Coupling Eye-Motion and Ego-Motion features for First-Person Activity Recognition

Keisuke Ogaki
The University of Tokyo
Tokyo, Japan
ogaki@iis.u-tokyo.ac.jp

Kris M. Kitani
Carnegie Mellon University
Pittuburgh, USA
kkitani@cs.cmu.edu

Yusuke Sugano, Yoichi Sato
The University of Tokyo
Tokyo, Japan
sugano, ysato@iis.u-tokyo.ac.jp

Abstract

We focus on the use of first-person eye movement and ego-motion as a means of understanding and recognizing indoor activities from an "inside-out" camera system. We show that when eye movement captured by an inside looking camera is used in tandem with ego-motion features extracted from an outside looking camera, the classification accuracy of first-person actions can be improved. We also present a dataset of over two hours of realistic indoor desktop actions, including both eye tracking information and a high quality outside camera video. We run experiments and show that our joint feature is effective and robust over multiple users.

1. Introduction

It has been shown in recent work, that various modalities of features can be used to effectively encode various ego-actions. In this paper, we focus on two types of modalities: (1) eye motion and (2) ego-motion, and show how the combination of these two modalities leads to better performance in recognizing *ego-actions*, i.e., actions captured in egocentric videos. Recent work has examined the usefulness of eye-motion [1] and ego-motion [6] *in isolation*, and have shown that they may be used successfully for recognizing different ego-actions. Building on the success of previous work, we show that by concatenating different feature modalities, we are able to improve classification performance.

It is known that a person's eye movement is a rich source of information for understanding a person's actions [7]. A sequence of eye motions, commonly measured by an *inside* looking camera, can reveal a person's focus of attention and even reveal our internal mental state. As such, eye movement analysis has been used widely in both clinical research, empirical psychology and neural science. It is notable that the use of eye-tracking in such studies has been used predominantly for *post-facto* analysis, *e.g.*, understanding how the eye moves for tasks such as reading, drawing or doing a jigsaw puzzle [7].

In contrast, we are interested in using eye-motion has a means of recognizing and classifying actions. Some previous work share a similar motivation with our study. Doshi *et al*. [3] used head pose and putative gaze locations (straight, right/left mirror, rear mirror) to predict lane changes. Courtemanche *et al*. [2] also used eye movement between predefined areas of interest (AOI) to recognize display interactions. While predefined spatial quantization is plausible for constrained tasks, more applications could benefit from a more general feature.

Recently, Bulling *et al*. [1] has shown that eye motion is a powerful feature for representing various first-person actions. They presented a saccade sequence quantization methodology that discretized eye motion into a symbolic sequence and extracted basic features from n-gram statistics. It was shown that eye motion is particularly well suited for ego-actions that require finer motor skills (*i.e*. office tasks). Figure 1 shows examples of eye movement trajectories for several office tasks.

On the other hand, vision-based techniques for understanding and recognizing ego-actions have focused largely on the use of *outside* looking cameras to capture information about the visual world, such as hand gestures, objects of interactions and ego-motion [12, 8, 5, 13, 10]. Recent work has also shown that a user's focus of attention on a rough macro-scale (*i.e*. head pose and detected objects or faces) can be used to model social interactions [4]. Kitani *et al*. [6] demonstrated in their recent work that global ego-motion has been shown to be a successful descriptor for human actions in sports.

While these macro motion-based approaches are well suited for dynamic body motion, there are also many tasks, such as office activities, which cannot be fully characterized and recognized by ego-motion alone.

Hence, it can be naturally seen that these two information sources are complementary. The *inside* looking camera

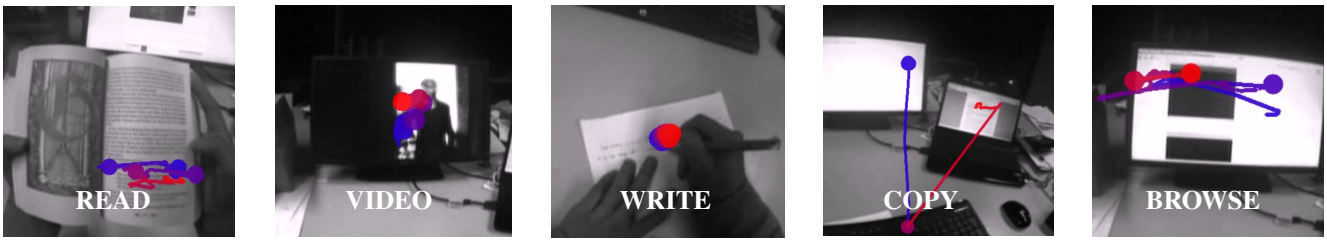


Figure 1. Eye movement trajectories for office tasks. Color of the trajectory represents time. Red is the current time and darker colors (blue) are past time steps.

tells us micro-level eye motion information, while the information about the outside visual world cannot be directly inferred from gaze data. The *outside* looking camera, conversely, tells us global ego-motion information, while it is very difficult to infer the internal state of the person by only using the egocentric visual information.

In this work we explore the joint use of eye movement and ego-motion, as an effective combination of feature modalities for encoding human activity. In particular, we show that by simply combining features we achieve an increase in classification performance. This indicates that using the an optimal combination of feature modalities can help to improve overall performance and the representative power of first-person ego-action analysis frameworks.

We summarize our contributions as follows:

- We show that the joint use of eye motion and ego-motion yields improvements in action recognition over those modalities used in isolation
- We present a labeled dataset of eye motion using an inside-out camera system for basic desk work activities with multiple subjects

2. Extracting inside-out motion features

Our goal is to model and detect primitive ego-action categories using first-person sensing. Recognizing primitive actions are important for understanding human activities because they can be used as building blocks to understand more complex high-level activities. While previous work has shown that characterizing eye motion is an important feature to use for understanding primitive ego-actions, we show that it is also important to characterize global ego-motion to better represent first-person actions. Here, we describe our method for extracting both eye-motion and ego-motion from our inside-out camera system.

We extract two types of sequential motion primitives s from eye motion and ego-motion. Eye motion can be roughly divided into two types of motion: *fixation*, where the eye focuses on a particular location and remains stationary, and the *saccade*, where the eye moves rapidly to scan the scene. With our inside camera, we detect saccade events

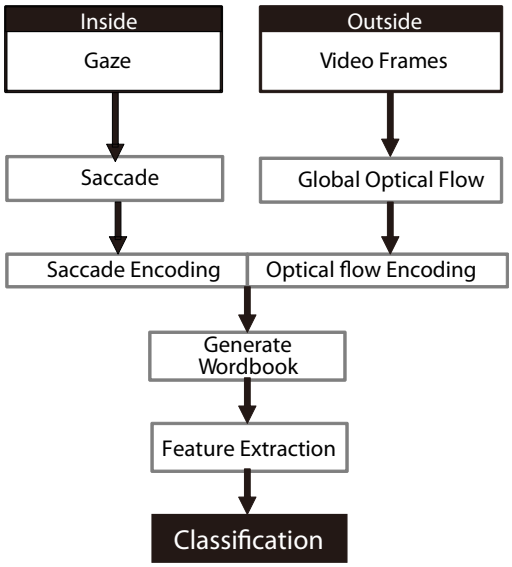


Figure 2. Flow chart for motion word features

to build a dictionary of sequential saccade primitives. Using the outside camera, we extract sequential motion primitives from ego-motion. What we call ego-motion here included both head motion and global body motion. For example, when we run, the outside camera moves up and down periodically. Figure 2 shows the overall system architecture for our prototype system. Our proposed method is constructed with two main processes: (1) motion-wordbook extraction and (2) saccade-wordbook extraction.

2.1. Extracting motion primitives

Here we describe how a motion sequence is quantized into a symbol string and an overview of this process is shown in Figure 3. Using the inside camera, we obtain the gaze coordinates $E = \{e_{x,t}, e_{y,t}\}_{t=1}^{T_E}$. Since the raw eye-tracking data is very noisy (due to blinking and tracking errors), we first smooth E with a median filter. Then following [1], use the continuous wavelet transform for saccade detection (CWT-SD) to compute the continuous 1-D wavelet values $C = \{C_x, C_y\}_{t=1}^{T_E}$ at a fixed scale α using a

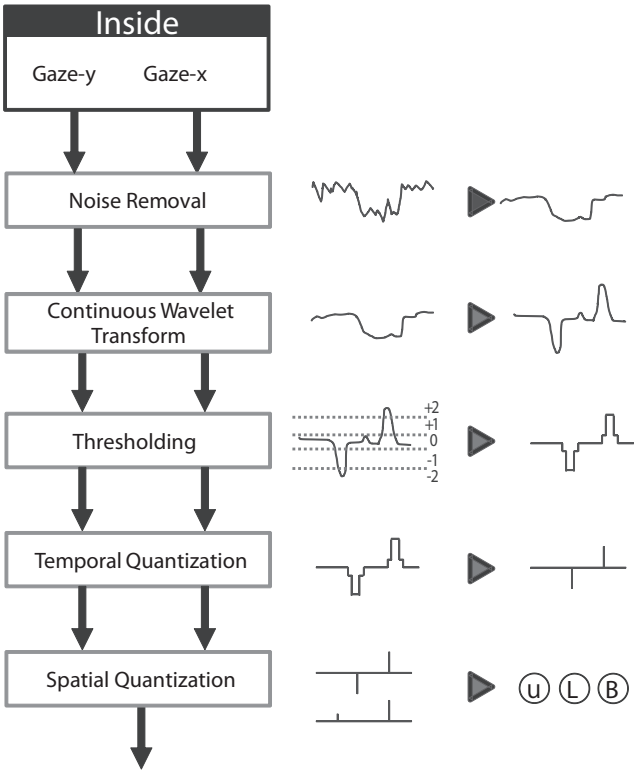


Figure 3. Flow chart for extracting saccade symbols.

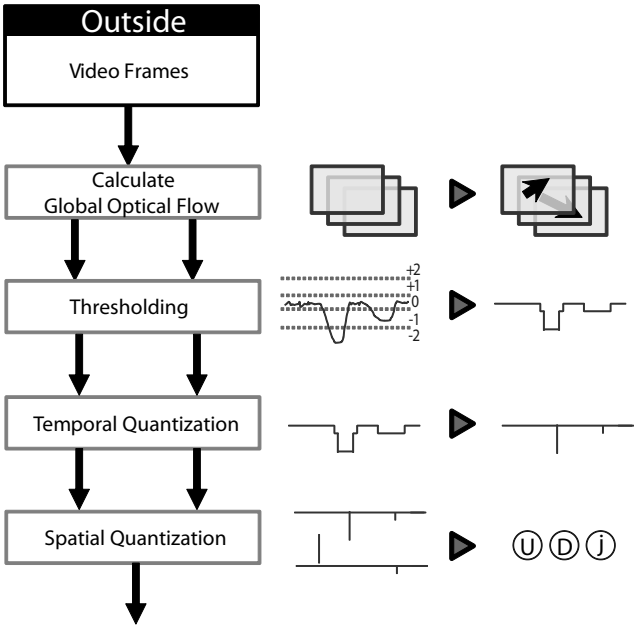


Figure 4. Flow chart for extracting ego-motion symbols.

Haar mother wavelet for both x and y gaze coordinates,

$$c_{x,b} = \frac{1}{\sqrt{\alpha}} \int \psi(\frac{t-b}{\alpha}) e_{x,t} dt, \quad (1)$$

$$\psi(x) = \begin{cases} 1 & (0 \le x < \frac{1}{2}) \\ -1 & (\frac{1}{2} \le x < 1) \end{cases} \quad (2)$$

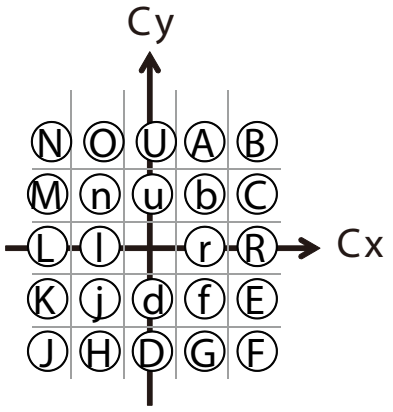


Figure 5. Motion quantization chart [1]. Lower case symbols represent small motion and uppercase symbols represent large motion.

where α is a scale parameter that depends on the sampling rate of the sensor (20 in our experiments, corresponding to about 80 ms). The parameter b is the time index. The values $c_{y,b}$ are calculate in the same manner. This process gives us a smoothed signal.

Next we quantize the motion sequence using the stepwise magnitude and direction. Two thresholds τ_{small} and τ_{large} are used to quantize the smoothed motion sequence C,

$$\hat{c}_{x,b} = \begin{cases} 2 & (\tau_{large} \le c_{x,b}) \\ 1 & (\tau_{small} < c_{x,b} \le \tau_{large}) \\ 0 & (-\tau_{small} \ge c_{x,b} \le \tau_{small}) \\ -1 & (-\tau_{large} < c_{x,b} \le -\tau_{small}) \\ -2 & (c_{x,b} \le -\tau_{large}) \end{cases} . \quad (3)$$

This quantization generates a discrete quantization over the joint space of magnitude and direction, as shown in Figure 5.

In a similar manner, we use the outside camera to obtain a sequence of global optical flow values $O = \{o_{x,t}, o_{y,t}\}_{t=1}^{T_O}$ and transform it into a symbols sequence. An overview of ego-motion quantization is shown in Figure 4. The global optical flow is computed by tracking corner points over consecutive frames and taking the mean flow in the x and y directions. We use the same quantization scheme (but with different magnitude thresholds) to generate a symbol string over ego-motion.

2.2. Statistical feature extraction

In our first step we quantized sequential motion primitives to generate a compact representation of motion in the form of a symbolic lexicon. In this second step, we extract statistical features over the lexicon as our motion descriptor. Using a sliding temporal window of size w centered at t, the symbol string $S_t = \{s_{t-w/2}, \ldots, s_t, \ldots, s_{t+w/2}\}$, is used to build n-gram dictionary, where s_t is a motion word. Then

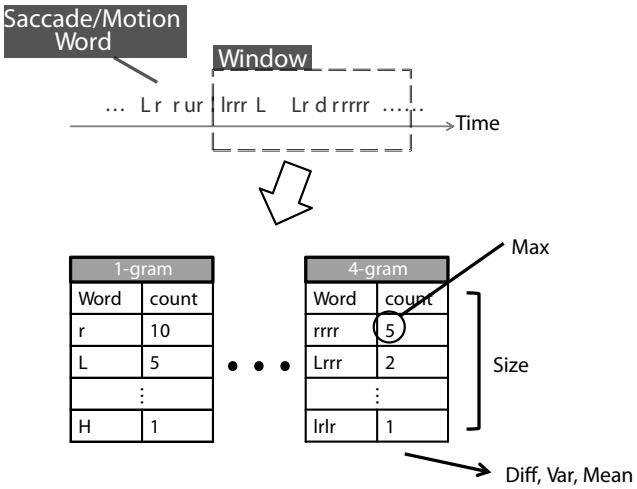


Figure 6. Extracting features from n-gram statistics. Given a quantized motion sequence, a set of statistical features (max, size, range, variance, average) are computed over the set of all n-grams. The same process is used for both the saccade and motion word features.

for this sequence of symbols S, a histogram over the dictionary is computed. Once this histogram has been generated, a feature vector $\mathbf{f}_t$ is computed by aggregating statistics over the n-gram histogram. Figure 6 shows the steps involved in computing the feature vector from a sequence of motion.

We calculate five feature for each sub-sequence of length n: (1) max-count, (2) average-count, (3) wordbook size (number of unique n-grams), (4) variance of counts and (5) range (difference between maximum and minimum count) are extracted from n-gram histogram. In our experiments we set the value of n to be four and this yields a 20 dimensional feature vector.

The temporal window size for the saccade feature is 3600 frames (roughly 15 seconds) and temporal window size for motion word is 900 frames (roughly 30 seconds). Optimal window sizes were determined so that classification performance is maximized.

3. Experiments

To evaluate our proposed method we perform experiments on a set of common daily office tasks. Our dataset includes the same five tasks (reading a book, watching a video, copying text from screen to screen, writing sentences on paper and browsing the internet) used in [1]. We recorded the actions of five subjects, who were instructed to perform each task for about two minutes. Tasks were performed in the following order: *read*, *video*, *write*, *copy* and *browse*. 30 seconds intervals of void class were placed between target tasks. To provide a natural experimental setting, the void class contains a wide variety of actions such as conversing, singing and random head motions.

The sequence of five actions was repeated twice to induce interclass variance. To assess robustness against scene changes, between the two cycles, the book used for *read*, video contents for watching a *video* task and the location of the physical workspace are changed to add more variations. The dataset consists of over two hours of data, where the video from each subject is a continuous $25 \sim 30$ minute video. Keyframes from the dataset are shown in Figure 1. We use a linear kernel support vector machine (SVM) as our classifier. In particular, we train a one-versus-all SVM to evaluate per class performance and a multi-class SVM to evaluate relative performance between classes. We compute the average precision as our global performance metric. For each subject, we produced two experiment sets. In the first set, the first cycle is used as training data and the second cycle is used as testing data. Adversely, the second cycle is used as training data and the seconde cycle is used as testing data in the second set. As a result we produced 60 dataset in total.

We used a commercial eye-tracking device (EMR-9 from NAC imaging technology) as our inside looking camera. Instead of the low-resolution view camera of the eye tracking device, an additional high-resolution camera (GoPro HERO 2 HD) was used as the outside looking camera.These two devices are synchronized by temporally aligning global optical flow vectors $O_H = \{\mathbf{o}_t^{(H)}\}_{t=1}^{T_H}$ of the GoPro camera and $O_E = \{\mathbf{o}_t^{(E)}\}_{t=1}^{T_E}$ of the EMR camera.Under an assumption that these two cameras are facing the same direction, the time offset is estimated so that a mean dot product between shifted frames becomes maximized.

3.1. Baseline methods

We performed four baseline experiments to measure the isolated performance of different feature modalities. The first baseline experiment uses the motion histogram (MoHist) proposed in [6]. This feature encodes instantaneous motion and period motion using Fourier analysis. The second baseline experiment uses the saccade word (SaWord) proposed in [1]. This feature encodes counts of partial trajectories (four frames in our experiments) of eye motion. The third baseline experiment uses a novel feature, called motion words (MoWord), in which uses the same quantization process as saccade words, but applied it to the average optical flow generated by ego-motion. This feature operated on a smaller temporal window compared to motion histograms [6] but takes into account higher order sequences of motion. The fourth baseline experiment uses the GIST feature [9], following [11], which captures the global visual context encountered while performing actions. We used 8 oriented filters at 4 different scales, over a 4×4 spatial grid. We also perform additional experiments with different concatenations of features to show how the combinations of different modalities affect performance.

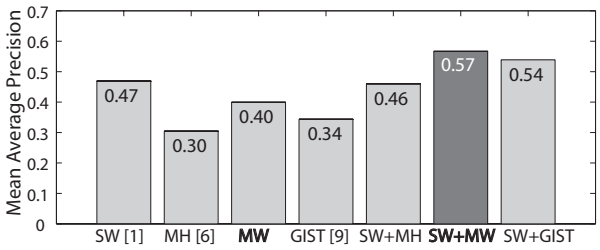


Figure 7. Mean average precision for each feature type. Proposed Method MoWord+SaWord (MW+SW) performs best of all features. SW, MH, MW mean SaWord, MoHist and MoWord

	SW	MH	MW	GIST
VOID	0.61	0.38	0.56	0.34
READ	0.53	0.35	0.29	0.43
VIDEO	0.22	0.18	0.24	0.25
WRITE	0.54	0.33	0.25	0.29
COPY	0.49	0.42	0.79	0.39
BROWSE	0.43	0.16	0.26	0.36

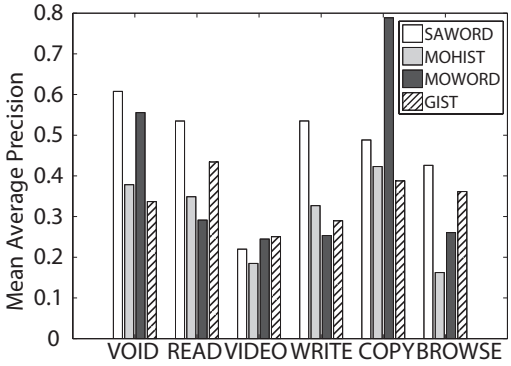


Figure 8. Baseline: Performance per action category. Mean average precision for each action for each feature type.

3.2. Comparison among all features

To evaluate the classification performance, we use the one-versus-all Support Vector Regression (SVR). We trained SVM with the labeled data in which frames in the target task are labeled as positive samples, while others are labeled as negative samples. As a performance measure, we calculated average precision for each unique combination of subject, task and experimental set. Mean average precision discussed below is arithmetic mean over subjects or tasks or experimental sets. Figure 7 shows average classification (mean average precision) performances of all features including baseline methods. Mean average precision is calculated over 60 experimental sets. It can be seen that our proposed MoWord+SaWord performs highest average classification performance.

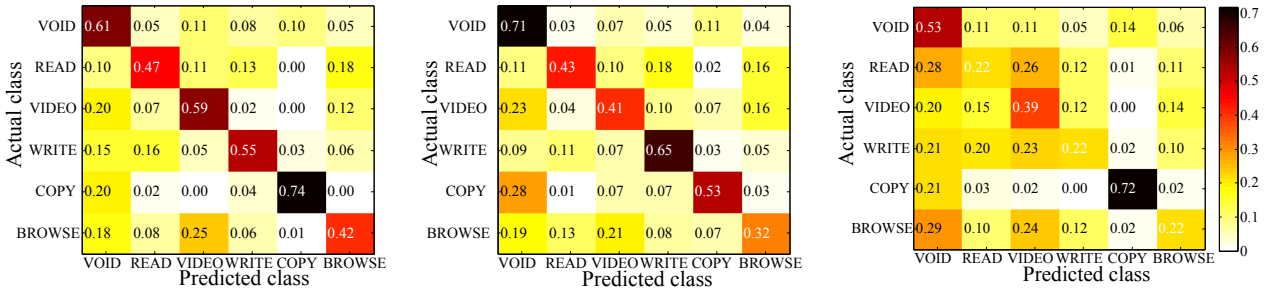
Among independent features, we observe that the saccade word feature based on eye movement has the highest classification performance with a mean average precision of 0.47. It is interesting to note, that while the saccade word feature has no access to the visual context, it is able to discriminate between various tasks better that ego-motion alone. Our proposed motion word feature performs second best, which indicates that ego-motion is also a discriminative feature. The motion histogram performs worse, which is expected since the feature was originally designed for large scale ego-motion and actions with periodicity.

Figure 8 shows the performance for each action category. The saccade word feature does particularly well on reading and writing tasks which have distinct eye movements due to the detailed nature of the task (*i.e*. eye scanning lines). The motion word feature outperforms all other features for the *copy* task (copying from screen to screen) due to the ego-motion induced by turning the head from screen to screen. As expected the motion histogram performs worse on detailed tasks like reading, writing, and watching a video, because the head is virtually still for much of the task.

Figure 9 shows the classification performance across bimodal combinations of features. Each mean average precision is calculated over 10 experimental sets. The average precision for each ego-action class if computed with a 1-vs-all SVM. Here we observe that the saccade word, when used with motion words (MoWord+SaWord) yields the highest average classification performance with an mean average precision of 0.57. In particular, we see a large improvement in performance for the *video* and *copy* actions. Since watching a video is defined over a joint feature space where the head is still and the eyes move over the screen region, the joint feature representation does a better job of encoding the action. Likewise, the action of *copy* is defined by large head motion followed by a specific eye motion pattern (scanning text) and is better described in the joint space. We also see a slight drop in performance for the *void* and *write* actions. Although the difference is small, it is possible that certain actions are defined predominantly by a single feature.

3.3. Multi-class classification experiments

To evaluate the cross-category performance, we used a multi-class SVM to compute a calibrated classifier response for each ego-action category. A visualization of the confusion matrix is given in Figure 10 to understand the nature of classification errors. Looking at Figure 10 (a) we can see that the *void* action and *copy* action have a high recall rate. In contrast, the lowest performing action is *browse* which is often confused as *void* (18%) or *video* (25%). The misclassification of *browse* as *video* is understandable, as both actions consist of looking at a screen with a relatively stable head position and they share similar eye motions. Similarly,



(a) MoWord+SaWord (b) SaWord (c) MoWord

Figure 10. Confusion Matrix

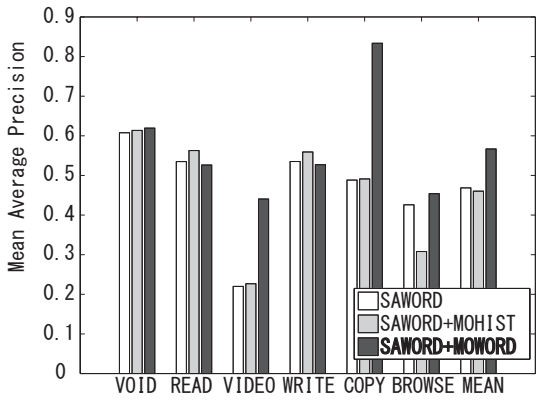


Figure 9. Mean Average Precision over subjects per action for bi-modal features. SAWORD+MOWORD is our proposed method. MEAN is calculated over all 6 tasks.

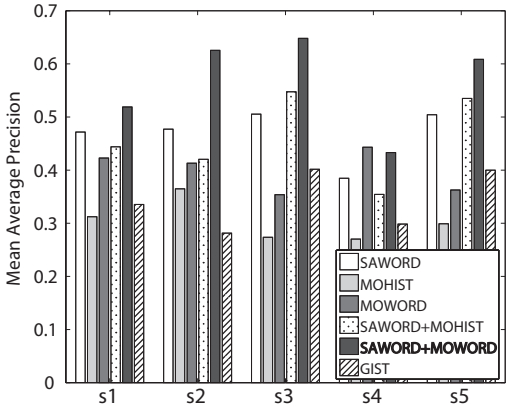


Figure 11. Average Precision *per subject* for various features. SAWORD+MOWORD is our proposed method.

read is confused as *browse* (18%) and *write* (13%). Again, this makes sense since all action involved the scanning of text. This may indicate that adding visual features (detection of a pen, hands or a screen) may help to disambiguate these actions.

3.4. Cross-subject performance

Now we show results performed across multiple subjects to show how the performance of features varies between subjects. Figure 11 shows the classification performance across different subject for our experiments. Notice that the relative performance between different feature types is the similar between subjects, with the exception of subject 4. We can see that the SaWord and SaWord+MoWord perform worse for subject 4 compared to the other subjects. This drop in performance was due to low-quality eye tracking for this user (*i.e.* subjects eyes were particularly hard to track). This result highlights that fact that classification performance is integrally linked to the low-level eye tracking performance. For most of the subjects, it can be clearly seen that our proposed method improves classification accuracy than existing methods.

4. Conclusion

We have presented an analysis of different feature modalities for ego-action classification. While previous works has focused on the independent use of eye motion or ego-motion, we have shown that the combination of eye motion features and ego-motion provides the best representation of indoor office work tasks. In our experiments, we have also shown that our joint eye motion and ego-motion feature is robust across multiple subjects and can be used to reliably detect ego-actions across different users. We believe that this exploration of multimodal features for ego-action representation is important in understanding the feature space covered by first-person actions and will serve as an impetus for future research along these lines.

References

[1] A. Bulling, J. Ward, H. Gellersen, and G. Troster. Eye movement analysis for activity recognition using electrooculography. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(4):741–751, 2011. 1, 2, 3, 4

[2] F. Courtemanche, E. Aďmeur, A. Dufresne, M. Najjar, and F. Mpondo. Activity recognition using eye-gaze movements and traditional interactions. *Interacting with Computers*, 23(3):202 – 213, 2011. 1

[3] A. Doshi and M. M. Trivedi. On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *IEEE Trans. on Intelligent Transportation Systems*, 10(3):453–462, 2009. 1

[4] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1

[5] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1

[6] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1, 4

[7] M. F. Land and B. W. Tatler. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press, USA, 2009. 1

[8] W. Mayol and D. W. Murray. Wearable hand activity recognition for event summarization. In *International Symposium on Wearable Computers*, 2005. 1

[9] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 4

[10] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1

[11] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *Proc. Workshop on Egocentric Vision, CVPR Workshops*, 2009. 4

[12] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998. 1

[13] L. Sun, U. Klank, and M. Beetz. Eyewatchme3d hand and object tracking for inside out activity analysis. In *Proc. Workshop on Egocentric Vision, CVPR Workshops*, 2009. 1