Desktop Action Recognition From First-Person Point-of-View

Minjie Cai, Feng Lu, *Member, IEEE*, and Yue Gao, *Senior Member, IEEE*

***Abstract*—Desktop action recognition from first-person view (egocentric) video is an important task due to its omnipresence in our daily life, and the ideal first-person viewing perspective for observing hand-object interactions. However, no previous research efforts have been dedicated on the benchmark of the task. In this paper, we first release a dataset of daily desktop actions recorded with a wearable camera and publish it as a benchmark for desktop action recognition. Regular desktop activities of six participants were recorded in egocentric video with a wide-angle head-mounted camera. In particular, we focus on five common desktop actions in which hands are involved. We provide original video data, action annotations at frame-level, and hand masks at pixel-level. We also propose a feature representation for the characterization of different desktop actions based on the spatial and temporal information of hands. In experiments, we illustrate the statistical information about the dataset, and evaluate the action recognition performance of different features as a baseline. The proposed method achieves promising performance for five action classes.**

***Index Terms*—Dataset, desktop action recognition, egocentric video, hand motion.**

I. Introduction

UNDERSTANDING human activities of daily living (ADL) [1] is an important research topic in computer vision. It has drawn much attention from both academia and industry in recent years. The ability to recognize human actions from video enables several important applications, including video surveillance, human computer interaction, and cognitive study. Most of the existing work focus on analyzing human body movements from third-person view video captured by fixed cameras. However, due to the limited field of view from fixed cameras, difficulties still exist in capturing all body parts involved in the action, especially those small parts, such as hands and fingers, without problems related to occlusion and resolution. This makes tasks, such as recognizing daily actions from the third-person view video very challenging, where complex hand-object manipulations are common and informative.

In contrast to conventional third-person capturing, egocentric video is recorded by a wearable camera from a first-person perspective [2], [3]. Egocentric video has become popular now due to the development in hardware, especially wearable cameras. By using wearable cameras, we can record our daily activities with great convenience. The captured video can either be reviewed for fun or used as input to vision-based systems for analysis. Especially, daily action recognition, which is challenging for conventional third-person methods, can be well handled by using egocentric approach. First, daily activities can be easily recorded without deploying fixed cameras in a tedious and time-consuming process. Therefore, egocentric datasets can include actions in more realistic scenarios with free body movements and complex object manipulation, compared to those standard datasets recorded by fixed cameras, such as KTH [4], where human body movements and patterns are greatly limited. Second, in the case of hand-object manipulation, both hands and objects are ensured to be visible with a sufficient resolution from the egocentric perspective. Since hand-object manipulation plays an essential role in daily activities, analysis using egocentric video can better utilize such information.

In this paper, we focus on the important task of recognizing desktop actions that commonly occur in our daily life. Desktop actions are defined as a set of actions performed by humans sitting around a desktop. We believe that a vision system with the ability to recognizing different daily desktop actions is important and can be applied to various applications, such as daily behavior analysis and study of working efficiency. However, there is no existing public dataset for egocentric desktop actions so far as we know. In this paper, we first collect a dataset to provide a benchmark for desktop action recognition in an egocentric paradigm. In particular, we focus on desktop actions that rely on hand-object manipulations, such as writing and reading. A head-mounted camera is used to record the hands and their interactions with the environment when performing different actions. Pixel-wise hand masks are provided to facilitate the analysis of hand, ground-truth action

Manuscript received June 16, 2017; revised December 9, 2017; accepted February 6, 2018. This work was supported in part by the Joint Funds of NSFC-CARFC under Grant U1533129, and in part by NSFC under Grant 61602020, Grant 61732016, and Grant 61671267. This paper was recommended by Associate Editor L. Shao. *(Corresponding authors: Feng Lu; Yue Gao.)*

M. Cai is with the Institute of Industrial Science, University of Tokyo, Tokyo 1538505, Japan (e-mail: cai-mj@iis.u-tokyo.ac.jp).

F. Lu is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the Beijing Advanced Innovation Center for Big Data-based Precision Medicine, Beihang University, Beijing 100191, China (e-mail: lufeng@buaa.edu.cn).

Y. Gao is with the Key Laboratory for Information System Security, Ministry of Education, Tsinghua National Laboratory for Information Science and Technology, School of Software, Tsinghua University, Beijing 100084, China (e-mail: kevin.gaoy@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2806381

annotations are provided at frame-level for evaluation of action recognition.

Based on the dataset, we provide baseline action recognition results by examining different feature representation, and further propose an effective feature representation via analysis on human hand. We extract variant hand features during hand-object manipulation, and evaluate their performance in recognizing daily desktop actions. We show that the information about hand shape and motion provide crucial cues for identifying common desktop actions, such as reading, writing, and typing. Even without concerning the manipulated objects, the problem can be well solved by using only hand features.

Primary contributions of this paper are as follows: 1) we propose an effective method to recognize desktop actions by using egocentric video captured by a wearable camera; 2) we show that by only using hand features, such a problem can be well solved. In particular, we propose several hand features to extract information about hand shape, position, and motion; and 3) we create a desktop action dataset, based on which we evaluate our method and show state of the art accuracy compared to conventional methods.

II. Related Works

As an important research topic, human action recognition has attracted research focus for decades and is related to many different domains, such as visual tracking [5], [6], human/salient object detection [7]–[9], event classification [10], pattern recognition [1], [11], [12] etc. Initial works on human action recognition mainly focus on simple and standard actions while no other objects are involved except the human body, such as walking and running. For instance, in [13]–[15], human body movement patterns are considered and full-body features are used for recognition. In contrast to these early works, which have achieved nearly perfect performances on standard dataset, such as KTH [4], recognizing daily actions that involves object manipulation is still far to be solved and has attracted increasing attention recently [1], [16], [17]. People manipulate objects as a natural part of daily actions, and therefore visual context, which tells us what is involved in the action, plays an important role in recognizing daily actions. Moore *et al*. [18] used object context in hidden Markov model for recognizing hand actions. Wu *et al*. [19] recognized activities based on temporal object usage, by using radio-frequency identification tagged objects with a dynamic Bayesian network model to jointly infer object labels and the most likely activity. Marszalek *et al*. [20] exploited the correlation between scene context and human actions to improve recognition rate. Yao *et al*. [21] used mutual context of objects and human poses to recognize actions in still images. Gupta *et al*. [22] used a Bayesian approach to analyze hand-object interaction by incorporating the contextual information of object evidence and hand trajectories. All these methods use static cameras fixed in the environment. It is therefore challenging for them to achieve high recognition performance for actions involving human-object interactions, where the relevant objects and body parts tend to be small and only partially visible. In contrast to these methods, we focus on recognizing daily actions from egocentric point of view.

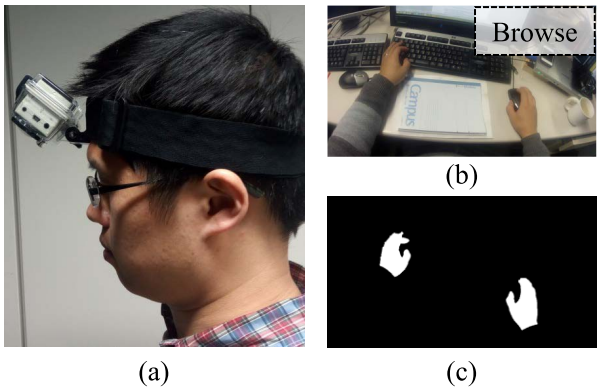


Fig. 1. Data collection and annotation. A head-mounted camera (a) is used to record desktop actions. A sample video image (b) is shown with action annotation. Pixel-level hand mask (c) is provided for the corresponding image.

Egocentric action recognition using wearable cameras has become popular recently due to advances of hardware technology [23]–[35]. An early work of unsupervised action segmentation and action recognition is done by Spriggs *et al*. [24]. In this paper, they trained a hidden Markov model with mixture of Gaussians output on the gist features [36]. In contrast to [24], most follow up works use object context to model different actions. Pirisiavash and Ramanan [28] trained an activity model with a fully labeled dataset, based on temporal object usage. Fathi *et al*. [29] also used object context by additionally considering eye gaze focusing on task-relevant objects. They utilized a wearable gaze tracking system that predicts regions of attention in the recorded scene. Different from their works, we only employ hand context to recognize daily actions without tedious object labeling and costly eye sensing. Other researchers also try to use eye and head movement patterns. Kitani *et al*. [26] proposed a fast, unsupervised approach to index action categories in sports video by using head motion vector. Ogaki *et al*. [27] combined head and eye movement patterns to recognize indoor desktop actions by using an inside-looking camera and an outside-looking camera similar to [29]. However, the global head movement contains much action-irrelevant information that does not help to discriminate between different actions. Instead, we examine hand motion patterns to encode human actions more effectively.

III. Desktop Action Dataset

A. Data Collection

Five common desktop actions of six participants (subjects) are recorded. The action categorization is based on empirical analysis of daily desktop activities. We asked one student and one secretary to record their activities around a desktop in a full work day without any instructions. The recorded videos are then examined to identify a set of actions with hand-object interactions that occur most frequently as the action categories used in our dataset. The details of the five actions are given in Section III-C. All subjects are right handed. During the data recording procedure, they wore a head-mounted camera

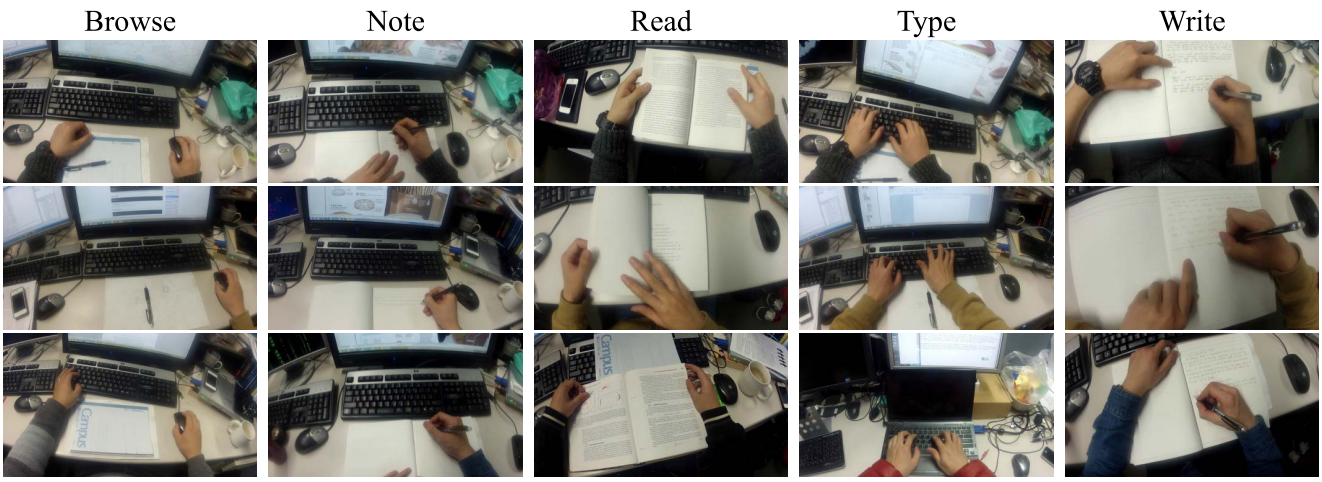


Fig. 2. Five common desktop actions recorded in the dataset. For each action category, three samples from different participants are presented to show the intersubject variation within the same action.

[shown in Fig. 1(a)]. We asked each subject to perform one type of action in front of a desk during each video recording. The same action was performed twice by each subject at different time. The recorded video are representative of the general set of desktop actions they perform in their daily life. In particular, we confirmed the hands were captured in the video.

The head-mounted camera (GoPro HD Hero2) used for recording actions consists of a high resolution video sensor (1920×1080 resolution, 30 FPS), a wide-angle fisheye lens (170° field of view), and a three-band head strap. This camera setup enables us to record the entire workspace of two hands and desktop objects in front of the body.

B. Data Annotation

Two annotation tasks were done along with the recorded video. One task is to annotate the actions. The second task is to annotate the hand pixels.

For action annotation, action label is provided for each video segment (the first frame and last frame) in which the defined action occurs. Fig. 1(b) shows one image sample with the tagged action class.

For hand annotation, we provide hand mask at pixel-level to facilitate the training of hand detector. To do this, we implement an interactive segmentation tool based on the GrabCut algorithm [38]. Given an image appearing on the tool interface, an user can choose to draw a "foreground" line on the hand region or a "background" line on the nonhand region. A hand segmentation mask is automatically generated based on the user input. This interactive process is iterated until the generated mask is satisfactory. Finally, a binary image is output in which the pixels of hand have value "1." One sample image of hand annotation is shown in Fig. 1(c).

C. Dataset Statistics

In this dataset, we provide original video data in MPEG-4 format, action annotation data at frame-level, and hand annotation data at pixel-level. A total of ten video sequences was recorded for each subject in which five different actions were performed twice. Each sequence lasts about 3 min and the total video data is over 3 h. The statistics of the dataset is summarized below.

1) *Size:* 60 video sequences, $324,000$ frames.
2) *Video Quality:* 1920×1080 resolution, 30 FPS.
3) *Number of Action Categories:* Five.
4) *Number of Hand Annotation:* 660 mask images.

For action annotation data, the start/end frames as well as the action class of each action instance are provided. The five action classes are: 1) browse; 2) type; 3) note; 4) write; and 5) read, as illustrated in Fig. 2. In browse, people use right hand to manipulate mouse while watching the screen. In type, people also watch the screen but with two hands typing on a keyboard. Note is similar to write while it differs in that user visual attention transfers between a notebook and the screen occasionally. In read, the hands are used to hold the book and flip to next page. These five actions represent common desktop actions in daily life and they all involve hand object interactions.

For hand annotation data, binary hand masks of 11 sampled images are provided for each sequence. In total, 660 images of hand annotation are provided (over five millions of hand pixels). To further facilitate the analysis of hand in desktop action recognition, we also provide hand probability maps for all video frames. The hand probability map has the same size with original images and the value of each pixel indicates the probability of the pixel belonging to the hand region. The details of how to generate the hand probability map will be introduced in Section IV-A.

To better illustrate the uniqueness of the proposed DesktopAction dataset, we also compared this dataset with other related datasets that involve hand-object interactions in Table I. MPII cooking activities dataset contains 65 different cooking activities recorded in high quality videos. Annotation of action categories and human pose on a subset of frames are provided. However, different from other datasets, actions were recorded under third-person camera view, and the overall body motion plays a more important role than the hand. In ADL

TABLE I
SUMMARY OF RELATED DATASETS WITH HAND-OBJECT INTERACTIONS

Dataset	Camera view	Resolution	Size	Action categories	Annotation
MPII Cooking [38]	third-person	1624×1224	8 hours	65 (cooking actions)	action label, human pose
ADL [29]	first-person	1920×1080	10 hours	18 (actions of daily living)	action label, object label
GTEA [35]	first-person	720×405	30 minutes	7 (cooking activities)	action label, hand mask
DesktopAction	first-person	1920×1080	3 hours	5 (desktop actions)	action label, hand mask

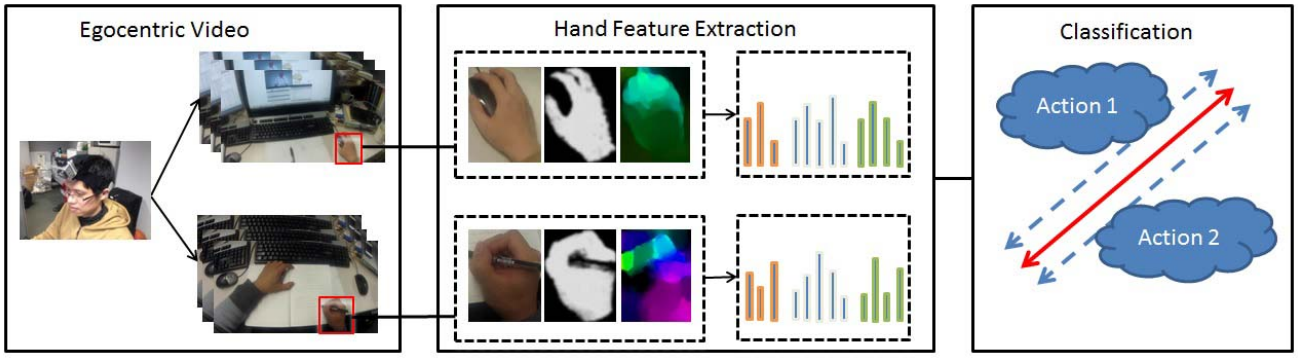


Fig. 3. Pipeline of desktop action recognition. We capture egocentric video, extract hand features, and use them for desktop action recognition.

dataset, long hours of daily activities were recorded with a body-worn camera. Objects play a central role in this dataset, therefore the object label as well as action label are provided. The Georgia tech egocentric activity (GTEA) Dataset contains seven types of daily cooking activities, also recorded from first-person camera view. Although both GTEA and DesktopAction datasets provide hand masks besides the action label, they are targeting at different action scenes (kitchen and office). Besides, the DesktopAction dataset has higher video quality and involves not only annotated hand masks but also hand probability maps for all the frames.

IV. RECOGNITION PIPELINE

In this section we present the pipeline of desktop action recognition that comprises three major stages as shown in Fig. 3. First, egocentric video capturing desktop actions (e.g., browsing and writing) are preprocessed for feature extraction. In particular, hand regions are detected by a multimodel hand detector and then cropped out (red bounding boxes in the left of Fig. 3) from the original video frames for extracting hand-related features. Then, various features are extracted as the representation of different actions at the second stage. Finally, the extracted features are used for training action classifiers and predicting unknown action classes.

A. Hand Region Segmentation

In order to extract reliable hand features, it is important to conduct accurate hand region segmentation first. In particular, we expect a pixel-level accuracy for hand region detection and segmentation in this paper. Performance of hand detection often suffers from large changes in illumination conditions because illumination changes induce fluctuations in the appearance of hands. In order to avoid such a problem and obtain satisfactory hand detection results under varied illumination conditions, a multimodel hand detector [39] was trained which is illumination-invariant as illustrated in Fig. 4.

The multimodel detector is trained by using a dataset containing multiple hand appearances under different illumination conditions. In practice, to train such an illumination-invariant detector, we first collect over 200 hand images taken under varied illumination conditions and backgrounds. For each hand image, hand regions are segmented and labeled by using an interactive segmentation tool based on graph cuts. Then, ten different global scene models are learned by using k-means clustering, with respect to the HSV histogram of each training image. In this manner, we obtain $p(c|\boldsymbol{g})$, which is a conditional distribution of a scene c given a specific HSV histogram $\boldsymbol{g}$. Then, a separate random forest regressor is trained for each global scene model, all of which compose a multimodel hand detector to handle different scene illumination. Finally, given the local feature $\boldsymbol{l}$ (encoding color, shape, and boundary information) of a pixel x and the HSV histogram $\boldsymbol{g}$ of the entire image, the probability of x belonging to a hand region is computed by marginalizing over different scenes c

$$p(x|\boldsymbol{l}, \boldsymbol{g}) = \sum_{c} p(x|\boldsymbol{l}, c)p(c|\boldsymbol{g}) \tag{1}$$

where $p(x|\boldsymbol{l}, c)$ is the output of the random forest regressor trained for global scene c and $p(c|\boldsymbol{g})$ returns the probability of scene c given $\boldsymbol{g}$ as described above.

Then, we can generate a pixel-wise hand probability map denoted as $\boldsymbol{M}_t$ at frame t under unknown illumination conditions with high confidence (as in the right column of Fig. 4). Note that the hand probability map $\boldsymbol{M}_t$ has the same resolution as original image and each element indicates the probability of being a hand pixel.

Given a test image, hand regions are segmented based on the corresponding hand probability map. Hand blobs are first obtained by binarizing the probability map with a threshold. To remove false segmented regions from the background, we

Fig. 4. Hand segmentation using multimodel hand detector.

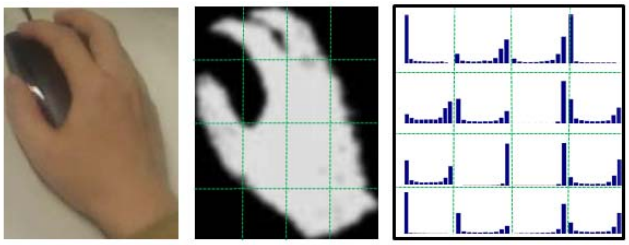


Fig. 5. Exacting hand shape feature via histogram of hand belief scores in each spatial division of hand region.

discard the blobs that are beyond certain predetermined area range. To remove the influence from skin colored arms due to different length of sleeves worn by the subjects, we adopt the method in [31]. Specifically, ellipse parameters (length of long/short axis, angle) are fitted to each hand blob, and the arm part is approximately removed by shortening the length of long axis to 1.5 times of the length of short axis. A fixed size bounding box is drawn by fixing the top-center of the bounding box to the top-center of the arm-removed hand blob. The size of the bounding box is determined heuristically for each video, based on the observation that the distance between the hands from the head-mounted camera is consistent throughout the video recording. Moreover, a temporal tracking method [40] is utilized to handle the case of two overlapping hands. Briefly speaking, the position and movement of each candidate hand region is stored and used in hand segmentation of the next video frame. Thus, two overlapped hands can be separated by using tracking information of each hand before overlapping.

There are three typical hand configurations in the egocentric scene, namely left hand, right hand, or hand intersection. For each image, we apply heuristic cues including centroid, orientation of principle axis and area of hand mask to determine the current configuration. Without loss of generality, we assume subjects are right-handed and only use right hand in the following discussion.

B. Feature Extraction

One typical way to represent an action is to use the visual context of the scene in which the action is performed. The appearance of the objects appeared in the background gives important clues about the possible actions being performed. For example, a book-like object in the center of the scene implies more possibility that the person is reading a book rather than typing on a keyboard.

Another important way of characterizing different actions is by the ego-motion which is unique in egocentric video. The ego-motion is concerning the body motion from the actor's perspective (also the camera wearer) and can be measured by the camera's motion. It depends on the camera's position that which body part's motion is captured. In the case of using a head-mounted camera, actor's head motion can be captured by estimating the optical flow of the scene. However, in desktop actions, the head motion is subtle and may not be sufficient to differentiate actions, such as "browsing" and "reading."

By leveraging the fact that active hands play a crucial role in desktop tasks, we propose to recognize desktop actions by extracting different types of informative hand features from egocentric video. The proposed hand features will be introduced in Section V.

C. Action Recognition

Given visual features extracted from video frames, discriminative action classifiers are desired to predict an action label for each frame. We train multiclass action classifier for the defined desktop actions. In particular, linear support vector machine (SVM) is trained for each action class using the extracted visual features.

For each action a, we train a binary classifier by finding the optimal model parameter $(\boldsymbol{w}, b)$ from labeled training data $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i$ is the extracted features from frame i and $y_i \in \{-1, 1\}$ is the label indicating whether that frame belongs to action a, by minimizing the objective function

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{2}$$

subject to

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i - b) \geq 1, \quad i = 1 \cdots n. \tag{3}$$

In practice, we utilize the implementation of LIBSVM [41] for training. Probability calibration [42] is conducted for each classifier in order to produce comparable scores.

During testing, each video frame is classified independently and assigned with an action class of which the classifier outputs the highest score.

V. Proposed Hand Feature

There are increasing interests on egocentric action recognition, which typically rely on object recognition or other visual cues. However, in this paper we argue that active hands themselves provide discriminative cues for desktop action recognition with a reasonable assumption that active hands involved in object manipulation are always visible in egocentric video. Therefore, we extract hand related features for accurate desktop action recognition from a first person perspective. These features are designed in accordance with several important egocentric cues for understanding object manipulation, including hand information of shape, position, and movement.

A. Hand Shape

The first cue we introduce here is the shape of the active hand. When manipulating an object, hand shape varies due to different object geometries and tasks. Therefore, it is reasonable to assume that hand shape provides discriminative

Fig. 6. Optical flow (middle) of egocentric video (left). Note the diverse motion in the hand region. Color map for flow visualization is shown in the rightmost.

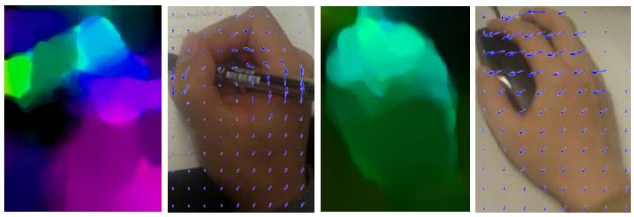


Fig. 7. Comparison of optical flows for different actions.

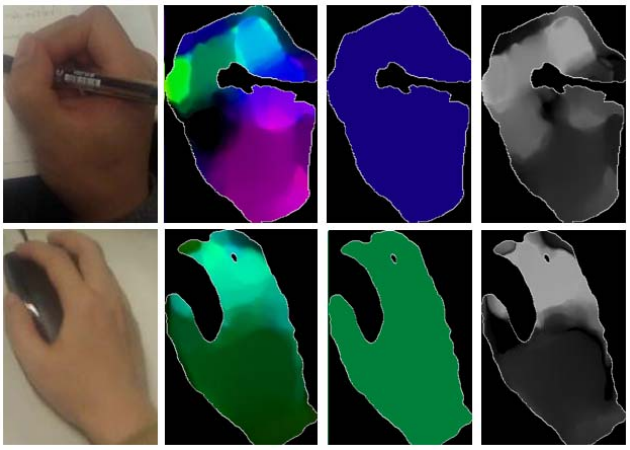


Fig. 8. Spatial optical flow variance. From left to right: segmented hand image, corresponding optical flow, mean flow, and the magnitude of flow variance.

information for recognizing different hand-object interactions. In order to describe hand shape, we extract histogram of hand shape as illustrated in Fig. 5. Inside the bounding box of the detected hand region, hand probability map is first evenly divided into 16 subregions denoted as $\boldsymbol{M}_t^{(i,j)}$, where (i, j) is the index of subregions. Then histogram of probability intensities is computed over $\boldsymbol{M}_t^{(i,j)}$. A 10-bins histogram denoted as $\boldsymbol{h}_t^{(i,j)}$ is generated for each subregion (in the right of Fig. 5). Finally, all histograms are concatenated to form a 160-D feature vector $\boldsymbol{h}_t$.

B. Hand Position

We notice that the hand position in egocentric video varies for different tasks. In the tasks which require a high head-hand cooperation, for instance, "writing," subjects need to fixate on the front end of the pen handled by his/her hand; while in the action "typing," subjects do not need to look at the keyboard, but the screen. In egocentric video, relative deviation of hands with respect to head orientation can be modeled by hand position since the head is always oriented to the center of the scene. Thus, the hand position can be used to model the head-hand cooperation and facilitates action recognition. Here, we use hand manipulation point $\boldsymbol{p}_t = [p_x, p_y]$ to represent hand position. Similar to Li *et al*. [43], manipulation point is defined as a point, where interaction of hand and object is most likely to occur. For a right hand, manipulation point usually locates at the left tip of the hand. For the case two hands intersect with each other, manipulation point usually centers around the intersection point.

C. Hand Motion

In this section, we investigate one of the most important features for action recognition: motion, by using optical flow estimation. Optical flow [44]–[46] measures local motion of pixels or keypoints, and previous research has shown that features based on optical flow are effective for action recognition. Unlike the conventional third person perspective scenario, egocentric video often involves irrelevant global motion at the background due to free head motion. Instead, motion of active hand provides rich information about complex hand-object interactions.

Taking Fig. 6 for an example, optical flow is almost uniform in the background, reflecting the fact that it is caused by global head movement. On the other hand, optical flow in the hand region is distinguishing because it is caused by the presently performed action (writing). Therefore, we propose only using optical flow detected in the active hand region to describe hand motion. This retains the most important information for action recognition and also greatly reduces the computational cost for motion estimation. In practice, we apply large displacement optical flow [47] between two consecutive frames for optical flow estimation.

Fig. 7 shows an example of different optical flows in hand regions due to different actions. Intuitively, the optical flow for writing forms a swirl centered at the pen, while in the case of browsing, the optical flow has more uniform directions and intensities. To effectively describe such differences, we introduce multiple flow-based features in the follows.

1) Inner Frame Flow Distribution: As described above, examples in Fig. 7 show that different actions result in different spatial distributions of optical flow intensities in the hand regions. To describe such a distribution, we extract the spatial histogram of hand optical flow (S-HoF). Note that the estimated optical flow is a combination of global motion (due to camera motion) and individual hand motion, while the latter one contains the valid information for our purpose here. Therefore, we first average the motion of the entire hand region, and then subtract it from the original hand motion at each pixel, as shown in Fig. 8. For the remaining local motion, we encode their magnitudes into a 10-bin histogram $\boldsymbol{s}_t$ as the inner frame flow distribution feature.

2) Temporal Analysis and Feature: The above inner frame flow distribution only concerns about spatial information in a single frame; while more complex motion pattern is encoded temporally. For instance, writing is related to continuously

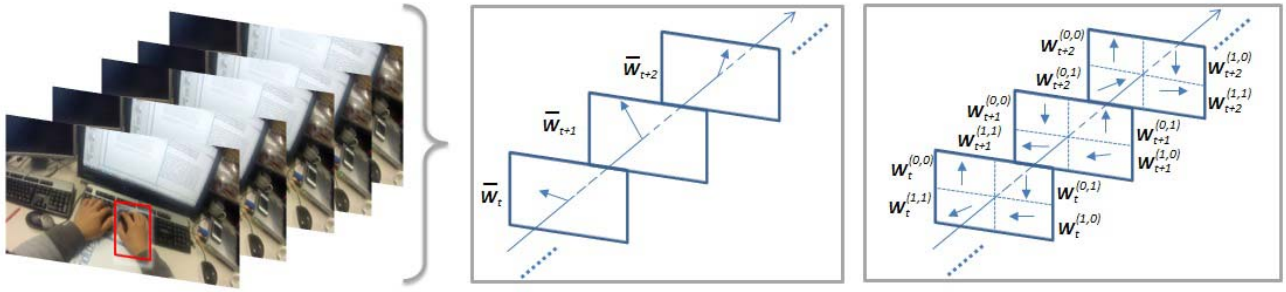


Fig. 9. Computing temporal hand feature. From left to right: cropped hand regions from input image sequence, average flow within hand region of each frame, flows averaged for each subdivision of hand region.

and repeatedly horizontal motions, while browsing with a mouse shows relatively random motions inside a certain desktop region. Moreover, camera motion also conveys information about action type for some tasks. It is therefore intuitive to think about using frequency analysis to describe temporal pattern of captured motion.

We extract the temporal histogram of the frequency component of hand optical flow (T-HoF). To this end, optical flow within entire hand region is averaged to get an average optical flow $\bar{\boldsymbol{w}}_t = [\bar{x}_t, \bar{y}_t]$ for each frame, as shown in the middle of Fig. 9. Discrete Fourier transform is then applied to both horizontal and vertical components of the average optical flow sequence, respectively

$$X_t(k) = \sum_{n=0}^{N-1} \bar{x}_{t+n} e^{\frac{-i2\pi kn}{N}} \qquad k = 0, \ldots, N-1 \tag{4}$$

$$Y_t(k) = \sum_{n=0}^{N-1} \bar{y}_{t+n} e^{\frac{-i2\pi kn}{N}} \qquad k = 0, \ldots, N-1 \tag{5}$$

where $N = 64$ is the sliding window size. The first 32 DFT coefficients for both horizontal and vertical components are concatenated to yield a 64-D motion histogram $\boldsymbol{f}_t = [X_t(0), \ldots, X_t(M-1), Y_t(0), \ldots, Y_t(M-1)]$ where $M = 32$ and t denotes the frame index.

3) Spatial–Temporal Analysis of Cumulative Motion: The proposed S-HoF and T-HoF features only take into account either spatial or temporal information of hand motion, while in this section we introduce a spatial–temporal descriptor of hand motion. The idea is that depending on the tasks, different subregions of the active hand show different intensities of motion. For instance, typing requires fast motion of all fingers, and in browsing only one finger is most active. Although such a difference looks random in only one frame, we notice that the cumulative motions during a time period show significant differences. Therefore, we compute cumulative motion for each hand subregion within a time period separately.

In particular, we extract spatial–temporal histogram of optical flow. As illustrated in the right of Fig. 9, we spatially divide the hand region into, e.g., four subregions and compute the average optical flow $\boldsymbol{w}_t^{(i,j)} = [x_t^{(i,j)}, y_t^{(i,j)}]$ for each subregion at frame t, where (i, j) denotes the subregion index. Then we sum up the absolute values of average optical flows' intensities for each subregion over 64 consecutive frames

$$\bar{x}_t(i,j) = \frac{1}{N} \sum_{n=0}^{N-1} |x_{t+n}^{(i,j)}| \qquad i,j = 0, \ldots, M-1 \tag{6}$$

$$\bar{y}_t(i,j) = \frac{1}{N} \sum_{n=0}^{N-1} |y_{t+n}^{(i,j)}| \qquad i,j = 0, \ldots, M-1 \tag{7}$$

where $N = 64$, and M denotes number of spatial division in each dimension which is set as 3 in our experiment. Note that this is done in horizontal and vertical directions separately. Finally, a $2M^2$ dimensional feature vector $\boldsymbol{c}_t$ is composed by the concatenation of cumulative motions in each subregion.

VI. Experiments

In this section, we would like to analyze the statistical properties of the provided dataset, and show the recognition performance of different features as a baseline for facilitating future research.

A. Feature Analysis

In the dataset, we have recorded five different actions for six subjects. Before showing the action recognition results, we first demonstrate the statistical properties of different desktop actions in order to have a better understanding of the dataset.

To demonstrate statistical properties of different actions, we draw feature distribution of all data samples on the embedded space. In particular, we depict the distribution on four different kinds of feature representation. The first is GIST feature [24] which encodes the scene context. The second is head motion histogram (MHIST) [26] which encodes the spontaneous and periodic properties of head motion. The last two features encode hand shape [spatial hand feature (HF-S)] and hand motion [temporal hand feature (HF-T)], respectively. For each feature representation, we perform principle component analysis (PCA) based on all the data samples and then embed the feature vector of each data sample into the three most important components.

Fig. 10 depicts the feature distribution on the embedded 3-D feature space. From scene context, it can be seen that the actions of browse and type can be easily separated from the actions of read, write, and note. However, the latter three actions are similar on this feature space (especially write and

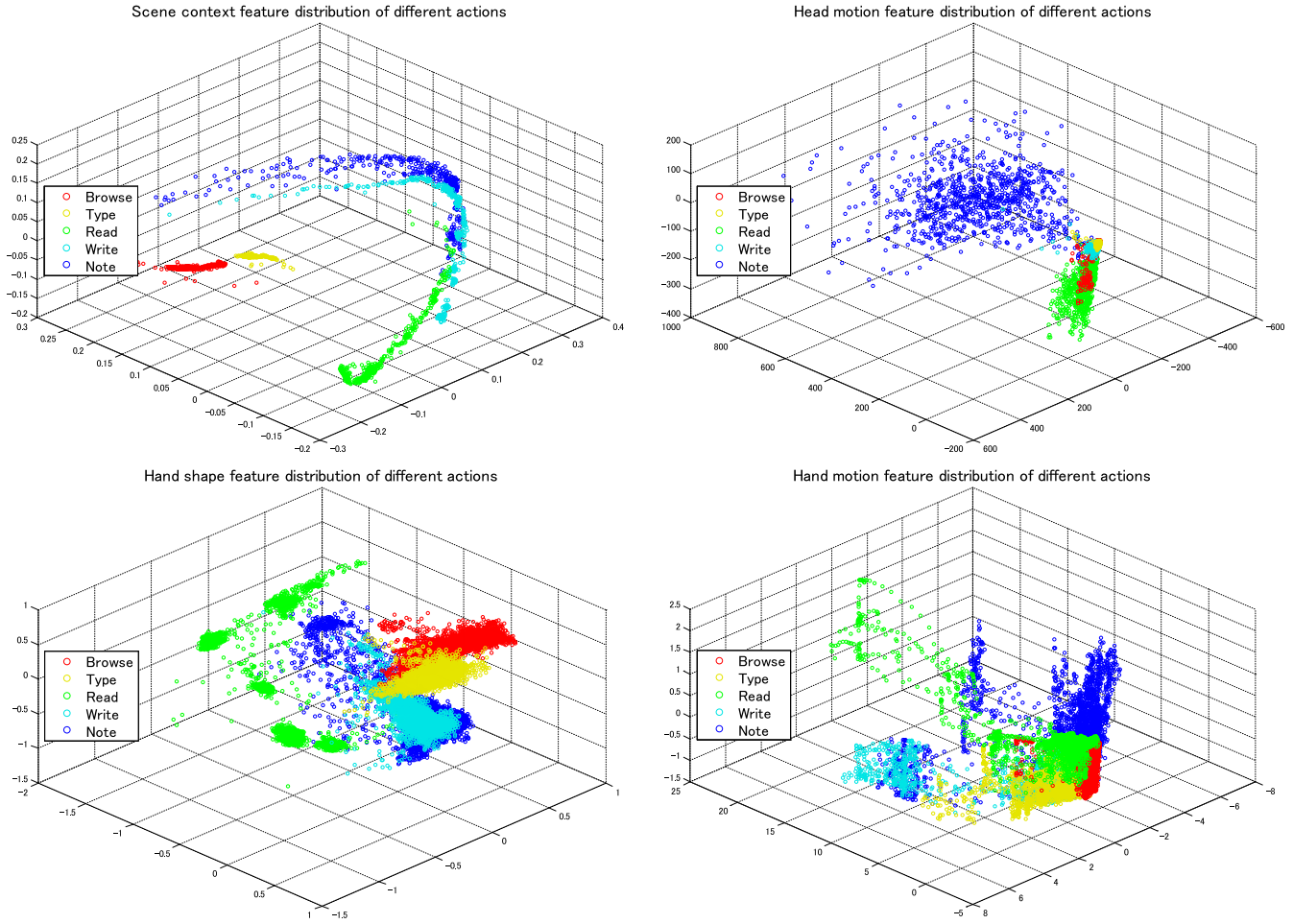


Fig. 10. Feature distribution of different actions. Feature vectors of all data samples are embedded into 3-D space using PCA. Four different kinds of feature representation (scene context, head motion, hand shape, and hand motion) are studied.

note), since the main object in the scene is a book or a notebook which have similar appearance. Feature distribution on head motion indicates that the head motion alone is not sufficient for distinguishing different desktop actions in which head motion is often very subtle. The most important observation, however, is from hand shape and hand motion. From hand shape, almost all actions are well separated, except for write and note which involve the same hand posture of holding a pen, while most data samples of the two actions are separated from hand motion. Most of all, the statistical properties of different actions shown in Fig. 10 provide us with intuitive understanding about different actions, facilitate the development of a reliable recognition system.

B. Accuracy

We perform action recognition at each individual frame. We use hand features that capture the hand shape, position, and motion patterns as described from Sections V-A to V-C. To evaluate the recognition performance, we train a multiclass action classifier for each subject using his/her training data labeled with five action categories. As a performance measurement of the subject-oriented classifier, we use a cross-subject validation method, namely, we test each subject-oriented classifier on all subjects' test data. We train action classifiers for each subject independently for three reasons: First, in egocentric vision, ideal classifier is not necessarily available by training over many subjects due to notable variance of action patterns between subjects. Second, feature consistence among subjects can be verified by comparing each individual classifier. Third, this allows to analyze the difference in motion patterns with respect to the same action done by different subjects.

For each train-test subject pair (A, B), we obtain a confusion matrix as a measurement of how the action classifier trained for subject A works for subject B. Visualization of all confusion matrices obtained from cross-subject validation is given in Fig. 11. The 6-by-6 plot is composed of confusion matrices for all train-test subject pairs. From the figure we can see a darker diagonal pattern, which shows high classification performance in the case that the classifier was trained and tested for the same subject. We can also find unsatisfactory results from the figure. For example, when applying the classifier trained for subject 3 to actions of subject 4, most of the actions (87% of browse, 91% of write, 69% of note) are misclassified to the action type.

We use classification accuracy, namely, proportion of correctly classified samples out of total test data, as an

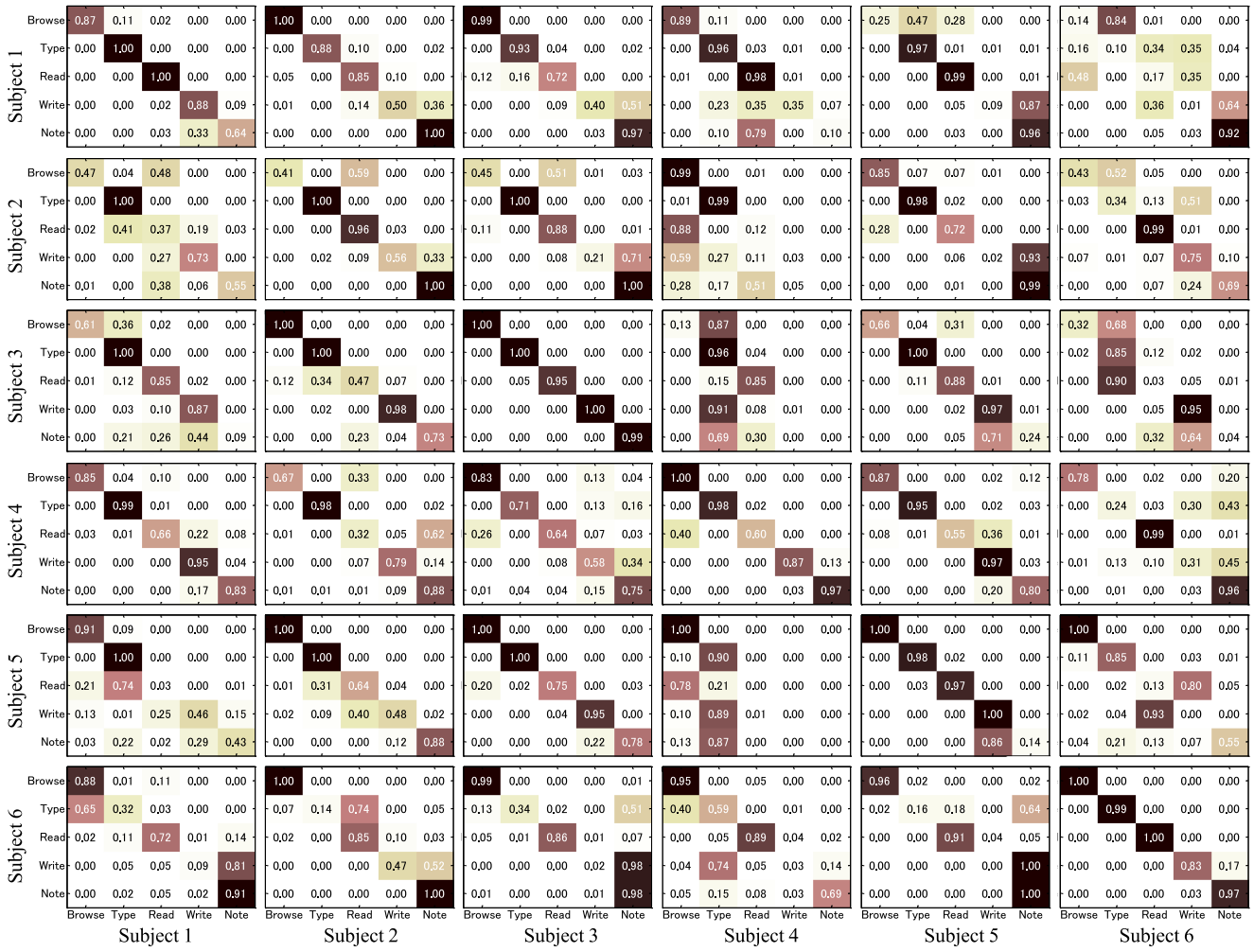


Fig. 11. Confusion matrices among all train-test pairs. Classifiers are trained for each subject indexed by row and tested on all subjects indexed by column. Within each confusion matrix, actual action categories are indexed by row, while predicted categories are indexed by column.

TABLE II
CLASSIFICATION ACCURACY OF ALL TRAIN-TEST SUBJECT PAIRS (NUMBERS IN THE LEFT 6 × 6 MATRIX) AND THE AVERAGE ACCURACY (NUMBERS IN LAST TWO COLUMNS)

Subject	S1	S2	S3	S4	S5	S6	Avg. (All)	Avg. (Self-excluded)
S1	0.87	0.84	0.82	0.64	0.65	0.26	0.68	0.64
S2	0.62	0.86	0.72	0.41	0.71	0.64	0.66	0.62
S3	0.68	0.82	0.98	0.37	0.74	0.43	0.67	0.61
S4	0.85	0.77	0.70	0.88	0.82	0.66	0.78	0.76
S5	0.57	0.80	0.89	0.36	0.81	0.50	0.66	0.62
S6	0.58	0.63	0.66	0.62	0.60	0.96	0.68	0.62
Avg.							0.69	0.65

performance measurement to quantitatively describe the results in each confusion matrix. The resulting numbers are given in Table II. By averaging accuracy among all subjects, we get two types of average accuracy, one of which considers recognition rates of all six subjects and reports an accuracy of 0.69, and the other one excludes the self-validation and the result is 0.65. Such a result is reasonably good and we provide with comparison with existing methods later. In the rest of this paper, we use self-excluded results if not mentioned.

To evaluate the feature consistence among subjects, we also report performance statistics (average and standard deviation of classification performance) of the proposed feature on different subjects and action classifiers, as shown in Fig. 12. It can be observed that the standard deviation of accuracy is high for almost all the subjects (except for S6), indicating significant variance of action patterns between subjects. Regarding individual variance in performing the same action, we can find that the motion patterns of browse and type among subjects are more consistent than other action categories, which can be seen from Fig. 12(b).

C. Comparison With Existing Methods

We compare the performance of our hand features with two baseline methods. The first method uses GIST features [24]

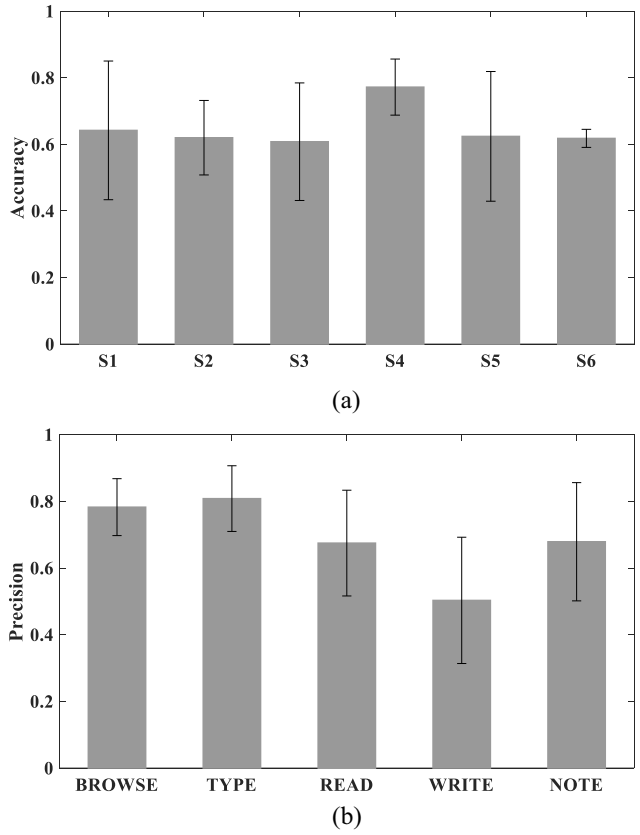


Fig. 12. Performance statistics (average and standard deviation of classification performance) of the proposed feature on (a) different subjects and (b) actions.

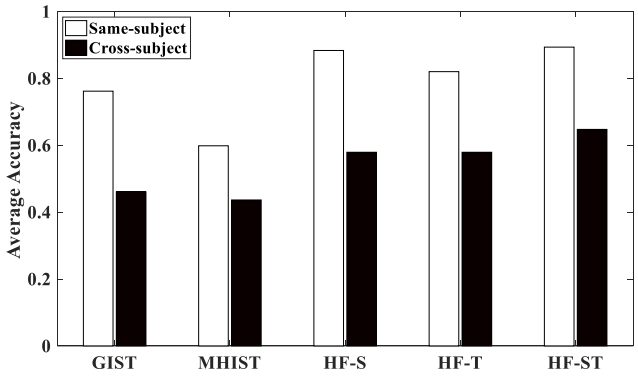


Fig. 13. Performance comparison of different features on two settings: same-subject and cross-subject.

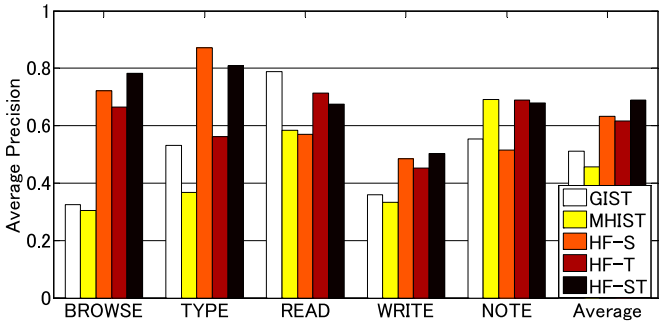


Fig. 14. Comparison of accuracy regarding different actions.

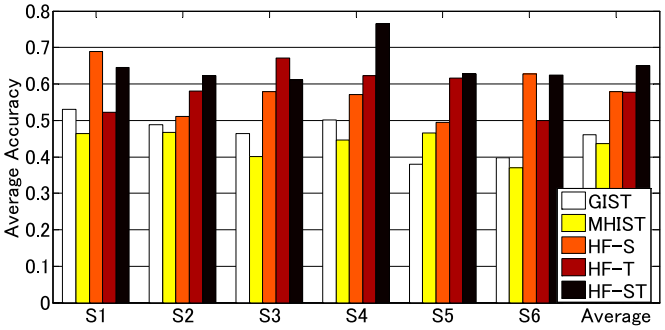


Fig. 15. Comparison of accuracy regarding different subjects.

which encodes the global visual context while performing actions. The second method uses MHIST proposed in [26], which is computed by encoding spontaneous motion and periodic motion using Fourier analysis, based on sparse optical flow of ego motion. These two features are commonly used in recent egocentric action recognition methods. In addition, to demonstrate how hand features work, we split our hand related feature into two parts: 1) HF-S and 2) HF-T, which encodes spatial information of hand shape/position and temporal information of hand motion patterns, respectively. Moreover, we use HF-ST to denote the proposed feature that combines HF-S and HF-T. Fig. 13 shows the average accuracy of different features calculated based on two settings. In the same-subject setting, the training and testing are conducted on the data of same subject. In the cross-subject setting, we use self-excluded average accuracy of cross-subject validation as described previously. Both spatial and temporal hand features clearly outperform existing baseline features. The combined feature (HF-ST) achieves the highest performance on both settings (0.89 and 0.65, respectively).

Classification performance is then studied regarding each action category as shown in Fig. 14. We extract classification precision of each category from the confusion matrix and average over all cross-subject pairs to calculate the average numbers. As for HF-S, it works pretty well for the type task, because the spatial hand shape is obviously different from those in other tasks. It performs relatively badly on write and note tasks due to the fact that similar hand shapes appear in these two tasks. On the other hand, MHIST and HF-T, which are features based on motion information, outperform other features for the note task (taking notes from screen to notebook) due to regular up-and-down global motion induced by changing the head pose to face either screen or notebook. However, MHIST performs badly for the browse task while our proposed HF-T still performs well. This shows the advantage of our method, because in browse there lacks global motion required by MHIST, while our proposed HF-T captures only hand motion and uses it effectively. Finally, our proposed combined feature HF-ST achieves good accuracy in most cases and outperforms others in average.

In Fig. 15 we show the performance regarding different subjects. For all subjects, our method clearly improves the classification performance compared to existing methods. Notice that performances of HF-S and HF-T may vary

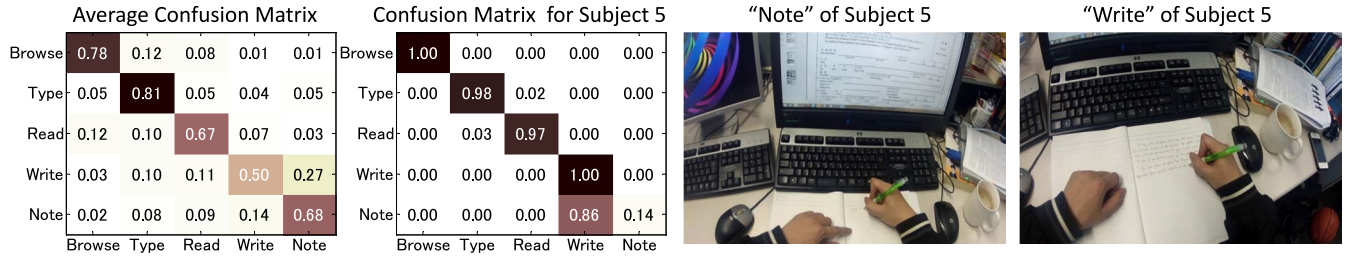


Fig. 16. Example of failure cases showing how individual variance in action performing influences the classification performance.

TABLE III
RESULTS FOR COMBINING ALL FEATURES

Feature	HF-ST	HF-ST (self-ex.)	ALL	ALL (self-ex.)
Accuracy	0.691	0.650	0.693	0.654

largely and it is not clear which is better. However, the combined feature HF-ST always shows stable performance, which demonstrates the necessity to consider both spatial and temporal hand information as we propose in this paper.

It is interesting to ask what will happen if we combine all features, i.e., GIST, MHIST, HF-S, and HF-T together. We did experiments and the results are summarized in Table III. There is no improvement in accuracy by adding other features to our hand features. This suggests that our hand feature is very efficient in the case of desktop action recognition from egocentric video. However, developing and combining other features to our features should be a promising future research topic.

D. Failure Cases

In this section, we study the limitations of our hand features by examining failure cases. Although the proposed method achieves good average classification performance, it still fails in several cases.

One important reason for false classification is due to individual variance in performing the same action. As shown in Fig. 16, in average, note and write tasks have a relatively high probability to be misclassified (27% of write are confused as note and 14% of note are confused as write). This is reasonable since the two actions share similar hand shapes and hand motion patterns. While in most cases (more than 50%) they can be correctly classified by head motion in note, the recognition rate for some subject is very low as shown in the confusion matrix for Subject 5. This is supported by the evidence that subject 5 did not make large head motion but relied on gaze movement to see the screen. One way to solve this problem is to add wearable sensors besides the wearable camera to estimate user gaze and help distinguish actions.

Bad video recordings, such as motion blur or hand occlusion could also make our system fail. Although head motion is used as an important egocentric cue in action recognition, it can bring big motion blur for head-mounted cameras and makes hand segmentation fail. Moreover, the hands are sometimes partially outside the visual field due to inappropriate camera setting, therefore causing problems for our hand-based features.

VII. DISCUSSION

In this section, we first discuss the connection between the feature distribution of different actions and the discriminability of different features. Then we discuss subjective differences in performing the same action and their influence on action classification performance.

As illustrated in the experiments, the features extracted from hand appearance and motion show notable advantage over other baseline features in discriminating between different desktop actions. The empirical results are consistent with the feature distribution of different actions as demonstrated in Fig. 10. The scene context feature like GIST is effective in differentiating actions with different background, however, insufficient for actions that have different motion. Head motion-based features like MHIST alone is incapable of discriminating between actions that differ in motion patterns of the hand. The proposed feature of HF-ST, which combines hand appearance and motion, separates the five actions very well from the embedded data space. More importantly, this kind of feature analysis could be used as a powerful tool in developing discriminative feature representation for a recognition system.

Although the proposed system achieves reliable action recognition performance for specific users, the performance degrades a lot in the cross-subject case (Fig. 13), indicating that individual variance in action performing has a big impact on the system performance. As shown in Fig. 12, while some actions (browse and type) among the five desktop actions are relatively consistent among different subjects, others are not. Taking read for example, the hand posture of holding a book and the temporal frequency of turning a page are different among subjects. The individual difference observed in the dataset suggests that a user-specific action classifier is more reliable than a general action classifier in desktop action recognition system. Furthermore, although system performance degrades in cross-subject validation, the results actually could be used to study the inherent diversity about an action. Just as we have analyzed in Section VI-D, the deviation of misclassification rate between write and note has guided us to discover the unique motion pattern of a subject in performing the action of note.

VIII. CONCLUSION

In this paper, we collect a dataset as a benchmark for desktop action recognition in egocentric paradigm. Video data of daily desktop activities is captured by a wearable camera from

a first-person perspective. Based on the dataset, we present a action recognition pipeline and provide baseline recognition results by examining different feature representation. In particular, we propose a novel feature representation by exploring information from hand appearance and motion. We design different hand features from egocentric video and analyze their discriminability in the embedded data space. We evaluate our method and compare it with existing egocentric video based methods. Experimental results show that our method achieves significantly better accuracy for both same-subject and cross-subject settings in the dataset.

As for future work, we plan to extend the current dataset to cover broader action categories under more general conditions. Besides, the methods for action recognition need to be further improved. One possible way is to incorporate temporal consistency to enhance the performance. Another way is to effectively combine hand information with global scene descriptors in an end-to-end framework (like convolutional neural networks).

REFERENCES

[1] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[2] T. Kanade and M. Hebert, "First-person vision," *Proc. IEEE*, vol. 100, no. 8, pp. 2442–2453, Aug. 2012.

[3] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 744–760, May 2015.

[4] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3. Cambridge, U.K., 2004, pp. 32–36.

[5] A. W. Smeulders *et al.*, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, Jul. 2014.

[6] L. Huang *et al.*, "Visual tracking by sampling in part space," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5800–5810, Dec. 2017.

[7] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[8] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[9] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, to be published. [Online]. Available: http://ieeexplore.ieee.org/document/7932195/, doi: 10.1109/TCSVT.2017.2706264.

[10] Y. Gao, H. Zhang, X. Zhao, and S. Yan, "Event classification in microblogs via social tracking," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, p. 35, 2017.

[11] Z. Zhang *et al.*, "Discriminative elastic-net regularized linear regression," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1466–1481, Mar. 2017.

[12] Y. Guo, G. Ding, and J. Han, "Robust quantization for general similarity search," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 949–963, Feb. 2018.

[13] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[14] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2003, pp. 726–733.

[15] A.-A. Liu *et al.*, "Multiple/single-view human action recognition via part-induced multitask structural learning," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1194–1208, Jun. 2015.

[16] J. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surveys*, vol. 43, no. 3, p. 16, 2011.

[17] L. Shao, L. Liu, and M. Yu, "Kernelized multiview projection for robust action recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 115–129, 2016.

[18] D. J. Moore, I. A. Essa, and M. H. Hayes, III, "Exploiting human actions and object context for recognition tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, 1999, pp. 80–86.

[19] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2007, pp. 1–8.

[20] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 2929–2936.

[21] B. Yao, A. Khosla, and L. Fei-Fei, "Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 1–8.

[22] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1775–1789, Oct. 2009.

[23] W. W. Mayol and D. W. Murray, "Wearable hand activity recognition for event summarization," in *Proc. Int. Symp. Wearable Comput. (ISWC)*, vol. 1, 2005, pp. 122–129.

[24] E. H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *Proc. IEEE Comput. Vis. Pattern Recognit. Workshops*, 2009, pp. 17–24.

[25] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 407–414.

[26] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 3241–3248.

[27] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato, "Coupling eye-motion and ego-motion features for first-person activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2012, pp. 1–7.

[28] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 2847–2854.

[29] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 314–327.

[30] T. Ishihara, K. M. Kitani, W.-C. Ma, H. Takagi, and C. Asakawa, "Recognizing hand-object interactions in wearable camera videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2015, pp. 1349–1353.

[31] M. Cai, K. M. Kitani, and Y. Sato, "A scalable approach for understanding the visual structures of hand grasps," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2015, pp. 1360–1366.

[32] M. Cai, K. Kitani, and Y. Sato, "Understanding hand-object manipulation with grasp types and object attributes," in *Proc. Robot. Sci. Syst. Conf. (RSS)*, 2016, pp. 1–10.

[33] M. Cai, K. M. Kitani, and Y. Sato, "An ego-vision system for hand grasp analysis," *IEEE Trans. Human–Mach. Syst.*, vol. 47, no. 4, pp. 524–535, Aug. 2017.

[34] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 287–295.

[35] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1894–1903.

[36] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[37] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 1194–1201.

[38] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[39] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 3570–3577.

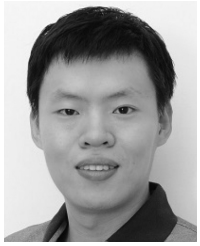
[40] A. A. Argyros and M. I. A. Lourakis, "Real-time tracking of multiple skin-colored objects with a possibly moving camera," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2004, pp. 368–379.

[41] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[42] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, 2000, pp. 61–74.
[43] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 3216–3223.
[44] S. Baker *et al.*, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, 2011.
[45] Y. Li, D. Min, M. S. Brown, M. N. Do, and J. Lu, "SPM-BP: Sped-up patchmatch belief propagation for continuous MRFs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4006–4014.
[46] J. Lu *et al.*, "Patchmatch filter: Edge-aware filtering meets randomized search for visual correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1866–1879, Sep. 2017.
[47] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

Minjie Cai received the B.S. and M.S. degrees in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2008 and 2011 respectively, and the Ph.D. degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2016.

He is currently a Post-Doctoral Researcher with the Institute of Industrial Science, University of Tokyo. His current research interests include computer vision and its applications on robotics and human–computer interaction.



Feng Lu (M'16) received the B.S. and M.S. degrees in automation from Tsinghua University, Beijing, China, in 2007 and 2010, respectively, and the Ph.D. degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2013.

He is currently a Professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing. His current research interests include computer vision and human computer interaction.

Yue Gao (SM'14) received the B.E. degree from the Department of Electronic Information Engineering, Harbin Institute of Technology, Harbin, China, and the master's and Ph.D. degrees from the School of Software, Department of Automation, Tsinghua University, Beijing, China.