Activity Recognition in Egocentric Life-logging Videos

Sibo Song[1], Vijay Chandrasekhar[2], Ngai-Man Cheung[1], Sanath Narayan[3], Liyuan Li[2], Joo-Hwee Lim[2]

[1]Singapore University of Technology and Design
[2]Institute for Infocomm Research, Singapore,
[3]Indian Institute of Science, Bangalore, India

Abstract. With the increasing availability of wearable cameras, research on first-person view videos (egocentric videos) has received much attention recently. While some effort has been devoted to collecting various egocentric video datasets, there has not been a focused effort in assembling one that could capture the diversity and complexity of activities related to *life-logging*, which is expected to be an important application for egocentric videos. In this work, we first conduct a comprehensive survey of existing egocentric video datasets. We observe that existing datasets do not emphasize activities relevant to the life-logging scenario. We build an egocentric video dataset dubbed LENA (Life-logging EgoceNtric Activities)[1] which includes egocentric videos of 13 fine-grained activity categories, recorded under diverse situations and environments using the Google Glass. Activities in LENA can also be grouped into 5 top-level categories to meet various needs and multiple demands for activities analysis research. We evaluate state-of-the-art activity recognition using LENA in detail and also analyze the performance of popular descriptors in egocentric activity recognition.

1 Introduction

With the increasing availability of wearable devices such as Google Glass, Microsoft SenseCam, Samsung's Galaxy Gear, Autographer, MeCam and LifeLogger, there is a recent upsurge of interest in lifelogging. Lifelogging is an activity of recording and documenting some portions of one's life. Typically, the recording is automatic using wearable devices. Lifelogging can potentially lead to many interesting applications, ranging from lifestyle analysis, behavior analysis, health monitoring, to stimulation for memory rehabilitation for dementia patients.

Much advancement has been made in the hardware design for life-logging devices. Figure 1 shows some of the life-logging wearable devices. For example, Microsoft SenseCam [1] (commercially available as Vicon Revue) and Autographer [2] are wearable cameras that incorporate numerous advanced sensors (accelerometer, ambient light sensor, passive infrared) to determine the appropriate time to take a photo. Google Glass is an augmented glass that can be worn

[1] `http://people.sutd.edu.sg/~1000892/dataset`

(a) Autographer (b) Google Glass (c) Lifelogger (d) SenseCam

Fig. 1: A variety of life-logging wearable devices.

throughout the day to record first-person view video (egocentric video) at 720p HD resolution. However, algorithms for analyzing life-logging data, especially videos, need further improvement. For example, Vicon Revue [3], a wearable camera used by many serious life-loggers, supplies only primitive software that allows simple photo navigation and manual captioning / labeling of individual photos. Note that lifelogging usually generates a large amount of data. For instance, it is common for a lifelogging camera like Autographer to capture over 1000 photos a day. Likewise, several hours of lifelogging videos may be recorded by Google Glass daily. Therefore, manual processing of life-logging data could be extremely laborious. Automatic analysis of lifelog is crucial for many applications.

In the context of automatic analysis of visual lifelog, we describe in this paper an effort to advance the field with the design of an egocentric video database containing 13 categories of activity relevant to life-logging applications. These videos are recorded using Google Glass and capture the diversity and complexity of different human daily activities in first-person view as shown in Figure 2. The dataset, dubbed LENA (Life-logging EgoceNtric Activities), can be used by the vision research community to develop or evaluate new algorithms for life-logging applications.

Compared with previously-proposed egocentric video databases, one particular feature of LENA is its hierarchical grouping of activities: top-level categorization represents broad classification of daily human activity, while second-level categorization represents finer activity distinction. We use the proposed LENA database to evaluate the performance of state-of-the-art activity recognition algorithms. We also compare the performance of algorithms with activities at top-level and second-level. This reveals the performance difference of state-of-the-art to recognize coarse and fine level human activities in the context of first-person-view video.

The rest of this paper is organized as follows. In Section 2 we survey egocentric video datasets. Our life-logging videos dataset is presented in Section 3. Approach for evaluation is explained in Section 4 and experiment results are in Section 5.

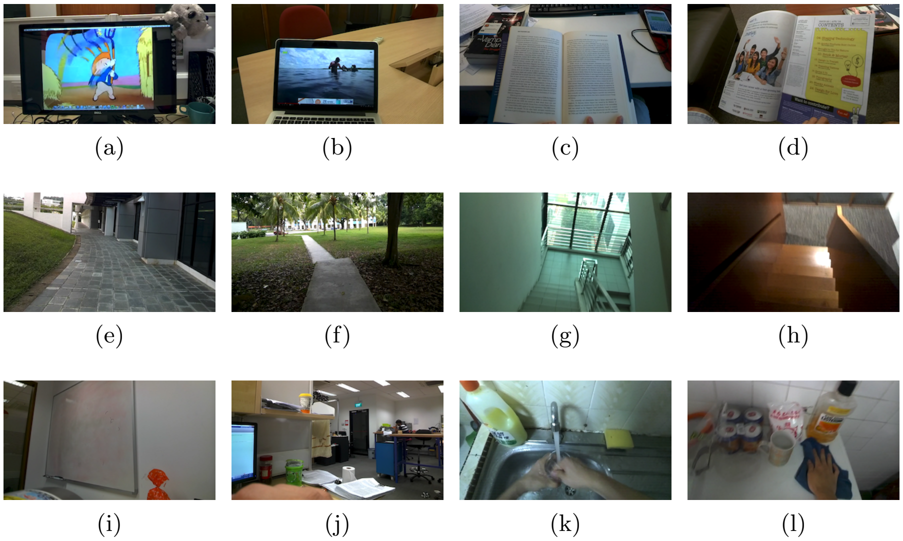


Fig. 2: Sample frames from our egocentric video database. The activities in frames are: (a),(b): watch videos, (c),(d): read, (e),(f): walk straight, (g),(h): walk up and down, (i),(j): drink, (k),(l): housework.

2 Survey of Datasets

We describe a survey of 12 existing egocentric video datasets in this section (see Table 1). We found that these datasets focus on different applications and have a large diversity of camera view-point, video quality, camera location, etc. For instance, dataset used in [4] focuses on object recognition, datasets of [5] and [6] consist of actions in kitchen, and datasets in [7] [8] [9] are mostly for social interaction. However, none of these focuses on comprehensive recording of activities related to life-logging.

2.1 Intel Egocentric Object Recognition Dataset

The Intel Egocentric Object Recognition Dataset (IEOR) [4] mainly focuses on recognition of handled objects using a wearable camera. It has ten video sequences from two human subjects manipulating 42 everyday object instances. However, the purpose of this dataset is to study object recognition in everyday life settings from an egocentric view instead of life-logging activities classification.

2.2 CMU-MMAC Dataset

The CMU Multi-Modal Activity Database (CMU-MMAC) database [10] contains multi-modal measures of the human activity of subjects performing the

Table 1: A list of existing egocentric video datasets.

	Dataset	No.	Activities	Comments
1	IEOR	10	Manipulate Objects	42 objects
2	CMU-MMAC	185	Cooking	5 recipes
3	GTEA	28	Food preparation	7 types of food
4(a)	UEC QUAD	1	Ego action like run, jump, etc.	11 simple action
4(b)	UEC PARK	1	Ego action like jog, twist, etc.	29 simple action
5	W31	31	Walking	From metro to work
6(a)	GTEA Gaze	17	Food preparation	30 kinds of food
6(b)	GTEA Gaze+	30	Food preparation	Around 100 actions
7	ADL	20	Food, hygiene and entertainment	18 indoor activities
8	UTokyo	5	Office activities	5 office tasks
9	FPSI	113	Social interaction	6 types of activities
10	UT Ego	4	Life-logging activities	11 events
11	JPL	57	Social interaction	7 types of activities
12	EGO-GROUP	10	Social interaction	4 different scenarios

(No.: Number of videos in each dataset.)

tasks involved in cooking and food preparation, for example, making brownies, pizza, sandwich, etc. The CMU-MMAC database was collected in Carnegie Mellon's Motion Capture Lab. Several modalities are recorded like video, audio, motion capture, etc. However, cooking alone is obviously not adequate to represent the diversity of life activities.

2.3 Georgia Tech Egocentric Activities Datasets

The Georgia Tech Egocentric Activities Dataset (GTEA) [5] consists of 7 types of daily activities, Hotdog, Sandwich, Instant Coffee, Peanut Butter Sandwich, Jam and Peanut Butter Sandwich, Sweet Tea, Coffee and Honey, Cheese Sandwich. The camera is mounted on a cap worn by the subject. The GTEA dataset focuses more on food preparation, so it is useful for recognizing objects and not so much for daily life activities classification.

2.4 UEC Datasets

The UEC Datasets [11] are actually two choreographed videos. The first video (QUAD) contains 11 different simple ego-actions, for instance, jump, run, stand, walk, stand, etc. The second video (PARK) is a 25 minutes workout video which contains 29 different ego-action categories such as pull-ups, jog, twist, etc. The actions are very fine-grained and more related to sports instead of life-logging activities.

2.5 W31 Datasets

The W31 Dataset [12] consists of 31 videos capturing the visual experience of a subject walking from a metro station to work. It consists of 7236 images in total. This dataset is collected to detect unplanned interactions with people or objects and does not contain other activities.

2.6 Georgia Tech Egocentric Activities Gaze(+) Datasets

The Georgia Tech Egocentric Activities Gaze(+) Datasets [6] consist of two datasets which contain gaze location information associated with egocentric videos.

The GTEA Gaze dataset is recorded by Tobii eye-tracking glasses. The Tobii system has an outward-facing camera that records at 30 fps rate and 480 × 640 pixel resolution. While, one problem of this dataset is that it only collects the meal preparation activity. There are 30 different kinds of food and objects in the videos. And the datasets includes 17 sequences of meal preparation activities performed by 14 different subjects. Each sequence takes about 4 minutes on average.

The GTEA Gaze+ dataset is collected to overcome some of the GTEA Gaze dataset's shortcomings. The resolution is 1280 × 960 and tasks are more organized. The number of tasks and objects used in this dataset are significantly bigger. It is collected from 10 subjects and each performs a set of 7 meal preparation activities. Gaze location at each frame is recorded. Each sequence takes around 10-15 minutes and contains around 100 different actions like pouring, cutting, mixing, turning on/off etc.

2.7 Activities of Daily Living Dataset

The Activities of Daily Living (ADL) Dataset [13] is a set of 1 million frames of dozens of people performing unscripted, everyday activities. The data is annotated with activities, object tracks, hand positions, and interaction events. The dataset is a 10 hours of video, amassed from 20 people in 20 different homes and recorded by chested-mounted cameras. The dataset is good for indoor activities classification. However, it does not involve any outdoor activities. From high level, it only has three categories: hygiene, food and entertainment.

2.8 UTokyo First-Person Activity Recognition Dataset

The UTokyo First-Person Activity Recognition Dataset [14] includes five tasks (reading a book, watching a video, copying text from screen to screen, writing sentences on paper and browsing the internet). Office activities of five subjects are recorded and each action is about two minutes. This dataset only records the office activities and five tasks.

2.9 Georgia Tech First-Person Social Interactions Dataset

The First-Person Social Interactions Dataset [9] contains day-long videos of eight subjects spending their day at Disney World Resort. The cameras are mounted on a cap worn by subjects. It is only a set of social interaction activities and recorded at Disney World Resort.

2.10 UT Egocentric Dataset

The University of Texas at Austin Egocentric (UT Ego) Dataset [15] contains four videos captured from head-mounted Looxcie cameras. Each video is about 3-5 hours long, captured in a natural, uncontrolled setting. The videos are recorded at 15 fps and 320×480 resolution. The videos capture a variety of activities such as eating, shopping, attending a lecture, driving, and cooking. While, the dataset is not easy and perfect for activities classification task because it needs to be cut into several clips and they may have different time duration.

2.11 JPL First-Person Interaction Dataset

The Jet Propulsion Laboratory (JPL) first-person Dataset [8] contains videos of interactions between humans and the observer. A GoPro2 camera is mounted on the head of our humanoid model and participants are required to interact with the humanoid by performing activities like shaking, hugging, petting, etc. Videos were recorded continuously during human activities and they are in 320×240 resolution with 30 fps. The limitation of this dataset is that camera is placed on the model instead of a real person. So some head motion and noise of egocentric view are removed. Another problem is that this dataset focuses on social interaction activities.

2.12 EGO-GROUP

The EGO-GROUP dataset [7] contains 10 videos, more than 2900 frames annotated with group compositions and 19 different subjects. There are four different scenarios in the dataset: laboratory, coffee break, party and outdoor. Similar to Georgia Tech First-Person Social Interactions Dataset and JPL First-Person Interaction Dataset, the video in EGO-GROUP dataset are collected in a more social way.

We have summarized the limitations for the popular egocentric video datasets when they are used for life-logging activities classification. In what follows, we describe our proposed life-logging dataset: LENA (Life-logging EgoceNtric Activities) to overcome these limitations.

3 Google Glass Life-logging Videos Dataset

Google Glass is a type of wearable technology with a camera and an optical display. It is relatively easier to collect egocentric videos using Google Glass than other wearable cameras. And for the existing dataset, the camera viewpoint varies considerably due to different kinds of cameras and also camera's positions (see Figure 1). The integrated camera in Google Glass makes it very similar to first-person view and also convenient to collect life-logging activities videos.

3.1 Dataset collection

The Google Glass Life-logging Dataset contains 13 distinct activities performed by 10 different subjects. And each subject record 2 clips for one activity. So each activity category has 20 clips. Each clip has a duration of exactly 30 seconds. The activity categories are: *watching videos*, *reading*, *using Internet*, *walking straight*, *walking back and forth*, *running*, *eating*, *walking up and down*, *talking on the phone*, *talking to people*, *writing*, *drinking* and *housework*. Sample frames in some of these categories are shown in Figure 2. Subjects have been instructed to perform the activities in a natural and unscripted way.

3.2 Video normalization

The original quality of Google Glass video is 1280 × 720 and the frame-rate is 30 fps. We also provide a version with dimension down-scaled to 430 × 240, to reduce the running time when needed. All the clips are processed with *ffmpeg* video library.

3.3 Characteristics

Firstly, LENA contains large variability in scenes and illumination. Videos are recorded both indoor and outdoor, with change in the illumination conditions (e.g., from morning to afternoon). Videos are collected from 10 different subjects and the activity videos are captured in an uncontrolled setting. There is also considerable variability for some activities like housework and reading. For housework recording, there are several different activities like washing cups and dishes, mopping, sweeping etc.

Secondly, we build a taxonomy based on the categories as shown in Figure 3. All 13 categories can be grouped into 5 top level types: *motion*, *social interaction*, *office work*, *food* and *housework*. This allows evaluation of new visual analysis algorithms against different levels of life-logging activity granularity. We believe that these characteristics can make LENA a very useful one for egocentric video research.

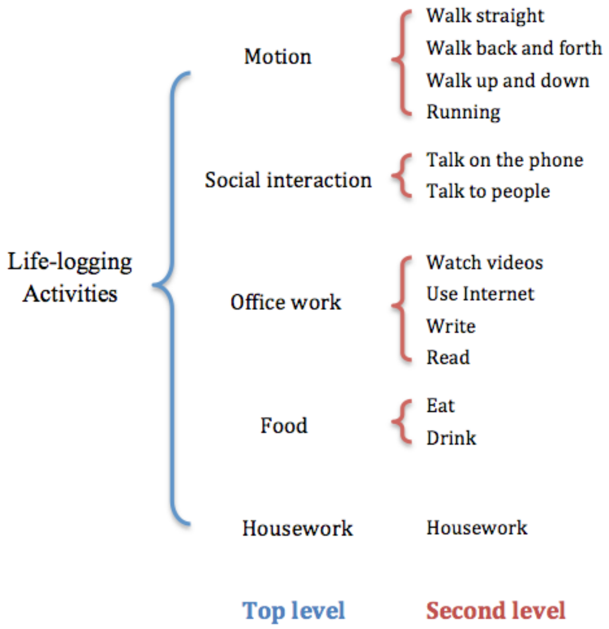


Fig. 3: Hierarchical grouping of life-logging activities

4 Activity Recognition Evaluation

We evaluate state-of-the-art trajectory-based activity recognition using our dataset. The dense trajectory approach we used has already been applied to third-person view action recognition in [16]. And object recognition is not performed for activities classification. We evaluate dense trajectories approach using LENA and trajectories are obtained by tracking densely sampled points using optical flow fields. Motion in frames and head movements are used in the recognition with this approach.

4.1 Trajectory features

Several kinds of descriptors (*HOG*, *HOF* and *MBH*) are computed for each trajectory. *Trajectory* is a concatenation of normalized displacement vectors. The other descriptors are computed in the space-time volume aligned with the trajectory. *HOG* (histograms of oriented gradients) focuses on static appearance information. And both *HOF* (histograms of optical flow) and *MBH* (motion boundary histograms) measure motion information. *HOF* directly quantizes the orientation of optical flow vectors. *MBH* splits the optical flow into horizontal and vertical components, and quantizes the derivatives of each component.

4.2 Fisher Vector encoding

We use Fisher vector to encode the trajectory features in the experiment. Fisher vector encodes both first and second order statistics between the video descriptors and a Gaussian Mixture Model (GMM). In our experiment, the number of Gaussians is set to $K = 256$ and randomly sample a subset of 256,000 features to estimate the GMM for building the codebook. The dimensions of the features are reduced by half using PCA. In particular, each video is represented by a $2DK$ dimensional Fisher vector, where D is the descriptor dimension after performing PCA. Finally, we apply power and *L2* normalization to Fisher vector.

The cost parameter $C = \mathit{100}$ is used for linear SVM and one-against-rest approach is utilized for multi-class classification. We use *libsvm* library [17] to implement the SVM algorithm.

5 Activity Recognition Experiment Results

In this section, we evaluate the performance of trajectory-based activity recognition with our Life-logging Egocentric Videos datasets and make a comparison using different descriptors. We apply the dense trajectory approach both on the 13 second-level categories and 5-top level categories.

5.1 Evaluation for fine-grained categories

The classification results on LENA using dense trajectory approach and Fisher vector encoding is reported in Table 2. The combined descriptors result in about 80% accuracy. Therefore, further improvement is desirable and this is the subject of future research. *HOF* and *MBH* descriptors result in better performance than *HOG*, as *HOG* only captures static information. Figure 4 shows the confusion matrix of combined descriptors on LENA.

Table 2: Comparison of all descriptors' accuracy for second-level (fine) categories.

Descriptor	*HOF*	*HOG*	*MBH*	*Trajectory*	Combined
Accuracy	76.38%	68.15%	78.04%	74.46%	81.12%

From Figure 4 we can see that performance of *walk up and down*, *read*, *use Internet* and *run* categories, are superior. While, for *talk on the phone*, *write*, *drink* and *eat* the performance is around 50%. Note that for *talk on the phone*, *drink* and *eat*, there are hardly any objects like phone, cup and snacks in the scene, as the videos are recorded in first-person view. Thus it is more difficult for recognition, especially using the *HOG* descriptor.

We also make a comparison among different descriptors. Figure 5 shows the performance of the 4 different descriptors on second level categories. Overall,

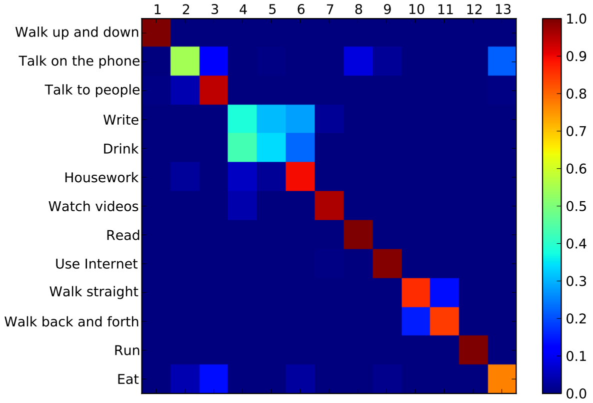


Fig. 4: Confusion matrix of combined descriptors for second-level (fine) categories.

HOF, *MBH* and *Trajectory* have similar results. While, performance on *HOG* descriptor is about 10% less than the other three. The *HOG* descriptor shows the worst performance which is 68.15%, especially for *write* and *drink* categories.

For combined descriptor, the accuracy of *drink* is 34% which is the lowest. Videos of *drink* action are often mis-classified into *write* and *housework*. The *MBH* descriptor alone obtains the best performance with LENA.

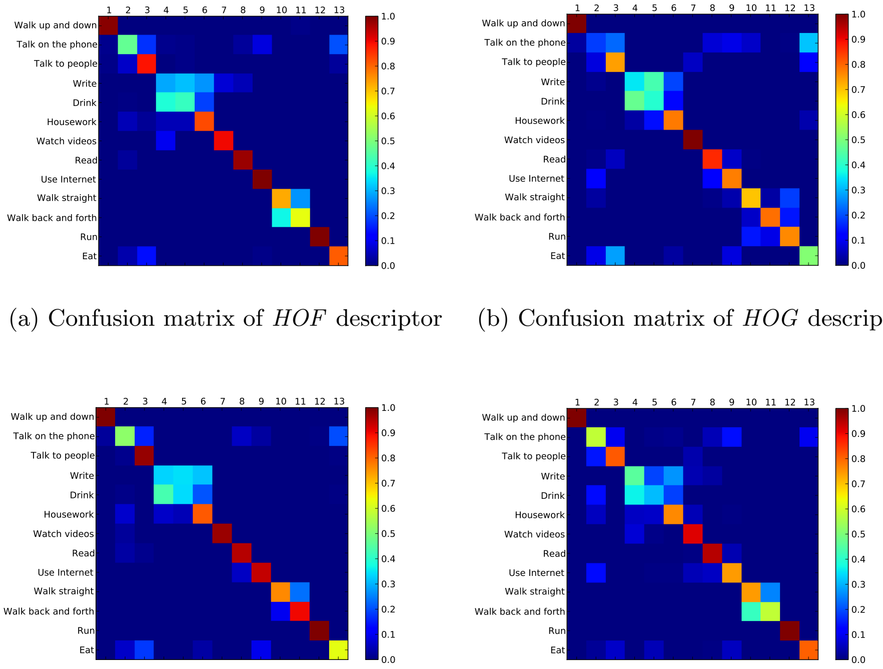
5.2 Evaluation for top-level (coarse) categories

One feature of LENA is that we have a hierarchical grouping of life-logging activities. Then we evaluate top-level categories using dense trajectory algorithm. The dataset we used is the same as the second level categories. We only changed the ground-truth of the dataset and trained one-against-rest SVM classifiers on the training set.

Table 3: Comparison of all descriptors' accuracy for top-level categories.

Descriptor	*HOF*	*HOG*	*MBH*	*Trajectory*	Combined
Accuracy	84.00%	77.42%	82.46%	84.50%	84.23%

In Table 3, the accuracy of every descriptor is much higher than results in Table 2. The performance has improved by about 10% for *HOF*, *HOG* and *Trajectory*. However, *HOG* still performs worst among all the descriptors. Figure 6 shows the comparison between top level and second level categories. The



(a) Confusion matrix of *HOF* descriptor (b) Confusion matrix of *HOG* descriptor

(c) Confusion matrix of *MBH* descriptor (d) Confusion matrix of *Trajectory* descriptor

Fig. 5: Confusion matrix of individual descriptor.

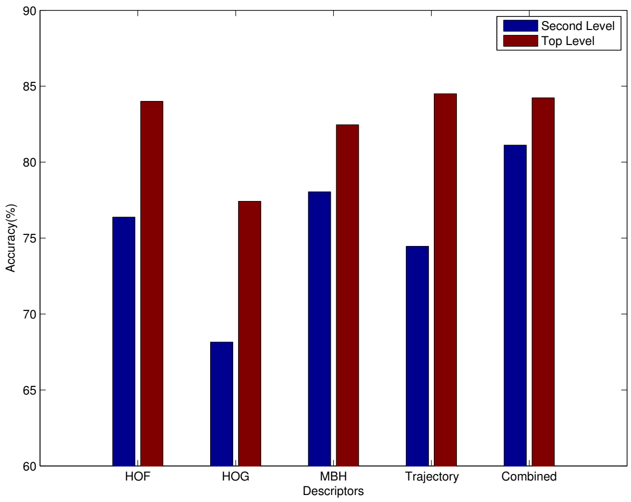


Fig. 6: Comparison between classification accuracy of top-level and second-level categories

accuracy of *HOG* descriptor has been improved most which is 8.67%. One reason could be that the difference among the top-level categories are much more salient than second level categories. We also observe that *HOF*, *Trajectory* and combined descriptor have almost the same results. *Trajectory* even has a slight higher accuracy than combined descriptor.

We also construct confusion matrices for top level categories classification in Figure 7. Interestingly, *motion* has nearly 100% accuracy. It is because the difference between *motion* and the other four activities are much more obvious. The most easily mistaken pair is *food* and *office work*. While, even the *food* activity which performs the worst has an accuracy of more than 50%. Overall, the trajectory algorithm does well on the top-level categories.

Confusion matrices of individual descriptor are also presented in Figure 8. From the figure we can see that *motion* has a high accuracy on every kind of descriptor. One the contrary, category *food* which contains *eat* and *drink* actions has the poorest performance. The low mis-classification rate of *motion* category is quite understandable, as in *motion* recording, there are always similar head motion and body movement. While, in *office work* and *food* recording, subjects usually do not move head and body sharply as they do in *run* and *walk straight* actions recording.

6 Conclusions

We summarize our contributions as follows:

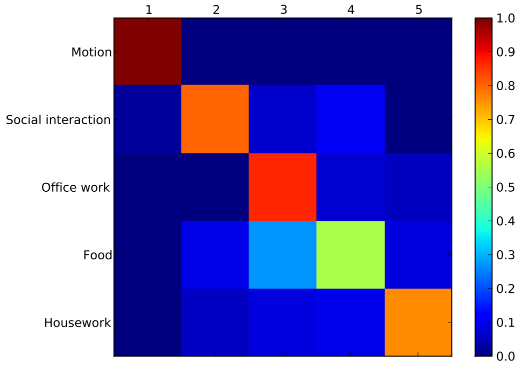
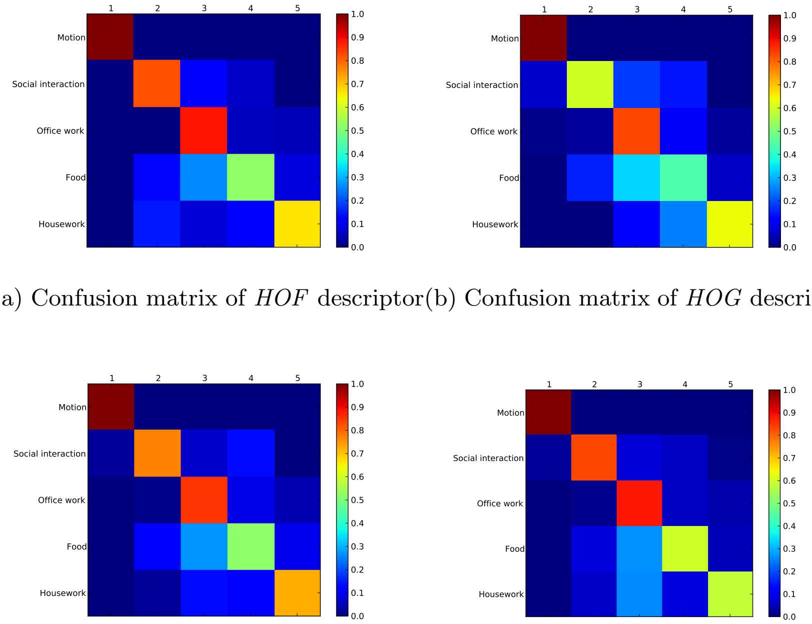


Fig. 7: Confusion matrix of combined descriptors for top-level categories.



(a) Confusion matrix of *HOF* descriptor

(b) Confusion matrix of *HOG* descriptor

(c) Confusion matrix of *MBH* descriptor

(d) Confusion matrix of *Trajectory* descriptor

Fig. 8: Confusion matrix of individual descriptor.

- We surveyed and discussed popular egocentric video datasets. Analysis of 12 existing datasets suggests that although they can address a variety of applications, their utility for daily life activities classification and analysis is inadequate in many ways.
- We presented LENA, an egocentric video dataset of life-logging activities. Recorded by Google Glass, the egocentric videos contain a variety of scenes and personal styles, capturing the diversity and complexity of life activities. The activities are organized into two levels of categorization, enabling research on coarse and fine-grained life activity analysis using a single dataset.
- We performed detailed evaluation of state-of-the-art activity recognition using LENA. We found that with dense trajectory approach the accuracy of activity recognition is around 80%. Thus, further research is needed on life-logging activity recognition. Furthermore, thanks to the two-level categorization structure, we were able to reveal the performance gap of state-of-the-art in recognizing activity at two different granularities. We also analyzed the performance of state-of-the-art descriptors (based on gradient, optical flow, motion boundary) in egocentric activity recognition.

We believe that LENA could be valuable for the research on egocentric activities recognition and classification, especially for life activities. Future work involves understanding the importance of global motion in egocentric activity recognition and using graph-theoretic approach for life activity recognition. In addition, while our focus in this paper is activity recognition, many other visual data analysis tasks such as automatic discovery of activity topics could be investigated using our proposed dataset.

References

1. : Microsoft research sensecam. http://research.microsoft.com/sensecam (2013)
2. : Autographer. http://www.autographer.com/ (2013)
3. : Vicon revue wearable cameras. http://viconrevue.com/ (2013)
4. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 3137–3144
5. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On, IEEE (2011) 3281–3288
6. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: Computer Vision–ECCV 2012. Springer (2012) 314–327
7. Alletto, S., Serra, G., Calderara, S., Solera, F., Cucchiara, R.: From ego to nos-vision: Detecting social relationships in first-person views. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2014) 580–585
8. Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 2730–2737

9. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 1226–1233
10. CMU: Multi-modal activity database. `http://kitchen.cs.cmu.edu/index.php` (2010)
11. Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3241–3248
12. Aghazadeh, O., Sullivan, J., Carlsson, S.: Novelty detection from an ego-centric perspective. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3297–3304
13. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2847–2854
14. Ogaki, K., Kitani, K.M., Sugano, Y., Sato, Y.: Coupling eye-motion and ego-motion features for first-person activity recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, IEEE (2012) 1–7
15. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: CVPR. Volume 1. (2012) 3–2
16. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103** (2013) 60–79
17. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27 Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.