
Peer reviewed version

Link to published version (if available):
10.1145/3041164.3041185

Link to publication record in Explore Bristol Research

PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via ACM at http://dl.acm.org/citation.cfm?id=3041185&CFID=918210373&CFTOKEN=32560132 . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms
ABSTRACT

In this paper we describe and evaluate an assistive mixed reality system that aims to augment users in tasks by combining automated and unsupervised information collection with minimally invasive video guides. The result is a fully self-contained system that we call GlaciAR (Glass-enabled Contextual Interactions for Augmented Reality). It operates by extracting contextual interactions from observing users performing actions. GlaciAR is able to i) automatically determine moments of relevance based on a head motion attention model, ii) automatically produce video guidance information, iii) trigger these guides based on an object detection method, iv) learn without supervision from observing multiple users and v) operate fully on-board a current eyewear computer (Google Glass). We describe the components of GlaciAR together with user evaluations on three tasks. We see this work as a first step toward scaling up the notoriously difficult authoring problem in guidance systems and an exploration of enhancing user natural abilities via minimally invasive visual cues.

CCS Concepts

• Human-centered computing → Mixed / augmented reality; HCI design and evaluation methods;

Keywords

Augmented Reality; Task Guidance; Eyewear computing. Assistive computing

1. INTRODUCTION

Within the mixed reality continuum, that goes from the virtual to the real [11], one of the most noble aims is arguably to augment people to gain extra skills on how to do something better or something for the first time.

The ability to feel able to go about and do anything while being supported by an assistive system has long been promised yet remains elusive. Consider being able to adjust your particular bike’s chain by easy to follow guidance; receiving training in front of a new machine that others (but not you) have repaired before; showing up in any previously unknown kitchen anywhere, and seamlessly being shown where the key utensils are. All simple and common instances of where widely available guidance can help.

The now wider availability of head-worn hardware systems which feature advanced sensing, see-through capabilities and even spatial positioning may appear to bring us closer to this goal. However, they also put significant focus on a largely overlooked aspect for mixed guidance systems: the authoring problem.

From the earliest systems such as Karma [5], guidance has assumed that 3D information is essential. But the reality is that 3D annotation is extremely hard to author and is so far only provided for well scripted tasks. Hardware limitations have in the past been blamed but the authoring problem is, we argue, a much more serious obstacle which results in objects, places and sequences of interactions having to be mostly known in advance.

The authoring bottleneck can be relieved somehow with in-situ annotation creation but to truly scale up the collection of key and nuanced information for any task, object, place and time, the spending of extra user time to author guides remains unconvincing.

Importantly, people is remarkably capable of following instructions as long as these are presented in a clear and intuitive manner. On the other side, delivering guidance information with the wrong type of visualization or with overly synthesized information can be confusing at best and undermining the user’s self confidence at worst. It is thus that the question of how best to guide remains important.

In this work we are concerned with making inroads into some of the above questions, namely 1) how to extract information in an unsupervised way for unscripted tasks so that guidance systems can start to scale up, 2) how to integrate a fully operational demonstrator of these ideas and 3) evaluate how such an automated learn-and-deliver guidance system supports people performing tasks.

The paper is organized as follows. In Section 2 we discuss guidance and authoring in MR/AR with emphasis on self-contained systems. Section 3 describes our approach starting with the description of the model of attention we use. Section 4 discusses the combination of the attention model with the video guide editor and object detector before in Section 5 we evaluate various aspects of the attention model performance including object discovery and multi user consistency. In Section 6 we evaluate GlaciAR with novice volunteers on three tasks before our discussion and conclusions.

2. GUIDANCE AND AUTHORING

We note that while there has been a large body of work on developing AR/MR systems for guidance, the vast majority of systems employ as part of their workflow pipeline an offline and supervised step for the authoring of the information to be displayed. This authoring can be from the earliest methods, text-based notes that appear when e.g. a tag appears in view [6] to more intricate 3D models that are meant to show assembly or repair instructions [17]. However as mentioned already, the authoring of the information to be displayed is non-trivial especially when we want to be able to
perform guidance for any object or process anywhere.

Making advances in ways in which guidance can be provided and the authoring bottleneck mitigated, will have a significant effect for MR/AR guidance.

A related system to ours is the Gabriel system [16] which also uses Glass to guide users on tasks. In that case tasks are 2D such as assembling figures with colored blocks or drawing. That system implements a standard guidance and monitoring pipeline where the task is pre-scripted offline, as well as it uses an offboard strategy for information comparison and storage. It also features a verification of the progress stages and does this via relatively well constrained image metrics. In GlaciAR, we do not enforce verification of a stage but rather concentrate in the arguably harder problem of automated capturing and delivery of information to inform the user about the task.

Other recent work has started to highlight the importance of automatically capturing workflows. In [13, 12], a method that uses image similarities captures and monitors workflows. The approach there compares incoming images with those stored for the same task and overlays them on a HMD. These overlays are a combination of mostly automated but also some manually authored information for the next step to follow. In [1], the workflow is captured by a rich array of on-body sensors and a semantic modeling of the task is used to keep track of the workflow. These approaches aim to alleviate the authoring problem somehow by trying to automate the extraction of relevant information as much as possible. In the case of [13, 12] this is done using 2D image metrics and the information to be displayed to users is an overlay on top of the object being part of the task.

On the other hand, some concepts such as Indirect Augmented Reality [4] have started to explore alternative MR/AR ways in which information can be delivered yet with less reliance on external positioning and associated hardware requirements.

Some other methods have extracted relevant task information from a variety of sources that include eye tracking, 3D mapping and positioning via SLAM and visual object appearance detection [4]. But such approaches are not demonstrated in real-time or on eyewear hardware and assume that many sources of information are possible to collect.

Our work can be seen as effort to improve in the above directions, but here we aim to push further the issues of what information to capture and how to display it, both in fully automatic ways. Furthermore, most of the works on AR guidance assume that overlaying of information is crucial for guidance, but often this information can either be hard to understand, jittery due to 6D positional inaccuracies or at least substantially invading the visual field of view of the user. Our approach exploits the apparent limitations of contemporary eyewear computers e.g. Google Glass with a limited field of view and side-located display. An opportunity to provide short, informative video snippets that have been automatically extracted from other users that have completed the task.

With GlaciAR, we develop a much more distilled and condensed authoring concept yet one that is amenable to be fully implemented onboard an eyewear system, in real-time and which allows for practical evaluation. The method we follow is illustrated in figure 1 and described in the next sections.

3. METHOD AND IMPLEMENTATION

GlaciAR is underpinned by three interlinked components:

- A model of user attention.
- The capture of video snippets around instances of attention.
- The detection of previously attended objects.

The overview of the system is shown in Figure 2. Two main sensors, the front facing camera and the inertial measurement unit feed information to the system. In GlaciAR, the module for attention determination is crucial since this is the mechanism used to make decisions of when to record or display information.

3.1 Attention detection using Glass

In other systems, eye-gaze has been useful as a source of attention determination via eye fixations [4]. That is, an angular velocity model for gaze fixations dictated when and where the person was paying attention. However in Glass there is no gaze tracker. We therefore use the work in [8] which estimates spatial and temporal attention from the onboard IMU unit. For completeness, we briefly describe the approach followed and we substantially expand on the
experimental evaluation of this attention model, and evaluate its usefulness within the overall guidance system.

It is important to highlight that we are interested in moments of attention where the user is about to or already interacting with something. This is therefore a subset of all potential moments of attention that a user may have. Yet, by defining our attention for the instances of object interactions we aim to cater for an important set of the moments when the user is doing something of relevance and or needs guidance. We thus define the head-motion attention in a similar way as attention is often defined in gaze tracking, that is, a threshold to determine an eye fixation based on eye angular velocity [14], is here replaced by a threshold on the angular acceleration and velocity of head motion. We define temporal attention as

$$T_{\text{attention}} = \begin{cases} \text{attending}, & \text{if } a \leq \tau \text{ and } \omega \leq v \\ \text{in motion}, & \text{Otherwise.} \end{cases}$$

(1)

Where $\tau$ is the relative head acceleration threshold and $v$ is the relative head angular velocity threshold for identifying whether the user is attending to something or not. But this is only the temporal attention model, i.e. the when the user is paying attention.

For spatial attention i.e. the where the user is looking at we opt for GlaciAR to use a fixed image location. This is backed by recent work that has investigated gaze fixations and that shows that for egocentric perception, when the user is interacting with things in the world, the location of where the user is fixating is concentrated around a small region in the image [9, 8]. To compute the centre of mass of egocentric gaze fixations on the Glass’ front camera image, we attach an eye gaze tracker to Glass and calibrate the location of where gaze is into it. With this information we can compute the gaze centre of mass for a number of people and tasks. We use images captured at 640x360 pixels instead of full resolution in order to reduce computational burden later and on these images the spatial attention point we use is located at coordinate (250, 189.5). The location toward the left side of the image is due to Glass’ camera being mounted on the right side of the head. Figure 3 shows the location of the fixed spatial attention coordinate. Note that the spatial attention region is only computed if the system is in the attending mode as per equation 1.

GlaciAR’s attention model is simple yet robust, requiring minimal computational burden and no image measurements. Extending the work of [8], in this paper we perform a more exhaustive evaluation on how useful this attention model is for automated capture of relevant information.

Figure 4 presents the user’s motion signals acquired from the Glass’ IMU including the relative acceleration (red) and relative angular velocity (blue) as the user is making a cup of tea. The periods of time when the user is paying attention at the tasks are highlighted as the cyan-shaded rectangles and correspond to moments of eye fixation. The detected user’s attention including manually selected ground-truth are shown in Figure 5. In this figure, the extracted attention using the optimal threshold values as $\tau = 3.0 \, \text{m/s}^2$ and $v = 0.5 \, \text{rad/s}$ compared to the ground-truth (top), and the user’s attention after applying a median filter (bottom).

4. AUTOMATIC GUIDE EDITING AND LINK WITH OBJECT DETECTOR

In GlaciAR, the attention model is the one that determines the when the person is doing something of relevance. This is based on the model described on the above section and is in contrast to other egocentric systems that use for example hand detection [10, 9] to indicate that something important is happening.

GlaciAR takes this approach for two reasons: first is that if the attention is linked to hands, the computational complexity required for the assessment increases substantially as hand detection in the

Figure 3: Spatial attention position (green point) and the area of interest (green box) acquired from Google Glass.

Figure 4: The IMU signal acquired from Google Glass during tasks performing of a user in a real environment setup.

Figure 5: The user’s attention over time extracted from the IMU signal (middle) using the threshold values $\tau = 3.0 \, \text{m/s}^2$ and $v = 0.5 \, \text{rad/s}$ compared to the ground-truth (top), and the user’s attention after applying a median filter (bottom).
wild is not trivial, and then, importantly a multitude of hand-eye coordination studies (e.g. see [7]) have shown that fixations (via gaze attention) precedes action by a good number of milliseconds. In this work, we thus hypothesize and evaluate how well a model of attention based on head motion can also be used to preempt interactions and thus use it for the task of extracting video snippets around moments of interest.

The approach is therefore to use the attention model instead of hand detection or any other environmental property as the director for video editing — video snippet extraction starts and stops automatically when the system enters and leaves the attending mode. The above approach can and indeed results in a number of videos captured as the user goes about interacting with objects. But to our advantage, most interactions with daily-living objects such as coffee machines, microwaves, car starting and similar are primarily driven by a reduced number or mostly a single way in which they can be interacted with. These tasks can include multiple steps such as to use a microwave “this button” needs pressing to open it, “this button” needs pressing for selecting the power option and “this knob” turned to select the time. All of these steps can and should form part of a single video guide on how to use the microwave.

The capturing of video snippets in the way GlaciAR works is not restrictive of collecting multiple ways in which the objects can be used. As per Figure 2 all collected videos are stored. For each one of these, the first frame for when the attention was detected and thus the object’s untouched state is used to train a textureless object detector. For this an area of interest (AOI) of size 200x200 pixels (see figure 3) around the spatial centre of attention (250,189.5) is cropped and a descriptor based on edge configurations extracted [3]. This detector is computationally lightweight [2], allows for multiple object detections, has invariance to scale, rotation and a degree of affine transformations and importantly runs entirely onboard Glass. All these contributes to reduced lag and robustness. In figure 6, example detections are shown. The detector runs from the information in the AOI estimated by the attention model which gates the image region and helps to keep computational demands low.

When GlaciAR is in training mode, the model of attention captures videos linked to attention periods and trains the detector all in real-time. In this way it can simply observe expert users performing tasks while collecting the relevant information for guidance. When GlaciAR goes into assistive mode, the attention model is used to indicate that the user is interested in the object being attended, this prompts the detector to try to match the current AOI with stored ones and if a sufficiently good match is found, the associated video guide that was extracted from the expert’s moment of attention is played on Glass. Note that since GlaciAR potentially captures one or more video guides for every expert in the training stage, the closest matching view when the novice is requiring guidance is displayed. This is the way in which novice users get guidance.

5. EXPERIMENTAL EVALUATION

We concentrate our evaluation first on the performance of the attention detection module since this is critical to all aspects in GlaciAR. One way in which we can evaluate its performance is via its ability to detect objects that the user has interacted with. To do this we follow the procedure used in [8] where we attach a wearable gaze tracker (ASL Mobileye) to Glass for it to serve as one source of ground truth. For this evaluation, we used 8 volunteers that were asked to wear the bundled device. This is a different group of users from the ones in the main results section. They were unaware of the type of information we were measuring to reduce bias on the data collected. The volunteers were then asked to interact with objects around a building (e.g. open this door, press this button, lift that telephone, etc) guided by an investigator that used a stick with a coloured “ping pong” ball attached to it to indicate which object to use. This coloured taget was important to both unambiguously guide the interaction and discard imagery of any distracting saccades not part of the instructions, as well as to ensure there is ground truth since no current eye tracking hardware delivers results every time. Images from the scene facing camera of the gaze tracker were recorded at 30fps and synchronized with the IMU data on Glass narrowed to within a single image frame.

5.1 Predicting interactions in advance

As discussed before, eye gaze has been used widely to estimate attention and used as a precedent to action. In essence, people eye-gaze to what is about to be manipulated [7]. Since GlaciAR uses a head motion-based attention model it is important to ascertain how much predictive power it has and how much in advance it will be able to detect hand-object interactions compared to eye-gaze. Recall that if the interaction is predicted in advance a video snippet can then be captured shortly before this interaction takes place and the object detector will be trained without hands occluding it.

Figure 7 illustrates the advance attention estimation. The figure shows an example of a user fixating at a microwave oven with eye-gaze, our head motion model then estimates attention and finally the microwave door is opened. Evaluating with a total of 91 hand-object of such interactions from the 8 users above mentioned, our approach can predict a hand-object interaction in advance on average 1.18(±0.47) seconds before, and only after about 0.60(±0.35) seconds of the attention estimated with gaze fixations (figure 8). This ability of prediction based on the attention model, that results in a sufficient margin (>1000ms in advance), enables the operation of GlaciAR as described.

5.2 Object discovery results

Object discovery is a good measure to test how well the attention module is performing. We calculate object discovery via the
another interesting question is what is the ratio of same-object dis-
5.3 Multi-user object discovery
ded stage. rates, but at the expense of a larger image area to process in the
nered objects are just outside the overlap criteria. In this test the
ared objects within the overlap region. Several of the failed to be discov-
ished and missed discoveries are marked in Figure 9, where the ground-truth (ping pong ball location) and the other centered at the attention estimation position. These two boxes were then compared using the standard PASCAL overlap criteria used for object discovery in Computer Vision, though note that no image processing is taking place here, all is driven by the IMU signal. We use a threshold of 30% overlap between the AOIs to identify a discovered object. We then declared the object as true-positive discovery if the overlap is satisfied for 10 consecutive frames. This mitigates unstable and outlier discoveries.

The recordings used for this evaluation add to about 80 minutes of interactions (8 users x 10min/user).

Examples of discovered and not discovered objects using the positions obtained from the spatial attention estimation are shown in Figure 9, where the ground-truth objects are within the blue coloured boxes, the successfully discovered objects are presented in magenta coloured boxes, and the missed discoveries are marked as magenta crosses (bottom row in figure). As can be seen, the discovered objects correspond with objects interacted with, and that are within the overlap region. Several of the failed to be discovered objects are just outside the overlap criteria. In this test the object discovery precision is 0.61 and the recall 0.56. These results are conservative and a more relaxed AOI would result in increased rates, but at the expense of a larger image area to process in the detection stage.

5.3 Multi-user object discovery
For an assistive system that aims to learn from multiple people, another interesting question is what is the ratio of same-object dis-

Figure 8: The distribution of the difference between detecting attention with eye gaze and head motion. The attention is detected on average only about 0.6(±0.35)s after that possible with gaze and overall 1.18(±0.47)s in advance of a hand-object interaction.

Figure 9: Some results of object discovery from 3 different participants. Top 3 rows are example successfully discovered objects (magenta squares) and the bottom row shows sample missed discoveries, where the bounding box overlap criteria is not satisfied. But note most are just outside it.

intersection of two AOIs with size 200 × 200 pixels one centered at the ground-truth (ping pong ball location) and the other centered at

Figure 10: Nine objects (rows) being discovered from 8 different users (columns) using our attention method as users explore various objects. Objects that failed to be detected by a specific user are labeled as a red dash. The worst performing discoveries are for objects that are interacted with only briefly e.g. doors (last rows). discovery across multiple users as well as what type of object is more or less likely to be discovered.

But some of the objects posed greater challenge. In particular, objects that are only briefly interacted with under naturalistic conditions such as a door handle as it is opened, are harder to be detected (bottom rows in figure 10). While other objects that are operated in the style of a printer’s panel, answering a telephone or opening a safe are more easily detected by all users (top rows in figure 10).

6. GUIDANCE EVALUATION
This section describes the evaluation of GlaciAR where tasks are automatically captured, edited and delivered to guide users. GlaciAR runs onboard a Google Glass Explorer Edition 2.0, such that all attention detection, snippet video editing, object detection, and video guide delivery are run in real-time and without user intervention.

The way in which GlaciAR is triggered for guidance is as the user walks up to an object where it has previously been trained. It detects such object and delivers the video guide closest to the detected viewpoint. This paper focuses on guidance evaluation and not the object detection performance which is subject to a separate analysis. There are thus two main evaluations. First of the attention model, then of the actual guidance performance.
6.1 Video Guide Authoring Evaluation

This section compares the effectiveness of GlaciAR in automatically extracting video guides versus video guides manually edited by expert users. This aims to demonstrate the extent of effectiveness of the attention-driven harvesting of information as well as the overall concept of video guidance from snippets. There are two conditions as follows.

- Automatically generated video guide: the video guides are automatically extracted by GlaciAR using the proposed model of attention.
- Expert generated video guide: the video guides are manually edited by experts, who indicate the start and end of the activities.

Our hypothesis before conducting the evaluation is that if the attention-driven model is useful, there will be little to no statistical difference between the performance achieved with the expert edited video guides and the ones automatically extracted by GlaciAR.

We had two operational sessions, training and testing. Three experts participated in the training session and 14 novice volunteers (eight females, six males) aged from 24 to 36 were recruited. None of the novice volunteers had participated in the our previous studies or possessed any prior experience with Google Glass.

The tasks involved an oscilloscope, an electric screwdriver and a sewing machine (Figure 11). Each one of these tasks have an increasing number of steps from three to five. Each expert performed every task, and then asked to watch and edit a video guide of the task by indicating the start and end of the task. In total, there were 9 videos (3 videos × 3 tasks) for each video guide condition. The task stages are described in Table 1.

(a) Oscilloscope (b) Screwdriver (c) Sewing machine

Figure 11: The objects used in the video guide evaluation.

During the testing session, each participant was asked to perform every task and their performance was recorded for analysis. After each task, the participant was also asked to fill out a NASA-TLX survey and an opinion feedback questionnaire. To be able to assess users close to our originally stated ambition of anywhere augmentation, the participants performed each task only once. Performing a task several times on the two different test conditions would have biased the evaluation, as the participants would already have been aware of the task after the first trial. The video guide condition, i.e. automatic or expert cut, was randomly assigned and the participant was not aware of which condition was being delivered. Each participant, therefore, performed all three tasks watching two video guide conditions, i.e. two automatic cut videos and one expert cut video, or one automatic cut video and two expert cut videos. Hence, each video authoring condition was used 7 times in every task.

6.1.1 Video Guide Authoring Results

The automatic generated video guides were automatically extracted by GlaciAR while the experts performed the task during the training session. After that, the experts were given their videos of entire activities in the training session and asked to indicate the start and end point in each video.

<table>
<thead>
<tr>
<th>Task</th>
<th>Video condition</th>
<th>Average video length (s)</th>
<th>Overlapping percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oscilloscope (Osc)</td>
<td>Automatic</td>
<td>16.3(±3.24)</td>
<td>89.71%</td>
</tr>
<tr>
<td></td>
<td>Expert</td>
<td>16.0(±2.65)</td>
<td>88.71%</td>
</tr>
<tr>
<td>Screwdriver (Scr)</td>
<td>Automatic</td>
<td>14.43(±3.20)</td>
<td>86.69%</td>
</tr>
<tr>
<td></td>
<td>Expert</td>
<td>16.33(±3.21)</td>
<td>85.97%</td>
</tr>
<tr>
<td>Sewing machine (Sew)</td>
<td>Automatic</td>
<td>21.87(±3.19)</td>
<td>89.27%</td>
</tr>
<tr>
<td></td>
<td>Expert</td>
<td>25.00(±3.60)</td>
<td>89.27%</td>
</tr>
</tbody>
</table>

Table 2: The average length (s) of the video guide generated by GlaciAR and expert cut, and the percentage of overlapping (%) of the automatic and expert authoring video guide condition.

Table 1: Task descriptions in the video authoring evaluation.

Figure 12 shows the testing scenarios and the screenshots taken in the three evaluated tasks.
Table 3: Success rate (%), average completion time (seconds) and average number of videos (times) needed for performing the tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>Video condition</th>
<th>Success rate</th>
<th>Completion time</th>
<th>No. of videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Osc</td>
<td>Auto</td>
<td>100</td>
<td>67.14(±9.67)</td>
<td>4.00(±1.00)</td>
</tr>
<tr>
<td></td>
<td>Expert</td>
<td>100</td>
<td>67.00(±12.17)</td>
<td>3.86(±0.38)</td>
</tr>
<tr>
<td>Scr</td>
<td>Auto</td>
<td>92.86</td>
<td>73.29(±20.30)</td>
<td>3.86(±0.69)</td>
</tr>
<tr>
<td></td>
<td>Expert</td>
<td>92.86</td>
<td>71.14(±22.26)</td>
<td>4.00(±1.15)</td>
</tr>
<tr>
<td>Sew</td>
<td>Auto</td>
<td>85.71</td>
<td>134.86(±48.22)</td>
<td>6.14(±1.95)</td>
</tr>
<tr>
<td></td>
<td>Expert</td>
<td>85.71</td>
<td>123.00(±34.93)</td>
<td>4.71(±0.76)</td>
</tr>
</tbody>
</table>

The results obtained via this objective evaluation follow the impressions captured by the NASA-TLX scores which are omitted for brevity.

6.1.3 The Participant’s Opinions

Participants’ feedback was also collected and actual quotes are presented in Table 4. These are separated into positive and negative feedback as well as additional opinions.

In the positive feedback, the participants overall felt that the video guide was convenient and easy to follow and highlighted they did not require an instruction manual and could perform the tasks by just following the video guide. For the negative feedback, some participants found the video guide was too small and quick and one participant was confused by the expert’s ego-motion in the video guide. In the screwdriver and sewing machine tasks, the expert’s hand covered parts of the object which blocked the view of details and caused them to skip one step in those tasks. Furthermore, the number of steps and details of the task affected the participants and prevented them from finishing the step of the task. As shown in the sewing machine task, only 4 out of 14 participants were able to perform the task perfectly.

Overall feedback shows that current implementation of GlaciAR is good for task sequences with three or four steps. More extended sequences or tasks with small details may cause the user to skip some steps and not finish the task completely. There are also cases of having to deal with occlusions made by the expert hand.

7. DISCUSSION AND CONCLUSION

In this paper, we develop and evaluate a new approach for automated capture and delivery of guidance within a self-contained eyewear computer.

The traditionally employed eye-gaze attention model is replaced by a head-motion attention model. By combining this attention model with a lightweight object detector, the system is able to de-
been considered ideal for other types of MR/AR formats such as those using 3D overlays.

It is thus somewhat surprising, yet encouraging that users were able to achieve the task with the small images on the Glass screen. In addition, the sensors and hardware used in GlaciAR are nowadays commonly available in mobile systems. GlaciAR uses an IMU, a small 2D display, and low computational visual requirements. Furthermore, as its operation requires no manual authoring and no synthesis or labelling, the results of the video guide authoring evaluation also suggest that the video guide produced by GlaciAR is as good as the one edited by experts. The results show that, for every task tested, there is no significant difference between the two video guide conditions.

In terms of directions for improvement, the participant comments offer some suggestions that are relatively easy to incorporate, but others would require an additional strategy. The participants, for example, reported that the view in the video guide was blocked by the expert’s hand, and people with glasses had problems watching the screen. Furthermore, the video guides for more complex tasks need to have strategies to help the user better follow the workflow, as demonstrated in the case of the sewing machine task. Recent work has proposed mechanisms to model and keep track of the workflow, as demonstrated in the case of the sewing machine task. Recent work has proposed mechanisms to model and keep track of the workflow [16, 13, 12, 1]. One approach, for example, that would improve this issue, is a mechanism to evaluate the user’s actions and workflow and highlight any missing steps.

GlaciAR aims towards that elusive concept of MR/AR smart eye-wear through which people can receive guidance to do any task anywhere and hence enabling cognitive augmentation. This aim imposes important challenges for conventional MR/AR systems, especially due to the problem of authoring and scalability. GlaciAR shows encouraging performance on three small-staged tasks and the discussion and results point to ways to improve the approach presented.

Further illustration on the way GlaciAR works is found on the accompanying video.

### 8. REFERENCES


