as FPP is higher than an FPP interaction being misclassified as TPP action, and the classification of interactions in FPP is better than the interactions in TPP. This is because the interactions in TPP can occur in multiple views while in FPP, they occur frequently in similar views. We can expect the FPP interactions to be bunched together in the Kernel space but the interactions from TPP may be scattered (due to view changes in TPP) and will require more number of training samples from different views to learn a better classifier. Also the interactions like using cellphone, typing are better observed in FPP. Hence the variation in performance for the two perspectives.

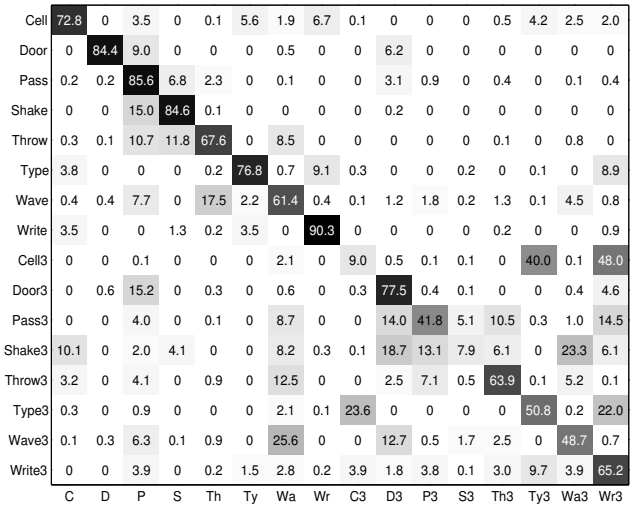

Figure 7. Confusion matrix for the proposed method evaluated on our dataset. Action labels suffixed with '3' represent third-person perspective.

5. Interaction detection

Detecting actions is necessary in long video sequences where multiple actions can occur in addition to periods of no activity occurrence. We use the improved trajectory features with Fisher encoding, as before, for detecting activities. We use the continuous videos of JPL interaction dataset for evaluation. The descriptors are computed in a sliding window method. The window length is 40 with overlap of 20 frames between successive window segments. The window duration of approximately 1.5 seconds is chosen to in order to effectively capture descriptors of short duration actions like waving. There are 57 videos with the number of activities varying between 1-3 in each video. The videos are divided into 12 sets and 6 each are used for training and testing. The experiment is run multiple times with

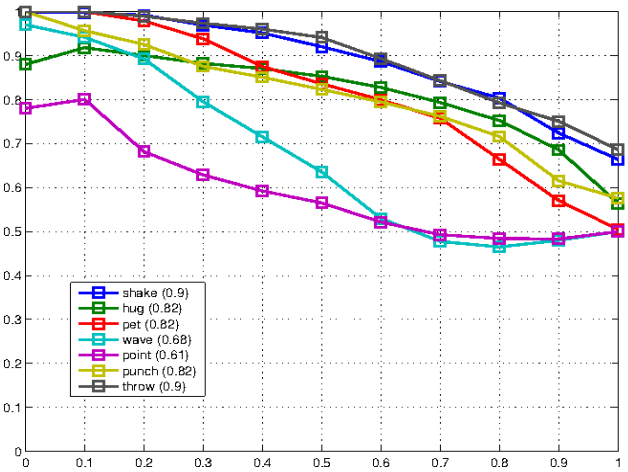

Figure 8. The average precision-recall curves for the 7 activities of the JPL interaction dataset are plotted. The mean precision for each class is given within braces in the legend. (Best viewed when zoomed).

random splits.

The average number of segments per class in the training videos is used in the detection stage. Initially, the probability for each segment belonging to a particular activity class are obtained using a *one-vs-rest* SVM. In the next step, the scores of nearby segments (number of segments depending on the average number of segments of the activity in the training videos) are added and used to detect the activity. The scores are sorted in decreasing order. As and when a particular segment is recalled by a detector, the segments nearby which contributed to its cumulative score are also given the same class label and considered as recalled. The precision-recall curves averaged over 100 runs for all the 7 classes are plotted in figure 8. The average precision for the dataset is 0.79. Ryoo and Matthies [14] report an average precision of 0.71 for the same. From figure 8, we observe that the 'point' and 'wave' activities have relatively lower average precision values (0.61 and 0.68). 'Point' action has less motion involved when compared to other activities and is similar to segments without interactions. The 'wave' action occurs within a short duration and is recalled with precision above 0.9 at lower recall but false positives increase as recall is increased. Both these actions have lower precision values with increasing recall.

6. Conclusion

In this work, we have evaluated the improved trajectories approach to interaction recognition in first-person videos captured in first and third person perspectives. An improvement, based on segmenting the trajectories according to motion magnitude and direction, to overcome the camera motion associated with head-mounted camera is discussed. Since an observer can actively be involved in an interaction or passively observe an interaction, in order to evaluate the method jointly from both perspectives, we introduce an interaction recognition dataset containing interactions from first-person and third-person perspectives. The method was also used to detect interactions in videos containing multiple interactions. Future work will involve capturing interactions simultaneously from first-person and third-person cameras and use the information from both in a unified manner to recognize interactions.

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *ICCV*, 2007. 2

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011. 5

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior a recognition via sparse spatio-temporal feature. In *VS-PETS*, 2005. 2

[4] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011. 2

[5] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012. 2

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

[7] I. Laptev and T. Lindberg. Space-time interest points. In *ICCV*, 2003. 2

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2

[9] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2

[10] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: action recognition through the motion analysis of tracked features. In *ICCV Workshop*, 2009. 2

[11] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010. 5

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

[13] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010. 2

[14] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013. 2, 5, 6, 7

[15] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, pages 1–20, 2013. 2

[16] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, 2013. 1, 2, 4, 5, 6

[17] S. Wu, O. Oreifej, and M. Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *ICCV*, 2011. 2