Transferring Expert Knowledge through Video Instructions

Michael Haslgrübler∗
Johannes Kepler University
Institute for Pervasive Computing
Linz, Austria
haslgruebler@pervasive.jku.at

Alois Ferscha†
Johannes Kepler University
Institute for Pervasive Computing
Linz, Austria
ferscha@pervasive.jku.at

Josef Heftberger
Fischer Sports GmbH
Ried im Innkreis, Austria
josef.heftberger@fischersports.com

ABSTRACT
Due to digitalization, companies which are not able to automated production, e.g. due to the fact that the production batch size is rather small, they are increasingly challenged by being not able to train personal for their production processes. Reason vary but a highly recognized fact is that training material provided does not include informal knowledge or expertise. Based on recent advances in cognitive psychology and instructional technologies, we investigate how different forms of video instructions convey process knowledge and informal expertise not in-situ but before the actual work is performed and we can measure this transformation process.

CCS CONCEPTS
• Human-centered computing → Ubiquitous and mobile computing systems and tools; • Computing methodologies → Perception; Activity recognition and understanding;

KEYWORDS
Eye tracking; learning; wearable and pervasive computing

ACM Reference Format:

1 INTRODUCTION
With the ongoing globalization competitive companies need to provide added value with product individualization – away from mass production → towards lot-size-1 products. Fully automated mass production systems however are incapable of providing the necessary adaptivity and human labor is rediscovered in industrial production [2]. Humans develop their expertise mostly through informal learning [3] in industrial professions. Progression from different levels of expertise takes however a significant amount of time and hence skilled laborers are hard to find on the one hand and close to retirement as most of them are baby boomer generations [6, 15].

In light of recent technology advancement, e.g. the quantified self-movement where people are looking for technology self-improvement or behavior change [16], the low cost of intelligent, smart or even cognitive technology provides a way to coach or train people so that they can faster acquire skills and excel in their professions. In contrast to traditional approaches these technology-driven approaches empower the learner to explore skill acquisition free from social fear or judgment in his own pace. Companies providing these self-learning technologies have the added value of (a) being interesting for newer generation (b) can train multiple workers at once without disturbing current production processes (as instructional personal would be missing elsewhere) (c) and have a highly adaptive workforce at their disposal which can easily switch between production of different products.

Technology based learning is in industrial production however still a vision and traditionally learning on the job and working alongside more experience workers are the two prominent techniques for learning [3]. In industrial production, i.e. for ski assembly workers, a three stage training process (a part of Collins 6 Stage Cognitive Apprenticeship Model [4]) is commonly used: (i) observation — observing another worker’s assembly, (ii) cooperation — working together on the assembly, (iii) monitoring — being observed while assembling. In the first phase, a less experienced or new employee observes his colleagues until he understands basic principles and common procedures of the task without actually doing any assembling himself. For the second phase, the trainee worker performs basic tasks along the side of his more experienced colleague. At final phase, the trainee works more or less autonomously, while under closed observation by his experienced colleague, which only intervenes when an error is about to happen.

In addition, due to material constraints, e.g. in a ski assembly process were specialized components are interwoven with glue or pre/post assembly work stations feed or need to be feded, it is essential that the whole assembly process is completed within a certain time-frame, putting quite a strain onto the whole learning processes and the learner itself and in turn also on any instructive technology to be successfully applied.

2 RELATED WORK
Learning Processes One of the underlying key principles for learning in such manufacturing setting is commonly known as Observational Learning (later refined as Social Learning Theory [1] and Social Cognitive Theory), which is governed by four internal
processes according to [1], namely: attention (noticing something worthwhile), retention (remembering in order to imitate), reproduction (ability to imitate) and motivation (willingness to perform). While the later process is commonly handled by paying people, the other three processes need to be successfully stimulated during the training phase of a new employee and by an instructive technology.

Gaze and Attention One of the most used inherent human behavior for observational learning is to follow human gaze [7]. Gaze does reveals the attention of a person, a fact which has also been successfully applied in training surgeons [17], where gaze reveals what experts find worthwhile. In addition, this is not a learned behavior but already present in infants [14]. While capturing gaze is a prominent topic in Human Computer Interaction [11] and work has already begun to built systems which also make use of its attention revealing facility [12], it is seldom applied to industrial work process.

Instructional Technology For industrial work processes [8] analyzed the effects of in-session assistance of projection, head-mounted displays, tablet, and paper instructions based on explicit knowledge using illustrations or annotations as means of conveying information.

While existing work [9] determined that illustration and text are superior to videos for technical documentations, they state that video based learning does come increasingly popular. We reason that this is due to the fact that it is much more easy to produce a video capturing implicit knowledge than to make that knowledge explicit by providing appropriate illustrations and annotations.

In [5] the usefulness of wearable head-worn camera as a learning tool was investigated for chemistry and could show that this new perspective is well liked whereas the viewing angle was the 2nd most liked reason after the explanatory effect of the camera view itself.

In light of this aforementioned works we do not investigate in-session instructions but investigate pre-session video based instructions. Whereas we analyze how different video representation forms, from wearable or stationary video devices, are able to transfer expert knowledge without any explicit instructional annotation.

Based on related work we derive the following three hypotheses:

- **H0** The participants trained will be affected by successive runs after instructional material is provided.
- **H1** The participants trained with 1st Person view, cf. Figure 1, will score better than 3rd Person view.
- **H2** The participants trained with 1st Person Gaze View will score better than 3rd Person view.
- **H3** The participants trained with 1st Person Gaze View will score better than 1st Person view.

3 SYSTEM DESIGN AND SETTING

Our system setting is located in a controlled laboratory environment where we rebuilt a ski assembly booth, cf. Figure 1c. We equipped an intermediate expert in ski assembly manufacturing with a wearable eye-tracking unit (SMI Eye Tracking Glasses 2.0) and placed an additional COTS camera in position similar to a human observer and recorded the assembly process of a ski which consisted of 17 parts of different colors, lengths, shapes and materials.

In a post processing step the recording of the eye tracking unit was split into a video without gaze annotation, corresponding to the 1st person view, cf. Figure 1a of an of the shelf action camera and a version with gaze annotated. This gaze annotated version showed the gaze of the expert using a clearly distinguishable circle representation, cf. Figure 1b and a spotlight effect where the pixel closed to the gaze where not or less darkened than those further away. This was done to prevent users to focus less on non-relevant information in the video.

4 EVALUATION

4.1 Experimental Setup

We recruited 18 Participants (12 Male, 6 Female) aged between 19 and 40 ($\mu$=24.94, $\sigma$=5.12) each of which had prior experience with videos in various form, cf. Figure 2. Denote that the experience with gaze annotated videos was almost none at all ($\mu$=1.78, $\sigma$=1.26) and also the experience with first person video was less common ($\mu$=3.17, $\sigma$=1.79) than watching normal – third person – videos ($\mu$=5.67, $\sigma$=1.46). Each participant was assigned by order of appearance into one of the three groups: 3rd person video (3P), 1st person video (1P) and gaze annotated video (GA). After a brief introduction of the setting, participants watched their respective instructional video for the first time and tried to perform the shown assembly procedure. Participants were also given the option to quit if they don’t know how to continue. Afterwards they watched the same video for the second time and performed the assembly procedure again. Finally, they filled out a questionnaire which included questions of demographics and prior experience, cf. Figure 2.

Denote that the layout of the parts to be assembled was the same as in the video instruction for each successive run and that all parts where to be used for a correct assembly procedure, hence there were no parts left when the assembly was completed.

From the 18 participants 7 (3P: 4, 1P: 1, GA: 2) did not complete the assembly procedure in the first round while all did complete the assembly procedure in the second run and in addition some made sequence errors or left out parts, cf. Figure 3. We identified the following criteria for assessing the instructional capabilities of the different video representation forms:

- **Parts [Count]** Parts used in the assembly
- **Sequence Errors [Count]** A Part is placed on a wrong other part
- **Left Out Errors [Count]** A Part was missing in the assembled product
- **Time [s]** Duration of Assembly Procedure

In addition we derived the following value for interpolating a prediction from incomplete data (e.g. no completed run 1).

- **Predicted Completion Time [s]** Expected Time for Completion till all parts are assembled
- **Predicted Error Rate [s]** Time Duration between Errors
- **Predicted Part Rate [s]** Time Duration for Part Placement

4.2 Improvement

First we investigated if the above mentioned criteria can be used to assess improvement. We defined improvement by two factors an increase in parts used and a decrease in errors made, subsequently
out of our 18 participants 5 (3P: 4, 1P: 0, GA: 1) showed no such improvement, from 1st to 2nd run. An overview of the operating figures split by improvement is provided in Figure 4.

For the quantitative analysis we used Analysis of variance (ANOVA) using the ez package [13] of R where we calculated Greenhouse-Geisser-corrected degrees of freedom and p-values using a Type-III ANOVA as we had a group imbalance.

For Parts we see that there is a dependence on improvement ($F_{1,16} = 7.2246, p = 0.0162, \eta^2 = 0.1747$) and run ($F_{1,16} = 20.8350, p = 0.0003, \eta^2 = 0.4089$) but no interdependence between the two unlike for Sequence Errors where there is only an interdependence effect ($F_{1,16} = 26.6398, p = 0.0001, \eta^2 = 0.2010$). Respectively Left Out Errors have the same dependence and values as Parts as they are the inverse of each other. Time is only depended on the run $F_{1,16} = 6.3377, p = 0.0229, \eta^2 = 0.1441$ and as expected not an indicator for improvement.

The predicted value, Predicted Completion Time are however dependent on Improvement ($F_{1,16} = 20.8105, p = 0.0003, \eta^2 = 0.3940$), Run ($F_{1,16} = 17.2803, p = 0.0007, \eta^2 = 0.3507$) and the interdependence of both ($F_{1,16} = 8.6698, p = 0.0095, \eta^2 = 0.2132$). Also, the Predicted Error Rate and Predicted Part Rate are dependent on Improvement ($F_{1,16} = 4.5730, p = 0.0483, \eta^2 = 0.1509$ and $F_{1,16} = 20.8105, p = 0.0003, \eta^2 = 0.3940$), Run ($F_{1,16} = 9.8775, p = 0.0063, \eta^2 = 0.1892$ and $F_{1,16} = 17.2803, p = 0.0007, \eta^2 = 0.3507$) and its interdependence ($F_{1,16} = 7.8678, p = 0.0127, \eta^2 = 0.1567$ and $F_{1,16} = 8.6698, p = 0.0095, \eta^2 = 0.2132$).

### 4.3 Instructional Material

As the above operating figures show that they are linked to improvement, we analyze them in conjunction with our instructive video group. We see that all but time, which is independent of all effects, dependent on the run: Sequence Errors ($F_{1,15} = 5.6842, p = 0.0308, \eta^2 = 0.0890$), Left Out Errors ($F_{1,15} = 11.7988, p = 0.0037, \eta^2 = 0.2870$), Parts ($F_{1,15} = 11.7988, p = 0.0037, \eta^2 = 0.2870$), Predicted Completion Time ($F_{1,15} = 6.3134, p = 0.0239, \eta^2 = 0.1421$), Predicted Error Rate ($F_{1,15} = 20.3020, p = 0.0004, \eta^2 = 0.3659$) and Predicted Part Rate ($F_{1,15} = 6.3134, p = 0.0239, \eta^2 = 0.1421$). Thus, we reason that our Null Hypothesis is proven. There was no dependence on the group or an interdependence between group and run.

In a first analysis we evaluated our 18 participants and found that when comparing 1st Person (1P) vs 3rd Person (3P); 3rd Person (3P) vs Gaze (GA) 1st Person (1P) vs Gaze (GA) only Predicted Error Rate
### DISCUSSION

The presented results represent no conclusive evidence to prove our hypothesis apart from the null hypothesis. If we however take into account our participants prior experience with our different video formats (3P: Highest, 1P: Average, GA: Lowest) and consider the improvement dropout from our users (3P: 4, 1P: 0, GA: 1) and the one which did not complete the first run (3P: 4, 1P: 1, GA: 2) we could argumentatively suggest that our hypothesis does have merit.

Additionally, some participants gave oral feedback, which we considered but were more individual opinions than representative, e.g. a group 1P person mentioned that they missed the overview (where things were happening), a GA group person mentioned that the eye annotation felt disturbing.

In any case the presented operating figures provide a valuable comparison mechanism to assess improvements on successive runs in assembly manufacturing scenarios and to assess the effect of assistive technologies.

### FUTURE WORK

While this current work was focus on providing instruction before session the plan to extend this approach to provide instructions in situ, where we are planning to automate the capturing process and combine them with our workflow detection system [10]. Thus, we would be able to automatically slice the video instructions into small pieces for each work step and be able to provided them in-situ.

### CONCLUSION

With ongoing digitalization in assembly factories and the high demand of lot size 1 products, companies are under pressure to adapt their product capabilities quickly. A major hindrance is that training skilled personal takes a significant amount of time and decreases production capabilities, as a skilled operator has to train new employees. With the broad availability of video capturing and distribution mechanism, workers can capture their often implicit knowledge and share it with others. In this work we investigated different capturing formats from 3rd person, 1st person videos to gaze annotated videos and investigated if they can be used for instructional purposes of assembly procedures and which operating
Transferring Expert Knowledge through Video Instructions

PETRA ’18, June 26–29, 2018, Corfu, Greece

8 ACKNOWLEDGMENTS

This work was supported by Attentive Machines (FFG, Contract No. 849976) and ProFuture (FFG, Contract No. 6112792).

REFERENCES


Figure 5: Operating Figures for User with different instructive material which improved and completed the first run

![Operating Figures](image-url)