immersivePOV: Filming How-To Videos with a Head-Mounted 360° Action Camera

Kevin Huang
kev_huang@sfu.ca
University of Toronto, Simon Fraser University
Canada

Jiannan Li
jiannanli@dgp.toronto.edu
University of Toronto
Canada

Maurício Sousa
mauricio@dgp.toronto.edu
University of Toronto
Canada

Tovi Grossman
tovi@dgp.toronto.edu
University of Toronto
Canada

ABSTRACT

How-to videos are often shot using camera angles that may not be optimal for learning motor tasks, with a prevalent use of third-person perspective. We present *immersivePOV*, an approach to film how-to videos from an immersive first-person perspective using a head-mounted 360° action camera. immersivePOV how-to videos can be viewed in a Virtual Reality headset, giving the viewer an eye-level viewpoint with three Degrees of Freedom. We evaluated our approach with two everyday motor tasks against a baseline first-person perspective and a third-person perspective. In a between-subjects study, participants were assigned to watch the task videos and then replicate the tasks. Results suggest that immersivePOV reduced perceived cognitive load and facilitated task learning. We discuss how immersivePOV can also streamline the video production process for content creators. Altogether, we conclude that immersivePOV is an effective approach to film how-to videos for learners and content creators alike.

CCS CONCEPTS

• **Applied computing** → **E-learning**; • **Human-centered computing** → **User studies**.

KEYWORDS

how-to video, online learning, YouTube, 360° video, POV

ACM Reference Format:
Kevin Huang, Jiannan Li, Maurício Sousa, and Tovi Grossman. 2022. immersivePOV: Filming How-To Videos with a Head-Mounted 360° Action Camera. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3491102.3517468

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517468

1 INTRODUCTION

Online instructional videos have become ubiquitous with the growth of video hosting websites and Massive Open Online Course (MOOC) platforms. A popular subcategory of instructional videos are "how-tos" for everyday physical tasks (hereafter referred to as *how-to videos*), such as cooking, shirt folding, and knot tying. How-to videos are commonplace on YouTube, and they provide an accessible way to learn practical tasks [47, 68]. These how-to videos invariably involve a demonstration of the task, often with narration [18].

However, there is considerable variance between how-to videos with respect to design choices, and many are filmed in ways that may not be conducive to viewer engagement and learning. The production of these videos remains guided primarily by the intuitions of content creators and existing videos, rather than documented principles [21]. An important design choice that differs between videos is the perspective from which a video is filmed. Surveying popular YouTube how-to videos, we find that the use of third-person perspective (3pp) camera angles remain prevalent, despite literature suggesting that this may hinder learning due to what has been termed the perspective effect [25, 26, 70]. While content creators also use first-person perspective (1pp) angles, these are often over-the-shoulder or static close-up shots which do not provide immersive viewpoints.

We were curious as to why more immersive viewing angles were not being used to film how-to videos. The point of view (POV) shot has its origins in cinema and has been used for decades, to show audiences what a given character is seeing "from their eyes" — their POV. This shot was frequently used by venerated film director Alfred Hitchcock, featuring in his films from as early as the 1920s [60]. Modern-day improvements in action-camera technology have opened up the possibility of filming 360° Virtual Reality (VR) video from a POV.

In this work, we investigate the effectiveness of this more immersive POV (hence, *immersivePOV*) compared to conventional 3pp and POV approaches. We filmed our immersivePOV videos using a head-mounted GoPro MAX 360° action-camera which could then be viewed in a VR headset, providing an eye-level POV while also giving the viewer 3 Degrees of Freedom (3DoF). The idea was that this would result in a "what you see is what you (would) get" (WYSIWYWG) quality, where what a viewer can see when learning

a task is similar to what they would see when attempting it. We hypothesized that this WYSIWYWG quality of immersivePOV would facilitate task learning and engagement. To evaluate this, we created how-to videos for two practical motor tasks motivated by our YouTube video survey: the folding of a shirt using the "ranger roll" method, and the tying of shoelaces with a "double slipknot". We simultaneously filmed these tasks from three perspectives — static 3pp, traditional POV, and immersivePOV — yielding three conditions. We conducted a between-subjects user evaluation where participants watched the task videos from one of the three conditions and then attempted to replicate the tasks. We then compared learning outcomes and user preferences between the conditions. We find evidence to suggest that immersivePOV how-to videos offer improvements over not only 3pp videos, but also traditional POV.

With this work, we add to an emerging literature that seeks to establish a framework that can better inform effective how-to video design [24]. Summarizing, we make the following contributions in this paper: 1) a survey and analysis of the filming perspective used for a sample of how-to videos on YouTube, and 2) a user study evaluation of our immersivePOV approach against two conventional approaches We show that immersivePOV reduces perceived mental effort for learning physical tasks. We discuss how immersivePOV offers advantages compared to not only 3pp because of the perspective effect, but also compared to traditional POV through what we term an "immersion effect". immersivePOV puts the viewer in the demonstrator's shoes, minimizing any adverse effects that could result from perceived demonstrator-viewer differences.

2 RELATED WORK

Recently, Mayer et al. [46] provided a historical overview of instructional videos, positioning the current Internet video age as the fourth phase along this timeline. Recent work has emerged studying various aspects of instructional video design. Most relevant to our work are (1) studies examining the use of perspective, and (2) studies examining the instructional application of immersive video. We also discuss (3) interfaces for learning from physical tutorials.

2.1 The Perspective Effect

The role of perspective in visuomotor processing has long been studied in cognitive science and psychology. Previous research in this area has shown that 1pp better facilitates the processing of visuomotor information compared to 3pp [36, 70, 71], and may help generate a learning stance [42]. Indeed, the mental transformations required to take the spatial perspective of others can be cognitively demanding [30, 37].

Recent work has examined the role of perspective in the learning of physical tasks from videos. Garland and Sanchez [26] compared participants' ability to learn how to tie various knots to a bar. Participants watched videos of a demonstrator tying each knot, seeing either a 1pp over-the-shoulder view or a 3pp across-the-table view. Notably, participants who watched the videos from 1pp were significantly faster and more successful at tying their knots than those who watched the videos from 3pp. Fiorella et al. [25] examined the effect of video perspective in a more procedural task involving the assembly of model circuit boards in 1pp and 3pp conditions. Participants were then tasked with assembling the boards on their own. Participants in the 1pp condition were both faster and more accurate in assembling their circuit boards. This "perspective effect" was found to be greater for more complex circuit boards compared to simpler ones. However, Boucheix et al. [12] found that evidence for a perspective effect was not as clear. In their study, nursing students watched a video of a complex medical hand procedure filmed either entirely from a 1pp over-the-shoulder view, entirely in a 3pp face-to-face view, or using a mix of both perspectives. Yet, participants who watched the procedure from a mix of both perspectives performed better than their 1pp-only or 3pp-only counterparts. The authors point out that these findings may be due to the medical procedure's dynamic nature, involving a larger task space with more body movement and changes in space than the knot tying studied by Garland and Sanchez [26]. Boucheix et al. [12] suggest that 1pp might instead be more effective for spatially constrained and focused tasks, and that the perspective effect would be less pronounced for more dynamic tasks.

In the above studies, all authors suggest that there is a need for further research looking at procedural tasks in different contexts. Moreover, we note that the 1pp viewpoints of the above studies on task learning were filmed in what we term "quasi-POV". That is, the authors recorded the videos from first-person viewpoints but using static shots [25], or from angles (such as over-the-shoulder) [12] which may not offer much in the way of viewer immersion.

2.2 Point-of-View & 360° VR Instructional Videos

Some studies have recently explored the use of action cameras to film POV instructional videos. Bright et al. [13] used an action camera to film chest-mounted and head-mounted instructional videos for psychomotor clinical skills. While the authors noted challenges involved in producing the videos, such as with sound quality and video stability, they expressed confidence in the overall instructional value of the produced videos. In another study, Fung [44] experimented with both a chest-mount and head-mount for university chemistry lab modules video demonstrations, and found that providing the videos improved student performance.

Meanwhile, early studies on instructional 360° VR videos revealed challenges concerning production quality [7, 34], as well as with their effectiveness as a medium for learning [58]. Yet, Rupp et al. [59] found that watching a 360° VR instructional video of the International Space Station in the newer Oculus CV1 headset resulted in increased subject-matter interest and better learning outcomes. Compared to identical 2D video content, 360° VR instructional video can be more positively received by students [69] and result in significantly higher engagement and lower distraction [29]. Furthermore, Yoganathan et al. [74] examined 360° VR video instruction for surgical knot tying, and found that medical practitioners who watched the procedure in 360° VR using a headset had significantly higher knot tying scores compared to those who watched on a laptop screen. These results are striking because of the constrained nature of the task and the task space, and the apparent lack of added benefit of watching the video in 360° VR in this context. Altogether, we conclude that the use of both POV

and 360° VR in instructional contexts shows promise and warrants further investigation.

2.3 Interfaces for Physical Tutorial Videos

Recent work has also examined the use of VR, Augmented Reality (AR), and other novel interfaces designed to help with learning from physical tutorials. Some AR interfaces facilitate real-time remote instruction [66] and self-training of physical movements [5]. However, the focus of this work is on how-to video instruction, and there are other interfaces which have recently been proposed which implement or augment existing how-to videos. Chi et al. [18] created a video editing system designed to help content creators segment instructional videos for physical tasks. Other work has looked to facilitate how-to video learning by generating interactive overviews for recipe videos [47], and by allowing for voice-based video navigation [16]. This recent work serves to highlight the growing HCI research focus on instructional video learning. However, we note that the focus of this literature has concerned various aspects of the post-production process rather than on video production itself (i.e. how videos are filmed).

3 VIDEO SURVEY

Among the existing literature, we noted a lack of investigation into how how-to videos were being popularly filmed. We conducted a survey of popular YouTube how-to videos to better understand the landscape, and here we contextualize trends that we observed with respect to the related work discussed above.

3.1 Rationale

While instructional videos can cover any educational content, such as programming tutorials [40] and other technical concepts [17], how-to videos are more restricted in scope. Specifically, we define how-to videos as videos which feature a complete demonstration of a practical motor task. How-to videos may include text or other graphics, and usually involve the demonstrator narrating their actions, either in real time or with a voice-over in post-production.

A prototypical model of a how-to video is that of a demonstrator performing a task while "talking" to the camera. We refer to this as a "show and tell" style of how-to video, where the demonstrator stands face-to-face with the camera (i.e. the viewer) to present a given task. This style is perhaps most closely associated with cooking videos by celebrity chefs such as Gordon Ramsay or Jamie Oliver [51, 55]. In these videos, the demonstrator is highly salient, and there is a cinematic quality to the video production as a whole. Multi-camera setups are often used, with frequent cutting between different camera angles and intricate post-production editing.

Our sense was that this show and tell style of how-to video was fairly widespread across YouTube, even among non-celebrity content creators. Although these videos tend to be less cinematic and involve primarily single-camera shots, the same show and tell paradigm is often used. This might be motivated by content creators' desire to increase face time or to imitate high-production videos, as part of larger concerns about reaching audiences and navigating the YouTube search algorithm [73]. The show and tell style is most similar to traditional lectures and to how people are taught in school, and is also arguably the most straightforward to film — these reasons might also contribute to the prevalence of the style.

A natural consequence of this show and tell style of filming is that it results in a predominantly 3pp view of the demonstration, which prior work has shown may not be conducive to learning such tasks. Crucially, our feeling was that this 3pp bias persists even in videos not filmed in this style, and in which demonstrators do not feature as prominently. To investigate this, we surveyed a variety of YouTube how-to videos across five fairly common physical tasks: folding clothes, tying knots, changing bike tires, making mayonnaise, and pruning bonsai trees. Similar tasks have also been featured in the related work outlined above. Studies exploring novel interfaces for tutorials have looked at recipe videos [47], bike repair [16, 68], and general DIY [18], while knot tying has featured prominently in aforementioned studies on the perspective effect [26] and 360° VR instruction [74].

	Subs	Views	Likes	Comments
"how to fold clothes efficiently"	1.53m	6.41m	93.5k	2.2k
"best knots to learn how to tie"	1.82m	3.66m	48.8k	1.6k
"how to change a bike tire"	1.55m	1.04m	8.67k	0.43k
"how to make mayonnaise"	2.86m	2.69m	30.0k	1.1k
"how to prune bonsai trees"	88.6k	148k	2.13k	0.15k

Table 1: Average number of subscribers, views, likes, and comments for each query.

3.2 Analysis

We created search queries for each of the five selected tasks, including the term "how to" in each query (see Table 1). The first author then entered these queries in a private YouTube browsing window so that the search results would not be affected by account search and watch history [31]. The top hits that were returned from each query were then compiled, without any filtering by the first author. We defined "top hits" as the top 12 videos that were returned for each query. While there are many factors that contribute to YouTube search results, the rationale for selecting these top videos was that the YouTube search algorithm displays the most relevant videos first [19], and that YouTube usually displays the top 12 videos before other search headers such as "People Also Watched". Average viewing and engagement statistics for videos from each query are shown in Table 1. We note that we do not distinguish between amateur/professional and celebrity/non-celebrity content creators in our analysis, as these lines have become increasingly blurred with the video recording capabilities of modern smartphones and with Internet culture. The focus of our analysis was instead on the use of perspective in the surveyed videos.

Accordingly, we analyzed and categorized each of the 60 videos (12 per query) based on the filming perspective used. Broad categories ranged from predominantly 3pp, to a roughly even mix of 3pp and 1pp, to predominantly 1pp. For videos filmed predominantly in either 3pp or 1pp, we further subcategorized each video by the predominant type of 3pp or 1pp shot used. Mixed videos used these different 3pp and 1pp shots, with roughly even frequency. An overview of our analysis is shown in Table 2.

	3pp		Mixed	1pp				
	FtF	AT		OtS	Close-up	Top-down	POV	360° VR
"how to fold clothes efficiently"	1	4	4	1	1	1	0	0
"best knots to learn how to tie"	4	0	3	0	5	0	0	0
"how to change a bike tire"	5	2	1	3	1	0	0	0
"how to make mayonnaise"	4	4	1	0	2	0	1	0
"how to prune bonsai trees"	3	1	1	5	2	0	0	0
	17	11	10	9	11	1	1	0
Total	**28**		**10**			**22**		

Table 2: Results from our video survey. The top 12 videos were compiled from each query, and categorized based on the predominant filming perspective and shot used. (FtF — face-to-face; AT — across-the-table; OtS — over-the-shoulder)

Our survey confirmed our hypothesis that there was a widespread use of 3pp among popular how-to videos. We found that nearly half (28/60) of the surveyed videos predominantly used 3pp. These included videos filmed in a show and tell style, with the demonstrator prominently featured and engaged with the camera in face-to-face (FtF) shots (see Figure 1). There were also videos where demonstrators did not feature as prominently, or at all, which were still filmed in 3pp. Such videos used an across-the-table (AT) angle, similar to previous perspective effect studies [25, 26]. While we retain this term used in prior work, we note that these shots do not have to literally be from across a table, but rather just from the viewpoint of an observer standing across from the demonstrator, looking down (Figure 2). We point out that videos categorized as Mixed also made use of 3pp shots, and indeed may have also been filmed in a show and tell style, while incorporating 1pp shots.

We did find a number of videos (22/60) that were filmed predominantly in 1pp. However, it is important to differentiate between 1pp shots as not all are created equal. One common 1pp shot is an over-the-shoulder (OtS) shot, where the demonstrator is filmed from just slightly behind and off to their side (over their shoulder), as seen in Figure 3. Another common shot is a close-up shot, showing a zoomed-in view of the demonstrator's hands, as shown in Figure 4a. Less common is a top-down shot, where the demonstrator and the task space are filmed from above (Figure 4b). As we noted earlier, both over-the-shoulder and close-up shots were used to film the task videos in prior studies examining the perspective effect [12, 25, 26]. Among our surveyed videos, we observed that over-the-shoulder shots can result in an occluded view of the task space. Close-up and top-down shots can offer a clearer view, but these can be from awkward or unnatural angles which are not WYSIWYG.

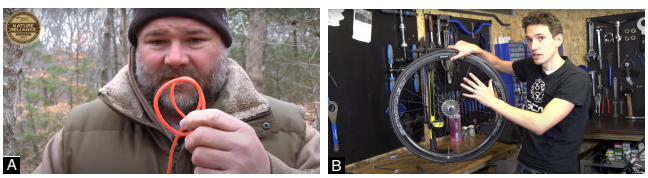

Figure 1: We came across a number of show and tell style demonstrations among our surveyed videos: (a) A demonstration of a clove hitch knot [57], (b) Changing a bike tire [48]

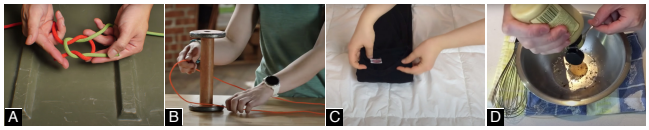

Figure 2: Example frames of videos surveyed. 3pp is often used even when demonstrators do not feature as prominently: (a) Tying a square knot [52], (b) Tying a trucker's hitch. A lower camera angle is used, but the demonstrator's face is not visible during the demonstration [56], (c) Folding a t-shirt [65], (d) Making mayonnaise [1]

We refer to these 1pp shots (over-the-shoulder, close-up, top-down) as being quasi-POV, which we differentiate from the head-mounted POV videos discussed in the Related Work. Our video survey revealed that the use of head-mounted POV was rare. In fact, we encountered only two videos featuring it. In one, a video on folding clothes with the "ranger roll" method, POV was mixed in with 3pp angles (see Figure 5a). The second video, on how to make mayonnaise, featured a nearly exclusive use of POV filmed in a continuous long take (Figure 5b). We found that the videomaker, chef Kenji-Lopez Alt ("Kenji"), regularly films how-to cooking videos in this head-mounted POV style. In these videos, Kenji films himself cooking various recipes from start to finish, usually in a single continuous take without edits — many videos are upwards of 20 minutes in length [4]. Kenji's videos have been well received, consistently receiving hundreds of thousands of views with a high number of likes. Looking at Kenji's videos we did find them to be engaging to watch, creating a greater sense of immersion and presence we felt was lacking from quasi-POV shots. However, we note that POV videos restrict the viewer to only seeing where the demonstrator is looking at a given point in time. While following along with the demonstrator's gaze might be helpful for task learning, we wondered if this might also result in less viewer agency and immersion than if the viewer were able to freely look around, as in 360° VR videos.

Indeed, 360° VR videos provide the viewer with 3DoF and we have seen that their use in instructional contexts has been promising. However, our survey did not reveal a single how-to video filmed in 360° VR. Looking beyond the surveyed videos, we found one 360° VR how-to video of a physical therapy demonstration which illustrates this. In the video [63], a cameraperson holds a 360° camera in their hands while standing perfectly still to film the physical demonstration, resulting in an awkward viewing angle and a lack of WYSIWYG. This kind of 360° VR video provides 3DoF but

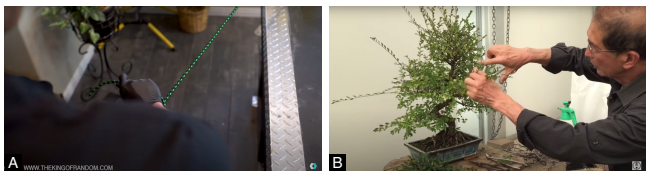

Figure 3: The over-the-shoulder is a common 1pp shot, but this can result in occlusions: (a) Tying a one-handed bowline knot [67], (b) Pruning a bonsai tree [11]

Figure 4: Other 1pp shots: a) A close-up shot of tying a figure-eight loop [50], b) A top-down shot. Note the inclusion of some graphics in this video [20]

is static and not filmed from an eye-level, similar to much of the related work on 360° VR instruction.

There were a number of findings from this preliminary survey that informed our evaluation. Although prior work suggests that filming task videos in 1pp rather than 3pp improves learning outcomes — the perspective effect — our survey revealed a widespread use of 3pp. While 1pp is also used, this is often from non-WYSIWYWG quasi-POV angles. The use of POV and 360° VR to film how-to videos may offer improvements, but these have individual limitations. We believe that immersivePOV combines the strengths of POV and 360° VR, and we set to evaluate this against the 1pp and 3pp that we observed in our survey. We focused our evaluation on the two tasks that we found had the highest amount of average views and engagement — knot tying and shirt folding.

4 IMMERSIVE POV

The recent development of more portable 360° action cameras has allowed for the filming of high-quality head-mounted 360° video filmed from a POV, which we believed would provide an immersive POV (*immersivePOV)*. We hypothesized that how-to videos could be effectively filmed in immersivePOV, and that immersivePOV videos viewed in a VR headset would provide a 3DoF eye-level POV viewing experience, and a *what you see is what you would (would) get (WYSIWYWG)* viewing experience.

We believed that these characteristics of immersivePOV would offer key improvements over existing approaches. Chief among these was the WYSIWYWG experience that results from putting the learner in the demonstrator's shoes and seeing from the demonstrator's eyes, with an eye-level 3DoF POV. In addition, we thought that the single long take nature of immersivePOV might also can make it easier to follow along a task without disruptive cuts between angles [43]. immersivePOV shows promise as an effective way for both producing and viewing how-to videos.

To film in immersivePOV, we experimented with mounting a GoPro MAX to a helmet. We found that mounting the camera at

Figure 5: POV: a) this video cuts between 3pp and POV [2], b) a long take POV video [4]

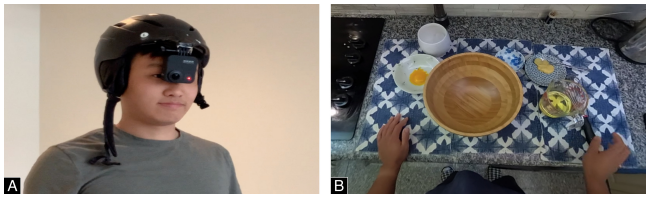

Figure 6: a) Our immersivePOV filming setup. b) The view from immersivePOV in a VR headset.

various points atop the helmet resulted in the viewpoint feeling too "tall", and above the eye-level of the demonstrator — this was also reported in a previous work examining viewer perceptions of various 360° camera placements [53]. We concluded that an overheard camera placement did not result in WYSIWYWG. Instead mounting the camera so that it was dangling roughly 5 cm in front of the demonstrator's eyes resulted in an eye-level view, and greater WYSIWYWG (see Figure 6a). Figure 6b shows an immersivePOV view, from an Oculus Quest VR headset, for a test video that we filmed demonstrating the making of homemade mayonnaise. Wearing the headset, the viewer can freely look around the kitchen and the task space with 3DoF. Video stabilization and horizon leveling [38] support on recent action cameras can help provide a smooth viewing experience without motion sickness.

5 EVALUATION

To test the effectiveness of immersivePOV how-to videos and examine its potential advantages, we conducted a user evaluation. Our evaluation was designed to examine the following research questions:

RQ1. Does immersivePOV lead to better task completion and learning outcomes than a static 3pp video, thus supporting the perspective effect theory?

RQ2. Does immersivePOV also lead to better task completion and learning outcomes than POV? In other words, we wanted to find out if the benefits of immersivePOV were also attributable to the immersive viewing experience and not just the 1pp filming angle.

5.1 Tasks

To investigate these questions, we first selected two motor tasks drawn from the video survey — (1) folding a t-shirt with the "ranger roll" method, and (2) tying a shoelace with a "double slipknot". In selecting the tasks, we reasoned that they needed to be neither too trivial nor too difficult to complete. At the same time, they had to be novel enough so that participants would likely not have prior exposure to them. Tasks also had to be objectively evaluable, with discrete steps and clear success/failure conditions. Owing to the remote nature of the study, task materials needed to be items that participants would likely have at home. Ultimately, we decided that the ranger roll and double slipknot fit all of the above criteria.

5.1.1 Ranger Roll. A ranger roll is a method of folding an article of clothing — in our case a crew neck t-shirt — into a self-enclosed roll. Our ranger roll was based off the ranger roll videos discovered in the video survey [2, 6], and consisted of five discrete steps:

(1) Folding the bottom of the shirt under 1-2 inches
(2) Folding the sides of the shirt over the middle
(3) Turning the shirt around
(4) Tightly rolling the shirt up to the end
(5) Wrapping the shirt into a pouch

Unlike other folding methods such as the *KonMari* fold [39], the ranger roll has discrete and discernable steps which made it ideal for evaluating task performance. An important characteristic of the ranger roll is that involves a 180° rotation of the shirt in order to roll it up. Object rotation did not feature in the motor tasks studied in the related work [25, 26], and how this rotation would be perceived between the video conditions was of interest to us.

5.1.2 Double Slipknot. A double slipknot for shoelaces — developed by Ian Fieggen and hence also known as "Ian's secure shoelace knot" [23] — is a method of tying shoelaces that is designed to be more secure than a conventional knot, such as the "rabbit-around-the-tree" method. Our double slipknot demonstration closely followed that of Ian Fieggen's [54], consisting of five discrete steps:

(1) Making a starting knot
(2) Making two loops
(3) Crossing the loops over, right-over-left (thus forming a hole)
(4) Threading the right loop through the front of the hole, and the left loop through the back of the hole
(5) Pulling on the loops

The double slipknot was selected to serve as a stand-in for the various knot tying demonstrations outlined in the video survey. In general, we expected that participants would find the double slipknot to be the more difficult task compared to the ranger roll. Despite both tasks being similar in length, we reasoned that knot tying would require more dexterity compared to folding up a shirt. We also believed the double slipknot to be the more perspective sensitive task, and we expected that participants in the 3pp condition might have a harder time with the double slipknot than the ranger roll.

5.2 Conditions

We filmed our own how-to videos for these tasks in immersivePOV, along with a POV baseline and a static 3pp control. That is, we produced three content-identical how-to videos for each task, yielding three conditions: 1) *a static 3pp video*, 2) *a POV video*, and 3) *an immersivePOV video* (see Figure 7).

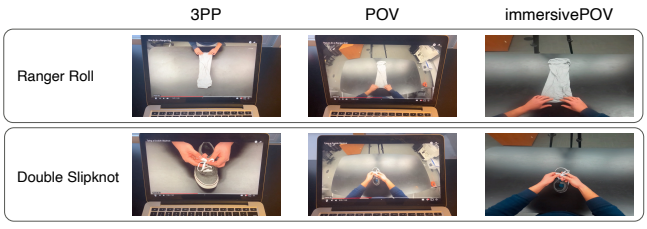

Figure 7: We filmed how-to videos for the ranger roll and double slipknot. Participants watched the videos either in: 3pp on a laptop, POV on a laptop, or immersivePOV in a VR headset

The static 3pp videos were filmed from an across-the-table (AT) angle, with only the demonstrator's hands visible. This angle was chosen because of its prevalence in the Video Survey, with 4/12 clothes folding videos being filmed entirely from an AT angle. While no knot-tying videos were filmed entirely in this manner, we note that the mixed-perspective videos used this AT angle (see Figure 2a and 2b). We also note that this was the filming angle used by Fiorella et al. [25] and Garland and Sanchez [26]. Furthermore, an AT angle hides the demonstrator's face — the visibility of a demonstrator, as with a FTF angle, would introduce new variables and potential confounds [9, 15, 62]. Meanwhile, the POV videos were in the style of Kenji-Lopez with a "pure" head-mounted POV view, and the immersivePOV videos were filmed in the manner described in Section 4.

5.3 Video Production

Both tasks were filmed in a lab environment, on the same black tabletop workspace (shown in Figure 7). Using the helmet mount, the demonstrator filmed himself performing both tasks with the GoPro MAX's 360° video mode in a single long take, while describing his actions in a conversational tone characteristic of how-to videos — this yielded the immersivePOV videos for both tasks. At the same time, an Apple iPhoneX mounted to a tripod from across the table filmed the 3pp videos for both tasks. Using the *Overcapture* editing tool of the GoPro app, we then created standard POV videos from the 360° videos. This resulted in content-identical videos corresponding to the three conditions: 3pp, POV, and immersivePOV. The videos for each task were then uploaded to YouTube with the same titles ("How to do a Ranger Roll" and "Tying a Double Slipknot") for the user evaluation. The videos were similar in length, with the ranger roll videos at 55s and the double slipknot videos at 63s.

5.4 Setup and Materials

The study was conducted during the COVID-19 pandemic and the resulting restrictions on in-person studies necessitated the conducting of the user evaluation remotely, by way of Zoom videoconferencing. With the evaluation conducted in this manner, measures were taken to introduce some controls across participants' various home environments. A screening form was sent out to prospective participants which went over the environment and materials required for the user evaluation: a sufficiently large table workspace, a crew neck cotton t-shirt, and a sneaker with laces.

5.5 Participants

Since the study was conducted remotely, we needed to recruit participants who had their own VR headsets for the immersivePOV condition. These participants (2 females, 6 males; M_{age} = 26.125, SD_{age} = 3.44) were drawn from the same population as the other participants, and there were no notable demographic differences. However, this placed a constraint on the total number of participants, and we felt that 8 participants per condition was the minimum viable number. This resulted in a total of 24 participants being recruited for the user evaluation (10 females, 14 males). Participants' ages ranged from 21 to 32, with a mean age of 24.3 years (SD = 2.5). None of the participants reported prior familiarity with either the

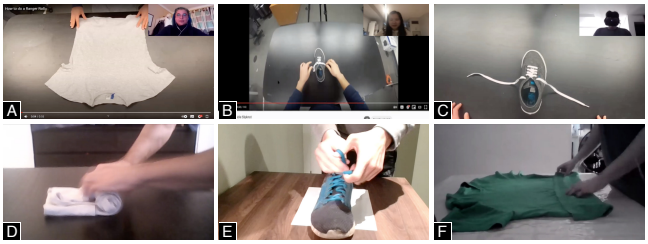

Figure 8: Participants watching a) the ranger roll in 3pp, b) the double slipknot in POV, c) the double slipknot in immersivePOV. Note the Oculus Quest headset that the participant is wearing here. Participants attempting d) the ranger roll, e) the double slipknot, f) the ranger roll. This participant has made the mistake of folding the bottom of the shirt up.

ranger roll or double slipknot tasks. All participants gave informed consent and were compensated $20 CAD for their participation.

5.6 Design and Procedure

We used a between-subjects design with three conditions, corresponding to the three video conditions: 3pp, POV, and immersivePOV (8 participants per condition). Participants with either an Oculus Quest or Oculus Quest 2 were recruited specifically for the immersivePOV condition. All other participants were randomly assigned to either the 3pp or POV conditions.

The procedure was the same for all three conditions. Consent forms were signed and submitted by participants before beginning the study, along with profile questionnaires. Each participant connected to a secure Zoom meeting hosted by the experimenter, where they were first given an overview of the study. The experimenter also checked with the participant to confirm that they had the prerequisite materials and workspace setup prepared, as specified in the screening form.

Participants watched either the ranger roll or double slipknot video first (task order was counterbalanced across conditions), in their assigned viewing condition. They were asked to share their screen, and were reminded that they could take as long as they needed to watch the video and in whatever fashion they wanted (e.g., pausing, watching multiple times, going forward and back), with the goal of replicating the task afterwards. Participants were not allowed to follow along with the videos to practice the given task while viewing. The immediate reason for this was to accommodate the VR condition, as participants using a VR headset would not be able to follow along with a task video. While we did trade off some ecological validity for internal validity — following along with how-to videos is common behaviour [16] — we note that the short length of our task videos would be more conducive to our watch-then-attempt methodology [28]. We note that this watch-then-attempt methodology was also used by Garland and Sanchez [26] and Fiorella et al. [25].

After viewing, participants set up the task materials and their laptop cameras so that the experimenter could see the task space, and then attempted the task (see Figure 8) — again they were told that they could take as long as they needed. Participants then filled out the post-task questionnaire. The process was then repeated for

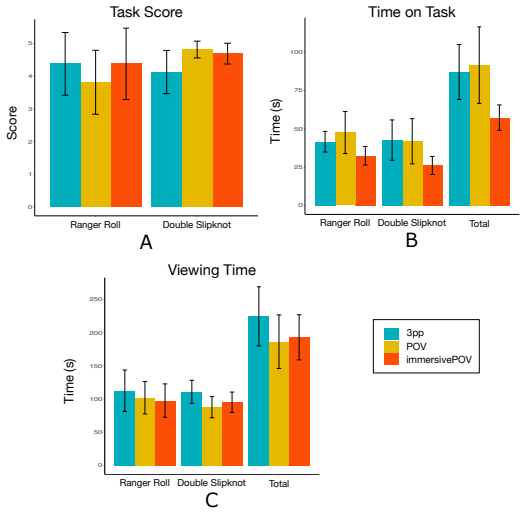

Figure 9: a) Mean task score for each group (maximum score = 5.0) b) Mean time on task for each group c) Mean viewing time for each group. Error bars represent 95% confidence intervals.

the second task, and the study concluded with a debrief session and guided interview. The studies lasted approximately 45 minutes.

6 RESULTS

We measured task performance, time spent attempting each task, and time spent watching each task video, along with playback behaviour (i.e., viewing timestamps). We compiled this data by analysing the screen recordings of each Zoom session in all condition. For the immersivePOV condition, we also asked participants to screencast the VR headsets to their computer screen. Finally, we gathered user preferences data for task difficulty, mental effort, helpfulness of the demonstrations, and helpfulness of the task videos, using a post-task questionnaire.

6.1 Task Performance

Each task had five steps and we scored each step using 1 point for completing a step, 0.5 points for an inaccuracy, -1 point for making a critical error. An inaccuracy was defined as a minor mistake that does not prevent the completion of the task, while a critical error was one which would make it impossible to complete the task correctly (e.g. failing to turn the shirt around for the ranger roll). The maximum attainable score on a given task was 5 points. As both the rubric and the steps were reasonably objective, scoring was done by one of the researchers. However, to mitigate concerns of bias, a second researcher scored 10 out of the 48 task recordings independently, which resulted in a high level of agreement (Cohen's kappa = 0.88) providing sufficient evidence for the reliability of the scores.

Across all conditions, 19/24 participants were able to successfully complete the ranger roll. Broken down per group, 7/8 participants were successful in the 3pp condition, 5/8 in the POV condition, and 7/8 in the immersivePOV condition. We found that the POV group

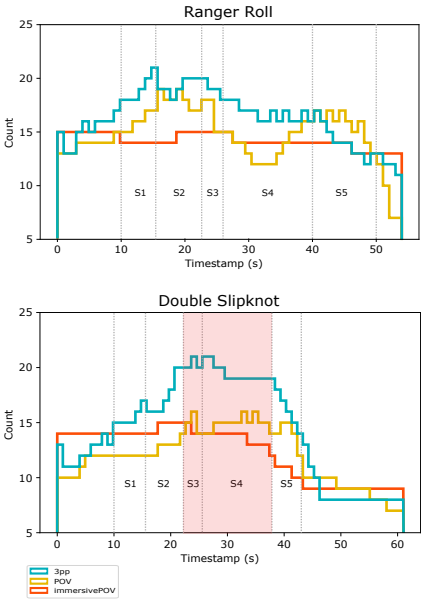

Figure 10: Histograms of group viewing behaviour for both tasks, indicating how frequently a given timestamp was viewed. Steps for each task are outlined, and steps where there were significant between-group differences are shaded red. a) Ranger Roll: S1 – Folding bottom of shirt under 1-2 inches; S2 – Folding sides of shirt over middle; S3 – Turning shirt around; S4 – Tightly rolling shirt up to end; S5 – Wrapping shirt into pouch. b) Double Slipknot: S1 – Making starting knot; S2 – Making two loops; S3 – Crossing loops over; S4 – Right loop through front, left loop through back; S5 – Pulling on loops

(M = 3.81, SD = 1.41) scored lower than both the 3pp (M = 4.38, SD = 1.38) and the immersivePOV (M = 4.38, SD = 1.58) groups, although a Kruskal-Wallis test by rank revealed that these differences were not significant, $\chi^2(2) = 1.08$, $p = .583$.

Participants were generally successful with the double slipknot. Across all conditions, 22/24 participants were able to successfully tie a double slipknot. Broken down by group 6/8 3pp participants, all 8/8 POV participants, and all 8/8 immersivePOV participants were successful. We found that task score was lowest for the 3pp group (M = 4.13, SD = 0.95) compared to the POV (M = 4.81, SD = 0.37) and immersivePOV (M = 4.69, SD = 0.46) groups (see Figure 9a), but a Kruskal-Wallis test revealed that the differences between the groups were not significant, $\chi^2(2) = 3.327$, $p = .190$.

While task performance is an important consideration, they do not present a complete picture of overall task learning. For that, it is also important to weigh how long participants spent performing the task, as well as how long they spent watching the task videos.

6.2 Time on Task

Time on task was measured as the time it took participants to complete the given task, or until they indicated that they were stuck and unable to finish. A one-way independent measures ANOVA revealed that total time on task (across both tasks) was significantly different between groups, $F(2, 18) = 3.937$, $p = .038$, $\eta^2 = .304$. A post-hoc Tukey's test revealed that total time on task was significantly lower among participants in the immersivePOV group (M = 57.29, SD = 11.15) compared to participants in the POV group (M = 91.57, SD = 33.74), $p = .047$ (Figure 9b). The difference between the immersivePOV and 3pp group (M = 87.0, SD = 24.2) was not found to be statistically significant, $p = .091$.

6.3 Viewing Time

Viewing time for the ranger roll video was the highest among the 3pp group (M = 112.5, SD = 44.8) (Figure 9c), but a one-way ANOVA revealed no significant differences between groups, $F(2, 21) = .307$, $p = .739$. The 3pp group also had the highest viewing time (M = 110.9, SD = 25.0) for the double slipknot video, although differences between groups were again not found to be significantly different $F(2, 20) = 2.015$, $p = .16$. Comparing total viewing time across both tasks also revealed no significant between-group differences, $F(2, 20) = 1.013$, $p = .381$.

However, a more nuanced picture emerged when examining participant viewing behaviour for the task videos. As we had recorded how participants interacted with the videos, we could construct timelines of their playback history for each video (e.g. *P6, Double Slipknot*: played from 0:00 - 0:59, 0:18 - 0:43). This allowed us to create histograms of each group's viewing behaviour, which we mapped onto the steps demonstrated in the videos (see Figure 10). Graphically, we can see a spike in the 3pp participants' viewing time around the middle of the double slipknot video. This part of the video from 0:22 - 0:37 corresponds to the steps of crossing the loops over each other, and feeding each loop through the front and back. A one-way ANOVA confirmed there were significant differences in viewing time between groups during this important section of the video, $F(2, 21) = 4.52$, $p = .023$, $\eta^2 = .301$. A post-hoc Tukey's test revealed that viewing time in this section among 3pp participants (M = 39.4, SD = 8.0) was significantly higher than that of participants in the immersivePOV group (M = 27.8, SD = 7.4), $p = .026$.

6.4 Questionnaires

Participants filled out the post-task questionnaire after completing each task (see Table 3). For the ranger roll, Kruskal-Wallis tests by rank revealed significant between group differences both for reported viewing mental effort, $\chi^2(2) = 8.941$, $p = .011$, and task mental effort, $\chi^2(2) = 7.994$, $p = .018$. Post-hoc Dunn's pairwise tests confirmed that immersivePOV participants reported investing significantly less mental effort in watching the ranger roll video compared to both 3pp and POV participants (immersivePOV-3pp: $p = .017$, immersivePOV-POV: $p = .028$). Reported task mental effort was also significantly lower among immersivePOV participants compared to both groups (immersivePOV-3pp: $p = .032$, immersivePOV-POV: $p = .030$).

For the double slipknot, Kruskal-Wallis tests again revealed significant between group differences for both reported viewing mental effort ($\chi^2(2) = 7.390$, $p = .025$) and task mental effort ($\chi^2(2) = 9.784$, $p < .01$). Post-hoc tests confirmed that immersivePOV participants reported significantly less viewing mental effort compared to 3pp participants ($p = .020$), but not POV participants ($p = .184$).

	Ranger Roll			*Double Slipknot*		
	3pp	**POV**	**immersivePOV**	**3pp**	**POV**	**immersivePOV**
The demonstrator's actions were clear	9 (0.25)	9 (1)	9 (0.25)	8.5 (1.25)	9 (0)	9 (1)
The instructions given by the demonstrator were helpful	9 (1)	8 (1.25)	9 (1)	7.5 (2.25)	8.5 (2)	9 (0)
Overall, I found the video was easy to follow	9 (1.25)	8 (1.25)	9 (1.25)	6.5 (3)	8.5 (1)	8 (1.25)
While watching the video, I could picture myself doing the task	8 (1)	9 (1.25)	9 (1.25)	8 (1.5)	7.5 (2)	9 (0)
Task Difficulty	2.5 (1.25)	2.5 (2.5)	1.5 (1.25)	5.5 (1.5)	3.5 (1.5)	4.5 (4.5)
Viewing Mental Effort *	5.5 (3)	5 (2.5)	3 (2)	6.5 (2.25)	5 (1.75)	2.5 (3)
Attempt Mental Effort *	5.5 (2.5)	5.5 (3)	2 (1.25)	6.5 (1.5)	4.5 (3.25)	3.5 (2.25)

Table 3: Results from the post-task questionnaire (Median, Interquartile Range). * denotes statistical significance, with red values significantly different compared to green.

immersivePOV participants also reported significantly less task mental effort compared to 3pp participants ($p < .01$), but again not compared to POV participants ($p = .124$).

Kruskal-Wallis tests revealed no significant between-group differences for all of the other questionnaire responses.

7 DISCUSSION

Heading into the user evaluation, our research questions centred around whether or not immersivePOV how-to videos would help facilitate task learning. Here, we discuss (i) evidence for a perspective effect, (ii) evidence for differences between the performance of the POV and immersivePOV groups, and then we (iii) discuss our findings in the context of our research questions.

7.1 Perspective Effect?

We had hypothesized that the 3pp participants would generally have more difficulty completing both tasks than their POV and immersivePOV counterparts. We expected that this would be more pronounced for the double slipknot than the ranger roll, as we believed the double slipknot to be the more perspective sensitive task. Prior work has shown evidence for a perspective effect with other knot-tying tasks [26], while tasks similar to the ranger roll have, to the best of our knowledge, not previously been studied.

In fact, the 3pp group performed better than expected on both tasks. Of the 3pp participants, 7/8 successfully completed a ranger roll while 6/8 successfully tied a double slipknot. 3pp participants did not score significantly worse, nor did they spend significantly more time on task than their 1pp counterparts. These findings do not at first seem to show strong evidence for a perspective effect, but a closer inspection reveals a more nuanced picture. As hypothesized, results appear to confirm that the double slipknot was a more perspective sensitive task than the ranger roll. 3pp participants tended to have a harder time completing the double slipknot than the ranger roll, with 3pp participants attaining the lowest double slipknot score despite scoring relatively well on the ranger roll. In addition, a closer analysis of viewing behaviour revealed that 3pp participants were taking significantly more time than immersivePOV participants to watch a key section of the double slipknot video where the two loops were crossed over (right-over-left), with one loop threaded through the front of the hole that is formed, and the other through the back (Steps 3 and 4, refer back to *Tasks*).

Tellingly, there were no significant differences in reported cognitive load between 3pp and POV participants for the ranger roll, with immersivePOV participants reporting significantly lower cognitive load than both of these groups. 3pp participants did in fact report significantly higher mental effort ratings (i.e. cognitive load) for the double slipknot compared to immersivePOV participants, while there were no differences between the POV and immersivePOV groups. The greater perspective sensitivity of the double slipknot was corroborated by debrief feedback from 3pp participants, with P11 noting that *"for the first one [double slipknot] I think if I had the video from my perspective it would be easier ... for the second [ranger roll] it didn't affect me as much"*. P12 also commented that *"with the second video [double slipknot] it was harder because the perspective was opposite to what I'd be seeing"*. While P12 noted that the perspective was of course also opposite for the ranger roll video, that task involved *"just less spatial coordination ... like it's a t-shirt you know"*. Other participants specifically commented on the challenge of watching and then replicating Steps 3 and 4 of the double slipknot: *"when you have the two loops and you cross them ... it's flipped for me"* (P2), *"when I was doing it it was backwards"* (P19).

Altogether, we found evidence of a perspective effect for the double slipknot, but not the ranger roll. We suggest that the concept of perspective sensitivity is an important consideration, serving to moderate the perspective effect in the case of the ranger roll. We suggest that tasks which involve a higher degree of demonstrator-relative direction might be more perspective sensitive, in line with work that has shown an egocentric bias in perspective taking [22, 64].

7.2 POV vs. immersivePOV

Perspective effect aside, a more striking finding was the underperformance of the POV group in comparison to the immersivePOV group. immersivePOV participants spent significantly less total time on task compared to their POV counterparts. As mentioned earlier, immersivePOV participants also reported significantly lower cognitive load for the ranger roll compared to POV participants. These findings bring to mind those by Yoganathan et al. [74], who found that doctors-in-training who watched the tying of a surgical knot in 360° VR were better able to tie the knots than those who watched a content-identical POV video.

The immersivePOV videos were notably well received by participants, especially with comments relating to the immersive WYSIWYWG quality of the videos. P5 commented that *"I could see like where my hands were supposed to be ... I was able to just easily follow with my hands"* and that he could *"imagine myself folding the shirt*

underneath" while watching the ranger roll video. Others echoed these sentiments as well, commenting about the videos that *"it seems like I'm the person doing it"* (P16), and that *"it was like [me] doing it"* (P10). While POV participants also commented on the WYSIWYWG quality of their videos, their feedback was somewhat more muted, and tended to be more along the lines of the POV perspective being helpful to them rather than immersive:*"it's useful to see what you would be seeing"* (P6), *"I think the up-down [POV] perspective was very helpful"* (P17).

We suggest that the increased immersion afforded by immersivePOV led to a greater WYSIWYWG quality, which in turn led to better task learning in immersivePOV participants. The idea that the immersion afforded by 360° VR educational videos might facilitate deeper learning, lead to a greater sense of presence [3], and more spatial awareness [43]. Here, we seem to find further evidence for an we might call an *immersion effect* of 360° VR, with immersivePOV providing advantages over POV. What is especially noteworthy is that here, as seen in the surgical knot tying study by Yoganathan et al., we observe this immersion effect despite the fact that participants did not make full use of the 3DoF afforded by VR. As was the case in Yoganathan et al., our tasks were constrained to a small tabletop task space such that participants had no need to, and indeed did not, have to look around in 360°. This suggests that the immersion effect is quite robust, and is not simply attributable to greater viewing DoF.

7.3 immersivePOV, redux

Overall, what emerged from the results was the effectiveness of the immersivePOV videos. Across both tasks, the immersivePOV group had the highest overall number of successes, as well as the highest overall performance score. immersivePOV participants spent the least total amount of time on task, suggesting greater task fluency — higher time on task was associated with participants' making false starts, hesitating, and taking time to recall steps. These differences could not be chalked up to differences in the perceived quality of demonstrator actions or instructions, for which there were no between-group differences. Therefore, we find evidence that the 1pp view of immersivePOV provides a viewing advantage for more perspective sensitive tasks (**RQ1**). We also find evidence for an immersion effect on task learning, with immersivePOV providing advantages over POV (**RQ2**). From these findings, we can conclude that immersivePOV is a promising medium with which to film how-to videos.

The picture that emerged was of the immersivePOV group being the most *efficient* at learning and then replicating the two tasks. We observed that the typical immersivePOV participant would watch a task video from start to finish once or twice, and then be able to complete the task fairly effortlessly, without hesitating or making false starts. immersivePOV participants tended to watch the videos in an interrupted and continuous manner, without the skipping back and forth that was characteristic of not only the 3pp group, but also the POV group. The immersivePOV condition seemed to induce a more continuous viewing behaviour than the other two videos. That is, immersivePOV participants tended to watch the immersivePOV videos in the spirit in which they were produced — as a single long take — to good effect. This seems to run counter to recent work which has suggested that long take videos might overload the viewer with information [45], and that instead segmenting videos into sections may improve learning in comparison [10]. Chi et al. [18] were motivated to create their aforementioned video editing system along these lines, with the authors noting that the long take nature of many amateur instructional videos could be "too long" and "include unnecessary or repetitive actions as well as mistakes" (p. 141). Here, we suggest that perhaps the long take nature of immersivePOV, and the continuous nature in which the immersivePOV participants watched the videos, actually has merit. Supporting this idea is Smith's Attentional Theory of Cinematic Continuity (AToCC) [61], which provides an overview of professional film editing techniques used to maintain a sense of visual continuity for the viewer — that is, to create a sense of flow in a scene which is in fact stitched together from different takes. Smith notes that "If a cut occurs that the viewers are not expecting, the visual transients created by the change from one shot to another will capture attention and a discontinuity will be perceived" (p. 9). While how-to videos are decidedly different from films — how-to videos are free of cinematic concerns such as plot development — we argue that this point about discontinuity might also be applicable to how-to videos. As we noted earlier (see *Video Survey*), the discontinuous cutting between different perspectives and camera angles is characteristic of a number of popular how-to videos. Perhaps the long take nature of immersivePOV, in conjunction with its WYSIWYWG quality, might aid with task learning.

As alluded to earlier, our immersivePOV videos remove the demonstrator from the picture and the implications of this bear further discussion. The prototypical show and tell style of filming how-to videos naturally involves a high degree of demonstrator presence and visibility, which has been the subject of some study in the instructional video literature. The Model-Observer Similarity (MOS) hypothesis [9] suggests that the more a viewer perceives themselves to be similar to the model, the greater the influence on feelings of self-efficacy. Early attempts to test this hypothesis within the context of instructional video learning by matching demonstrator age [32] and gender [33] have not found strong support for the hypothesis. In fact, Hoogeheide et al. [32] found that adolescents who learned from adult demonstrators attained better learning outcomes than adolescents who learned from adolescent demonstrators. Indeed, questions remain about the extent to which a demonstrator's appearance might affect viewer engagement and learning outcomes. We suggest that these effects may be adverse in nature, and demonstrator-viewer differences might activate stereotypes and "stereotype threats" which adversely affect engagement and learning [15, 62]. Perhaps instead putting the viewer in the demonstrator's shoes with immersivePOV enhances viewer confidence and perceived self-efficacy [8].

8 LIMITATIONS AND FUTURE WORK

The COVID-19 pandemic presented challenges and limitations to the present study. We acknowledge that the remote nature of the study inevitably introduced noise to our user evaluation. Although the experimenter was able to see each participant's environment and ensure there were no critical differences in setup, the fact remained that each participant's environment and materials were

slightly different. The remote study also made it difficult to recruit VR participants, as we needed to recruit those who owned their own Oculus Quest headsets. This group proved to be the limiting factor on the number of participants recruited, as we could only recruit as many participants in the non-VR groups as there were in the VR group. We also acknowledge that we introduced sampling bias for the immersivePOV group by recruiting VR headset owners. This demographic is often thought to skew a certain way, although differences between non-owners and owners may not be as appreciable as believed [35].

The tasks that we selected for the experiment were also constrained by the remote setup. Our use of the ranger roll and double slipknot tasks were in part motivated by the fact that t-shirts and sneakers were commonplace items that any would-be participants would likely have. While it was good that the tasks were novel and unfamiliar to participants, they were arguably too uncomplicated, and this may have dampened between-group differences and the perspective effect. We had experimented with filming a double-pulley task, as well as the making of mayonnaise, but these tasks were ultimately discarded due to the impracticality of having these tasks be attempted by participants in an at-home environment. As such, the generalizability of our findings should be considered in future work.

It could also be argued that the asynchronous manner in which we had participants watch the video and then attempt the task afterwards is not always representative of how how-to videos are utilized [16]. We acknowledge that in certain scenarios it may be more natural to follow along with the video. This preference for synchronous viewing was pointed out by a number of participants, and was also the focus of a recent paper on the use of pausing in the synchronous viewing of how-to videos [68]. As such, our results should be considered most applicable for asynchronous viewing, and further studies could investigate synchronous usage. Another limitation to acknowledge is that no follow-up retention tests were done. We note that such tests were also absent from prior work [25, 26]. Evaluating learning retention after some period of time (i.e. a day or a week) might be an important dimension of how-to video learning that could be an avenue for future work.

Future work looking at the effectiveness of immersivePOV should also explore different tasks. We were initially motivated to include the mayonnaise task because of its lack of perspective sensitivity and its quality of being a "symmetrical" task — the perspective from which a viewer watches someone mix ingredients in a bowl does not seem particularly important. It might be interesting to further explore this idea of task symmetry, and how this characteristic could affect the perspective effect or the immersion effect. More dynamic and longer tasks should also be investigated, for which the 3DoF afforded by immersivePOV might be utilized to greater effect. A number of immersivePOV participants expressed that they believed that immersivePOV would be a useful medium in which to film longer procedures, such as lab safety demonstrations. The potential application of immersivePOV in more formal educational settings, such as for engineering practicals (e.g. filming the assembly of a circuit board), seems like a promising future direction.

Future work could look at ways of making immersivePOV a synchronous viewing experience, namely through some incorporation of immersivePOV video into AR. We acknowledge that this study did not make use of AR, even though it may seem like an appropriate choice. AR has shown great promise for the guiding of physical tasks in real-time [49, 72]. Previous work has also looked into incorporating instructional videos into AR for synchronous task learning. However, these approaches have either involved complicated video processing [27], or the use of picture-in-picture video display [41] which may be limited in its effectiveness [14]. AR technology is also still relatively limited with respect to field-of-view and resolution compared to VR, and VR headsets are more widely accessible than AR headsets. We point out that our VR immersivePOV approach requires only access to a VR headset and YouTube VR to be used.

9 CONCLUSION

In this paper, we identified that many popular how-to videos on YouTube continue to be filmed using techniques that prior work has shown might not be conducive to viewer learning. Inspired by YouTube content creators as well as recent work on the instructional use of GoPro POV and 360° VR, we developed an approach for filming *immersivePOV* how-to videos. After conducting a user evaluation, we find support for the effectiveness of our immersivePOV approach, both for facilitating task learning as well as for video production. We believe that the WYSIWYWG quality and long take nature of immersivePOV makes it a promising medium for filming how-to videos.

We believe that immersivePOV could provide advantages for how-to video producers. immersivePOV can greatly streamline video production by allowing for point-and-shoot filming, bypassing the need to think about setting up and adjusting camera angles, as well as removing the need for additional camera-persons. Filming in a single long take also eliminates the need for editing in post-production, a process that can be time-consuming and technically demanding. Our immersivePOV videos were undoubtedly easy to produce, and present-day 360° action cameras represent a considerable improvement on a technology that was fairly inaccessible just a few years ago [34]. We foresee that this technology will improve in the coming years, and we believe that this will make the filming and viewing of immersivePOV videos even better. To better accommodate immersivePOV video production, we suggest that future action cameras should be small and light, incorporate voice control, and potentially be integrated into camera glasses. In a practical use case, we can imagine an individual using immersivePOV as a time-efficient way to demonstrate a fairly complex procedure, and then share that video with a colleague or coworker. We see promise for the use of immersivePOV video in other educational settings, and the potential for its incorporation with mixed reality task learning interfaces.

ACKNOWLEDGMENTS

This research was supported in part by the National Sciences and Engineering Research Council of Canada (NSERC) under Grant IRCPJ 545100 - 18.

REFERENCES

[1] French Cooking Academy. 2015. How To Make A French Mayonnaise Sauce (in just a few minutes). Video. Retrieved from https://youtu.be/AHGpyrRRlx0?t=63.
[2] MILE30 Adventures. 2016. How to Pack your Clothing Efficiently - Army Roll Method. Video. Retrieved from https://youtu.be/fuD-ZZydsVg?t=81.

[3] Tanja Aitamurto, Shuo Zhou, Sukolsak Sakshuwong, Jorge Saldivar, Yasamin Sadeghi, and Amy Tran. 2018. Sense of presence, attitude change, perspective-taking and usability in first-person split-sphere 360 video. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[4] J. Kenji-Lopez Alt. 2021. How to Make Mayonnaise. Video. Retrieved from https://youtu.be/9TnIeYc2CWU?t=73.

[5] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 311–320.

[6] armygringo. 2012. How to Fold T-Shirt for Vacation (Ranger Roll). Video. Retrieved from https://www.youtube.com/watch?v=so93nqxZLjM.

[7] Chetan Arora and Vivek Kwatra. 2018. Stabilizing first person 360 degree videos. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1405–1413.

[8] Albert Bandura. 1997. Self-efficacy: The exercise of control. (1997).

[9] Albert Bandura and Vilayanur S Ramachaudran. 1994. Encyclopedia of human behavior. *New York: Academic Press* 4 (1994), 71–81.

[10] Nicolas Biard, Salomé Cojean, and Eric Jamet. 2018. Effects of segmentation and pacing on procedural learning by video. *Computers in Human Behavior* 89 (2018), 411–417.

[11] Herons Bonsai. 2017. How to Prune a Chinese Elm Bonsai Tree EASY!!! Video. Retrieved from https://youtu.be/MzxwsfFHA-8?t=172.

[12] Jean-Michel Boucheix, Perrine Gauthier, Jean-Baptiste Fontaine, and Sandrine Jaffeux. 2018. Mixed camera viewpoints improve learning medical hand procedure from video in nurse training? *Computers in Human Behavior* 89 (2018), 418–429.

[13] Peter Bright, Bill Lord, Helen Forbes, Florin Oprescu, Nigel Barr, Terri Downer, Nicole M Phillips, Lauren McTier, Vilma Simbag, and Kristel Alla. 2015. Expert in my pocket: creating first person POV videos to enhance mobile learning. In *THETA 2015: Create, connect, consume-innovating today for tomorrow: Proceedings of the 2015 The Higher Education Technology Agenda Conference*. THETA, 1–15.

[14] Yuanzhi Cao, Xun Qian, Tianyi Wang, Rachel Lee, Ke Huo, and Karthik Ramani. 2020. An exploratory study of augmented reality presence for tutoring machine tasks. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[15] Felix Chang, Mufan Luo, Gregory Walton, Lauren Aguilar, and Jeremy Bailenson. 2019. Stereotype Threat in Virtual Learning Environments: Effects of Avatar Gender and Sexist Behavior on Women's Math Learning Outcomes. *Cyberpsychology, Behavior, and Social Networking* 22, 10 (2019), 634–640.

[16] Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to design voice based navigation for how-to videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[17] Rimika Chaudhury and Parmit K Chilana. 2019. How learners engage with in-context retrieval exercises in online informational videos. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*. 1–10.

[18] Pei-Yu Chi, Joyce Liu, Jason Linder, Mira Dontcheva, Wilmot Li, and Bjoern Hartmann. 2013. Democut: generating concise instructional videos for physical demonstrations. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 141–150.

[19] YouTube Creators. 2017. How YouTube Search Works. Video. Retrieved from https://www.youtube.com/watch?v=gTrLniP5tSQ&.

[20] DaveHax. 2013. How to Fold a Shirt in Under 2 Seconds. Video. Retrieved from https://youtu.be/uz6rjbw0ZA0?t=23.

[21] Björn B de Koning, Vincent Hoogerheide, and Jean-Michel Boucheix. 2018. Developments and trends in learning with instructional video. *Computers in Human Behavior* (2018).

[22] Nicholas Epley, Boaz Keysar, Leaf Van Boven, and Thomas Gilovich. 2004. Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology* 87, 3 (2004), 327.

[23] Ian Fieggen. 2020. *Ian's Secure Shoelace Knot.* https://www.fieggen.com/shoelace/secureknot.htm

[24] Logan Fiorella and Richard E Mayer. 2018. What works and doesn't work with instructional video.

[25] Logan Fiorella, Tamara van Gog, Vincent Hoogerheide, and Richard E Mayer. 2017. It's all a matter of perspective: Viewing first-person video modeling examples promotes learning of an assembly task. *Journal of Educational Psychology* 109, 5 (2017), 653.

[26] TB Garland and Christopher A Sanchez. 2013. Rotational perspective and learning procedural tasks from dynamic media. *Computers & Education* 69 (2013), 31–37.

[27] Michihiko Goto, Yuko Uematsu, Hideo Saito, Shuji Senda, and Akihiko Iketani. 2010. Task support system by displaying instructional video onto AR workspace. In *2010 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 83–90.

[28] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. 41–50.

[29] Cuan M Harrington, Dara O Kavanagh, Gemma Wright Ballester, Athena Wright Ballester, Patrick Dicker, Oscar Traynor, Arnold Hill, and Sean Tierney. 2018. 360 Operative videos: A randomised cross-over study evaluating attentiveness and information retention. *Journal of surgical education* 75, 4 (2018), 993–1000.

[30] Mary Hegarty and David Waller. 2004. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence* 32, 2 (2004), 175–191.

[31] YouTube Help. 2021. *Manage your recommendations and search results.* https://support.google.com/youtube/answer/6342839

[32] Vincent Hoogerheide, Sofie MM Loyens, and Tamara van Gog. 2016. Learning from video modeling examples: Does gender matter? *Instructional Science* 44, 1 (2016), 69–86.

[33] Vincent Hoogerheide, Margot van Wermeskerken, Hilke van Nassau, and Tamara van Gog. 2018. Model-observer similarity and task-appropriateness in learning from video modeling examples: Do model and student gender affect test performance, self-efficacy, and perceived competence? *Computers in Human Behavior* 89 (2018), 457–464.

[34] Sam Kavanagh, Andrew Luxton-Reilly, Burkhard Wüensche, and Beryl Plimmer. 2016. Creating 360 educational video: a case study. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*. 34–39.

[35] Jonathan W Kelly, Lucia A Cherep, Alex Lim, Taylor Doty, and Stephen B Gilbert. 2021. Who Are Virtual Reality Headset Owners? A Survey and Comparison of Headset Owners and Non-Owners. *PsyArXiv* (2021).

[36] Rachel Louise Kelly and Lewis Wheaton. 2013. Differential mechanisms of action understanding in left and right handed subjects: the role of perspective and handedness. *Frontiers in psychology* 4 (2013), 957.

[37] Klaus Kessler and Lindsey Anne Thomson. 2010. The embodied nature of spatial perspective taking: embodied transformation versus sensorimotor interference. *Cognition* 114, 1 (2010), 72–88.

[38] Abe Kislevitz. 2019. MOST STABLE CAMERA EVER? GoPro MAX HyperSmooth + Horizon Leveling. Video. Retrieved from https://www.youtube.com/watch?v=I4ZD8wnbehg.

[39] Marie Kondo. 2020. The KonMari Fold | Basics. Video. Retrieved from https://youtu.be/IjkmqbJTLBM.

[40] Bridjet Lee and Kasia Muldner. 2020. Instructional video design: Investigating the impact of monologue-and dialogue-style presentations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[41] Gun A Lee, Seungjun Ahn, William Hoff, and Mark Billinghurst. 2020. Enhancing first-person view task instruction videos with augmented reality cues. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 498–508.

[42] Robb Lindgren. 2012. Generating a learning stance through perspective-taking in a virtual environment. *Computers in Human Behavior* 28, 4 (2012), 1130–1139.

[43] Andrew MacQuarrie and Anthony Steed. 2017. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In *2017 IEEE Virtual Reality (VR)*. IEEE, 45–54.

[44] FUNG Fun Man. 2016. Seeing through my lenses: a GoPro approach to teach a laboratory module. (2016).

[45] Richard E Mayer. 2014. Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. *The Cambridge handbook of multimedia learning* 16 (2014), 345–370.

[46] Richard E Mayer, Logan Fiorella, and Andrew Stull. 2020. Five ways to increase the effectiveness of instructional video. *Educational Technology Research and Development* (2020), 1–16.

[47] Megha Nawhal, Jacqueline B Lang, Greg Mori, and Parmit K Chilana. 2019. VideoWhiz: Non-Linear Interactive Overviews for Recipe Videos.. In *Graphics Interface*. 15–1.

[48] Global Cycling Network. 2015. How To Change A Bike Tyre. Video. Retrieved from https://youtu.be/sGdu4fkrQ9M?t=132.

[49] Ohan Oda, Carmine Elvezio, Mengu Sukan, Steven Feiner, and Barbara Tversky. 2015. Virtual replicas for remote assistance in virtual and augmented reality. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 405–415.

[50] The Weavers of Eternity Paracord. 2019. 7 ESSENTIAL Knots EVERYONE Should Know! Video. Retrieved from https://youtu.be/MeEcMWGzdLc?t=321.

[51] Jamie Oliver. 2013. How to make mayonnaise with Jamie Oliver. Video. Retrieved from https://youtu.be/ypQuZX5MVsI?t=56.

[52] InnerBark Outdoors. 2016. 7 Essential Knots You Need To Know. Video. Retrieved from https://youtu.be/3X8drKsdf5E?t=25.

[53] Kevin Pfeil, Pamela Wisniewski, and Joseph J LaViola Jr. 2019. An Analysis of User Perception Regarding Body-Worn 360° Camera Placements and Heights for Telepresence. In *ACM Symposium on Applied Perception 2019*. 1–10.

[54] ProfessorShoelace. 2015. Ian's Secure Shoelace Knot tutorial - Professor Shoelace. Video. Retrieved from https://www.youtube.com/watch?v=1RbaIo4VdbA&t=110s.

[55] Gordon Ramsay. 2019. Gordon Ramsay Demonstrates Basic Cooking Skills | Ultimate Cookery Course. Video. Retrieved from https://youtu.be/FTociictyyE?t=147.

[56] REI. 2019. Best Knots for the Outdoors || REI. Video. Retrieved from https://www.youtube.com/watch?v=erHTpIUFA_g&t=185s.

[57] Nature Reliance. 2014. Top Five Useful Knots for camping, survival, hiking, and more. Video. Retrieved from https://youtu.be/ABIRlz-qxSI?t=528.

[58] Michael A Rupp, James Kozachuk, Jessica R Michaelis, Katy L Odette, Janan A Smither, and Daniel S McConnell. 2016. The effects of immersiveness and future VR expectations on subjec-tive-experiences during an educational 360 video. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 2108–2112.

[59] Michael A Rupp, Katy L Odette, James Kozachuk, Jessica R Michaelis, Janan A Smither, and Daniel S McConnell. 2019. Investigating learning outcomes and subjective experiences in 360-degree videos. *Computers & Education* 128 (2019), 256–268.

[60] Daniel Sallitt. 1980. Point of View and Intrarealism in Hitchcock. *WIDE ANGLE-A QUARTERLY JOURNAL OF FILM HISTORY THEORY CRITICISM & PRACTICE* 4, 1 (1980), 38–43.

[61] Tim J Smith. 2012. The attentional theory of cinematic continuity. *Projections* 6, 1 (2012), 1–27.

[62] Claude M Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology* 69, 5 (1995), 797.

[63] The Naked Surgeon. 2020. How to reduce a TMJ dislocation. Video. Retrieved from https://www.youtube.com/watch?v=BwZqEORKX3M.

[64] Andrew DR Surtees and Ian A Apperly. 2012. Egocentrism and automatic perspective taking in children and adults. *Child development* 83, 2 (2012), 452–460.

[65] Chasing the Look. 2015. How to Fold Clothes to Save Space & Prevent Wrinkles. Video. Retrieved from https://youtu.be/PlBv0BsUFfU?t=60.

[66] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating remote instruction of physical tasks using bi-directional mixed-reality telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 161–174.

[67] TKOR. 2018. 6 Must-Know Survival Knots. Video. Retrieved from https://youtu.be/Qf_TZgBWE0A?t=137.

[68] Sylvaine Tuncer, Barry Brown, and Oskar Lindwall. 2020. On Pause: How Online Instructional Videos are Used to Achieve Practical Tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[69] Frank Ulrich, Niels Henrik Helms, Uffe Poulsgaard Frandsen, and Anne Vollen Rafn. 2019. Learning effectiveness of 360° video: experiences from a controlled experiment in healthcare education. *Interactive Learning Environments* (2019), 1–14.

[70] Stefan Vogt, Paul Taylor, and Brian Hopkins. 2003. Visuomotor priming by pictures of hand postures: perspective matters. *Neuropsychologia* 41, 8 (2003), 941–951.

[71] Rui Watanabe and Takahiro Higuchi. 2016. Behavioral advantages of the first-person perspective model for imitation. *Frontiers in psychology* 7 (2016), 701.

[72] Matt Whitlock, George Fitzmaurice, Tovi Grossman, and Justin Matejka. 2019. AuthAR: Concurrent Authoring of Tutorials for AR Assembly Guidance. (2019).

[73] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, gatekeeper, drug dealer: How content creators craft algorithmic personas. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[74] S Yoganathan, DA Finch, E Parkin, and J Pollard. 2018. 360 virtual reality video for the acquisition of knot tying skills: a randomised controlled trial. *International Journal of Surgery* 54 (2018), 24–27.