Integration of Experts' and Beginners' Machine Operation Experiences to Obtain a Detailed Task Model

AUTHOR(S):

CHEN, Longfei; NAKAMURA, Yuichi; KONDO, Kazuaki; DAMEN, Dima; MAYOL-CUEVAS, Walterio

CITATION:

CHEN, Longfei ...[et al]. Integration of Experts' and Beginners' Machine Operation Experiences to Obtain a Detailed Task Model. IEICE Transactions on Information and Systems 2021, E104.D(1): 152-161

ISSUE DATE:

2021-01

URL:

http://hdl.handle.net/2433/262917

IEICE TRANS. INF. & SYST., VOL.E104–D, NO.1 JANUARY 2021

PAPER

Integration of Experts' and Beginners' Machine Operation Experiences to Obtain a Detailed Task Model

Longfei CHEN[†a)], ***Nonmember***, **Yuichi NAKAMURA**[†b)], **Kazuaki KONDO**[†c)], ***Members***, **Dima DAMEN**[††d)], ***and*** **Walterio MAYOL-CUEVAS**[††e)], ***Nonmembers***

SUMMARY We propose a novel framework for integrating beginners' machine operational experiences with those of experts' to obtain a detailed task model. Beginners can provide valuable information for operation guidance and task design; for example, from the operations that are easy or difficult for them, the mistakes they make, and the strategy they tend to choose. However, beginners' experiences often vary widely and are difficult to integrate directly. Thus, we consider an operational experience as a sequence of hand–machine interactions at hotspots. Then, a few experts' experiences and a sufficient number of beginners' experiences are unified using two aggregation steps that align and integrate sequences of interactions. We applied our method to more than 40 experiences of a sewing task. The results demonstrate good potential for modeling and obtaining important properties of the task.

key words: *egocentric vision, hotspots, gaze, dynamic alignment, task modeling, operation difficulty*

1. Introduction

In recent decades, the video tutorial has become increasingly popular for people that want to acquire knowledge and skills. It provides the flexibility of time and place, in addition to learning efficiency in a cost-effective manner [1]. To relieve the large effort of manual content-making, many studies have explored automatic guidance authoring using experts' experiences recorded through actual work or demonstrations [2]–[7]. The emergence of wearable devices, for example smart glasses and active cameras, makes such recording easy in a human-centric manner, which can be referred to as first-person vision/view (FPV) or egocentric vision [8], [9]. It provides an intuitive and involving perspective – what the wearer sees is what you get – with less occlusion and flexibility of views [10].

In this research, we focus on using FPV to acquire a comprehensive task model for guidance on the operation of a machine, such as a printer, rice cooker, or DIY tool. Recordings of operational experiences are potentially valuable resources; for example, for directing users, particularly beginners; to select appropriate candidates for the next action; for avoiding possible mistakes; and for aiding recovery from errors. However, experts' experiences are often insufficient for the resource of learning resources for beginners. Experts often choose efficient and quick approaches, which may be difficult for beginners; experts do not have difficulties that beginners often encounter; and experts often skip the confirmation of results that are already familiar to them.

To overcome this problem, we consider a framework for integrating beginners' operational experiences in addition to experts' experiences into a unified operation model. Beginners' experiences are supplemented to fill the gap between experts' experiences and the actual requirements for guidance. In this framework, difficulties arise from the diversity of behaviors of beginners. Unlike experts' efficient operation behaviors, beginners often make mistakes, or perform unnecessary operations or redundant trials; they sometimes devise an easier approach or a new order of performing operations; or they do not complete tasks that are too difficult for them. Our approach to managing such varieties is as follows: We first automatically summarize experts' experiences into the baseline model, which is a sequence of symbolized hand–machine interactions that correspond to the crucial operation locations on a machine, that is, *hotspots* [11]. We then integrate beginners' experiences into the baseline model and obtain a unified model. We applied the above method to the operational experiences of a tabletop device: a sewing machine. Our experiments demonstrated good potential for modeling the operation and acquiring important properties of the operation.

This paper is organized as follows: Related works are introduced in Sect. 2. The basic idea of summarizing and integrating operational experiences is explained in Sect. 3, and actual methods are presented in Sects. 4 – 6. The experimental results are presented in Sect. 7.

2. Related Works

Many studies have investigated video-based user guidance in a variety of applications, for example, office work [2], [6], cooking [3]–[5], and farm work [7]. These systems provide guidance based on recorded expert experiences. Hamada et al. developed a cooking navigation system [4], in which a cooking process is decomposed into action units (i.e., ingredients, actions, and time), and multimedia-based guidance is provided. Doman et al. [3] synthesized a multimedia cook-

Manuscript received June 27, 2019.
Manuscript revised June 19, 2020.
Manuscript publicized October 2, 2020.
[†]The authors are with the Academic Center for Computing and Media Studies, Kyoto University, Kyoto-shi, 606–8501 Japan.
[††]The authors are with the Department of Computer Science, University of Bristol, Woodland Road BS8 1UB, UK.
a) E-mail: chenlf@ccm.media.kyoto-u.ac.jp
b) E-mail: yuichi@media.kyoto-u.ac.jp
c) E-mail: kondo@ccm.meida.kyoto-u.ac.jp
d) E-mail: dima.damen@bristol.ac.uk
e) E-mail: walterio.mayol-cuevas@bristol.ac.uk
DOI: 10.1587/transinf.2019EDP7180

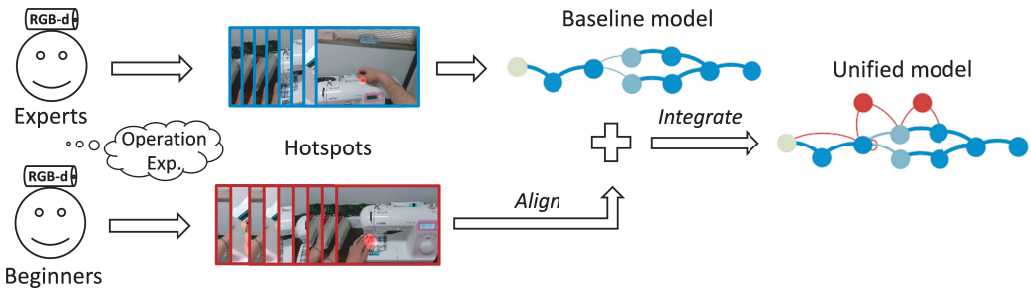


Fig. 1 Framework of capturing and modeling machine operational experiences through egocentric vision. Experts and beginners wear an RGB-D camera that records their operation process. A baseline model is built with experts' experiences first, then beginners experiences are aligned and integrated into a unified model.

ing recipe by composing a dataset of video clips from cooking shows. Zhuo et al. [12] developed a wearable cognitive assistant system. The guiding instructions are generated from downloaded YouTube tutorial videos, and indexed by their titles and descriptions. The automated acquisition of a task model from experts' behaviors has been investigated to reduce the burden of the manual collection of guidance data. Dima et al. proposed a method of automatically integrating multiple experts' experiences and providing video guidance through a Google Glass [2], [6]. Chen et al. [11] proposed a method for automatically extracting temporal interactions using hand shape and touch. However, to date, there have been few attempts to systematically use beginners' experiences for guidance. We need intensive studies of possible methods and verification of their advantages.

By contrast, recording and analyzing beginners' behaviors have been studied for skill evaluation. Zhang et al. reported a video-based evaluation of skills in surgical training using motion features, assuming that a newer behavior pattern demonstrated more skill compared with an older behavior pattern [13]. Doughty et al. proposed a supervised deep ranking model to determine skills in video records in a pairwise manner [14]. To compare and evaluate behaviors, temporal features, such as spatio-temporal interest points [13] and a two-stream CNN [14], are used to determine the correspondence between two or more experiences. Although the integration of beginners' experiences has not been investigated, these works provide good directions for the analysis and assessment of beginners' experiences.

In analyzing behaviors, gaze and attention can provide significant information about users, for example, for operability, user skill, and task difficulty. Land and Hayhoe [5] analyzed eye movement when making tea and a sandwich, which indicates that the eye provides information on an "as needed" basis. The gaze searches and locates objects to formulate memory, leads the hand to approach objects, guides the operation process, and checks the results. Peltz et al. [15] investigated the coordination of the eye, hand, and head in a block-copying task, which manifests in a temporary synergistic linkage. The general coordination is that the eye leads, followed by the head, and then the hand. Attention cues are used in the aforementioned works [2], [6] for guidance data acquisition. We expect that a comparison of gaze and attention among experts and beginners will provide detailed information about a task.

3. Task Modeling Using Beginners' Experiences

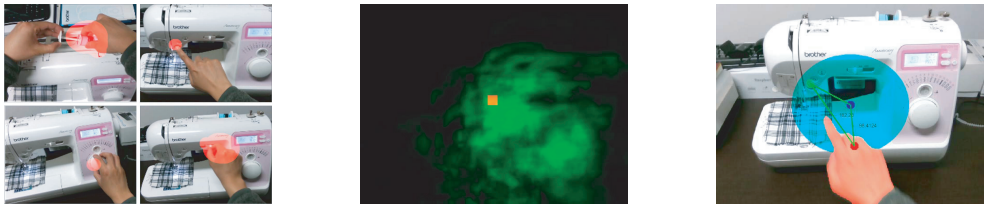
Learning to master a skill consists of more than simply following the experts' rules. The beginner's learning process is described in [16] as "contingent on concept formation and the impact of fear, mistakes, and the need for validation." Item searching and result confirmation are essential aspects for novice learning of concept formation and assimilation [17]. Accordingly, the approaches most experts take are not sufficient for beginners. Guidance for beginners may require the following functions.

1. Support diverse approaches suitable for beginners that experts do not normally choose.
2. Provide sufficient information and details that experts often skip or ignore for efficiency.
3. Provide the properties of operation steps, such as assessment of difficulty or possibility of failure.

The aim of this work is to enrich a task model with beginners' experiences to meet the above functions because beginners' experiences are good resources for this purpose. Possible means of performing the task can be covered if a sufficient number of beginners' experiences can be gathered in addition to those of experts. Some of them may also be easier methods suitable for beginners. Beginners' common mistakes and failure cases can provide good hints to guide beginners by recommending easier approaches or by alerting beginners to avoid similar failures. Beginners tend to pay more attention to the results of their actions and perform each step more slowly than experts, which may provide data that is easy to understand for guidance.

Figure 1 shows our framework for integrating beginners' experiences. It is composed of two steps because building a network of interactions directly from the diverse experiences of beginners is difficult. First, the baseline model is composed using experts' experiences which are less diverse and include fewer unnecessary portions than beginners' experiences. Next, beginners' experiences are one by one aligned and integrated into the baseline model to obtain a unified model.

We expect that the unified model provides the follow-



(a) (b) (c)

Fig. 2 (a) Examples of detected hotspots. (b) Accumulated touches (*green*) through all steps visualized relating to the center of sight (*red*) of an egocentric camera [21]. (c) Distances among the locations of the hand (*red*), gaze center (*blue*), and hotspot (*green*).

ing features. The behaviors that commonly appear in multiple users' experiences manifest statistical significance, that is, indispensability, substitutability, and the probability of choice. Unnecessary behaviors and errors are also included as less frequent interactions of beginners. The properties of each step, including difficulty, can be estimated by beginners' behaviors.

4. Baseline Model

4.1 Task and Interaction

We focus on machine operation tasks in which operators are sitting in front of a table manipulating machines. We chose a sewing machine as a typical example. The sewing task usually comprises a sequence of physical operations on machine surfaces performed by hands, for example, *push buttons, seize a lever, rotate a knob, and grab the cloth*, which have a relatively quick tempo without a substantial waiting time. The task is sufficiently complex and the included interaction patterns are sufficiently diverse to represent everyday machine operations. More importantly, the operations can typically be performed in a degree of freedom (DoF), that is, several steps can be substituted or their orders are changeable.

To describe and integrate experiences, we consider an operational experience as a sequence of hand–machine interactions. Hotspots and their interaction patterns can appropriately summarize the semantics of the interactions, that is, *where*, *when*, and *how* an interaction occurs in a machine operation.

Hotspots are automatically obtained from FPV through the detection of a hand and its contact points with the machine. To identify hotspots throughout FPV, a global map, that is, the sewing machine surface, is prepared first, and then hotspots are located on the map through the estimation of the camera pose. The interaction at each hotspot is classified by the hand shape. Figure 2 (a) shows an example of hotspot detection results, whose details are provided in [11].

4.2 Train Baseline Model

Sequences of temporal interactions at hotspots enable us to characterize an operational experience and make correspondences among experiences. To integrate the arbitrariness and redundancies of interactions, a hidden Markov model (HMM) is adopted to obtain a baseline model from experts' experiences. A left-to-right model is trained with all experts' interaction sequences using the Baum–Welch algorithm [18]. The training data is denoted as:

$$E = e_1^n, \quad e_k : o_1^m = \{o_1, o_2, \ldots, o_m\}. \tag{1}$$

Here, E is the set of n experiences as training samples, and each experience e_k is symbolized by a sequence of hotspots. Then, for any hidden state s_i with more than one observation, a replacement subnet of hidden states is created as follows:

$$s_i \rightarrow subnet : [s_i, s_{i+1}, \ldots, s_{i+m-1}]^T, \tag{2}$$

where m is the number of observations in s_i. The subnet is retrained with all the observations from s_i, and replaces the original state in the model. Here, the observations from s_i is a set of observations belonging to this state that are extracted from all training samples, which can be denoted as:

$$\{o_{(s_i)}\} \in e_1^n. \tag{3}$$

An example of subnet creation is shown in Fig. 8.

Through the above process, the observation ambiguity of the model is eliminated, that is, each hidden state only outputs a single observation. Thus, the HMM can hold the DoF of the task, that is, *alternative* and *order-changeable* interactions are allocated to separate state transition branches. Additionally, the difficulty of determining the optimal number of states for training the HMM is relaxed by applying subnets with adaptive configurations.

5. Integration of Interaction Sequences

A dynamic alignment approach is adopted to integrate experiences. Each beginner's interaction sequence is aligned to the baseline model to obtain corresponding state transition paths in the model, and then a unified model is obtained by adding all beginners' state paths and observations to the baseline model.

5.1 Alignment of Interaction Sequences

The alignment between a beginner's interaction sequence

and the baseline model can be defined as follows:

$$\begin{aligned}\hat{A} &= \arg\max_{A} \ Pr(A, O|\,\Theta) \\ &= \arg\max_{a_1^T} \prod_{t=1}^{T} Pr(a_t, o_t|\,\Theta),\end{aligned} \tag{4}$$

where $O \in \mathbb{R}^T$ is the observation (interaction) sequences of a beginner, $A \in \mathbb{R}^T$ is the assigned path of the corresponding hidden states for the observations in the baseline model, and a_t and o_t are the elements in the state and observation sequence, respectively. Θ is the parameter of the baseline model.

We assume that the operating procedures of the task typically have an inherent *forward order* with certain variations (DoFs) in several steps. Therefore, the alignment of a current interaction depends on the alignment position of the previous interaction, which is similar to the time alignment problem in speech recognition [19]. We adopt the HMM-based word alignment model proposed in [20]:

$$\begin{aligned}Pr(a_t, o_t|\,\Theta) &= Pr(a_t, o_t|\,a_1^{t-1}, o_1^{t-1}, \Theta) \\ &= Pr(a_t, o_t|\,a_{t-1}, \Theta) \\ &= \sum_{a_{t-1}} p(o_t|a_t) * p(a_t|a_{t-1}) * p(a_{t-1}|\,\Theta).\end{aligned} \tag{5}$$

To perform dynamic alignment, a recursion formula is used:

$$\begin{aligned}Q(t) &= \max_{a_t} \ Pr(a_t, o_t|\,\Theta) \\ &= \max_{a_t} \ [p(o_t|a_t) * p(a_t|\hat{a}_{t-1})] * Q(t-1)\end{aligned} \tag{6}$$

Then we have

$$\hat{a}_t = \arg\max_{a_t} \ p(o_t|a_t) * p(a_t|\hat{a}_{t-1}), \tag{7}$$

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

In experiments, we assume that any hidden state can be the starting point (a_1) of alignment. We use dynamic time warping (DTW) to compare a beginner's sequence with the baseline model. The actual process is as follows:
First, we determine the best match for the observation sequence in the baseline model to the beginner's observation as follows:

$$\hat{\omega}_E = \arg\min_{w_E}(\mathbb{E}(w_E - w_B)), \tag{8}$$

where w_B and w_E are the corresponding observation sequences of beginners' interactions and the baseline model after warping, respectively, and $\mathbb{E}$ is the Euclidean distance. The best-match expert's sequence is represented by $\hat{\omega}_E$.
Then, the first index of the same observation between two sequences is derived and we consider the starting point for alignment as the hidden state that corresponds to this observation as follows:

$$a_1 = s_E^{(\hat{k})}, \quad \text{where} \quad \hat{k} = \arg\min_{k}(\hat{\omega}_E^{(k)} == w_B^{(k)}), \tag{9}$$

where $\hat{k}$ is the index of the first-matched observation and s_E is the hidden state path that corresponds to $\hat{\omega}_E$, which is derived using the Viterbi algorithm.

All the possible state transitions $p(a_t|a_{t-1})$ are first added to the baseline model when preparing the alignment. The probabilities of jump transitions are adaptively set according to the jump width, where the maximum width of the forward or backward jump of the transition is set to three states [20] by referring to the task DoF. For the newly 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 an Operation Interaction Sequence

Input: beginner's observation sequences $O\{o_1, o_2, \ldots, o_N\}$ for alignment, the baseline model M (prior π, emission matrix E, transition matrix T, state number m), small constant probability δ ($\ll 1$), and DoF of the task $\mathbb{D}$.
Output: best state transition path $A\{a_1, a_2, \ldots, a_N\}$ corresponds to O.

```
i. Prepare for alignment:
for i = 1 to m do
  (a) add self-transition:
      T1(i, i)+ = δ;
  (b) add forward transitions (dynamic value based on forward-jump width):
  for f = 1 to 𝔻 do
      T1(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
      T1(i, i − b)+ = 1/8 * δ/b;
  end for
end for
ii. Start alignment:
Initial state a1 ← DTW(O, M);
for t = 2 to N do
  seqt ← [ot−1 ot]; πt(at−1) ← 1;
  path ← Viterbi(Et−1, Tt−1, πt, seqt);
  if path exist then
      at ← path(end);
      Tt ← Tt−1;
  else
    (d) add new hidden state:
      m ← m + 1; at ← m;
      Tt ← Tt−1(at−1, at) = δ;
      Et ← Et−1(at, ot) = δ;
  end if
end for
```

5.2 Integration for Unified Model

Using the alignment mentioned above, state transitions that correspond to each beginner's observation sequence are obtained. They are added to the baseline model with a constant pseudo-probability $\delta(\ll 1)$. Repeating the above process for all beginners, beginners' and experts' experiences are integrated into the unified model.

Commonly appearing hidden states and transitions

among experts and beginners are considered as *essential interactions* to the task. New transitions and hidden states correspond to, for example, *new methods*, *repeated interactions*, *order-changeable interactions*, or *missing interactions*.

Most experts' interactions in the unified model manifest higher frequencies than beginners' newly added interactions, which indicates that they are typically more creditable. However, if multiple beginners share common interaction patterns, then the probabilities of the corresponding paths may increase, which manifests their importance for beginners.

6. Properties of Operations

6.1 Resource for Estimating Properties

The unified model obtained through the above process is improved from the baseline model in accordance with points of 1 and 2 claimed in Sect. 3. The model covers more diverse approaches to perform a task.

Beginners tend to take more time for each interaction than experts and intensively watch the target of the operation, which is expected to make operation records more comprehensible. Concerning point 3, that is, the assessment of each operation step, there are several metrics that may directly demonstrate the advantage of the unified model. One possible measurement is the time spent on performing each step of a task. A comparison of how much time is spent by beginners or experts could provide information on the difficulty of the operation. Another metric is the number of failures. However, it is not easy to measure failures only by observing FPV without subjective introspection. We alternatively consider redundant or unnecessary interactions as signs of difficulty, which is explained below. Additionally, gazing properties are also expected to provide rich information on the characteristics of operations.

6.2 Redundant Operations

Redundant behaviors, e.g., mistakes, unnecessary trials, or repeating the current operation, often appear in beginners' experiences. The frequency of such behaviors increases when the operator's skill level is low or the operation is difficult. For the assessment of difficulty, we calculate two indices from the acquired unified model:
(i) the repetition of interactions is an indicator of unexpected results, mistakes, or perplexed states, which are closely related to the difficulty of an operation;
(ii) the frequency of uncommon operations appearing before common operations.
Uncommon or exceptional interactions are typically unnecessary interactions that are closely related to mistakes or trial and error. Conversely, easy and tractable interactions with the task are expected to be adopted by multiple users if we gather a sufficient number of samples. Thus, we manually set the threshold of occurrence frequency as 10% among all experiences to distinguish between common and uncommon behaviors in the unified model in our experiment. The correlation between the above indices and subjective difficulty were investigated in our experiments.

6.3 Gaze Properties

The relations of hands and gaze properties around hotspots can be used to characterize behaviors, that is, how a user's gaze and hands cooperatively or independently move around the operation location. The general coordination is that the eye leads, followed by the head, and then the hand. Previous studies have reported that an operator's gaze typically precedes an operation action by a fraction of a second and it leads the hand to trigger the action [5], [15], [22]. Considering this characteristic, we define the combination of behaviors of "pure-gazing (saccade/fixation), hand-approaching, and operating" as a basic operation unit (OU), and an experience can be divided into a sequence of such units. The pure-gazing period is considered to be in-between the end of the previous physical contact and the moment when the hand goes into the sight of FPV, while the operator searches or locates the target. The hand-approaching period can be considered as the period between the moment the hands just appear and the moment at which the operation begins, while the operator leads his/her hands to the hotspot with/without confidence. Subsequently, the operating period is the period in which physical touches occur.

We expect that temporal features in those periods characterize each operation step well. For example, difficult interactions require a relatively longer time for both the pure-gazing and hand-approaching periods, whereas a shorter time suggests that operators perform actions without much thinking or planning. We also expect positional features to be good clues to the characteristics of each step. We can measure the distances between the target of the gaze, hands, and hotspot for each OU, as shown in Fig. 2 (c). We can clearly say that difficult interactions require concentration on the operation area, which requires small distances among the gaze, hand, and hotspots. Conversely, some easier interactions do not require much attention, which causes the gaze to leave the hotspot earlier.

To measure the gaze, preceding studies have demonstrated strong correlations between gaze and head movement in an egocentric operation environment [6], [23]. We simply consider the average of the operator's initial gaze location as the reference of the attention location during the operation process, and use head movement to approximate the attention shift. For calibration, we lead the operator's gaze using a red point and ask the operator to position the point at the center of his/her view as accurately as possible by adjusting the head-mounted camera installation. To compensate for the bias between the center of sight and the actual attention location, the minimum distance between the center of sight and the hotspot is subtracted to obtain the distance calculation for each OU.

Table 1 Accuracy of temporal interaction detection and alignment.

Interaction (groundtruth)			Detected (F-score)		Alignment		
Total	Essential	Unessential	Expert [11]	Our	Essential	Aligned	Acc.
678	468	210	0.854	0.71	289	276	95.5%

7. Experimental Result

7.1 Environment and Parameters

The sewing task was designed with 11 essential interaction operations, which included four pairs of order-changeable steps (total DoF is $2^4 = 16$). Forty-three records of a sewing machine operation task were gathered from 16 participants. Two records were performed by a professionally skilled expert, whereas the remaining 41 were from beginners with various skill levels. The participants were only instructed with the task requirements before starting, for example, "please get the sewing machine prepared, and sew the cloth with thread pattern A and speed B", and then they were asked to perform the task without other restrictions. In total, 678 interactions occurred in the experiences, of which 69% were essential interactions. The remaining interactions were unnecessary interactions, mistakes, and other noise.

The recording device was a head-mounted RGB-D camera with both color and depth resources at 30 fps (the actual aligned depth and color fps reduced to around 20 with real-time saving to hard disk). Recordings were stopped when the participants finished the entire process or failed halfway. To detect hotspots and interaction patterns, the same parameters as those in [11] were adopted, that is, the depth threshold for detecting valid touches was ±7 mm and the temporal window size for clustering touches was 0.25s.

For the integration, beginners' aligned states and observations were directly added to the experts' baseline HMM with an equal constant probability δ, and then the parameter matrices of the HMM were normalized to ensure that all the probabilities were between zero and one. The ground truth of temporal interactions, that is, the sequence of hotspots and interaction patterns for how the operator actually processed the task, was manually annotated by each participant. The ground truth of alignment for each interaction of beginners' experiences to the baseline model was provided by an expert who viewed all experiences.

7.2 Detected Interactions

Table 1 shows the accuracy of interaction detection. The recall, precision, and F-score of temporal interaction detection at hotspots for all experiences were 0.62, 0.84, and 0.71, respectively. Compared with the case of using only experts' experiences [11], the decrease in the recall is significant, and was caused by beginners who performed redundant or unnecessary touches more frequently than experts. Additionally, beginners' hotspots were sometimes difficult to match to the corresponding location on the global map because of differences in the viewing angle and position, particularly when they were searching, attempting the operation, or confirming results. Typical examples are shown in Figs. 6 (a) and (b). The time-saving behaviors of experts, for example, operating quickly or without looking at the current operating location, also led to detection failure. For example, Fig. 6 (f) shows an example in which an expert pushed the power button without looking because he/she already knew where the button was, whereas a beginner needed to look at the button first (as shown in Fig. 6 (g)).

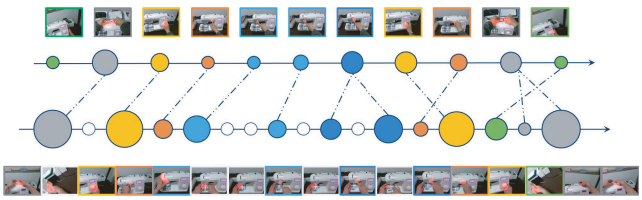
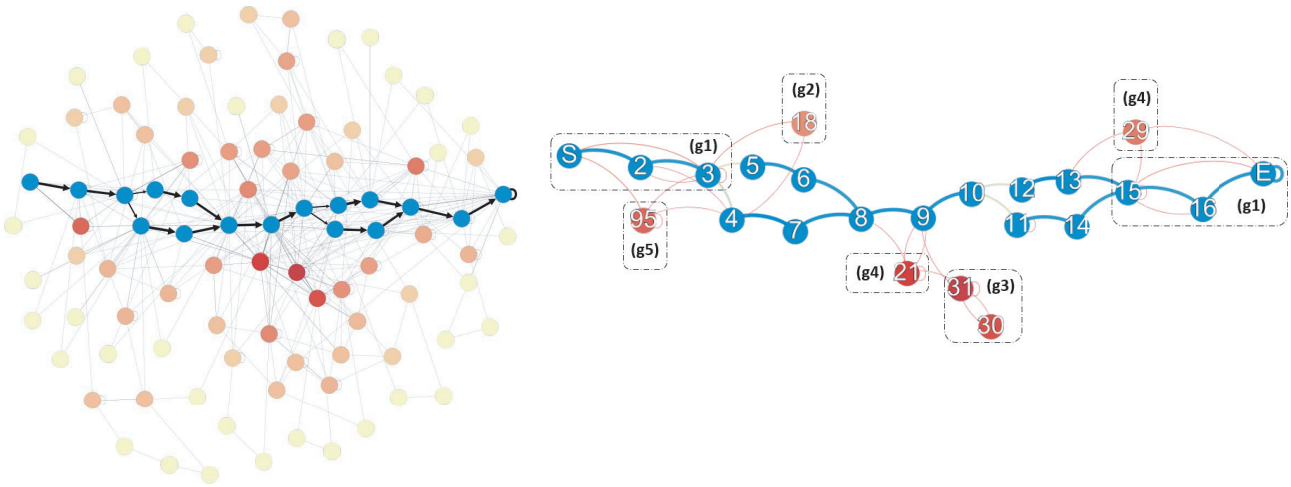


Fig. 3 Example of alignment between the experiences of an expert (top) and a beginner (bot). The size of the dots indicates the temporal duration of the interaction and the color represents different patterns, where *white dots* are interactions that newly appeared in the beginner's experience. The expert performed the task without any redundant interactions, whereas the beginner had new and repeated interactions. The equivalent interactions and the order-changeable interactions in the beginner's sequence were successfully matched to those of the expert's sequence, whereas the newly introduced interactions were located correctly among the aligned interactions.

The overall accuracy for the alignment of beginners' essential interactions was 95.5%, as shown in Table 1. An example of alignment between an expert and beginner is shown in Fig. 3. Regarding failures, two beginners' experiences were misaligned to the operational graph because the initial states of the two samples were incorrectly located on the baseline model. The reason is that they performed many wrong trials before the initial step. Another beginner's sample was successfully aligned only for the first half because it was missing many essential steps in the latter half of the task process.

The model size growth through integration is illustrated in Fig. 4 in terms of the number of hidden states. The initial baseline model of experts' experiences contained 17 states. With the beginners' experiences integrated, the number of states increased to 95. Note that after 33 experiences were integrated, the growth slowed down; it seemed to be almost saturated after 40 experiences were integrated.

Figure 5 (a) shows the full HMM model after integration. It contains all the common behaviors and diversity of beginners and experts. The blue nodes represent the baseline model, and we can see that a variety of interactions colored in red to yellow were added by the integration. Figure 5 (b) shows a graph that contains only the common interactions that occurred in multiple experiences with high probability ($>= 4\delta$). Thus, it can be regarded as a "summary" of



(a) Full (b) Summery

Fig. 5 Models for integrating all experts' and beginners' experiences of the sewing task. The expert baseline model is shown in *blue*, and the added beginner hidden states and transitions are shown in *red*. The saturation of the node indicates its sum of in-out transition probabilities, (a) Full model after the integration of all experiences. (b) High-probability states and transitions (>= 4δ).

Table 2 Semantic meaning of beginner–expert differences.

(subgraph) States		Semantic Meanings
(g1)	$s_S \rightarrow s_3 \rightarrow s_2$ $s_{16} \rightarrow s_{15} \rightarrow s_E$	New ways: *"new orders of achieving several procedures"*
(g2)	s_{18}	Common mistakes: *"operating wrong places"*
(g3)	s_{30}, s_{31}	Confirm: *"seize the cloth to confirm it's moving orientation, speed, and fixation"*
(g4)	s_{21}, s_{29}	Unnecessary: *"hand put on the cloth panel (rest or support the other hand)"*
(g5)	s_{95}	Other noise: *"trials before the starting procedure (i.e., power on)"*

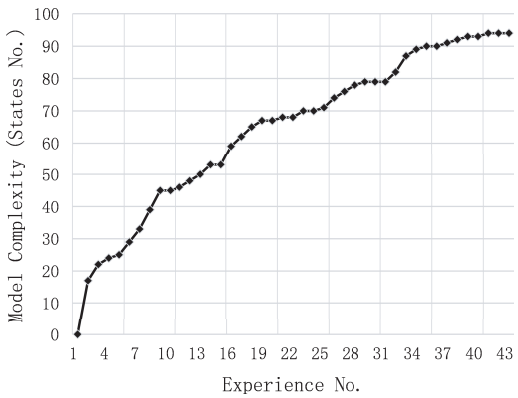


Fig. 4 Hidden state number growth of the unified model by integrating beginners' experiences.

the model. This graph includes the primary differences between the baseline model and the beginners' experiences, that is, where experts and beginners frequently chose different methods. (g1) shows new approaches and (g2) shows a common mistake. (g3), (g4), and (g5) show results in confirmation behaviors, unnecessary interaction, and other noisy trials, respectively. The detailed semantic meanings are explained in Table 2. Figure 6 (c) shows an example of a mistake: a beginner attempted to pull the cloth out after the needle was still down, which was not possible. Figure 6 (d) and (e) show an example of confirmation behaviors. The expert was watching the sewing process without any additional action (d), whereas a beginner was feeling the moving orientation/speed of the cloth by hand (e).

7.3 Estimation of Properties

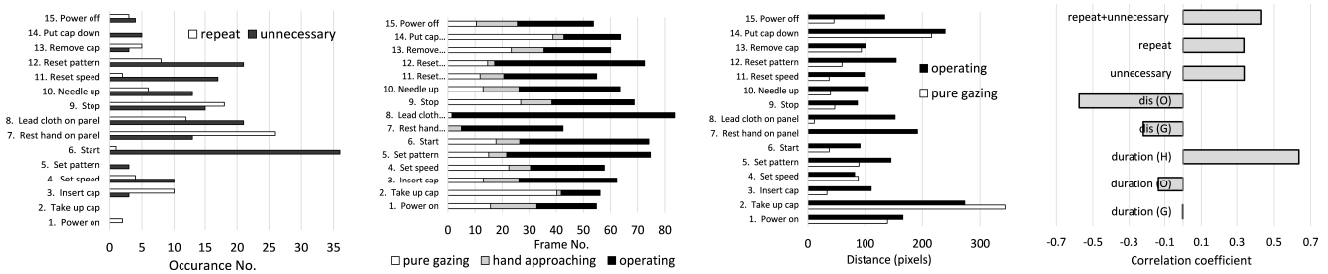
For the assessment of each operation step, we focus on the properties of 15 common interactions in the summary shown in Fig. 5. As potential clues for the properties of each operation step, such as difficulty, we consider, redundant operations, duration of each interaction, and gazing behaviors, as mentioned in Sect. 6. As the ground truth of difficulty, three experts and three beginners were asked to rate each interaction, then the difficulty scores were normalized between zero and one (from the easiest to the most difficult).

Figure 7 (a) shows the frequency of redundant interactions, that is, the repetition of the current operation, and uncommon interactions before a common interaction. Figure 7 (d) shows their correlation to the subjective difficulty. The combination of the two types of redundant operations has a higher positive correlation than either of them; which indicates that a difficult step tends to cause both types of features simultaneously. Misdetection, that is, false positives, occurred in a few cases (e.g., interaction 7 and 8), in which the operators continuously touched a relatively large

(a) Redundant touches (b) Extreme view angle (c) Mistake

(d) E: watch sewing (e) B: feel sewing (f) E: power on (g) B: power on

Fig. 6 Examples of expert (E) and beginner (B) operation behavior comparisons. (a, b) Difficult scenarios for hotspot detection and (c) an operational mistake (trying to pull out the cloth but failed because of the needle is still down). (d, e) Different confirmation behaviors of the expert (pure gazing) and the beginner (feeling the cloth's moving speed and direction). (f, g) Lack of details in the expert's behavior, which were supplemented by the beginner's behavior.



(a) Redundant Operations (b) Duration (c) Gaze–Hotspot Distance (d) Correlation of Features to Difficulty

Fig. 7 Features for describing the properties of each interaction, and their correlations with the user-rated interaction difficulty. (a) Accumulated occurrence of redundant operations for each common interaction. (b) Average duration of each period of interaction. (c) Gaze–hotspot distance in the operating period. (d) Correlations between each of the above features and subjective estimation of operation difficulty, where G, H, and O denote the pure-gazing period, hand-approaching period, and operating period, respectively.

area with moving hands. We need to improve the accuracy of interaction detection, which is future work.

Figure 7 (b) shows the average duration of each OU period for the above common interactions. As shown in Fig. 7 (d), the duration of the hand-approaching period has a high positive correlation with the difficulty, which indicates that the longer the time in which the hand is approaching the next hotspot, the more difficult the operation is. However, the duration of pure-gazing and operating periods do not demonstrate a significant correlation. The duration of the hand-approaching is closely related to the hesitation or confidence of the operator; whereas the duration of the operating period is mainly based on the interaction pattern itself. For instance, some interaction patterns, such as *lead cloth*, naturally require a longer time than others, such as *push buttons*.

To illustrate the operator's gaze distribution in operations, we calculate the average gaze–hotspot distance for each interaction, as shown in Fig. 7 (c). Figure 7 (d) shows that the gaze–hotspot distance in the operating period has a strong negative correlation with difficulty. Difficult interac-

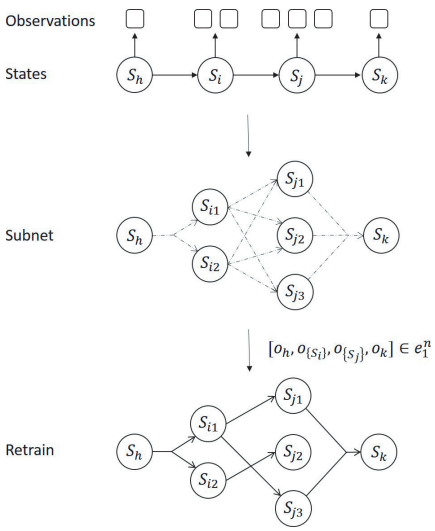


Fig. 8 An example of subnet creation. Initial model after training with experts' experience samples (top). For any state outputs more than one observation, stretch the state into vertical states and create transitions (middle). Possible result of the model after re-training with the observations corresponding to those states from all training samples (bottom).

tions require gaze concentration in the operation area. Conversely, the further the center of the gaze is located from the hotspot during operation, the easier an operation is. This is significant for steps such as *take up cap* or *rest hand on panel*, in which attention is almost unnecessary during operation. However, the gaze–hotspot distance during the pure-gazing period does not demonstrate such a close relationship to the operation difficulty because operators may not concentrate on the hotspot when searching around in the pure-gazing period.

Although Fig. 7 (d) shows the correlation to difficulty, the detected properties have different meanings, as explained in the following observations. Some operation steps, such as *start* and *set pattern*, caused unnecessary operations but little repetition occurred. An action, for example, *push the button*, is not difficult in itself; however, locating or finding the correct target is difficult. Hence, users repeated unnecessary trials at other operation locations. Moreover, some operations, for example, *stop*, were repeated frequently. This is typically because of the operators' behavior of testing the function of a hotspot. The duration of the hand-approaching was a clue to the hesitation or confidence of the operator. Short pure-gazing and hand-approaching periods indicated actions performed without much thinking or planning. The gaze–hotspot distance in the operating period was also a strong clue to the difficulty of an operation step. These facts suggest the importance of analyzing the diversity of those behaviors and how they arise, which could be useful for understanding the user's operational activities.

8. Conclusion

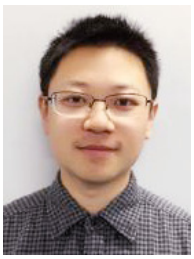
In this paper, we proposed a framework for automatically modeling a hand–machine operation task from FPV records from experts and beginners. By integrating beginners' and experts' machine operational experiences, we obtained a unified model that contained diverse approaches to interactions and samples of errors. Moreover, the temporal and spatial features of the gaze, hand, and hotspot of each operation step provided operators' behaviors and the properties of operation closely related to the difficulty of each step. The experiments demonstrate that the alignment and integration methods used were sufficient for supplementing experts' experiments, for example, by providing easy approaches that are suitable for beginners. They also proved that gaze behaviors and redundant operation behaviors provide good clues for an operation's properties, such as difficulty.

As future work, we need to gather a variety of hand–machine operation examples and verify the suitability of our framework for actual applications. The design of actual guidance systems is also necessary for improving the framework. Both the baseline model and the unified model are expected to be effectively used for guiding operators from beginners to those slightly-below-expert level.

References

[1] V. Arkorful and N. Abaidoo, "The role of e-learning, advantages and disadvantages of its adoption in higher education," International Journal of Instructional Technology and Distance Learning, vol.12, no.1, pp.29–42, 2015.
[2] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. Mayol-Cuevas, "You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video," BMVC, p.3, 2014.
[3] K. Doman, C.Y. Kuai, T. Takahashi, I. Ide, and H. Murase, "Video cooking: Towards the synthesis of multimedia cooking recipes," International Conference on Multimedia Modeling, vol.6524, pp.135–145, Springer, 2011.
[4] R. Hamada, J. Okabe, I. Ide, S. Satoh, S. Sakai, and H. Tanaka, "Cooking navi: assistant for daily cooking in kitchen," Proceedings of the 13th annual ACM international conference on Multimedia, pp.371–374, ACM, 2005.
[5] M.F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?," Vision research, vol.41, no.25-26, pp.3559–3565, 2001.
[6] T. Leelasawassuk, D. Damen, and W. Mayol-Cuevas, "Automated capture and delivery of assistive task guidance with an eyewear computer: the glaciar system," Proceedings of the 8th Augmented Human International Conference, pp.1–9, ACM, 2017.
[7] R. Tatsuta, D.T.D. Phuong, Y. Kajiwara, and H. Shimakawa, "Guidance of farming works to improve efficiency considering physical behavior," Proceedings of the 9th International Conference on Machine Learning and Computing, pp.28–32, ACM, 2017.
[8] S. Mann, K.M. Kitani, Y.J. Lee, M. Ryoo, and A. Fathi, "An introduction to the 3rd workshop on egocentric (first-person) vision," 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.827–832, IEEE, 2014.
[9] T. Kanade and M. Hebert, "First-person vision," Proceedings of the IEEE, vol.100, no.8, pp.2442–2453, 2012.
[10] A. Betancourt, P. Morerio, C.S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," IEEE Transactions on Circuits and Systems for Video Technology, vol.25, no.5, pp.744–760, 2015.
[11] L. Chen, Y. Nakamura, K. Kondo, and W. Mayol-cuevas, "Hotspot modeling of hand-machine interaction experiences from a head-mounted rgb-d camera," IEICE TRANSACTIONS on Information and Systems, vol.E102-D, no.2, pp.319–330, 2019.
[12] Z. Chen, L. Jiang, W. Hu, K. Ha, B. Amos, P. Pillai, A. Hauptmann, and M. Satyanarayanan, "Early implementation experience with wearable cognitive assistance applications," Proceedings of the 2015 workshop on Wearable Systems and Applications, pp.33–38, ACM, 2015.
[13] Q. Zhang and B. Li, "Relative hidden markov models for video-based evaluation of motion skills in surgical training," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.37, no.6, pp.1206–1218, 2015.
[14] H. Doughty, D. Damen, and et al., "Who's better, who's best: Skill determination in video using deep ranking," arXiv preprint arXiv:1703.09913, vol.1, no.2, p.3, 2017.
[15] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," Experimental brain research, vol.139, no.3, pp.266–277, 2001.
[16] S.E. Dreyfus, "The five-stage model of adult skill acquisition," Bulletin of science, technology & society, vol.24, no.3, pp.177–181, 2004.
[17] B.J. Daley, "Novice to expert: An exploration of how professionals learn," Adult education quarterly, vol.49, no.4, pp.133–147, 1999.
[18] Z. Ghahramani, "An introduction to hidden markov models and bayesian networks," Hidden Markov models: applications in computer vision, pp.9–41, World Scientific, 2001.

[19] F.J. Och and H. Ney, "A comparison of alignment models for statistical machine translation," Proceedings of the 18th conference on Computational linguistics-Volume 2, pp.1086–1090, Association for Computational Linguistics, 2000.
[20] S. Vogel, H. Ney, and C. Tillmann, "Hmm-based word alignment in statistical translation," Proceedings of the 16th conference on Computational linguistics-Volume 2, pp.836–841, Association for Computational Linguistics, 1996.
[21] Intell RealSense SR300: https://software.intel.com/en-us/realsense/sr300.
[22] T. Foulsham, "Eye movements and their functions in everyday tasks," Eye, vol.29, no.2, pp.196–199, 2015.
[23] Y. Li, A. Fathi, and J.M. Rehg, "Learning to predict gaze in egocentric video," Proceedings of the IEEE International Conference on Computer Vision, pp.3216–3223, 2013.



Longfei Chen is now a Ph.D. student in the Department of Electrical Engineering, Kyoto University. He received B.E. and M.E. in electrical engineering from Sichuan Agricultural University and Sichuan University in 2011, 2014, respectively. His research interests are on computer vision and human-computer interaction.

Yuichi Nakamura received B.E., M.E., and Ph.D. degrees in electrical engineering from Kyoto University, in 1985, 1987, and 1992, respectively. From 1990 to 1993, he worked as an instructor at the Department of Electrical Engineering of Kyoto University. From 1993 to 2004, he worked for Institute of Information Sciences and Electronics of University of Tsukuba, Institute of Engineering Mechanics and Systems of University of Tsukuba, as an assistant professor and an associate professor, respectively. Since 2004, he has been a professor of Academic Center of Computing and Media Studies, Kyoto University. His research interests are on computer vision, multimedia, human-computer and human-human interaction including distance communication, and multimedia contents production.

Kazuaki Kondo received his M.E. and Ph.D. degrees from Osaka University in Japan. He became a research associate at Osaka University in 2007, an assistant professor at Kyoto university in 2009, and a lecturer in 2015. He was awarded the Kusumoto award in 2002. His research interests are computer vision and intelligent support on human communications. He is a member of IEICE.

Dima Damen received the BSc degree (2002) in computer science from Birzeit University, MSc (2003) and PhD (2009) degrees in computer vision from the University of Leeds, United Kingdom. Currently Associate Professor (Reader) at the University of Bristol. Dima's research interests are in the automatic understanding of object interactions, actions and activities using static and wearable visual sensors. Dima co-chaired BMVC 2013, is area chair for BMVC (2014-2019), associate editor of Pattern Recognition (2017-) and IET Computer Vision (2014-). She was selected as a Nokia Research collaborator in 2016, and as an Outstanding Reviewer in ICCV17, CVPR13 and CVPR12. She is a member of the IEEE and the BMVA.

Walterio Mayol-Cuevas received his Ph.D. in 2005 from The University of Oxford. And his BSc from The National University of Mexico (UNAM) in 1999. Since 2015 he is Professor at the Department of Computer Science, University of Bristol. His interests span Computer Vision, Robotics and Mobile Computing. Currently directs projects and dissertations on topics such as assistive systems, handheld robotics, automated learning from observation, visual mapping and novel visual sensors. Was general co-chair of BMVC 2013 and General Chair of IEEE ISMAR 2016.