16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Hotspots Integrating of Expert and Beginner Experiences of Machine Operations through Egocentric Vision

Longfei Chen[1] , Yuichi Nakamura[1] , Kazuaki Kondo[1], Dima Damen[2], Walterio W. Mayol-Cuevas[2]
Academic Center for Computing and Media Studies, Kyoto University[1]
Department of Computer Science, University of Bristol[2]
yuichi@media.kyoto-u.ac.jp[1], {Dima.Damen}{Walterio.Mayol-Cuevas}@bristol.ac.uk[2]

Abstract

Beginners can often provide useful information regarding machine operations, e.g., how difficult a specific operation is or how beginners deal with difficulties. However, their experiences oftentimes widely vary, and it is difficult to summarize these diverse experiences in an extensive way. To try and solve this problem, this study focused on developing a framework for integrating beginners' and experts' experiences into a unified operation model. A baseline model was first obtained based on hand–machine interactions automatically extracted from experts' egocentric vision records. Then, the beginners' behaviors were integrated by dynamically aligning them to the baseline model. Through this process, an integrated model based on the experiences of a wide range of users was achieved. We applied our method to the operation experiences of two tabletop devices, an IH heater and a sewing machine. The results show good potentials in modeling the common and different behaviors among experts and beginners.

1 Introduction

With the flourishing development of online educational resources like Khan Academy and YouTube, increasingly more people are sharing their skills and experiences through videos at anytime and anywhere. Many studies have explored automatic guidance for guiding novices in residential or working environments such as in an office [3] or kitchen [4]. If a novice follows each step of the expert's process, there will be a high probability of success.

Nowadays, with the emergence of wearable devices, e.g., smart glasses and active cameras, recording experiences in a more human-centric way, i.e., through first person vision (FPV) or egocentric vision, is possible. An FPV experience can be more intuitive, as it provides what the wearer is seeing with less occlusion [8], and because of this, the FPVs of experts are expected to serve as good learning material for novices.

However, unlike manually-made video tutorials, expert's naturally-recorded experiences are often insufficient as learning materials for the following reasons:

1. Actions not appropriate for beginners: Experts often perform tasks quickly without explicit checking or confirmation, which are essential for novices [1]. They tend to skip checking the results, because they already familiar with most of the possible outcomes. Similarly, they often divert their attention to the next target while working towards the current target, i.e., they prepare for the next step in an optimized manner.

2. Not providing the easy way out: Experts often choose a method that is efficient, which may require skillful behaviors that are not easy for beginners.

3. Not providing enough diversity: Teaching manuals are expected to cater to the differences of personal knowledge, conditions, and environments. However, catering to such disparities is difficult when only expert behaviors are considered.

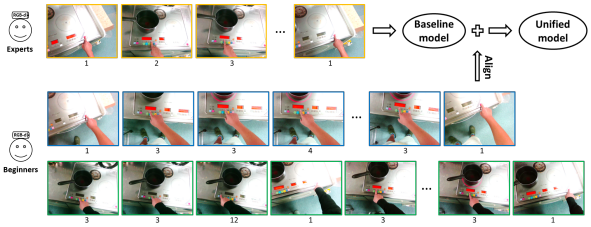


Figure 1: The framework of capturing and modeling machine operation experiences through egocentric vision. The example shows that an expert and two beginner wear a RGB-D camera recording their experiences using an IH heater. The number below each shot is the index of the hotspots.

To ameliorate the above problems, beginner experiences were integrated with expert experiences. Utilizing beginner experiences has the following advantage: they tend to pay more attention on the results of each action and perform each step slowly and carefully [10], which may provide better learning materials when successful; and they may also discover an easier method that is more suitable for other novices.

2 Related works

Several studies have investigated user guidance in daily tasks using expert experiences [3, 4, 9]. Reiko et al. developed a cooking navigation system [4], wherein a cooking process is decomposed into Action Units, and multimedia-based guidance is provided. Zhuo et al. [9] developed a wearable cognitive-assistant system. The guiding instructions are generated from downloaded tutorial videos and indexed by their titles and descriptions. These systems are well-designed guidance tools. However, their guidance schemes are relatively static,

e.g., the order of explanation is fixed, and the data required for guidance are manually prepared.

To reduce the burden of manual data collection, the automated acquisition of guidance data has been investigated. Dima et al. proposed a method of automatically integrating multiple experts' experiences and providing video guidance through a Google Glass [3]. This method recognizes task-relevant-objects and their modes-of-interactions by user's attention fixation. Chen et al. [2] built models for machine operation tasks by automatically extracting the temporal interactions using hand shape and touch.

Although the expert behaviors are good guidance resources, they have some insufficiency (mentioned in previous section). Thus far, to our knowledge, there are no published studies utilizing beginner-related data for automated guidance generation. However, there are some preceding works that compare two or more experiences. Zhang et al. reported a video-based evaluation of skills in surgical training, assuming that a newer behavior pattern demonstrated more skill compared to an older behavior pattern [5]. Hazel et al. proposed a supervised deep ranking model to determine skills in video records in a pairwise manner [6].

To compare and evaluate the behaviors, temporal features such as spatio-temporal interest points [5] and a two-stream CNN [6] are used for finding the correspondence between two or more experiences, that is, sequences of actions. However, these features cannot identify every step of a task step and lack semantic explanations. Attention cues are used in the above mentioned works [3] for guidance-data acquisition, and Chen conjectured that a hotspot could be a better feature for this purpose [2]. Thus, a method for automatically integrating both expert and beginner experiences based on hotspots was investigated, and hidden Markov models (HMMs) were used to model the temporal structures of the experiences.

3 Key Idea

In this work, we focus on common everyday machines, such as printers, rice cookers, and DIY tools. Operating these machines is not easy for first-time users, and detailed instructions are often required when using advanced functions. Here, we chose an IH heater to represent flat 2D devices with control panels, while choosing a sewing machine to represent 3D tabletop machines.Those tasks comprise a sequence of hand–machine interactions, e.g., *push buttons, seize a lever, rotate a knob.* Such tasks are complex enough and the included interaction patterns are diverse enough to represent everyday-machine operations. More importantly, both of the tasks can be done with a degree of freedom (DoF), which is, several interactions can be substituted by others or their orders can be changeable.

The convenience of FPV for recording operation experiences is utilized, aiming at automatically extracting features from the experiences and summarizing them to create an operation model. Obtained models are expected to properly describe the task process and be used for guidance applications and behavior prediction.

The aim of this work is to enrich a task model with beginners experiences as well as experts experiences, which are often insufficient for creating an extensive task model. The challenge lies in dealing with the beginners' noisy or incomplete data and integrating them.

In order to integrate experiences from FPV videos, we need descriptions that can be compared with one another. For this purpose, we use *hotspots*, i.e., crucial locations on a machine, and *interaction patterns* that can properly summarize the actions, i.e., *where*, *when*, and *how* an hand–machine interaction happens.

There are also difficulties that arise from the arbitrariness and redundancies in beginners behaviors, such as *unnecessary/missing* interactions, *mistakes*, *D*oF of interactions, and individual *disparities* of dealing with a given task. To deal with these variations, a dynamic alignment approach was adopted. First, a baseline temporal model was composed with only experts experiences. Then, the alignment between each beginner's experience to the baseline model is calculated and added to the baseline model. Repeating this process for all beginners experiences, the final unified model with enriched information is obtained.

4 Interaction Detection and Integration

As illustrated in Figure 1, the FPV experiences are firstly recorded with an RGB-D camera attached to the user's head. Then, the following processes are applied.

4.1 Hotspots Detection

Valid interactions, i.e., physical contacts, are detected from FPV records utilizing the distance between hand and machine. Each location of such interaction is considered to be a hotspot, and each hotspot is stored with its interaction pattern. The sequences of such temporal interactions on hotspots enable us to characterize and make correspondences between FPVs. To locate the hotspot from FPV to global location, global maps of the machine such as an IH heater and a sewing machine surfaces are constructed beforehand, and the FPV frame is matched to the global maps by estimate the camera pose and 2D registration. Details are given in [2].

4.2 Baseline Model

HMM is adopted to obtain the baseline model from expert experiences. First, a left-to-right HMM model is trained with all expert interaction sequences. The hidden state number is chosen to be the average length of all expert samples. After training, for any hidden state s_i with more than one observation, a replacement subnet is created, as follows:

$$s_i \rightarrow subnet : [s_i, s_{i+1}, \ldots, s_{i+m-1}]^T \quad (1)$$

where m is the number of observations in s_i. Then, the subnet is re-trained with all the observations from s_i among all the samples.

Through this process, the observation ambiguities of states are eliminated, i.e., each hidden state only outputs a single observation. Thus, the HMM model can express the essential DoF of the task, e.g., *alternative* and *order-changeable* actions in separate state transition branches. Additionally, the difficulty of determining the optimal number of states for training HMM is relaxed by subnets with adaptive configurations.

4.3 Finding Alignment

The alignment between a beginner's interaction sequence and the baseline model can be defined as:

$$\begin{aligned}\hat{A} &= \arg\max_{A} \quad Pr(A, O|\ \Theta) \\ &= \arg\max_{a_1^T} \prod_{t=1}^{T} Pr(a_t, o_t|\ \Theta)\end{aligned} \tag{2}$$

where O is the observation (interaction) sequences of the beginner with length T, and A is the assignment of the corresponding hidden states path for the beginner's interactions by the baseline model, respectively; and Θ is the parameter of the baseline model.

We assume the operation procedures of the task have inherent forward orders with certain DoFs. Therefore, the alignment of a current interaction has dependence on the alignment position of the previous interaction, which is a similar problem to the time alignment problem in speech recognition. We adopt the HMM-based word alignment model proposed in [7]:

$$\begin{aligned}Pr(a_t, o_t|\ \Theta) &= Pr(a_t, o_t|\ a_1^{t-1}, o_1^{t-1}, \Theta) \\ &= Pr(a_t, o_t|\ a_{t-1}, \Theta) \\ &= \sum_{a_{t-1}} p(o_t|a_t) * p(a_t|a_{t-1}) * p(a_{t-1}|\ \Theta)\end{aligned} \tag{3}$$

For alignment, a recursion formula can be used:

$$\begin{aligned}Q(t) &= \max_{a_t}\ Pr(a_t, o_t|\ \Theta) \\ &= \max_{a_t}\ [p(o_t|a_t) * p(a_t|\hat{a_{t-1}})] * Q(t-1)\end{aligned} \tag{4}$$

then we have:

$$\hat{a_t} = \arg\max_{a_t}\ p(o_t|a_t) * p(a_t|\hat{a_{t-1}}) \tag{5}$$

where all the $p(o_t|a_t) \in \{0, 1\}$ by creating subnets.

In experiments, it is assumed that any hidden state can be the starting point of alignment. We utilize Dynamic Time Warping (DTW) to compare a beginner sequence with the baseline model. The process is described as follows: first, we find the best-match expert's observation sequence as:

$$\begin{aligned}(w_B, w_E) &= DTW(O_B, O_E), \\ \omega_E &= \arg\min_{w_E}(\mathbb{E}(w_E - w_B))\end{aligned} \tag{6}$$

where $O_B \in \mathbb{R}^I$, $O_E \in \mathbb{R}^J$ are the observation sequence of the beginner and an expert, respectively. And w_B and w_E are the corresponding sequences after warping, $\mathbb{E}$ is the Euclidean distance. The best-match expert sequence is represented by ω_E. Then, the first index of the same observation between the matched sequences $\hat{k}$ is derived, the hidden state corresponds to the first-matched observation from the baseline model is adopted as the starting point (a_1) of alignment:

$$\hat{k} = \arg\min_{k}(\omega_E{}^{(k)} == w_B^{(k)}), \quad a_1 = s_E^{(\hat{k})} \tag{7}$$

where s_E is the hidden states to ω_E in the baseline model, which can be derived by Viterbi algorithm.

In addition, we assumed the jump width for alignment same as [7], in which the maximum length of the forward or backward jump for operations is set to 3 hidden states by considering the task DoF. The transition probabilities $p(a_t|a_{t-1})$ in jump situations are adaptive set according to jump width. For the new appearing behavior patterns in beginners experiences, new states are added to the baseline model during alignment. The detailed alignment algorithm is shown in Algorithm 1.

Algorithm 1 Dynamic Alignment for Operation Interactions

Input: beginner's observation sequences $O\{o_1, o_2, \ldots, o_N\}$ for alignment, the baseline model (expert's HMM) M (prior π, emission matrix E, transition matrix T, state number m), probability constant δ, and the DoF of the task $\mathbb{D}$.
Output: best state-transition path $A\{a_1, a_2, \ldots, a_N\}$ corresponds to O.

```
for i = 1 to m do
    (a) add self-transition:
        T_1(i,i) += δ;
    (b) add forward-transitions (dynamic value based on forward-jump width):
    for f = 1 to 𝔻 do
        T_1(i, i+f) += 8 * δ / f;
    end for
    (c) add backward-transitions (dynamic value based on backward-jump width):
    for b = 1 to 𝔻 do
        T_1(i, i-b) += 1/8 * δ / b;
    end for
end for
Initial state a_1 ← DTW(O, M);
for t = 2 to N do
    seq_t ← [o_{t-1} o_t]; π_t(a_{t-1}) ← 1;
    path ← Viterbi(E_{t-1}, T_{t-1}, π_t, seq_t);
    if path exist then
        a_t ← path(end);
        T_t ← T_{t-1};
    else
        (d) add new hidden state:
        m ← m + 1; a_t ← m;
        T_t ← T_{t-1}(a_{t-1}, a_t) = δ;
        E_t ← E_{t-1}(a_t, o_t) = δ;
    end if
end for
```

4.4 Integration to a unified model

After alignment, a sequence of states corresponding to a beginner's observation sequence can be acquired. Then we can integrate the beginners' interactions and

the baseline to a unified model, i.e., the *extensive model*. All the different patterns of behaviors in beginners' experiences, i.e., *new methods, repeating interactions, order-changeable interactions, missing interactions*, are assigned with an equal constant probability $\delta(\ll 1)$ to be added to the baseline model.

Repeating this process for all beginners' state transition sequences, the network of the extensive model is obtained. The expert's behaviors manifests much higher probabilities than beginners', which indicate more creditable behaviors. However, if multiple beginner interactions share common patterns, the probabilities of the corresponding paths increase, i.e., common beginner behavior patterns are taken into account.

5 Experiment

5.1 Experimental Environment

26 records of nine people using an IH heater and 37 records of thirteen people performing a sewing machine operation task were gathered, respectively. 2 records of each task were performed by a professionally skilled expert, while the rest are from beginners with varying skill levels. The water boiling task using the IH heater is designed with 4 essential interaction operations, which includes only fixed procedures; while the sewing task is designed with 11 essential interaction operations, which included 4 pairs of order-changeable steps (total DoF is $2^4 = 16$). Color Markers are adopted for locating the hotspots on the 2D global map of the IH heater, because it has almost textureless surface, which brought difficulty in finding enough local features for image registration.

The recording device was a head-mounted RGB-D camera (Intel RealSense SR300) with 30 fps for both color and depth resources. The participants are only instructed with the task requirements before starting, e.g., "please boil the water" or "please get the sewing machine prepared, and sew the cloth with thread pattern A and speed B", then they can perform the task freely. Recordings were stopped when the participants finished the whole process or failed halfway.

For detecting hotspots and interaction patterns, the same parameters as in [2] were adopted, i.e., the depth threshold for detecting valid touches is ± 7 mm. For the integration, the expert baseline model was first directly trained; then, each beginner observation was integrated by assigning the small constant probability $\delta = 0.01$, and adding it to the original paths and stats of the baseline model. The parameter matrices of HMM were normalized after all the samples were integrated to ensure that all the probabilities lay between 0 to 1.

The ground truth of temporal interactions for evaluating the results was manually constructed for each participant. The ground truth of alignment of experiences is done by an expert who manually viewed all experiences and aligned each interaction to the baseline model. To evaluate the accuracy of interaction detection, the focus is placed on the essential interaction patterns which have appeared in both expert and beginner behaviors; the individually occurring patterns are considered to be noise and are ignored.

Table 1: The accuracy of temporal interaction Detection and Alignment.

	R	P	F	Essential	Aligned	Acc.
IH	0.76	0.91	0.83	103	96	93%
SW	0.67	0.88	0.76	252	239	94.8%

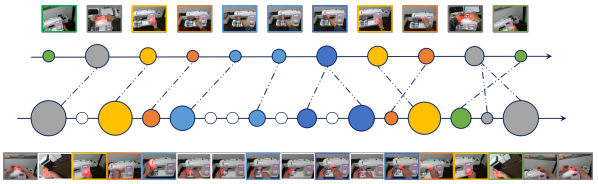


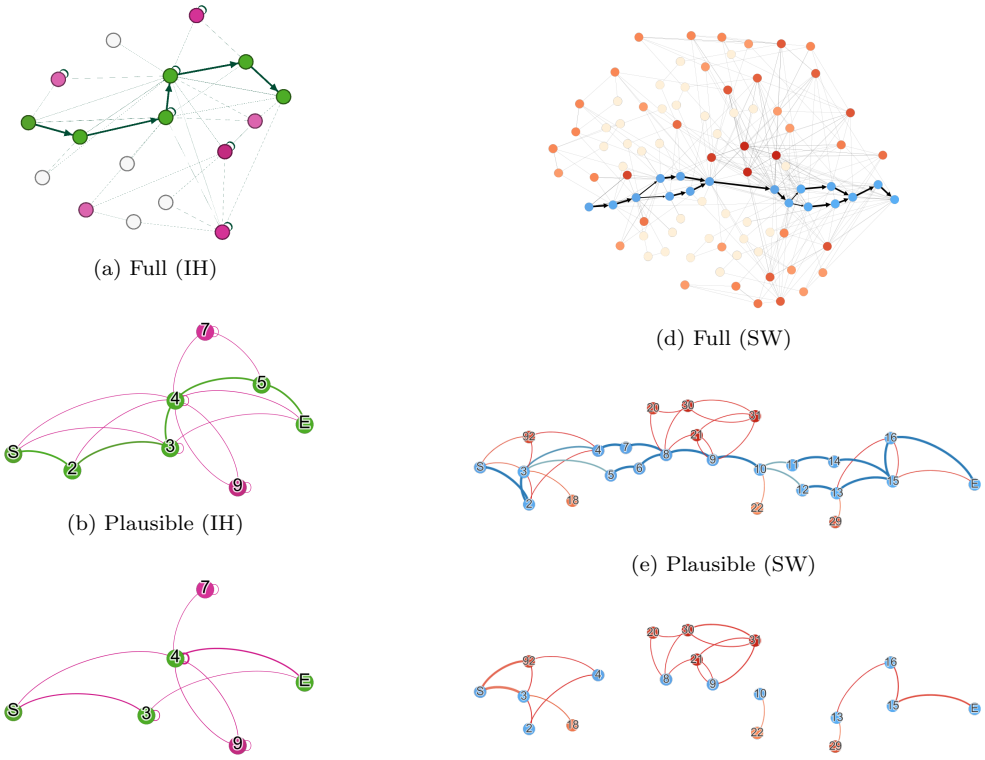
Figure 2: An example of alignment between the sewing machine operation experiences of an expert (top) and a beginner (bot). The size of the dots indicates the duration of each temporal interaction, and the color indicates the different patterns (where the *white dots* are new appeared interactions).

5.2 Results

We adopted $F_1 - score$ to evaluate the accuracy of interaction detection. The overall results are shown in Table 1. The F-score of detecting temporal interactions for all the experiences of the IH heater task and the sewing task are 0.83 and 0.76, respectively. The main reason for degrading the *Recall* is that beginners more frequently performed redundant or unnecessary touches than experts. Redundancies occasionally made it difficult to locate essential interactions with a hotspot. Additionally, beginners hotspot sometimes could be difficult to be matched to the corresponding location of the global map, due to the differences in the viewing angle and position, or the bad light condition. Typical examples are shown in Figure 4 (b - d).

The overall alignment accuracy of all essential interactions in all experiences is 94.3%. Two of the incorrect samples were entirely misaligned because the initial states were incorrectly located, when the beginners performed many wrong trials before the correct initial procedure. Another sample was successfully aligned for only the first half, as it was missing many essential steps in the second half due to the misdetection of temporal interactions. An example of alignment between an expert and a beginner in the sewing task is shown in Figure 2. The expert performed the task without any redundant interactions while the beginner had several unessential interactions. The equivalent interactions and the order-changeable steps (interactions) in the beginner's sequence were successfully matched to expert's, while the beginner-introduced interactions are located among the essential ones.

Figure 3 shows the results of integration. For the



(a) Full (IH) (d) Full (SW)

(b) Plausible (IH) (e) Plausible (SW)

(c) Differences (IH) (f) Differences (SW)

Figure 3: The models for integrating expert and beginner experiences for tasks of IH heater (IH) and sewing machine (SW). The expert baseline model is shown in *green (blue)* while the added beginner hidden states and transitions are shown in *purple (brown)*. The saturation of the nodes indicate the sum of In-Out transition probabilities of the nodes. (Top) The full model after integration of all experiences. (Middle) The high probability states and transitions ($> 3\delta$). (Bot) The differences between beginner and expert experiences.

Table 2: The semantic meaning of beginner-expert differences.

Types	Task	States Transitions	Semantic Meanings
new ways	IH:	$s_S \rightarrow s_4 \rightarrow (s_9) \rightarrow s_4 \rightarrow s_E$.	*"beginners discover a completely new way of achieving the task"*
	SW:	$s_3 \rightarrow s_2 \rightarrow s_4$, $s_{13} \rightarrow s_{16} \rightarrow s_{15}$.	*"beginners discover new orders of achieving several procedures"*
confirm/ supportive	SW:	s_{20}, s_{21}, s_{22}, s_{30}, s_{31}.	*"beginners rest hands on the cloth to feel the speed of sewing, or seize the cloth to confirm whether it can be pulled out or not"*
unnecessary	IH:	s_3 *self-repeating*, $s_4 \rightarrow s_7$.	*"beginners select the heat patterns repeatedly to choose the best one', "beginners try to cancel the procedure by* s_7 *but not working"*
common mistakes	IH:	$s_S \rightarrow s_3 \rightarrow s_E$.	*"jump from* s_S *to* s_3 *due to missing detection of one essential procedure in several experiences"*
	SW:	s_{18}, s_{29}.	*"beginners performed some trials before some certain procedures (push some wrong places on the machine surface)"*
other noises	SW:	s_{92}.	*"beginners performed some noisy operations before the starting step of the task"*

IH task and sewing task, the baseline model of experts contain 6 and 17 states, respectively, including manually added "start (S)" and "end (E)" states. The 4 states in IH task model ("S" and "E" excluded) show the one to one correspondence with the 4 essential interaction steps designed by the task. The 15 states in sewing task baseline model contains two branches, which covers 4 different routines (DoF) of achieving the task, with 11 interaction steps for each routine. The After the integration, as illustrated in Figure (a), (d), the number of states increased to 17 and 92, the variations of the model are largely increased. Figure (b) and (e) show the states and transitions with high probability ($> 3\delta$), which contains only common operations performed by multiple users, and the personal variations are removed; thus it becomes a more "plausible" description of the task. Figure (c) and (f) show the expert-beginner differences, such as the new states and transitions that appear only in the beginner' experiences. The detailed semantic meanings are manually extracted and shown in Table 2.

6 Discussion

From the acquired models and the manually-extracted semantic meanings of the expert-beginner differences, we can see the possible benefits of supplementing with beginner experiences .

Duration of explanation As illustrated in Figure 2, the duration of a beginner's interactions are generally longer than that of an expert's. Thus, a beginners video records may provide details that could otherwise be missed in the quick actions of an expert.

Providing Variations From Table 2, *new ways* and *different orders* of achieving the task are detected in beginner experiences. This does not only provide a full description of the task, but also contributes to the prediction of possible behaviors of an user.

Supplemental Explanations Moreover, the time-saving behaviors of experts, e.g., operating at great speed or without looking at the operating place at hand, lead to the detection failure of some steps. For example, the expert pushed the power button without looking as shown in Figure 4 (e) because he/she already knows where the button is, while the beginner needed to look at the button first as shown in (f). Beginner's experience provides a better explanation in this sense.

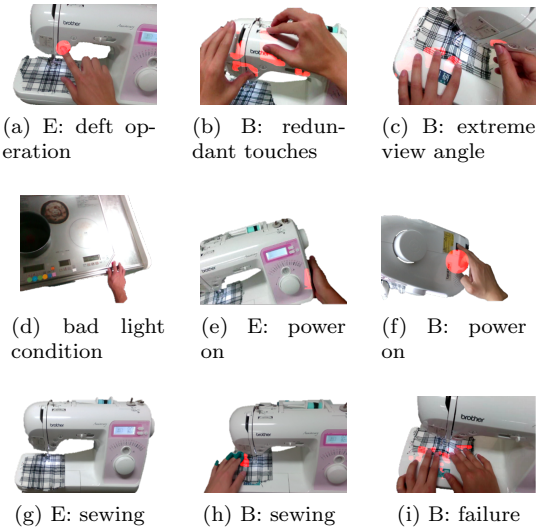
Furthermore, *confirmation* behaviors in beginners' experiences, e.g., watching the result of an action, provides good information of how a task is processed. Figure 4 (g), (h) shows confirmation behaviors different between experts and beginners. The expert is watching (g) the sewing process without any additional actions, while the beginner is checking the moving direction and speed of cloth by hand (h).

Common Mistakes Common mistakes that frequently appear in beginners' behaviors can be used to warn the user before he/she makes any actual mistake. In Figure 4(i), the beginner tried to pull out the cloth after the needle was up, which was not possible; the cloth would have torn if it was forcibly pulled. The incorrect choices of beginners can help better guide a user and provide a better understanding of the operations.

Limitations: However, the model integrates every single behavior of the beginners with an equal probability, thus the positive behaviors (e.g.,*new ways*) and negative behaviors (e.g., *mistakes*) to the task manifest both as common behaviors, which may cause difficulties for applications such as *guidance generation*.

7 Conclusion

In this work, a framework was proposed for automatically aligning beginners' and experts' machine operation experiences and integrating them to a extensive model. The performed experiments show that the utilized alignment and learning methods are sufficient for distinguishing and extracting useful information. By gathering beginners and experts machine operation experiences with this model, it is possible to provide an extensive description of a task, which can be used as good material for guidance generation/prediction, product design, etc. There is room for future work. In this paper, only physical hand–machine interactions were considered. Other important aspects of user behaviors, such as attention or sound/speech, should also be investigated.



(a) E: deft operation (b) B: redundant touches (c) B: extreme view angle

(d) bad light condition (e) E: power on (f) B: power on

(g) E: sewing (h) B: sewing (i) B: failure

Figure 4: Examples of expert (E) and beginner(B) operation behaviors comparisons. (a - d) The easy and difficult situations of hotspots detection. (e, f) Lack of details in the experts behavior while supplemented by the beginner's. (g, h) The confirmation behavior of experts and beginners. (i) A common mistake.

References

[1] SE Dreyfus, The five-stage model of adult skill acquisition[J]. Bulletin of science, technology & society, 2004.

[2] L Chen, Y Nakamura, et al. Hotspot modeling of hand-machine interaction experiences from a head-mounted RGB-D camera[J]. IEICE Transactions, 2019.

[3] D Damen, et al. You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video[C]. BMVC 2014.

[4] R Hamada, J Okabe, et al. Cooking navi: assistant for daily cooking in kitchen[C]. ACM MM 2005.

[5] Q Zhang, B Li, Relative hidden Markov models for video-based evaluation of motion skills in surgical training[J]. PAMI 2015.

[6] H Doughty, et al. Whos Better, Whos Best: Skill Determination in Video using Deep Ranking, CVPR'18.

[7] S Vogel, H Ney, C Tillmann, HMM-based word alignment in statistical translation[C]. COLING 1996.

[8] A Betancourt, et al. The evolution of first person vision methods: A survey[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015.

[9] Z Chen, L Jiang, Early implementation experience with wearable cognitive assistance applications[C].WearSys 2015.

[10] BJ Daley, Novice to expert: An exploration of how professionals learn[J]. Adult education quarterly, 1999.