Delving into Egocentric Actions

Yin Li Zhefan Ye James M. Rehg
School of Interactive Computing, Georgia Institute of Technology
{yli440, zye32, rehg}@gatech.edu

Abstract

We address the challenging problem of recognizing the camera wearer's actions from videos captured by an egocentric camera. Egocentric videos encode a rich set of signals regarding the camera wearer, including head movement, hand pose and gaze information. We propose to utilize these mid-level egocentric cues for egocentric action recognition. We present a novel set of egocentric features and show how they can be combined with motion and object features. The result is a compact representation with superior performance. In addition, we provide the first systematic evaluation of motion, object and egocentric cues in egocentric action recognition. Our benchmark leads to several surprising findings. These findings uncover the best practices for egocentric actions, with a significant performance boost over all previous state-of-the-art methods on three publicly available datasets.

1. Introduction

Understanding human actions from videos has been a well-studied topic in computer vision. The recent advent of wearable devices has led to a growing interest in understanding egocentric actions, i.e. analyzing a person's behavior using wearable camera video, otherwise known as First-Person Vision (FPV). Since an egocentric camera is aligned with the wearer's field of view, it is primed to capture the first person's daily activities without the need to instrument the environment. Knowledge of these activities facilitates a wide range of applications, including remote assistance, mobile health and human-robot interaction.

Despite the tremendous effort on understanding actions in a surveillance setting [1, 35], it remains unclear whether previous methods for action recognition can be successfully applied to egocentric videos. Our first observation is that egocentric video includes frequent ego-motion due to body movement. This camera motion can potentially hamper the motion-based representations that underlie many successful action recognition systems. In contrast, state-of-the-art egocentric action recognition methods [6, 27, 7] rely mainly on an object-centric representation for discriminating action categories. However, most of these works did not test motion-based representations on a common ground, e.g. separating the foreground motion from the camera motion. Thus, a systematic evaluation of motion cues in egocentric action recognition remains missing.

What makes egocentric videos different from surveillance videos? The key is not simply that a camera is moving, but rather that the movement is driven by the camera-wearer's activities and attention. In a natural setting, the camera wearer performs an action by coordinating his body movement during an interaction with the physical world. The action captured in an egocentric video contains a rich set of signals, including the first person's head/hand movement, hand pose and even gaze information. We consider these signals as mid-level egocentric cues. They usually come from low-level appearance or motion cues, e.g. hand segmentation or motion estimation, and are complementary to traditional visual features. These mid-level egocentric cues reveal the underlying actions of the first person, yet have been largely ignored by previous methods for egocentric action recognition.

We provide an extensive evaluation of motion, object and egocentric features for egocentric action recognition. We set up a baseline using local descriptors from Dense Trajectories (DT) [36], a successful video representation for action recognition in a surveillance setting. We then systematically vary the method by adding motion compensation, object features and egocentric features on top of DT. Our benchmark demonstrates how these choices contribute to the final performance. We identify a key set of practices that produce statistically significant improvement over previous state-of-the-art methods. In particular, we find that simply extracting features around the first-person's attention point works surprisingly well. Our findings lead to a significant performance boost over state-of-the-art methods on three datasets. Figure 1 provides an overview of our approach. Materials for reproducing our results can be found in our project website.[1]

Our work has three major contributions: (1) We propose

[1] www.cbi.gatech.edu/egocentric

a novel set of mid-level egocentric features for egocentric action recognition, and demonstrate that how they can be combined with low-level features to effectively improve the performance. (2) We provide the first systematic evaluation of motion, object and egocentric features in egocentric actions. Our benchmark shows how different features contribute to the performance. (3) Our study identifies a key set of ingredients that are critical to the performance. These best practices are shown to provide significant performance boosts on existing datasets. In addition, our findings contain valuable insights for understanding egocentric actions.

Our findings, derived from a large set of experiments, can be summarized into three parts: (1) Motion cues, with an explicit model of camera movement, can provide comparable results with the state-of-the-art methods that use object-centric features. This result is surprising and challenges the prevailing view that motion features are less reliable in egocentric videos. (2) Object cues, even simple visual features, when combined with foreground regions, can significantly improve the performance in object related actions. This supports the argument for the importance of object-centric representations. (3) Egocentric cues, when combined with motion and object cues, can provide a further large increase in performance. The performance gap indicates that mid-level egocentric cues are crucial for egocentric action recognition. We also discuss issues regarding implementation details and existing benchmarks.

2. Related Work

2.1. Action Recognition

There is a large body of literature on action recognition in computer vision. Recent surveys can be found in [1, 35].

Local spatial-temporal features have been the most prevalent over the past few years [19, 36]. Laptev [19] introduced the Space-Time Interest Point (STIP) by extending 2D Harris corner to 3D. Wang et al. [36] propose to densely sample feature points and track them using optical flow. Multiple descriptors, including HOG [3], HoF [4], MBH [36] or Cuboids [5] can be computed around the interest points, followed by a bag-of-features representation for action recognition. These spatial-temporal descriptors aim to find key features that are relevant to actions. The framework was shown to be robust to challenging scenes and achieved the state-of-the-art performance on major benchmarks. However, they have not been fully explored in the egocentric setting, due to the significant ego-motion.

There has been growing interest in using mid-level features for action recognition. Fathi and Mori [8] propose to construct mid-level motion features from optical flow using AdaBoost. Raptis et al. [28] extract action parts by forming clusters of trajectories. Recognition is then formulated as matching a subgraph of parts to a template. Tian et al. [34] extend the deformable part model to 3D volumes and learn 3D spatio-temporal parts for action detection. Jain et al. [11] demonstrate that mid-level discriminative patches can be mined from video, and used for classification as well as building correspondences between videos. Most recently, Mathe and Sminchesescu [23] demonstrate promising results for recognizing actions by sampling local descriptors from a predicted saliency map. While most of the previous work focus on a surveillance setting, we show that egocentric video provides rich signals about the camera wearer and these signals can be used as mid-level features for understanding the first-person's actions.

2.2. First-Person Vision

By taking advantage of a point-of-view camera, there have been several recent advances in First-Person Vision (FPV) [14], also known as egocentric vision, such as video summarization [20], video stabilization [16], object recognition [29] and action and activity recognition [33, 6, 27].

We focus on egocentric action and activity recognition in this paper. Spriggs et al. [33] address the segmentation and classification of activities using a combination of egocentric videos and wearable sensors. Fathi and Rehg [9] propose to differentiate egocentric actions by modeling the change of the states of objects and materials in the environment. Moreover, several papers [6, 9, 27] reported that local spatial-temporal features (such as STIP [19]) often fire at locations irrelevant to an action due to the camera motion and therefore perform poorly. A possible solution is an object-centric representation proposed by Pirsiavash and Ramanan in [27]. However, directly applying local features in egocentric videos is problematic, as most of them implicitly assume a static camera. We provide a systematic study of these local features, and demonstrate that they perform surprisingly well if the camera motion is removed.

In contrast, the camera motion generated by the first person can be a useful cue for understanding egocentric actions. For example, Kitani et al. [15] track camera motion using sparse optical flow, and encode the motion of tracked points into a histogram, which is used to cluster egocentric video content and discover egocentric actions. Ryoo and Matthies [31] combine a global motion descriptor with local motion cues for interaction-level action recognition. In fact, egocentric videos embed a rich set of mid-level egocentric cues, including body movement, hand pose and location and even gaze information. Our previous work [22] models the coordination of hand, head and eye, which is then used to predict egocentric gaze from hand pose and head movement. Fathi et al. [7] propose to select key visual features around egocentric gaze for action recognition. However, these mid-level cues had never been evaluated carefully in egocentric action recognition. We explore novel approaches for encoding these features, and provide a study of how

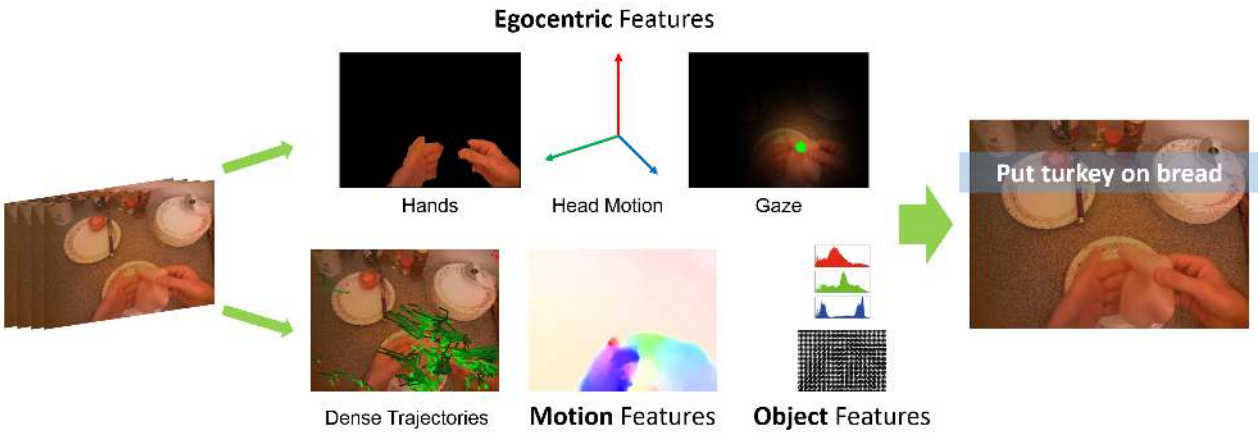

Figure 1. Overview of our approach. We propose to combine a novel set of mid-level egocentric cues with low-level object and motion cues for recognizing egocentric actions. Our *egocentric* features encode hand pose, head motion and gaze direction. Our *motion* and *object* features come from local descriptors in Dense Trajectories, with motion compensation using head motion. We design a systematic benchmark to evaluate how different types of features contribute to the final performance, and seek the best recipe using motion, object and egocentric cues. Our findings significantly advance the results in all our benchmarks.

different egocentric cues affect the performance.

3. Motion, Object and Egocentric Cues

We give a brief description of the motion, object and egocentric features used in our paper. Different encoding schemes for the egocentric features are also discussed.

3.1. Local Descriptors for Motion and Object Cues

Our method is built upon the pipeline of DT [36], which has not been fully explored in egocentric video. The success of DT lies in its dense tracking strategy using optical flow, and the combination of multiple descriptors aligned with the trajectories. Dense sampling ensures that key visual information is captured by the trajectories. The feature set is designed for different aspects of an action, including trajectory shape, 2D image boundary, motion direction and motion boundary. Each descriptor in DT thus includes its spatial-temporal **trajectory** and the **features** along the trajectory. We further separate the features into motion features and object features.

Motion Features: Motion is the inherent nature of an action. DT captures motion information by 1) trajectory features of the shape of a trajectory; 2) Histogram of Flow (HoF) as the local motion pattern; 3) Motion Boundary Histogram (MBH) using the gradient of optical flow split into vertical (MBHy) and horizontal directions (MBHx), as the shape of moving foreground objects.

Object Features: Object information is crucial in egocentric settings, as many of the actions are based on interactions with objects. DT include Histogram of Oriented Gradient (HOG), which encodes the 2D image boundaries. We further augment DT with histogram of LAB color and Local Binary Patterns (LBP) along the trajectory, capturing color and texture information. We deliberately choose our object features as low-level descriptors to keep our pipeline simple.

3.2. Egocentric Cues

We introduce our egocentric cues, which are used as mid-level features, and show how they can be inferred from an egocentric video.

Hand Pose and Movement: The first person's hand pose, location and movement are directly related to their interaction with objects. Accurate segmenting and tracking of hands in egocentric video remains an open problem [21]. Our previous work [22] demonstrated that egocentric hands can be abstracted by the concept of a manipulation point. A manipulation point is a $2D$ point in the image, defined as a control point where the first person is most likely to manipulate an object using his or her hands. It can be obtained by analyzing the shape of the hands in each frame, and is relatively robust to hand segmentation performance. We apply textonBoost with CRF [32] for hand segmentation, and match the boundaries to different templates under four hand configurations as in [22].

Head Movement: Head movement is important for egocentric actions. The egocentric camera is aligned with the first-person's head direction, and thus the head motion is captured by the camera motion. We model the camera motion by matching sparse interest points between successive frames using ORB [30]. To ensure a homogeneous distribution of interest points for robust motion estimation, we divide the image plane into a grid, trying to extract at least 3 points per cell. We also delete interest points that lie on the hand mask. We then fit a homography for head

motion using RANSAC [10].

Gaze Direction: As we sense the visual world through a series of fixations, our gaze reveals important information about our goals. Gaze points often lie on objects that are relevant to the task we are performing, since gaze is used to coordinate actions [18]. In egocentric gaze, the point of regard is represented by a $2D$ image point in each frame. In our experiments, we use a wearable eye tracker to obtain gaze measurements.

Encoding Head and Hand Movement: Head motion and hand movement are complementary to visual features. We, therefore, propose to directly encode the head motion and the trajectory of manipulation points as separate feature channels. We also experimented with encoding the trajectory of gaze points, yet found only negligible improvement.

3.3. Egocentric Cues Improve Local Descriptors

The key challenge for using local descriptors in egocentric video is that they often fire at locations that are irrelevant to the current action. This is mainly due to camera motion and background clutter. In addition to directly exploiting egocentric cues, we show that they can be used to produce meaningful local descriptors for egocentric actions. This is done by motion compensation and trajectory selection.

Motion Compensation: Camera motion compensation is important for egocentric actions. Several recent efforts addressed the issue in a surveillance setting by either stabilizing the input video [25] or compensating the optical flow [37, 12]. The latter has been proved to be effective for action recognition [37, 12]. Thus, we adapt the technique from [37]. We directly subtract camera motion from dense optical flow field and reject trajectories with a small motion. We did not back-warp a future frame as in [37]. Warping is less reliable when a large camera motion in egocentric video is approximated by a homography.

Motion compensation has two major effects. First, it helps to select trajectories on foreground regions that move differently from the camera motion. Secondly, it generates more reliable motion features that exclude the ego-motion from the dense optical flow field. Note that our implementation is different from [37]. Our version uses ORB features homogeneously distributed on the image plane. It only requires dense optical flow to be computed once, and is more efficient with comparable results.

Trajectory Selection: Gaze points index key locations that are discriminative for actions. Fathi et. al [7] proposed a simple heuristic by only encoding visual features around gaze points. In this case, egocentric features provide a weak spatial prior of an action. We experiment with selecting local descriptors by their trajectories in the vicinity of both manipulation point and gaze point. Trajectory selection drives local descriptors to focus on egocentric actions, by filtering out descriptors with irrelevant trajectories, e.g. trajectories due to the background clutter. It also improves efficiency as fewer descriptors are used for recognition.

4. Egocentric Action Recognition

We now describe our approach to combine object, motion and egocentric features for action recognition. We discuss the details of our implementation and benchmark, followed by our results and findings. Our results demonstrate a significant performance boost on three existing datasets.

4.1. Method and Implementation

Feature Extraction: Our method shares a similar pipeline with [37]. We track feature points using DT in an input video, using a time window of 6 frames. Note the trajectory length is shorter than [37], as many of the egocentric actions only last for a few seconds. We extract a set of local descriptors aggregated along the trajectories. Each descriptor consists of 7 feature channels, including trajectory features, MBHx, MBHy, HoF, HoG, LAB color histogram and LBP. We use 8-neighbour comparison for LBP and quantize three color channels (LAB) separately into 8 bins each. Each trajectory is further divided by $2 \times 2 \times 3$ grids and histograms of features within each grid are concatenated. The final dimensions of the descriptors are 12 for Trajectory, 96 for HoG and LBP, 108 for HoF, 192 for MBH and 288 for Color. Other parameters of DT are kept the same as in [36]. We also extract egocentric features at each frame, including head motion parameters (8D homography) and hand manipulation point (2D).

Fisher Encoding: We encode all descriptors using Improved Fisher Vector (IFV) and concatenate the result vectors. IFV [26] has been shown to outperform other encoding methods in action recognition [24]. IFV is obtained by soft quantization of the projected descriptor of dimension D using a Gaussian Mixture Model (GMM) of K components. Zero, first and second order differences between each descriptor and its Gaussian cluster mean are calculated, and weighted properly by the Gaussian soft-assignments and covariance. They are then averaged into an unnormalized fisher vector. We further take a signed square root of its scalar components and normalize the vector with a unit l_2 norm [26]. The result is a fisher vector of the dimension $(2D + 1)K$. For all of our experiments, we first perform PCA to reduce the input feature dimensions by 50%, followed by a GMM with $K = 50$ using 200K randomly sampled descriptors. To eliminate randomness in clustering, all results are obtained by averaging over 5 runs.

Classification: We concatenate the IFVs from different features into the final representation of the video. We train a linear SVM over the final FV for action recognition. The SVM parameter C is selected by leave-one-subject-out

	Task	Mounting	Resolution	FPS (Hz)	Duration (hours)	# Subjects	# Action Categories	# Action Instances	Other Sensors
GTEA [6]	Action	Head	1280*720	30	0.6	4	71(61*)	525(456*)	N/A
GTEA Gaze [7]	Action	Head	640*480	30	1	14	40(25*)	331(270*)	Gaze
GTEA Gaze+ [7]	Action	Head	1280*960	24	9	6	44	1958	Gaze
UCI ADL [27]	Activity	Chest	1280*960	30	10	20	18	364	N/A
JPL Interaction [31]	Activity	Chest	320*240	30	0.4	N/A	7	94	N/A
EgoAction [15]	Action Discovery	Head	840*480	30	0.7	N/A	N/A	N/A	N/A
UT Ego [15]	Summary	Head	480*320	15	20	4	N/A	N/A	N/A

Table 1. Comparison between existing FPV datasets. GTEA Gaze+ is the largest egocentric action dataset in terms of the number of action categories and instances. We choose datasets captured by head-mounted cameras (GTEA, GTEA Gaze and GTEA Gaze+) to benchmark our method. (*Only a subset of the actions is benchmarked in previous work [9, 22, 7].)

Figure 2. Sample frames from our bechmark (GTEA, GTEA Gaze, GTEA Gaze+ dataset), which are captured by head mounted cameras. These datasets focus on recognizing object manipulation tasks during meal preparation activites. While GTEA and GTEA Gaze are collected in a controlled lab environment, GTEA Gaze+ is collected in a real-world kitchen setting with complex backgrounds.

cross-validation on the training set on GTEA (best C=40) and GTEA Gaze+ (best C=60). We manually set $C = 60$ for GTEA Gaze, where cross-validation is not feasible.

Implementation Details: We also implement spatial FV (SFV) [17] and data augmentation [2]. Data augmentation is done by mirroring the videos horizontally in both training and testing. Our final classification results are given by the averaged score between the original video and its mirrored version. We find that SFV and data augmentation consistently improve the performance, and include them in all methods across experiments.

4.2. Datasets and Baselines

We utilize three datasets for our experimental evaluations: GTEA, GTEA Gaze and GTEA Gaze+. They are publicly available and include action annotations. Each action consist of a verb and a set of nouns, such as "put turkey (on) bread". We summarize the statistics of the datasets in comparison to other FPV datasets in Table 1. We choose these datasets because 1) they are designed for egocentric action and activity recognition; 2) they are captured by head-mounted cameras, and therefore contain a rich set of egocentric cues. Sample frames of these three datasets are shown in Figure 2.

Results have been reported on GTEA and GTEA Gaze in [9, 7, 22, 6]. However, these findings can not be directly compared in order to properly understand the performance. This is because (1) the action annotation of GTEA does not include all actions that start with the verb "put", which biases the benchmark; (2) Results in [7, 22] are reported over a subset of all actions in GTEA Gaze, again, missing all "put" actions; (3) GTEA Gaze+ includes over 900 categories in total, yet most of them happen only 1-2 times and no previous result has been reported; (4) No cross-validation is performed for the reported results. Thus, the generalization error estimate may not be accurate.

We establish the first rigorous baseline through a comprehensive performance evaluation. This is done by (1) re-annotating GTEA dataset to reinstate actions that include the verb "put"; (2) reporting leave-one-subject-out cross-validation results on both the old and new lists of categories on GTEA; (3) reporting benchmark results on both the partial and full lists of GTEA Gaze; (4) redefining the list of action categories on GTEA Gaze+ by requiring each action to occur at least twice for each subject, which results in 44 action classes with 1958 instances; (5) providing leave-one-subject-out cross-validation results on GTEA Gaze+; (6) generating a set of common baselines for all three datasets

	GTEA (61) Fixed Split	GTEA (61) Cross Valid	GTEA (71) Cross Valid	GTEA Gaze (25) Fixed Split	GTEA Gaze (40) Fixed Split	GTEA Gaze+ (44) Cross Valid
STIP	32.9	31.1	25.3	26.3	23.8	14.9
Cuboids	11.2	12.5	13.3	20.1	20.6	22.7
DT	33.0	34.1	32.9	34.2	34.1	42.4
IDT	39.8	42.5	40.5	41.3	27.7	49.6
M	37.3	39.6	38.7	40.3	27.5	45.6
O	56.7	53.9	55.0	42.5	28.2	53.4
O+M	56.9	56.1	55.2	43.2	29.5	56.3
Ego Only (E)	15.3	16.3	16.5	19.9	17.4	22.3
O+M+E	59.4	55.9	55.7	44.5	32.0	56.7
O+M+E+H	61.1	59.1	59.2	53.2	35.7	**60.5**
O+M+E+G	N/A	N/A	N/A	**60.9***	**39.6***	**60.3***
M+E+H	40.8	43.1	42.3	47.6	30.3	53.2
O+E+H	**66.8**	**64.0**	**62.1**	51.1	35.1	57.4
M+E+G	N/A	N/A	N/A	44.1*	33.1*	51.3*
O+E+G	N/A	N/A	N/A	53.4*	34.1*	57.7*
State-of-the-art	39.7 [9]	N/A	N/A	32.8 [22](47.0*) [7]	N/A	N/A

Table 2. Our results are grouped into four parts, with all numbers in percentages. The first group (row) includes the baselines of STIP, Cuboids, DT and IDT. In the second group, we compare motion (M) and object (O) features. Note our motion features are a subset from IDT with trajectory features, HoF, MBHx and MBHy. The third part focuses on incorporating egocentric features. We consider direct encoding of egocentric cues (E), as well as feature extraction around an attention point given by hand (H) or gaze (G). In the fourth part, we explore the combination of motion (M) and object (O) features with the attention point by hand (H) or gaze (G). By systematically varying different components, we uncover the ingredients of success for egocentric action recognition and significantly advance the state-of-the-art. (*Results that utilized the measurement of human gaze using eye tracking glasses)

and then comparing to our method. We also supplement the datasets with 2.5K hand masks. These masks are used to train our hand segmentation pipeline. The action annotations together with hand masks are publicly available at our project website.[2]

Our baselines include STIP [19], Cuboids [5], DT and Improved DT (IDT) [36, 37]. Note that we supplement IDT with our head motion estimation, which provides slightly better results in first person videos. We also include results from [9, 7, 22]. Our results are obtained by adding motion compensation (IDT [37]), object features and egocentric features on top of DT. We report average class accuracy as the benchmark criterion. For efficiency, we resize the videos into 320×240 for GTEA Gaze and GTEA Gaze+ dataset. We use the rectified frames for GTEA from [6] and resize the video to 360×203. We also reduce the frame rate by half for all datasets. Using a higher resolution or frame rate was found to have negligible impact on the performance.

4.3. Results and Findings

To facilitate comparison to previous work, we provide benchmark results for 5 different settings: (1) GTEA dataset with old labels using the same training and testing split as in [9, 6]; (2) GTEA dataset with old labels using leave-one-subject-out cross-validation; (2) GTEA dataset with new labels and leave-one-subject-out cross-validation; (3) GTEA Gaze dataset with the same action categories and training testing split in [7, 22]; (4) GTEA Gaze dataset with all action categories using the same training testing split in [7, 22]; (5) GTEA Gaze+ dataset with leave-one-subject-out cross-validation.

In particular, we divide the features into three parts and benchmark them separately: (1) **M**otion features obtained by concatenating FVs from trajectory features, MBHx, MBHy and HoF; (2) **O**bject features by concatenating FVs from HoG, LAB color histogram and LBP; (3) **E**gocentric features by concatenating FVs from head motion and manipulation point. We also denote H and G as selecting local descriptors using manipulation point and gaze point. Our results are summarized in Table 2. The best results are highlighted in bold.

Imbalanced Data: We notice that both GTEA and GTEA Gaze have very few number of instances ($3 \sim 4$) for many categories. More precisely, the distribution of instances within each category is highly imbalanced. For example, in GTEA, while the action of "take bread" has 28 instances, 33 out of the 71 categories have less than 5 instances. This can produce misleading results [13], as missing one instance in these "sparse" categories can impose a large penalty on average class accuracy. There is

[2] www.cbi.gatech.edu/egocentric

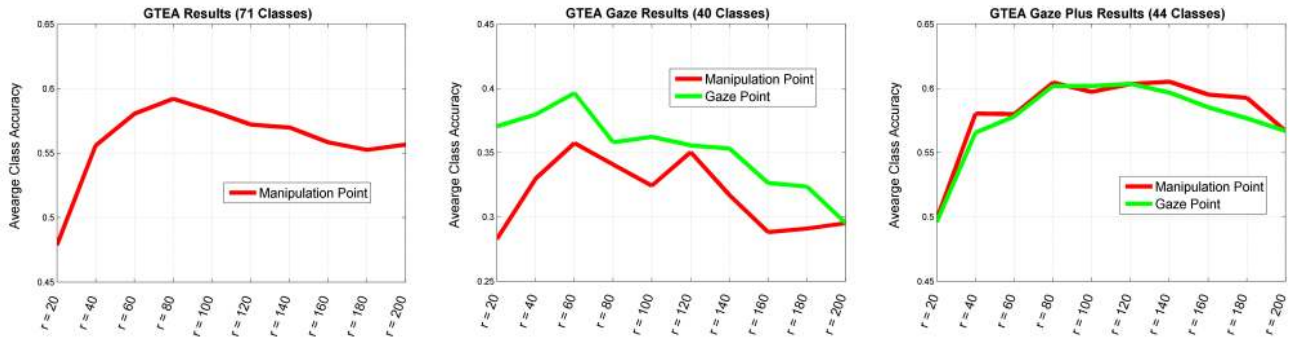

Figure 3. We encode features within a circle of raduis r in pixels around either a manipulation point (red) or a gaze point (green) for egocentric action recognition. These three plots show the impact of region size on the recognition accuracy. The baseline accuracy at $r = 200$ is given by encoding all local descriptors within the video.

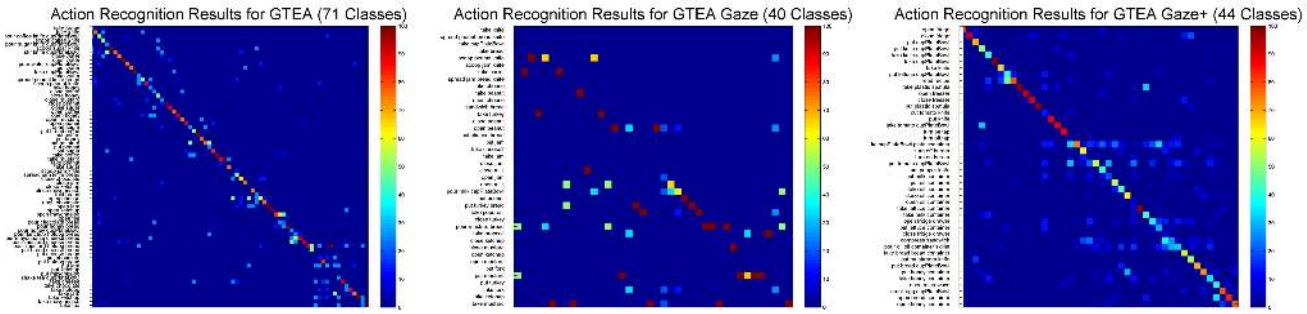

Figure 4. Confusion matrix of our method (O+M+E+H) on three datasets. Action categories are sorted based on decreasing number of instances. Our results are centered at the diagonal on GTEA and GTEA Gaze+. Our method achieves a performance boost of 27.0% in GTEA, 13.9% in GTEA Gaze and 10.7% in GTEA Gaze+ over the state-of-the-art methods [9, 7, 37].

no good single measurement to resolve this issue. Random undersampling or oversampling is a possible solution, but they may miss important instances.

The same issue also holds for GTEA Gaze+, yet to a lesser extent. With more instances, GTEA Gaze+ has a median number of 25 instances per category in comparison to 8 (GTEA) and 5 (GTEA Gaze). Therefore, methods are less likely to get penalized in average accuracy by missing a few instances. Moreover, GTEA Gaze+ is collected in a real kitchen setting with a higher resolution, while both GTEA and GTEA Gaze are captured at a lab environment. In this paper, we report results on all datasets. However, *we highly recommend that future works focus on evaluations with GTEA Gaze+*, as GTEA and GTEA Gaze are preliminary and incomplete.

Motion Compensation: Traditional action recognition methods without explicit camera motion compensation do not work well. STIP, Cuboids and DT all performed poorly, in comparison to state-of-the-art method. In our first experiment, we added motion compensation. IDT with our head motion estimation and motion features significantly improves the results on all datasets (fourth row in Table 2), except for GTEA Gaze with 40 classes. This is due to imbalanced data as we examine the confusion matrix. While we expect better results by removing the camera motion, it is a bit surprising to find that IDT already provides comparable results with state-of-the-art methods.

Object vs. Motion: We proceed by supplementing IDT with object features. We compare three different settings as shown in the second group of Table 2: (1) IDT with motion features (M) along the trajectories as baseline; (2) IDT with object features (O) along the trajectories; (2) IDT with both object and motion features (O+M) . Even with simple object features, the results are surprisingly good, outperforming all previous state of art results and motion features by a large margin. The results justify the argument that object cues are crucial in understanding egocentric actions.

We also notice that the trajectories given by IDT provide a rough location of foreground objects. Extracting object features along these trajectories is equivalent to extracting features on the foreground moving regions, which is similar to [6]. Our object features encode which object the first person is interacting with, and are therefore useful in recognizing egocentric actions. Combining object and motion features (O+M), however, only leads to marginal improvements, in comparison to using only object features.

Egocentric Cues: We further leverage egocentric features (E) in our method. These features are obtained by en-

coding the first-person's head motion and hand movements. Using only egocentric features, we achieve a reasonable performance comparable to Cuboids. We combine egocentric features with motion and object features (O+M+E), and only observe a slight improvement over all datasets. Directly encoding egocentric cues is not effective. Head motion is less discriminative for fine-grain actions. For example, taking a slice of bread and taking a peanut butter jar will result in very similar head motion. Moreover, hand movement is already encoded by the local motion features.

Trajectory Selection: In addition, we select descriptors based on their trajectories using manipulation or gaze point (O+M+E+G/H). We only encode a subset of the trajectories, which lie within the vicinity of an "attention" point, defined by a circle of radius r. The radius is defined by the minimum of the $2D$ distances between each point on the trajectory and the attention point in the corresponding frame. We vary the radius of the local region, plot the classification accuracy on all datasets in Figure 3 and report the best results in Table 2 (third group). We observe peaks along the curves. With a small region of radius equal to 60 pixels, roughly occupying 20% of the image area, our method is able to achieve a consistent performance boost from 2% to 16% over all datasets. This strategy is also very efficient as many fewer descriptors are encoded. The result indicates that "attention" points, e.g. gaze or manipulation points, provide a strong prior of where an action occurs.

In GTEA Gaze, the performance gap between manipulation points and gaze points is large. Again, we find that this result is dominated by categories with a few instances. In GTEA Gaze+, this gap is small. In fact, the manipulation point has shown to be effective for gaze prediction [22]. While current evidence cannot support the replacement of gaze points, we confirm that the concept of manipulation point is a powerful tool for egocentric action recognition. We also notice a plateau around the peaks of r, which suggests that our method is relatively robust to the measurement error of manipulation or gaze points.

Object vs. Motion Revisited: We further analyze which cue is more important with the selected descriptors. We benchmark object and motion features with the best radius (O/M+E+H/G) in the fourth group of Table 2. Constraining the features within a salient region improves the baseline performance of encoding object or motion features over the whole video. Moreover, object and motion features are complementary towards the final performance, except on GTEA, where object features are clear winners. This is largely due to the fact that this dataset used the same object instances in all actions under an ideal illumination.

Confusion Matrix: Our final results with O+M+E+H/G outperform previous results by a large margin. We improve the performance by 27.0% in GTEA, 13.9% in GTEA Gaze and 10.7% in GTEA Gaze+, in comparison to the state-of-the-art [9, 7, 37]. However, as we discussed in the beginning of the section, the single average class accuracy is not a proper measure for imbalanced data. We include the full confusion matrices as a source of additional insight. We sort the action categories with decreasing number of instances, and report confusion matrix using our best combination (O+M+E+H) on all three datasets in Figure 4. Our method is able to get most of the categories correct, except on GTEA Gaze. The result on GTEA Gaze is worse due to the mixture of low video quality, imbalanced samples and insufficient training data.

4.4. Best Practices

Based on our experimental results, we recommend the combination of O+M+E+H for egocentric action recognition. We summarize and briefly explain the best practices.

- Motion compensation is important. It leads to more reliable motion features, and can identify foreground regions for meaningful object features.
- Object cues are of crucial importance for understanding egocentric actions.
- Using an "attention" point (manipulation/gaze point) to guide feature encoding works surprisingly well. A manipulation point derived from hand shape serves as a good approximation to the actual gaze point.

5. Conclusion and Future Work

In this paper, we propose a novel set of mid-level egocentric cues, and demonstrate how they can be combined with low-level motion and object features for egocentric action recognition. In addition, we establish a rigorous benchmark baseline, and provide an extensive study of how object cues, motion cues and egocentric cues contribute to egocentric action recognition. Our method achieves a significant performance boost in three major benchmarks. We identify three key components for performance: motion compensation, object features over foreground regions and the use of an attention point to guide feature extraction. Our work provides insight into egocentric actions, and motivates us to continue to explore principled approaches for modeling these egocentric cues.

Acknowledgment: This research was supported by grant U54EB020404 awarded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative www.bd2k.nih.gov. This research is also partially supported by Intel Science of Technology Center for Pervasive Computing (ISTC-PC).

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, Apr. 2011. 1, 2

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014. 5

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006. 2

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, 2005. 2, 6

[6] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, pages 407–414, 2011. 1, 2, 5, 6, 7

[7] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*. 2012. 1, 2, 4, 5, 6, 7, 8

[8] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, 2008. 2

[9] A. Fathi and J. M. Rehg. Modeling actions through state changes. In *CVPR*, 2013. 2, 5, 6, 7, 8

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4

[11] A. Jain, A. Gupta, M. Rodriguez, and L. Davis. Representing videos using mid-level discriminative patches. In *CVPR*, 2013. 2

[12] M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013. 4

[13] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245–251. IEEE, 2013. 6

[14] T. Kanade and M. Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012. 2

[15] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 2, 5

[16] J. Kopf, M. F. Cohen, and R. Szeliski. First-person hyper-lapse videos. *ACM Transactions on Graphics (TOG)*, 33(4):78, 2014. 2

[17] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with fisher vectors for image categorization. In *ICCV*, 2011. 5

[18] M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25 - 26):3559 – 3565, 2001. 4

[19] I. Laptev. On space-time interest points. *International Journal of Compuer Vision*, 64(2-3):107–123, Sept. 2005. 2, 6

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

[21] C. Li and K. M. Kitani. Model recommendation with virtual probes for egocentric hand detection. In *ICCV*, 2013. 3

[22] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 2, 3, 5, 6, 8

[23] S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *ECCV*. 2012. 2

[24] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, 2013. 4

[25] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollar. Exploring weak stabilization for motion feature extraction. In *CVPR*, 2013. 4

[26] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*. 2010. 4

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

[28] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012. 2

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

[30] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. In *ICCV*, 2011. 3

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

[32] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009. 3

[33] E. H. Spriggs, F. De la Torre Frade, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *IEEE Workshop on Egocentric Vision, CVPR 2009*, 2009. 2

[34] Y. Tian, R. sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013. 2

[35] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, Nov 2008. 1, 2

[36] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013. 1, 2, 3, 4, 6

[37] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 4, 6, 7, 8