SKILL MEASUREMENT VIA EGOCENTRIC VISION IN WETLAB

Licheng Yu, Yin Li, James M. Rehg

College of Computing, Georgia Institute of Technology
lyu40@gatech.edu

ABSTRACT

With the development in egocentric vision, skill measurement has been recently proposed as a novel topic in this emerging field. In this report, we record the the experimenter's first-person videos in Wet labratories (wetlab), and measure his/her operative skills. Specifically, given the videos of expert and amateur, we analyze their head motions, hands motions, eye-hand coordinations and key-moment actions. Accordingly, we investigate the cues that are discriminative between amateur and expert. The analytical results show that during the experiment expert is more likely to involve faster head/hands movements, more consistent eye-hand manipulation and more stable pipetting actions compared with amateur.

1. INTRODUCTION

With the advent of wearable cameras, such as Google Glass and SMI glasses, there have been increasing research interests in egocentric vision. In egocentric vision, we aim to analyze the captured first-person video. In comparison to that traditional camera system, the first-person vision has several advantages — consistent view point, high quality recorded video and little occlusion for objects. Based on these advantages, there have seen several promising applications: Fathi *et al.* [1, 2] tried to understand daily actions in the kitchen environment, Li *et al.* [3] predicted gaze points via statistical head-hands-eye relations and applied it to saliency detection and activity recognition.

In this report, we focused on measuring the operative skills of the experimenter in wetlab. Wetlab is a laboratory where chemicals, drugs, or other biological matter are handled in liquid solutions or volatile phases. The experimenters often need to operate the specialized pipettes, droppers and tubes in wetlab. These operations usually require delicate manipulation and rich experience. As a result, the success of an experiment is highly dependent on the experimenter's skills. Our goal is to monitor the experimenter's operating process from his/her first-person perspective, measuring his/her experiment skills and giving the confidence of his/her experiment results.

Before realizing the long-term goal, we currently focus on a simpler case — skill measurement. We investigate egocentric cues that are possible to discriminate the expert from amateur. Specifically, we find the following ones:

- Expert has faster head motion than amateur.
- Expert needs less time doing the same experiment step than amateur.
- Expert has faster hands motion than amateur.
- Expert's hands appear more ofthen than amateur.
- The hand-eye coordination of expert is more consistent compared with amateur.
- Expert can perform the action of pipetting with a more stable and consistent pattern.

2. EXPERIMENT SETTING AND SUBJECTIVE EVALUATION

2.1. Data Collection

The data were collected as follows: The experimenters were required to put on a wearable camera (glasses), doing the experiment according to the protocols shown on a monitor. We recorded the videos on three types of experiments — mixing, PCL and pipetting. Among them, mixing is relatively easier and pipetting is relatively more challenging. For each type of experiment, we have a pair of videos of expert and amateur. Thus, we have six videos in total, the average length of which is ten minutes and the frame rate is 24 fps. In Table 1, we present more details about the experiment setting, including the steps, corresponding frames and time needed for both amateur and expert.

2.2. Subjective Evaluation

Given the video segments in Table 1, we performed subjective evaluations on them. The evaluations were conducted on two types of videos: 1) without gaze information. 2) with gaze information. We randomly picked up a pair of segments from Table 1 and let subjects to provide subjective comparison. With the help of 8 subjets on the 27 pairs, we collected 48 responses on who is more professional and why he/she is more professional. The evaluation results are

Table 1. Experiment setting.

Mixing	time$_{amateur}$	frame$_{amateur}$	time$_{expert}$	frame$_{expert}$
act1 place items	00:03-00:40	721:1795	00:03-00:03	3250:4000
act2 label	01:15-02:34	1796:370	02:47-04:32	4001:6538
act3 add water	02:3503:22	3710:4853	04:33-05:10	6539:7438
act4 glycerol	03:23-06:12	4854:8930	05:11-07:32	7439:10854
act5 measure	06:13-10:39	8931:15344	07:33-11:20	10855:16327
act6 add water	10:40:-11:18	15345:16292	11:21-11:51	16328:17071
act7 mix	11:19-12:07	16293:17450	11:52-12:12	17072:17578

PCL	time$_{amateur}$	frame$_{amateur}$	time$_{expert}$	frame$_{expert}$
act1 prepare	00:51-01:24	1227:2010	00:55-01:13	1330:1750
act2 add solution	01:25-05:15	2011:7567	01:14-03:29	1751:5030
act3 centrifuge	05:16-05:40	7568:8160	03:30-03:50	5031:5530
act4 wait	05:41-06:34	8161:9470	03:51-04:56	5531:7115
act5 get tubes	06:35-07:10	9471:10308	04:57-05:29	7116:7900
act6 add solution	07:11-08:08	10309:11718	05:30-06:12	7901:8942
act7 centrifuge	08:09-08:26	11719:12154	06:13-06:24	8943:9220
act8 wait	08:27-09:24	12155:13547	06:25-07:31	9221:10827
act9 get tubes	09:25-09:53	13548:14252	07:32-07:45	10828:11150
act10 add solution	09:54-11:27	14253:16480	07:46-08:26	11151:12150
act11 centrifuge	11:28-12:02	16481:17285	08:27-08:40	12151:12500
act12 wait	12:03-13:05	17286:18846	08:41-09:49	12501:14144
act13 get tubes	13:06-13:23	18847:19280	09:50-11.67	14145:14680

Pipetting	time$_{amateur}$	frame$_{amateur}$	time$_{expert}$	frame$_{expert}$
act1 write	00:37-00:46	880:1110	01:26-01:35	2070:2275
act2 open lids	00:47-01:02	1111:1495	01:42-01:56	2460:2775
act3 add solution	03:00-03:23	4331:4870	02:10-02:25	3125:3490
act4 add solution	03:27-03:41	4970:5300	02:28-02:45	3550:3970
act5 add solution	03:42-04:00	5301:5770	02:47-03:11	4020:4580
act6 add solution	04:57-05:08	7130:7400	03:44-03:48	5375:5465
act7 add solution	05:23-05:33	7750:7980	03:58-04:03	5730:5833

shown in Table 2. As can be observed, there are 31 correct judgements on the experimenter's skilfullness. Moreover, the subjective judgements becomes more reliable with the difficulty of an experiment, as we got 8/15 (53%) accuracy on mixing and 8/12 (67%) accuracy on pipetting. From the 48 responses, we also collected 41 reasons supporting their judgements. They are mainly divided into four types—speed-oriented, hand-oriented, action-oriented and gaze-oriented, among which the speed-oriented reasons account for a significant proportion. Specifically, they are based on quicker picking/pipetting/dumping actions and faster head motion.

3. HEAD MOTION AND HANDS MOTION

Based on the subjective evaluations, we get the hypothesis that the movement of head and hands are possible cues for discriminating expert's and amateur's skills. Thus, in this section, we analyze the importance of head motion and hands motion.

3.1. Head Motion Analysis

3.1.1. Hypothesis and Problem Setting

In wetlab setting, the camera is mounted on the experimenter's head which continuously captures the scene in the front of him/her, thus the head movement can be roughly estimated via the global motion vector of the captured images. Considering there are substantial motion in egocentric videos, we apply Large Displacement Optical Flow (LDOF) [4] to estimate the motion field between each two consecutive frames. We denote the optical flow at the k-th frame as the follows,

$$w_k = (u_k, v_k), \tag{1}$$

where u_k, v_k are the displacements along x, y-axis respectively. Accordingly, the optical flow magnitude field for each

Table 2. Subjective evaluation: 8 testers on 25 video comparisons.

Experiment	#True	#False	#Unknown
Mixing	8	5	2
PCL	15	3	2
Pipetting	8	1	3

Reason type	#reasons	Details
Speed-oriented	23	faster actions, e.g., dump, pipette
Hand-oriented	6	steady, less trembling
Action-oriented	7	more or less actions
Gaze-oriented	5	gaze fixation

frame is $m_k = \sqrt{u_k.^2 + v_k.^2}$. Then, the head motion is computed as its median,

$$\text{head}_k = \text{median}(m_k). \tag{2}$$

3.1.2. Statistical Analysis

Given the head displacement of each frame, we first plot the distribution of head_k, $k = 1, ..., N$, where N is the total frame number for some video segment. We observe that most of the distributions are heavy-tailed, following the exponential distributions which are shown in Fig. 1(a). For simplicity, we use the mean of exponential distribution to describe the average head motion,

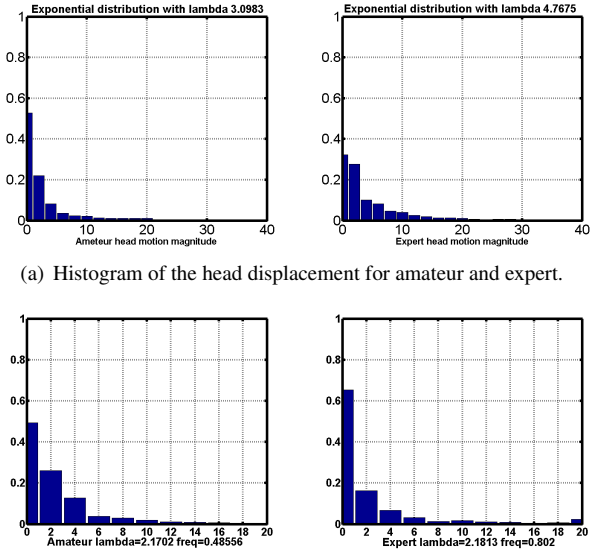
$$\lambda = \frac{\sum_{k=1}^{N} \text{head}_k}{N}. \tag{3}$$

Here, λ is also the parameter for the exponential distribution. Typically, the larger value of λ indicates faster head motion on average.

In Table 3[1], we compute λ's and the time durations for all pairs of video segments. We observe there are 15 out of 23 cases (65.2%) that experts' average head motion are faster than amateurs', and 21 out of 23 cases (91.3%) that expert needs less time than amateur for the same experiment step.

We then analyze the discrimination power of λ on random clips of the videos. Considering the videos are of different lengths for amateur and expert, we extract the same portion from the videos for comparison. Analyzing on 1000 random clips, we get 59% cases of λ_{expert} greater than $\lambda_{amateur}$ for Mixing; 77% cases of λ_{expert} greater than $\lambda_{amateur}$ for PCL; and 96% cases of λ_{expert} greater than $\lambda_{amateur}$ for Pipetting. For each experiment, we further plot the histogram of λ for the 1000 clips in Fig. 2(a, b, c). As can be observed, the discrimination power of head motion is increased with the difficulty of an experiment.

[1]PCL act1, act4 and act12 are discarded since there are no experiment-related actions.



(a) Histogram of the head displacement for amateur and expert.

(b) Histogram of the hands displacement for amateur and expert.

Fig. 1. Distribution of motion magnitude.

The above statistical results demonstrate the assumption that expert involves more faster head movements than amateur doing the same experiment, especially for the difficult ones. Moreover, expert generally needs less time to finish it.

3.2. Hands Motion Analysis

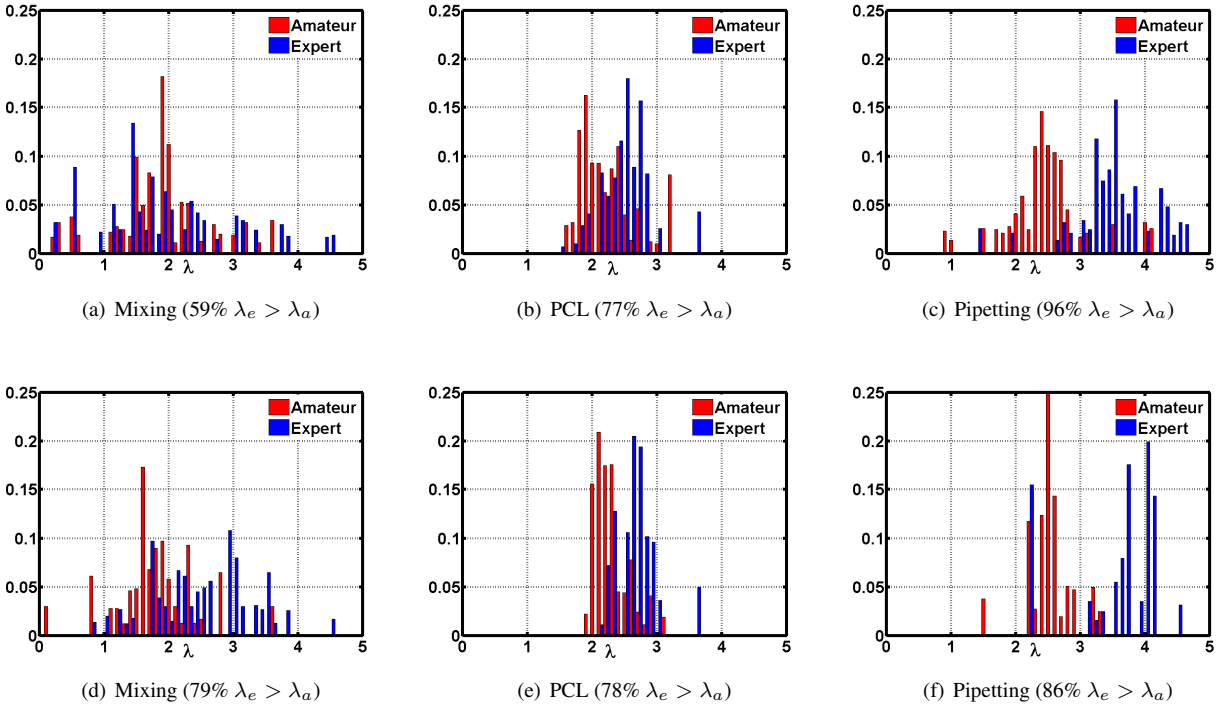
3.2.1. Hypothesis and Problem Setting

In addition to head motion, hands motion is also a crucial cue in Section 2. Intuitively, the expert should manipulate objects more efficiently than amateur, in which case, expert has faster hands motion than amateur. In addition, expert might take less time looking for tools or looking at monitors, and take more time working on experiments. This makes the expert's hands appear more often than amateur's.

We first apply the algorithm in [5] to segment hands. Specifically, the hands are modeled using color, texture, shape and global appearance features, and then we extract 210 images with hands and 60 images without hands to train the model. Some of the segmentation results are shown in Fig. 3. It is observed that some objects with similar skin colors might be wrongly segmented as hands. But overall, the method in [5] achieves encouraging segmentation performance on the wetlab videos.

3.2.2. Statistical Analysis

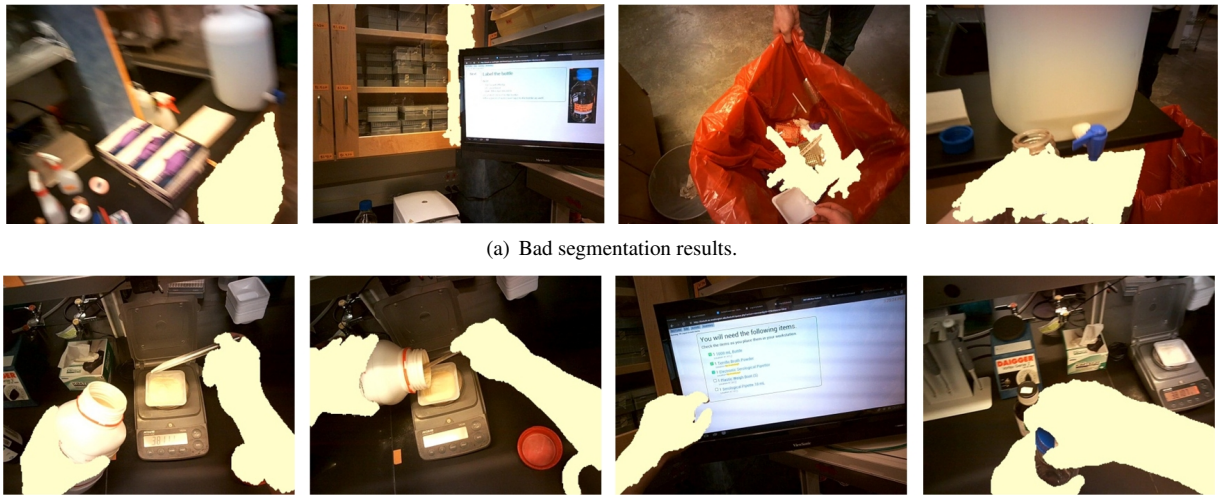
We denote the segmented hands mask at the k-th frame as m_k, then the hands motion can be estimated by computing the median of the optical flow field w_k within m_k. Similar to Section 3.1, we still observe that most histograms of the hands displacement are exponentially distributed, as shown



(a) Mixing (59% $\lambda_e > \lambda_a$) (b) PCL (77% $\lambda_e > \lambda_a$) (c) Pipetting (96% $\lambda_e > \lambda_a$)

(d) Mixing (79% $\lambda_e > \lambda_a$) (e) PCL (78% $\lambda_e > \lambda_a$) (f) Pipetting (86% $\lambda_e > \lambda_a$)

Fig. 2. Histograms of head motion λ in Figure (a, b, c), and histograms of hands motion λ in Figure (d, e, f).



(a) Bad segmentation results.

(b) Good segmentation results.

Fig. 3. Hands segmentation results using the algorithm in [5]. The segmented hands are painted with carnation color.

in Fig. 1(b). Thus, we can also simply use the mean λ to represent the hands motion,

$$\lambda = \frac{\sum_{k=1}^{N} \text{hand}_k}{N_h}, \tag{4}$$

where N_h is the number of frames containing hands. In Table 4, we list the values of $\lambda_{amateur}$ and λ_{expert} for each experiment step, where there are in all 16 out 23 cases (69.6%) with λ_{expert} larger than $\lambda_{amateur}$. We then analyze the discrimination of hands motion λ on random clips. After computing λ on the 1000 random clips for both amateur and expert, we plot the histograms of λ in Fig. 2(e, d, f). It can be observed that there are 79% cases of λ_{expert} greater than $\lambda_{amateur}$ for Mixing; 78% cases of λ_{expert} greater than $\lambda_{amateur}$ for PCL; and 86% cases of λ_{expert} greater than $\lambda_{amateur}$ for Pipetting.

Table 3. Head motion analysis. The 1st column indicates the steps, the 2nd-3rd columns show the mean of head displacement, and the 4th-5th columns show the time needed for each step.

Mixing	$\lambda_{amateur}$	λ_{expert}	$t_{amateur}$(s)	t_{expert}(s)
act1	3.8	**5.32**	35	**25**
act2	3.0	**4.76**	**63**	84
act3	2.3	**2.7**	38	**30**
act4	1.4	**1.7**	135	**113**
act5	**1.92**	1.50	213	**182**
act6	**2.56**	2.19	32	**24**
act7	3.36	**4.51**	38	**16**

PCL	$\lambda_{amateur}$	λ_{expert}	$t_{amateur}$(s)	t_{expert}(s)
act2	1.76	**1.89**	185	**109**
act3	**1.80**	1.75	19	**16**
act5	2.86	**3.70**	27	**26**
act6	1.53	**1.98**	47	**34**
act7	**2.18**	1.91	14	**9**
act9	**2.68**	2.09	23	**10**
act10	**2.61**	1.68	74	**33**
act11	**2.38**	2.01	26	**11**
act13	2.70	**4.14**	**14**	17

Pipette	$\lambda_{amateur}$	λ_{expert}	$t_{amateur}$(s)	t_{expert}(s)
act1	1.21	**1.66**	7	**6**
act2	1.43	**1.84**	12	**10**
act3	**2.20**	1.65	18	**12**
act4	1.56	**2.32**	11	**14**
act5	1.63	**3.34**	26	**18**
act6	1.10	**1.62**	11	**4**
act7	0.99	**1.92**	10	**5**

Table 4. Hands motion analysis. The 1st column is the step, the 2nd-3rd columns show the mean hands displacement, and the 4th-5th columns show the frequency of hands appearance.

Mixing	$\lambda_{amateur}$	λ_{expert}	$f_{amateur}$	f_{expert}
act1	3.40	**4.14**	0.35	**0.61**
act2	2.24	**3.96**	**0.68**	0.52
act3	2.17	**2.18**	0.48	**0.80**
act4	1.30	**3.67**	**0.68**	0.22
act5	**1.78**	1.43	0.80	**0.85**
act6	**3.60**	1.98	0.48	**0.87**
act7	3.89	**5.03**	0.61	**0.72**

PCL	$\lambda_{amateur}$	λ_{expert}	$f_{amateur}$	f_{expert}
act2	1.96	**2.28**	**0.74**	0.72
act3	2.09	**2.14**	0.67	**0.89**
act5	2.84	**3.63**	0.71	**0.77**
act6	1.79	**2.22**	0.80	**0.81**
act7	**2.68**	2.34	0.79	**0.93**
act9	2.98	**3.35**	0.78	**0.94**
act10	**2.58**	1.69	0.71	**0.81**
act11	**2.97**	2.40	0.78	**0.87**
act13	2.27	**4.04**	0.66	**0.86**

Pipette	$\lambda_{amateur}$	λ_{expert}	$f_{amateur}$	f_{expert}
act1	**2.13**	1.97	0.62	**0.66**
act2	1.58	**1.92**	**0.56**	0.34
act3	**2.17**	1.74	0.83	**0.95**
act4	1.38	**2.26**	0.67	**0.97**
act5	1.23	**3.62**	0.71	**0.86**
act6	0.79	**1.63**	0.92	**1.00**
act7	1.07	**2.06**	0.94	**1.00**

Second, we estimate how often hands appear during each experiment step. The frequency of hands appearance is estimated using N_h/N. We list the frequencies in the 4th and 5th columns of Table 4. As can be seen, there are 19 out of 22 cases (86.4%) that expert's hands appear more often than amateur's.

Based on the above analysis, we conclude that expert tends to be more focused on experiments, with faster hands motion and more frequent appearance of hands.

4. EYE-HAND COORDINATION ANALYSIS

Eye-hand coordination has been studied in many psychology and vision literatures. The fact that eye gaze and hands movement are generally coupled was applied in [2] for recognizing daily actions. Land discovered that eye fixation may precede hand movement by a fraction of second [6]. A more regular, rhythmic pattern of eye, head and hand movements for natural tasks was found in [7]. However, the coordination of eyes and hands is variable under different situations [8]. In this section, we analyze the relation between the manipulation point and gaze point in wetlab.

4.1. Manipulation point

Instead of modeling hands with various pose templates, we introduce manipulation point by analyzing the hand shapes. The manipulation point is defined as a control point where the first person is mostly likely to manipulate the object using his/her hands [3]. For example, for a single left hand, manipulation usually falls on the right tip of the hand; for a single right hand, manipulation usually happens on the left tip of the hand; for two intersected hands, the manipulation point is generally around the intersecting part; for two separated hands, a manipulation point usually falls on the middle of the two hands. These four examples can be seen in Fig. 4, where the manipulation points are marked with red circles. In Fig. 4, we also add to each frame the ground truth of gaze point which is marked with green crosses. The relation between the manipulation point and gaze point was studied in [3], where

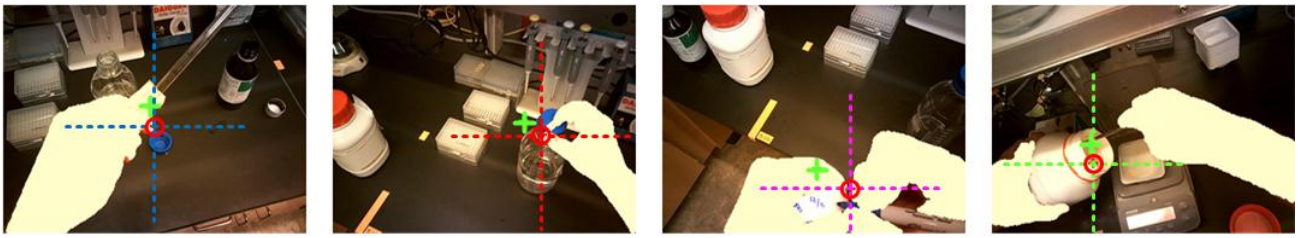


Fig. 4. Manipulation point and gaze point for the case of left hand, right hand, intersected hands and separated hands. The manipulation point is marked with red circle, and the gaze point is marked with green cross.

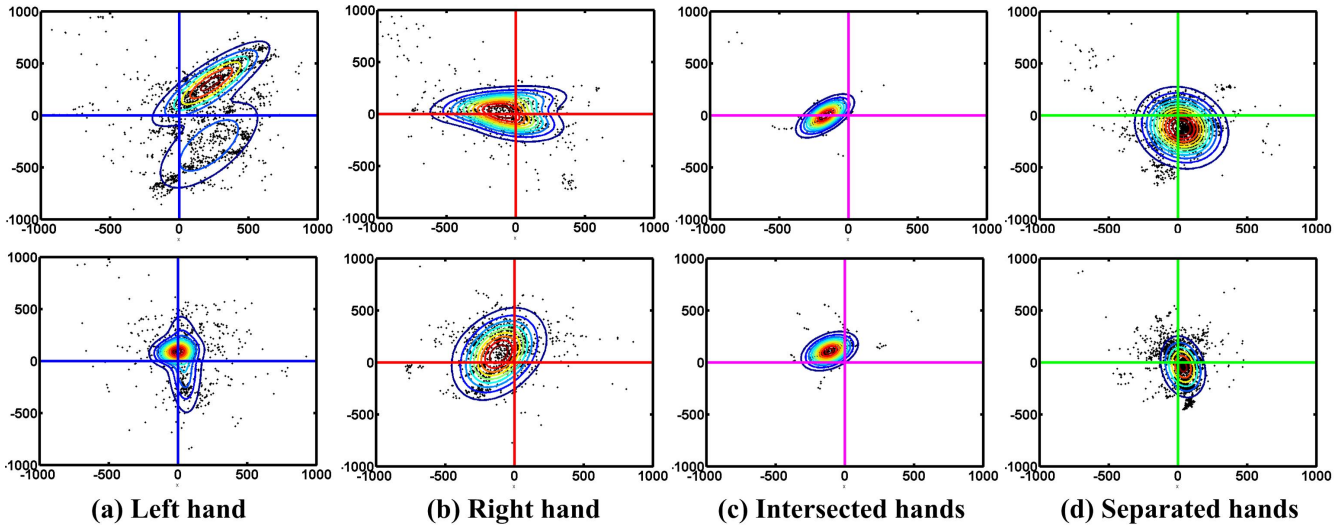


Fig. 5. Aligned gaze density map, where the first row is for amateur and the second row is for the expert. We align the gaze points into hand's coordinates by selecting the manipulation points as the origin, and projecting the gaze point into the new coordinate system every frame. We then plot the density map by averaging the aligned gaze points across all frames within the dataset. Higher density clusters can be found around the manipulation points.

the spatial locations of the two points are usually consistent as people tend to look at the object they are manipulating.

4.2. Statistical Analysis

Intuitively, the expert's hands should coordinate better with his/her eyes than the amateur. To validate this assumption, we align the gaze points into the hand's coordinates where the manipulation points are set as the origins, and plot the density maps in Fig. 5[2]. We then fit the gaze points density using Gaussian mixture models (GMM) with the number of clusters equal to 2. As shown in Fig. 5, higher density clusters can be found around expert's manipulation points, which demonstrates the expert can better coordinate eyes and hands than amateur.

[2]We only plot the manipulation-gaze density map for Mixing, as the ground truths of gaze points in PCL and Pipetting are interfered because of the eye color or hardware issue.

5. ANALYSIS ON KEY MOMENT—PIPETTING

5.1. Hand Pose for pipetting

There are many operations for conducting experiments in wetlab. Among them, the key moment is manipulating pipette to add liquid solutions into tubes — pipetting. In this section, we analyze the video sequences involving the pipetting action only. From the given six videos, we observe that there are in all four types of poses used for pipetting, as shown in Fig. 6. Among them, the pose in Fig. 6(a) appear more often than others. Thus, we extract the sequences with the pose template in Fig. 6(a), and investigate the coordination between the tube and pipette. From Fig. 7(a)(b), we observe that amateur is accustomed to lean the pipette onto the tube lid during pipetting. This action would make some angle between the principle axes of tube and pipette, and thus might leave liquids onto the wall of the tube. On the contrary, the expert is able to hang the pipette with the same angle of tube's principle angle. In this case, the principle axes of tube and pipette are

Fig. 6. Different hand poses for pipetting.

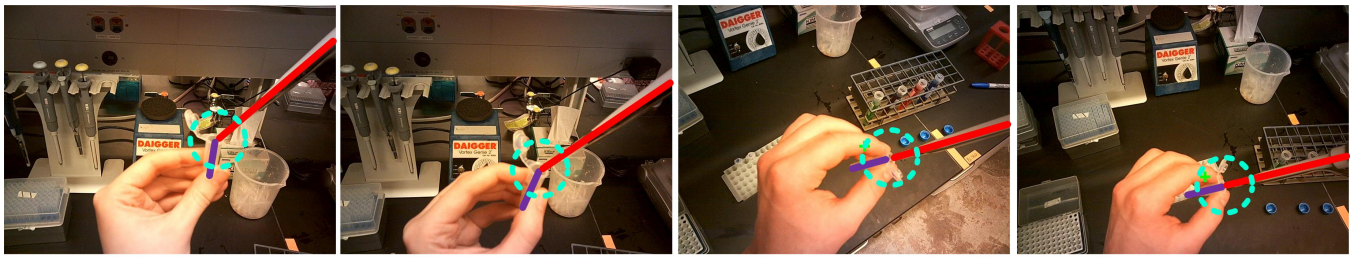


Fig. 7. Different poses of handling pipettes and tubes during pipetting, where Figures (a)(b) are the amateur's hand poses, and (c)(d) are expert's hand poses. The red line marks the principle line of pipette, and the purple line marks that of tube. It can be observed while the principle axes of tubes and pipettes are intersected with some angle, the two axes are well aligned for expert.

Table 5. Time needed for pipetting.

	$\text{frame}_{amateur}$	frame_{expert}	$\text{t}_{amateur}$(s)	t_{expert}(s)
pipetting1	4771:4825	3410:3450	2.25	**1.67**
pipetting2	5157:5204	3889:3927	1.96	**1.58**
pipetting3	5620:5662	4510:4545	1.75	**1.46**
pipetting4	7192:7366	5405:5435	7.25	**1.25**
pipetting5	7795:7910	5766:5804	4.79	**1.58**
variance	–	–	5.67	**0.03**

well aligned, as shown in Fig. 7(c)(d). So the liquid from expert's pipette can be directly pipetted into the bottom of tube. Second, we compute the time for amateur's and expert's pipetting action in Table 5[3]. It can be observed that expert can finish this action far more quickly than amateur. Besides, the time variance for expert is also smaller, which highlights the stability and consistency of expert's operation.

6. CONCLUDING REMARKS

Given the six videos, we conduct analysis on head motion, hands motion, eye-hand coordination and key-moment of pipetting action. We summarize the observation into the following six cues:

- Expert is more likely to involve faster head motion and hands motion compared with amateur.
- The time needed for expert doing each experiment step is shorter than amateur.
- Expert can manipulate objects faster than amateur
- Expert's hands appear more often finishing each experiment step.
- Expert can coordinate eyes and hands better than amateur.
- Expert has more stable and consistent pattern for pipetting than amateur.

Future work includes further validating these cues with more data, and introducing object detection and action recognition into this project.

[3]The actions for pipetting are extracted from act3 to act7 of Pipetting in Table 1.

7. REFERENCES

[1] Alireza Fathi, Ali Farhadi, and James M. Rehg, "Understanding egocentric activities," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 407–414.

[2] Alireza Fathi, Yin Li, and James M. Rehg, "Learning to recognize daily actions using gaze," in *Computer Vision–ECCV 2012*, pp. 314–327. Springer, 2012.

[3] Yin Li, Alireza Fathi, and James M. Rehg, "Learning to predict gaze in egocentric vision," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013.

[4] Thomas Brox and Jitendra Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 3, pp. 500–513, 2011.

[5] Cheng Li and Kris M Kitani, "Pixel-level hand detection in ego-centric videos," in *Computer Vision and Pattern Recognition, 2013 IEEE Computer Society Conference on*. IEEE, 2013.

[6] Michael F Land and Mary Hayhoe, "In what ways do eye movements contribute to everyday activities?," *Vision research*, vol. 41, no. 25, pp. 3559–3565, 2001.

[7] Jeff Pelz, Mary Hayhoe, and Russ Loeber, "The coordination of eye, head, and hand movements in a natural task," *Experimental Brain Research*, vol. 139, no. 3, pp. 266–277, 2001.

[8] Michael F Land, "The coordination of rotations of the eyes, head and trunk in saccadic turns produced in natural situations," *Experimental Brain Research*, vol. 159, no. 2, pp. 151–160, 2004.