Characterizing Time Spent in Video Object Tracking Annotation Tasks: A Study of Task Complexity in Vehicle Tracking

Amy Rechkemmer [*1], **Alex C. Williams** [2], **Matthew Lease** [2,3], **Li Erran Li** [2]

[1] Purdue University, [2] AWS AI, Amazon, [3] The University of Texas at Austin
arechke@purdue.edu, {acwio, matlease, lilimam}@amazon.com

Abstract

Video object tracking annotation tasks are a form of complex data labeling that is inherently tedious and time-consuming. Prior studies of these tasks focus primarily on quality of the provided data, leaving much to be learned about how the data was generated and the factors that influenced how it was generated. In this paper, we take steps toward this goal by examining how human annotators spend their time in the context of a video object tracking annotation task. We situate our study in the context of a standard vehicle tracking task with bounding box annotation. Within this setting, we study the role of task complexity by controlling two dimensions of task design – *label constraint* and *label granularity* – in conjunction with *worker experience*. Using telemetry and survey data collected from 40 full-time data annotators at a large technology corporation, we find that each dimension of task complexity uniquely affects how annotators spend their time not only during the task, but also before it begins. Furthermore, we find significant misalignment in how time-use was observed and how time-use was self-reported. We conclude by discussing the implications of our findings in the context of video object tracking and the need to better understand how productivity can be defined in data annotation.

Introduction

As machine learning continues to advance, we see its rise in both the number of domains it is impacting and the overall complexity of the challenges it solves (Jean et al. 2016; Kube, Das, and Fowler 2019). As such, the training data that is fundamental to these advancements must similarly be able to keep up in both scope and complexity, requiring complex and domain-specific annotations to further this growth. Though much of the literature on facilitating complex data annotation has been focused on independently contracted workers on platforms such as Amazon Mechanical Turk and Upwork (Bernstein et al. 2010; Retelny et al. 2014; Doroudi et al. 2016; Dow et al. 2012), we are seeing an increase in commercial data labeling services in which full-time annotators are trained and recruited either to work on internal facing annotation or contracted out to handle other organizations' data annotation needs (Joshi 2019). One prominent example of this is within the context of video object tracking annotation, which involves annotating objects to be tracked across video footage. The implications of this type of complex annotation span from augmented reality to autonomous vehicle training, attracting customers such as Meta, General Motors, and the U.S. government to enlist these annotation services (Barnett 2023). However, such complex annotation not only requires very large amounts of labeled data, but is also often time-consuming and tedious due to precision requirements. Therefore, facilitating this process through a greater understanding of annotator productivity has significant value to a variety of stakeholders.

In understanding how to define and optimize data annotator productivity, we can draw parallels to the existing literature on the productivity of "knowledge workers", workers whose output is comprised of the creation and transformation of knowledge. Unlike earlier forms of labor, whose productivity was easier to measure in terms of quantifiable outputs (Das and Shikdar 1999; Shikdar and Das 2003), understanding what it means to be productive as a knowledge worker is not always straightforward. Given this challenge, work in this space typically relies on how time is spent and self-reported perceptions of time and productivity as measured outcomes, focusing on developing tools and interventions that assist in tracking time and observing workers' perceptions of their own productivity (Kim et al. 2016, 2017; Hiniker et al. 2016; Whittaker et al. 2016; Williams et al. 2018). Another avenue that has been explored involves asking workers to regularly reflect on how they are spending their time through daily diary entries or a regular check-in (Kim et al. 2019; Guillou et al. 2020; Meyer et al. 2019).

Though data annotators and traditional knowledge workers both work in complex spaces and primarily utilize mental resources, data annotators are largely disregarded in the productivity literature. Vondrick, Patterson, and Ramanan conducted user studies on how workers perform video object tracking tasks, but this work a) studies microtask crowd workers rather than data annotators, b) does not include a fine-grained study of how time is being spent during annotation, and c) does not include the impact of modern assistive annotation tooling. More recently, there has been effort to better understand data annotators' feelings towards their work (Wang, Prabhat, and Sambasivan 2022), contributing towards better understanding data annotator well-being and productivity in a more holistic way, but lacking an under-

[*]Work completed during an internship at Amazon.

standing of the time spent in complex annotation tasks.

In this paper, we present findings from an exploratory study of 40 full-time data annotators engaged in a video object tracking annotation task that focuses on vehicle tracking under different levels of label constraint and label granularity. We first find that annotators' interface activities, specifically video sequence navigation and zooming, are influenced by both labeling constraints and labeling granularity. Second, we find that experienced annotators spend a larger amount of time editing and deleting their annotations, indicating that annotators may become more meticulous in their labeling as they accumulate more experience. Experience aside, we find that annotators, on average, identify zooming activity and video playback as the two most time-consuming activities related to the study's task. Finally, we observe that 68% of our annotators spent a portion of their task time referencing the task's annotation guidelines not only before the task began, but also continuously throughout the task. We conclude by considering the limitations of telemetry and self-reported time estimates as methods of accurately tracking time, rethinking how we can develop measures that capture time more holistically without being invasive. We further discuss how our work sheds light on the next steps of defining what productivity looks like in data annotation as well as limitations of the work.

Related Work

Our research aims to characterize how time is spent in the context of video object tracking annotation tasks. In this section, we review relevant literature that spans video object tracking, the emerging work practice of data annotation, and the use of time in information work.

Video Object Tracking

Video object tracking is a sub-field of computer vision that focuses on the localization of relevant objects in video sequences. Object tracking for 2D video sequences is typically facilitated with bounding box annotations that contain objects of interest or with semantic segmentation that tightly fits to the object's non-linear shape. Video object tracking is widely applicable to a myriad of 2D video domains, including surveillance, sports analytics, and video editing (Wu et al. 2022). Similarly, the widespread deployment of advanced sensing systems has expanded the relevance of object tracking to 3D domains, such as autonomous driving and augmented reality, in which tracking is performed on temporal three-dimensional data (e.g., LiDAR) (Yao et al. 2020). A recent report estimates that the video analytics market will grow to more than $22 billion by 2029, indicating that the area of video object tracking will only become more prominent as time progresses (Fortune Business Insights 2021).

Modern video object tracking is a highly interdisciplinary area of research. The foundation of video object tracking is based on a wealth of methods and techniques that automate the process of object localization and vary in their approach (Yilmaz, Javed, and Shah 2006). As many of these vision-based techniques fail to generalize to all possible use-cases, human annotation remains the gold standard for ensuring accurate and precise localization of objects (Anjum, Lin, and Gurari 2021; Vondrick, Patterson, and Ramanan 2013; Yuen et al. 2009). This is further supported by the notion that data labeling service providers (e.g., Amazon Sagemaker Ground Truth[1]) offer video object tracking as a key annotation service for their customers. In general, the area remains an active area of interest for researchers across computer vision, machine learning, human computation, and human-computer interaction.

Researchers have explored a large number of solutions for improving the efficiency of video object tracking annotation tasks. For example, video keyframe interpolation is a technique that asks human annotators to create labels for objects of interest on a subset of frames in a given video sequence and automatically localizes objects of interest on all remaining frames using an interpolation algorithm or model (Wang et al. 2004). The technique not only relies heavily on several characteristics of the video data to perform correctly (e.g., video frame rate, predictability of object trajectories, etc.), but also introduces new challenges for researchers, such as keyframe selection (Kuznetsova et al. 2021). Alongside techniques for reducing the amount of data, prior research has introduced methods that allow annotators to apply labels to regions of 2D space in an imprecise fashion (Bai et al. 2009; Veksler, Boykov, and Mehrani 2010).

Productivity in Knowledge Work

Knowledge work can be defined as the efficient utilization of intellectual resources to accomplish tasks, solve problems, and generate valuable outcomes in a professional setting. Historically, productivity is measured by comparing an individual's output to an expected target output. However, recent studies suggest that the productivity of knowledge workers is best defined by the individual rather than uniformly applying a measure of productivity to a group of people in aggregate (Kim et al. 2019). Researchers generally agree that there is no uniform definition or conceptualization of productivity in knowledge work as each task context is unique.

Understanding Time Spent in Knowledge Work. Researchers in human-computer interaction have taken strides in understanding how time can be leveraged as a measure of productivity. A large number of commercially available software tools allow knowledge workers to understand how they spend their time across their computer applications (e.g., RescueMe, ActivityWatch, Apple's ScreenTime). Prior research has demonstrated how the time measurements collected by these applications translate to observed or perceived productivity (Williams et al. 2018). Studies also illustrate how these systems can facilitate self-reflection and support people in spending their time in a more desirable fashion (Whittaker et al. 2016). It should be noted that spending time productively is an assessment that, as with measures of productivity, can only be defined and assessed from the perspective of the individual (Guillou et al. 2020).

Understanding and Supporting Worker Productivity. Worker productivity has been, and continues to be, a centerpiece of human computation research. A breadth of studies

[1]https://aws.amazon.com/sagemaker/data-labeling/

Constraint Type	Constraint Description
Local	Is this object a motorized vehicle in a valid object class?
Local	Are all of this vehicle's brake lights currently visible in this frame?
Local	Is this vehicle in the same lane as the ego vehicle or at most one lane over?
Local	Is this vehicle moving in the same direction as the ego vehicle?
Global	Does this object have its brake lights on at least once at any point in this video?

Table 1: The labeling constraints considered in our experimental design, as well as type of constraint.

have examined how the productivity of workers on Amazon Mechanical Turk is, for example, affected by software-driven interruptions (Williams et al. 2019), task design interventions for well-being (Rzeszotarski et al. 2013), and tools to increase pay (Savage et al. 2020). A small number of studies have examined crowd workers' time use through the lens of multitasking (Lascau et al. 2019) or platform design (Lascau et al. 2022). Other studies of crowd work collect information that describes how time is spent at the macro-level (e.g., total task time) rather than the micro-level (Siu, Guzdial, and Riedl 2017). Though the labor performed in human computation and crowd work has yet to be formally categorized as "knowledge work", researchers have voiced a number of challenges that workers experience in the emerging profession of "data annotation" (Miceli, Schuessler, and Yang 2020; Wang, Prabhat, and Sambasivan 2022). However, unlike most knowledge work settings, data annotators often lack the ability to self-define their own productivity as a result of goals defined by their managing organization that dictate their individual work.

Study Design

To better understand how time is spent during the preparation and completion of video object tracking (VOT) annotation tasks, we conducted a controlled study with full-time annotators in a large-scale data labeling organization. Here, we describe the task in which our study was conducted, the research questions that motivated our study, and the methodology used to address our questions in this context.

Task Design

Our study aims to better understand how annotators spend their time throughout VOT annotation tasks. An important caveat of time-use, particularly in software applications, is that findings are heavily intertwined with the task's complexity (Kiani et al. 2019). Task complexity has been, and continues to be, a thriving area of research in human computation because it involves subjective task properties and a variety of individual factors related to the individual's experience, knowledge, and skillset (Campbell 1988). To better understand how time-use varies across a broader set of contexts, we manipulate our study's task design on the basis of two key dimensions that serve as proxies for task complexity and have been studied extensively (i.e., by alternative names) in other contexts (Yang et al. 2016):

1. *Label Granularity*: A task design dimension that describes the specificity of an object label.
2. *Label Constraint*: A task design dimension that describes the prerequisites for labeling an object.

In support of using these dimensions, we defined a binary spectrum that allows us to map each task design dimension with one of two values: *"low"* and *"high"*. Our conceptualization of label granularity mirrors definitions from prior work that described labels as being either coarse-grained (e.g., Animal, Vehicle) or fine-grained (e.g., Dog, Plane) (Chen et al. 2018). We refer to "coarse-grained labels" as having low label granularity while we refer to "fine-grained labels" as having high label granularity. Similarly, our conceptualization of label constraint draws from prior work that describes the various types of global and local constraints that are prerequisites for labeling an object (Zhang et al. 2012). We refer to a near-zero number of constraints as having "low" label constraint while having a larger number of constraints would be classified as "high".

Vehicle Tracking

We ground our study in a common object tracking setting: *Vehicle Tracking*. Vehicle tracking is a type of VOT in which human annotators are tasked with labeling vehicles in video sequences (Zheng 2015). As the efficacy of automated techniques for vehicle tracking can vary significantly based on the quality of video data, vehicle tracking has been recognized as a valued area of application and practice for human annotation (Schubert, Richter, and Wanielik 2008). Speaking to its longitudinal relevance, data annotation for vehicle tracking has only become more prominent with the rise of autonomous vehicles, whose market is projected to grow substantially (Bridgelall and Stubbing 2021).

Experimental Conditions

We designed and conducted a between-subjects 2×2 factorial design study. Each condition in the study is mapped to a specific task design that is jointly dictated a specific Label Granularity (i.e., "low" or "high") value and a specific Label Constraint value (i.e., "low" or "high") as follows:

- *Condition 1*: Low Granularity X Low Constraint
- *Condition 2*: Low Granularity X High Constraint
- *Condition 3*: High Granularity X Low Constraint
- *Condition 4*: High Granularity X High Constraint

Tasks completed with a "low" value for label granularity were asked to label and track vehicles with only one generic label (i.e., "vehicle") while tasks created with the "high"

Figure 1: Task interface.

value were asked to label and track vehicles with five possible labels of higher specificity (i.e., "Car", "Van", "Bus", "Truck", and "Motorbike"), which reflect the most common vehicular labels in publicly available datasets (Agarwal and Suryavanshi 2017). This means that annotators who only differed on level of label granularity were expected to generate all of the same annotations, with the only difference being the level of label specificity assigned to each annotation. Tasks completed with a "low" value for label constraint were only asked to label all motorized vehicles (that belonged to a valid object class, if the corresponding label granularity was high), while tasks with the "high" value needed to consider all of the constraints found in Table 1 before deciding whether to create an annotation.

Research Questions

The following main research questions motivate our study:

- **Q1:** How do characteristics of the VOT task and of the worker impact the time spent performing annotation?
- **Q2a:** How do characteristics of the VOT task and of the worker impact how workers perceive time spent performing annotation?
- **Q2b:** How do characteristics of the VOT task and of the worker impact how workers perceive time spent performing activities that assist with their annotation?

Due to the fact that our mechanism for capturing how time is spent is limited to actions performed in the annotation task environment, we can only explore perceived time spent performing assistive activities rather than a more objective measure of time spent.

Task

We facilitate our study using one of the most common types of VOT for vehicle tracking contexts: *labeling vehicles in 2D video sequences with bounding box annotations*.

Task Interface. Prior research suggests that an individual's familiarity with software can dramatically affect how they spend time using it (Kiani et al. 2019). In an effort to limit any such tooling bias, we employ a commercial annotation interface that is already being used by study participants in their everyday work to complete VOT annotation tasks. Importantly, the interface provides native support for the collection of interface telemetry data, which we detail further when describing our data collection paradigm.

As shown in Figure 1, the annotation interface provides a core set of functionality that is generally shared across modern VOT annotation interfaces. The interface's interactive functionality is primarily facilitated by four modes that dictate how user input (e.g., mouse and keyboard interaction) translates to an interface action. The interface allows only one mode to be active at a time and facilitates mode-switching with a button toolbar at the bottom of the screen. The interface's default mode is the *Annotation* mode, which maps mouse clicks on the video frame canvas to initiate the creation of bounding box annotations. As users click and hold the mouse on the canvas, the interface will draw a bounding box that resizes the box until the mouse is released. The interface uses a similar click-and-drag technique when the *Panning* mode is active, allowing the user to reposition the portion of the video frame that is visible. In contrast, the *Zoom-In* and *Zoom-Out* modes facilitate zoom functionality mapping their respective actions using only the mouse-down event.

The annotation interface provides several functionalities that can be invoked regardless of the active mode. For example, the interface facilitates video playback with a set of standard controls (i.e., Play, Pause, Next Frame, and Previous Frame) as well as a slider that facilitates rapid video scrubbing. The interface also allows the modification and deletion of bounding box annotations, at any time, using conventional direction manipulation techniques (e.g., clicking-and-dragging edges) as well as keyboard shortcuts (e.g., press-

ing "Delete" removes the selected annotation). To ease the burden of tracking objects across frames, the interface provides two toolbar buttons – *Copy-to-Next* and *Copy-to-All* – that facilitate copying annotations, in place, from the current video frame to the next frame or from the current frame to all subsequent video frames, respectively. The interface also provides a *Predict Next* toolbar button that copies the annotation and then attempts to adjust and extrapolate its positioning for correctness on the subsequent frame.

Task Dataset. Prior studies of human annotators in object tracking contexts use publicly available datasets to facilitate their research (Agarwal and Suryavanshi 2017). Recent reports suggest that these datasets may be flawed as their notion of "ground truth" may be polluted with error (Northcutt, Athalye, and Mueller 2021). Furthermore, as their re-use across experiments continues, it remains unclear whether biases arise from repeatedly seeing the same task or data. We therefore designed our study's vehicle tracking task to use 2D video sequence data from an internal proprietary dataset of 100 video sequences. We designed each task to show a single video sequence that was recorded at 30 frames-per-second and was composed of 1,000 individual video frames. Each video sequence was filmed in an Asian country and included a variety of locations that were primarily urban, such as city streets and highways. All sequences were filmed from the point-of-view of the "ego" vehicle, which is actively driving on the road at the time of data collection.

A total of 10 video sequences were selected for use in our study through random sampling. All sampled sequences are representative of the dataset's environmental diversity, including a range of different types of vehicles and a range of different driving scenarios (e.g., two-way traffic vs. one-way traffic, multi-lane vs. single lane, etc.).

Data Collection

We address our research questions using a variety of qualitative and quantitative data that was collected before, during, and after our study. We now describe the techniques that facilitated this data collection in detail.

Pre-Study Survey. We deployed an online survey at the beginning of our study to collect information about our annotators. We first inquired about participants' demographics and their prior experience with annotation as a career, prior experience with bounding box annotation, and prior experience with video object tracking tasks. After encouraging participants to briefly open the task interface in a new browser tab, we then inquired about participants' estimates of how much time they will spend on various activities during the task in the form of Likert-based agreement questions (e.g., "I will spend a significant amount of time creating annotations in the interface.").

Interface Telemetry Data. We analyze interface telemetry data collected during the completion of the study task. Each telemetry event includes an event type, an event name, and a timestamp of when the action occurred. All event names map directly to the user action that motivated the creation of the telemetry event (e.g., manually creating a bounding box by hand). We simplify our analysis of this data by examining it through the lens of aggregated count data rather than individual event logs. The types of collected telemetry events can be found in Table 2.

Event Type	Event Name	Analysis Type
Label	Create - Manual **(m.create)**	Count / Time
	Create - Copy **(c.create)**	Count
	Create - Predict **(p.create)**	Count
	Edit **(edit)**	Count / Time
	Delete **(delete)**	Count
Navigate	Play Video **(play.v)**[2]	Time
	Previous Frame **(f.back)**	Count
Zoom	Zoom In **(zoom.i)**	Count
	Zoom Pan **(zoom.p)**	Count
Window	Suspend Window **(w.halt)**	Count
	Resize Window **(w.size)**	Count

Table 2: The types of telemetry we considered for our analysis, broken up into larger event types, and also including whether the analysis of that event consisted of event counts or durations of time spent performing that event.

Post-Study Survey. After task submission, we administered a post-study survey that inquired about participants' perceptions of how time was spent on activities both before the task (e.g., reviewing instructions) and during the task (e.g., creating annotations, playing the video, etc.). The survey also inquired about the use of resources that were used to assist them with the task (e.g. information from coworkers).

Procedure

Upon recruitment, each study participant was assigned to one of four possible task design conditions alongside one of ten possible videos sampled from the dataset. Our assignment mechanism was performed such that no two participants shared the same combination of condition and video assignments. Following assignment, study participants were provided with a PDF file that described instructions for completing the task and several participant-specific URLs to each of the surveys. Participants were then instructed to begin the pre-study survey and and start recording their screen. After completing the pre-study survey, study participants were asked to set a one-hour timer to mark their allotted task time and were subsequently permitted to work on the task. Once the one-hour timer was finished, participants were asked to submit the task that they had been assigned regardless of how much progress they had made. Afterwards, participants were asked to complete the post-study survey and upload the recording of their study activity to a shared cloud-based storage directory.

[2]This metric was calculated based on the Load Frame telemetry event. The Load Frame event appeared both in between traditional frame navigation forward or backward and also when the video was being played through. Time spent playing through the video was calculated by recording the time between successive Load Frame events with no other event types in between.

Dep.	Ind.	β	*Std.Error*	z	*Adj. p*	Dep.	Ind.	β	*Std.Error*	z	*Adj. p*
m.create	*exp.*	-0.360	0.246	-1.467	0.277	**f.back**	*exp.*	0.272	0.243	1.119	0.380
	gran.	0.319	0.232	1.373	0.310		*gran.*	-0.298	0.232	-1.287	0.336
	con.	-1.338	0.233	-5.741	$< 0.001^{***}$		*con.*	0.619	0.232	2.672	0.043^{*}
c.create	*exp.*	0.835	0.501	1.668	0.206	**zoom.i**	*exp.*	0.188	0.236	0.796	0.536
	gran.	0.838	0.478	1.754	0.182		*gran.*	-0.710	0.225	-3.153	0.013^{*}
	con.	0.320	0.478	0.670	0.594		*con.*	-0.476	0.225	-2.115	0.103
p.create	*exp.*	0.214	0.660	0.324	0.797	**zoom.p**	*exp.*	-0.492	0.398	-1.236	0.338
	gran.	-0.677	0.630	-1.075	0.393		*gran.*	1.321	0.356	3.708	0.002^{**}
	con.	0.527	0.630	0.837	0.524		*con.*	-0.427	0.346	-1.235	0.338
edit	*exp.*	0.368	0.195	1.891	0.164	**w.halt**	*exp.*	0.538	0.245	2.192	$0.096^{\dagger}$
	gran.	0.133	0.186	0.716	0.578		*gran.*	-0.013	0.237	-0.055	0.957
	con.	-0.284	0.186	-1.529	0.259		*con.*	0.056	0.237	0.238	0.834
delete	*exp.*	2.136	0.435	4.910	$< 0.001^{***}$	**w.size**	*exp.*	-0.668	0.377	-1.772	0.182
	gran.	-0.916	0.421	-2.178	$0.096^{\dagger}$		*gran.*	-0.757	0.337	-2.247	$0.096^{\dagger}$
	con.	0.534	0.394	1.356	0.310		*con.*	0.901	0.344	2.619	0.043^{*}

Table 3: Regression models for the impact of worker experience (*exp.*), task granularity (*gran.*), and task constraint (*con.*) as predictors of counts of the telemetry events referenced in Table 2. Adjusted p-values have been corrected for false discovery rate. *, **, and *** represent the statistical significance levels of 0.1, 0.05, 0.01, and 0.001, respectively.

Dep.	Ind.	β	*Std.Error*	z	*Adj. p*
m.create	*exp.*	-0.116	0.371	-0.313	0.797
	gran.	0.339	0.352	0.964	0.460
	con.	-1.558	0.352	-4.425	0.001^{**}
edit	*exp.*	< 0.001	< 0.001	-2.944	0.0377^{*}
	gran.	< 0.001	< 0.001	-0.488	0.721
	con.	< 0.001	< 0.001	1.895	0.173
play.v	*exp.*	-0.001	0.001	-1.139	0.380
	gran.	< -0.001	0.001	-0.346	0.797
	con.	-0.004	0.002	-2.444	$0.085^{\dagger}$

Table 4: Regression models for the impact of worker experience (*exp.*), task granularity (*gran.*), and task constraint (*con.*) as predictors of the duration of time spent on the telemetry events referenced in Table 2. Adjusted p-values have been corrected for false discovery rate. *, **, and *** represent the statistical significance levels of 0.1, 0.05, 0.01, and 0.001, respectively.

Recruitment Methodology and Participants

Data annotation is becoming increasingly more prominent as an emerging information work profession. We therefore modeled our recruitment methodology on the basis of recruiting participants that are *already* engaged in full-time data annotation work. Recruitment was facilitated using an internal mailing list of full-time employees that are engaged in data annotation work.

Using this recruitment methodology, we recruited a total of 40 full-time annotators (20M/20F) based in India. All study participants reported having prior experience with the study's annotation interface, ranging from less than 2 months to more than 2 years. Our participants' age ranges were near-equally split between 25-34 (53%) and 18-24 (47%). 23 participants (58%) reported having less than 2 months of experience with VOT tasks. Although condition was assigned to participants randomly, we found that the proportion of low experience to high experience was generally balanced across all four conditions, with each condition having either 3 high/7 low experience participants or 4 high/6 low experience participants. The equipment used by participants was standardized, and all 40 participants indicated that they used a single screen during this study.

Results

In this section, we present findings for each of our study's research questions. We start off by analyzing how task complexity impacts time spent performing annotation as measured by our telemetry (**RQ1**). We then look to our pre-study and post-study survey responses in order to examine how task complexity impacts how workers perceive time spent while both performing annotation (**RQ2a**) and performing activities that assist with annotation (**RQ2b**).

The dimensions of task complexity that we consider our independent variables for this analysis are label granularity, label constraint, and worker experience. Much like label granularity and label constraint, we defined "high" and "low" levels for the worker experience variable and assigned a value to a participant based on their own self-reported experience with VOT tasks[3]. Although the majority of our participants fell into the least experienced category as presented in our question (i.e., "Less than 2 months"), we chose not to make the split there, even though it would have more appropriately divided the participants into equal groups, because we felt that 2 months of experience was too low. We instead define the split between "low" and "high" worker experience at 6 months, such that a participant with less than that is considered low experience and a participant with more than that

[3]Our participants' answers to the three experience questions in the pre-study survey were found to be highly correlated, so we chose experience with VOT tasks to be our experience indicator as it was the most specific.

is considered high experience. With this split, we ended up with 14 participants with "high experience" and 26 participants with "low experience."

RQ1: Understanding Time Spent

To understand how time was spent by our participants, we analyzed the telemetry events detailed in Table 2 either as event counts or durations of time spent by our participants, depending on the *Analysis Type* of the event. In order to investigate the impacts of label constraint, label granularity, and worker experience on the telemetry events recorded, we ran generalized linear models (GLMs). During early investigations into creating generalized linear mixed models (GLMMs) with video ID as the random effect, we found no indication of it playing a significant role, so we opted for more simplified GLM models instead. Model distributions were selected based on best fit per AIC score and residuals.

For predicting counts, we used a negative binomial distribution, although specifically for the counts of *Zoom In*, *Zoom Pan*, and *Delete* events, a zero-inflated negative binomial distribution was used due to the number of participants who did not perform these events. For predicting time durations spent on the *Edit* and *Play Video* events, we ran GLMs with a gamma distribution, and for the time durations spent on *Create - Manual*, a Tweedie distribution was used instead. For all model outputs across both count and time analysis, p-values were corrected for false discovery rate (FDR) using the Benjamini and Hochberg method (Benjamini and Hochberg 1995). The output of the models analyzing event counts and time durations can be found in Table 3 and Table 4, respectively.

For our subsections on the impacts of label constraint, label granularity, and worker experience, we start by looking at the results of the GLM models presented in the tables. For the findings with levels of significance, we present the mean (*mean*) and standard error of the mean (*sem*) values of the counts and time durations to provide context. Finally, we present the findings of differences in generated state diagrams based on transitions between telemetry event types as exhibited by participants.

General Annotator Behavior. Across all 40 of our participants, the actions performed most were Previous Frame (*mean count*: 694.75), Zoom In (*mean count*: 464.65) and Edit (*mean count*: 379.83, *mean time*: 726.66 seconds). Each participant, on average, also manually created 18.83 annotations, deleted 58.63 annotations, and spent 196.67 seconds (3.28 minutes) playing through the video.

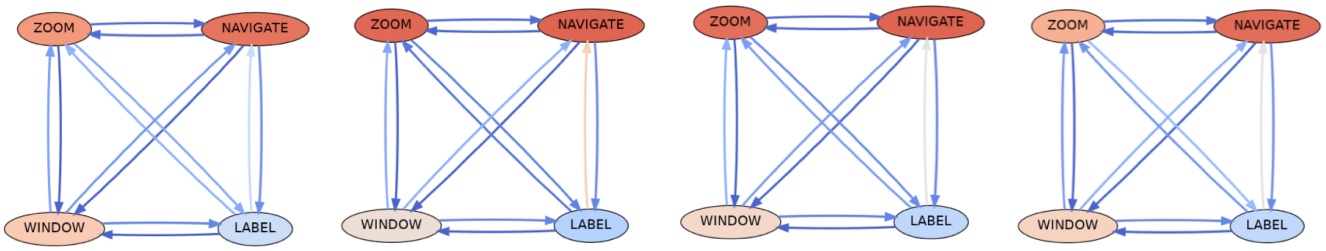
The Impact of Label Constraint. Starting off with the impact of label constraint on how our participants spent their time, we find that those assigned to a low constraint condition not only created significantly more Create - Manual annotations (*mean*: 32.45, *sem*: 10.30) than those assigned to a high constraint condition (*mean*: 5.20, *sem*: 1.10), but they also spent significantly more time in total creating manual annotations (**LC** *mean*: 156.52 seconds, *sem*: 71.55 seconds; **HC** *mean*: 15.16 seconds, *sem*: 2.68 seconds). However, we see no significant difference in terms of number of or time spent on Create - Copy annotations, Create - Predict annotations, edits, or deletions.

In terms of navigation while annotating, we see evidence that participants assigned to a high constraint condition navigated backwards through frames (*mean*: 961.00, *sem*: 312.07) more frequently than those assigned to a low constraint condition (*mean*: 428.50, *sem*: 140.15). Our results also suggest that participants with a high constraint condition may have spent more time playing through the video (**LC** *mean*: 102.58 seconds, *sem*: 28.17 seconds; **HC** *mean*: 290.77 seconds, *sem*: 57.99 seconds). We did not find any difference in terms of number of zoom events performed, but initial analysis suggests that those with a high constraint condition resizing the window of their annotation more frequently (*mean*: 1.80, *sem*: 0.93) than those with a low constraint condition (*mean*: 0.45, *sem*: 0.11). Looking closer into this finding, however, we see that it was largely driven by a single worker assigned to the high constraint condition.

Lastly, we look to the state diagrams of workers in low and high constraint conditions shown in Figure 2a and Figure 2b, respectively. The most notable difference in patterns of behavior between the two groups is that participants assigned to a high constraint condition had a much higher probability of transitioning from a Label event directly to a Navigate event (*transition probability*: 0.63) as compared to those assigned to a low constraint condition (*transition probability*: 0.39). This was also the most likely transition from a Label event for those with a high constraint task. Instead, participants assigned to a low constraint task were most likely to transition from a Label event to another Label event (**LC** *transition probability*: 0.40; **HC** *transition probability*: 0.33), but they were also more likely to transition from a Label event to a Zoom event (*transition probability*: 0.21) than those assigned to a high constraint task (*transition probability*: 0.05).

Altogether, our findings suggest that annotators given a VOT task with a high level of constraints spend more time navigating through the task (including moving backwards to see previous frames and playing through the video) and more quickly jump back into navigation events after performing labeling events. Annotators given a VOT task with a low level of constraints, however, spend more time manually creating annotations and may need to perform more actions after labeling before they can move on to the next frame. This aligns with our intuition that a low constraint task will require more annotations to be made per frame, including more manual annotations on the first frame before assistive tools can be used for subsequent frames and more labeling activities that may need to be done before one can move to the next frame. In terms of navigation, the presence of global constraints in high constraint tasks seems to also require annotators to skip around through the video in order to determine whether objects meet constraints, leading to greater play time and backwards navigation.

The Impact of Label Granularity. Looking now into the impact of label granularity, we find that those assigned to a low granularity condition zoomed into the annotation canvas significantly more (*mean*: 659.30, *sem*: 144.21) than



(a) Low Constraint (b) High Constraint (c) Low Granularity (d) High Granularity

Figure 2: State diagrams showing participants' likelihood of transitioning from one type of event to another in both low constraint and high constraint conditions. Telemetry events in each corresponding event type are seen in Table 2. The diagrams are heat maps reflecting transition probabilities, and the coloring of event type nodes reflect the probabilities of self-loops.

those assigned to a high granularity condition (*mean*: 270.0, *sem*: 62.44). On the other hand, our participants assigned to a high granularity condition performed significantly more zoom pan events (*mean*: 94.15, *sem*: 46.73) than those assigned to a low granularity condition (*mean*: 16.35, *sem*: 6.35). Looking closer into this finding on zoom pan events, we see that the majority of workers from both granularity conditions did not perform any zoom pan events, but the 7 high granularity participants who did performed significantly more zoom pans (*mean*: 269.00) than the 8 low granularity participants who did (*mean*: 40.88).

For label events, we find some evidence that our participants assigned to a high granularity condition may perform more deletions (*mean*: 90.35, *sem*: 54.19) than those with a low granularity condition (*mean*: 26.90, *sem*: 11.91). Looking closer into this finding, we see that only half of the low granularity participants performed any deletions at all while 17 out of the 20 high granularity participants performed at least one deletion event. As a brief look at window events, we see some evidence that participants assigned a low granularity condition may resize the window more frequently (**LG** *mean*: 1.70, *sem*: 0.93; **HG** *mean*: 0.55, *sem*: 0.17). However, this finding appears to be driven a single participant rather than a trend between the two groups.

Finally, we consider the state diagrams of workers in low and high granularity conditions shown in Figure 2c and Figure 2d, respectively. Most notably, we see that participants assigned to a low granularity condition were more likely to perform successive zoom events (*transition probability*: 0.88) as compared to those assigned to a high granularity condition (*transition probability*: 0.74). Although both groups were most likely to transition to a zoom event from a previous zoom event, the group with high granularity showed a greater probability of transitioning from a zoom event to a label event (*transition probability*: 0.23) than the group with low granularity (*transition probability*: 0.10).

Altogether, we find that differences in label granularity largely impact the zoom events that annotators perform. Annotators with a low label granularity task appear to perform more zoom in events and be more likely to perform successive zoom events. On the other hand, annotators with a high label granularity task appear to perform more zoom pan events and be more likely to transition to the next label after a zoom event. This may correspond to annotators with high granularity needing to be zoomed into a greater number of annotations in order to determine which label they fall into, causing them to zoom pan between different annotations in order to keep that level of magnification. Annotators with low granularity, however, may only need to zoom in to specific objects that they have trouble determining boundaries for, potentially increasing the number of zoom in and out events rather than panning around the canvas. Lastly, annotators with high granularity may also be more likely to perform label deletions, suggesting that many of these deletions may be related to an incorrect label for the annotation chosen rather than the bounding box of the annotation itself.

The Impact of Worker Experience. Comparing our participants with different levels of VOT annotation experience, we find that although there is no significant difference in terms of number of edits performed, those with a higher level of experience spent significantly more time editing their annotations (*mean*: 905.30 seconds, *sem*: 177.35 seconds) than those with lower experience (*mean*: 548.03 seconds, *sem*: 51.89 seconds). We also see that participants with higher levels of experience performed significantly more delete events (*mean*: 131.43, *sem*: 76.42) than those with lower experience (*mean*: 19.42, *sem*: 7.13). Looking closer into this finding, we see that lower experience participants were more likely to perform at least one deletion (77%) compared to higher experience participants (50%), but higher experience participants who performed deletion events did significantly more (*mean*: 262.86) than participants with lower experience (*mean*: 25.25).

Looking into window events, we also see that participants with higher levels of experience may have suspended the annotation window more often (*mean*: 12.93, *sem*: 1.76) than those with lower levels of experience (*mean*: 6.12, *sem*: 1.55). We also see that 100% of participants with high experience suspended the annotation window at least once, while the same was true for 77% of low experience participants. State diagrams comparing transition probabilities between event categories for low and high experience participants

were also generated, but there were no notable differences found between the groups.

Altogether, we see evidence that high experience annotators may be more meticulous when it comes to their work than low experience annotators. We see this in the finding that our participants with high experience spent more time editing their annotations, and we partially see this in the interesting split down the middle of our high experience annotators where one half of them performed an incredibly high number of annotation deletions. For the other half of our high experience participants who performed no deletions, this may potentially be an indication of higher confidence with the task. Due to the limitations of our telemetry, we do not know how our high experience participants who were more frequently suspending the annotation spent this time, but a hypothesis is that they were more frequently checking task instructions than our participants with low experience.

RQ2: Understanding Perceived Time Spent

To understand how time was perceived to be spent by our participants, we analyze the following pre-study and post-study survey questions. In the pre-study, participants were asked to estimate how much time they would spend on various annotation activities and assistive activities before and during the VOT task with the following prompt: "I will spend a significant amount of time _____ in the interface." These statements were presented in the form of an agreement Likert question on a scale from 1 to 5, where the blank was filled in with an annotation activity (e.g., editing annotations, playing the video, etc.).

In the post-study, participants were asked a Likert-style question analogous to the one in the pre-study, except the wording of the statements was changed from "I will spend" to "I spent." In addition, participants were asked additional Likert-style questions on a scale from 1 to 5 related to how frequently they engaged in assistive activities during the task (e.g., looking back at annotation instructions, discussing the task with others, etc.), questions related to the types of assistive activities they performed and resources they referenced before and after the task, and how long they estimated to have spent performing assistive activities before starting the task. We also asked participants to rank annotation and assistive activities based on their perception of time spent.

We start this section first by discussing general behavior seen across all participants in terms of how they ranked their time spent, as well as the assistive activities and resources they engaged with. For analyzing the impact of task complexity on perceived time spent on annotation (**RQ2a**) and assistive activities (**RQ2b**), we compared the Likert responses between groups for both the pre-study and post-study questions by performing Mann-Whitney U-tests (Mann and Whitney 1947), a non-parametric test for examining differences between two unpaired groups. Mann-Whitney U-tests were also used to compare estimated time spent performing assistive activities across groups.

General Behavior. When asked to rank their annotation related activities, with 1 being the most time-consuming and 10 being the least, our 40 participants indicated the following ordering in their post-study survey: **1)** zooming/panning (*mean rank*: 3.40), **1)** playing the video (*mean rank*: 3.40), **3)** performing activities that improved their understanding of the task *before* working on it (*mean rank*: 3.78), **4)** creating annotations (*mean rank*: 3.90), **5)** editing annotations (*mean rank*: 4.50), **6)** performing activities that improved their understanding of the task *while* working on it (*mean rank*: 5.23), **7)** reviewing annotations (*mean rank*: 5.25), and **8)** deleting annotations (*mean rank*: 6.55). It's important to note that we don't have any sort of objective time or count estimates for a few of these activities (i.e., performing activities to improve understanding and reviewing annotations), meaning that collecting perceptions is currently the only way to measure them. It is also interesting to note that even though we found that our average participant spent more than 3 times the amount of time editing annotations as they did playing through the video, editing was on average ranked 5th in terms of perceived time consumption as opposed to playing through the video, which tied for 1st.

When asked whether they had spent time performing any activities before starting with the annotation task, 36 of our participants (90%) indicated that they had. The most common activity performed was reading through the guidelines they had been provided (indicated by 27 of the 36 who performed these activities), but 2 participants also spent time playing through the VOT task video, and another 3 chatted with a colleague or manager before starting on the task. The most common amount of time spent by our participants on these activities was between 5 and 10 minutes (indicated by 12 out of the 36). When asked about resources referenced while working on the task, 26 of the 40 participants (65%) mentioned referencing the annotation guidelines, 4 participants referenced physical notes or printouts, and 3 participants mentioned having conversations with others about the annotation task while working on it. 13 of the 40 participants (32%) used no resources during annotation.

The Impact of Task Complexity on Annotation and Assistive Activities. We compare the responses of how our participants perceived to spend their time across the levels of label constraint, label granularity, and worker experience separately. We also analyzed the responses from each of the questions mentioned individually. Altogether, we found no significant differences between the levels for any of the dimensions of task complexity. In sum, we found no impact of task complexity on either how our participants expected to spend their time on annotation activities prior to engaging in the task or how they perceived to have spent their time on annotation and assistive activities after engaging in the task.

Conclusions and Discussions

Our research takes strides in understanding not only how annotators spend their time in the completion of VOT tasks, but also how they estimate and perceive their spent time. First, we find that annotators' interface activities, specifically video sequence navigation and zooming, are influenced both by label granularity and by label constraint. Second, we find that experienced annotators spend a larger amount of time editing and deleting their annotations, indicating anno-

tators may become more meticulous in their labeling as they accumulate more experience. Experience aside, we find that annotators, on average, underestimate the amount of time they will spend editing annotations threefold. Finally, we observe that 68% of annotators spent a portion of their task time referencing the task's annotation guidelines not only before the task began, but also continuously throughout the task. We now discuss the implications of our findings and conclude with limitations of our study.

Rethinking How We Capture Time Spent

Our study demonstrates how annotators can spend time performing task-related activities beyond the boundaries of an annotation interface. Our work therefore reveals clear limitations of data collection systems, such as telemetry instrumentation, that assume activities will be observable. In reality, there exist a number of activities that are inherently invisible, such as computer activities beyond the browser tab and non-digital activities beyond the computer. Without appropriate context, differentiating between idle time and "invisible" activities is intractable.

Although our survey questions on perceptions of time provided some valuable information, such as the types of resources that participants engage with during annotation, we found no indication that perceptions of time can be used as a proxy for actual time spent. In fact, we found evidence of large discrepancies between actual time and perceived time spent, most notably in time spent editing being underestimated across participants by a factor of 3. A number of reasons for this exist. For one, our method of capturing perceptions may have been flawed (see **Limitations and Future Work**), but properly capturing estimations of time may be challenging in general. Estimating exact amounts of time can be difficult when the actions are so interleaved, and less rigid language prompting time estimation can be open to interpretation. For instance, a "significant amount of time" spent on an activity could be considered either in relation to the total time spent or the time that one normally expects the action to take.

With these limitations in mind, we consider more accurate means of capturing time spent. Recording both the screen with the annotation interface and the annotator's environment would yield significant improvements, but this solution brings up clear ethical questions of when productivity tracking becomes surveillance and the impact of such supervision on annotator well-being. Limiting to screen recording helps reduce this concern, but analyzing such recordings is subjective and tedious, not allowing for feedback on the fly. This analysis, however, could help with identifying likely context behind unknown patterns of behavior captured through telemetry. Lastly, embedding materials that annotators frequently use for assistance within the interface can capture time spent that would normally be unknown.

Productivity in Professional Data Annotation

Our research sheds light on the importance of defining what it means to be productive in data annotation work, showing that how time is spent is only part of the story. We would expect annotators with more experience in VOT tasks to have learned more productive patterns of behavior over time that result in more efficient labeling, but our findings potentially suggest otherwise. However, understanding productivity goes beyond how quickly one can go from annotation creation to navigating to the next frame, requiring an understanding of the quality of annotations. We consider the possibility that the extra time editing and additional deletions done by our high experience annotators may be more productive if it leads to an increase in quality above some unknown threshold that results in stakeholder satisfaction. One example of where this may come into play is when accounting for "jitter". Therefore, looking into the impact of task and worker characteristics on annotation quality is a necessary step towards understanding productivity.

Though the estimates measured in our study do not directly correspond to productivity, our findings suggest that productive patterns of behavior likely vary depending on complexity of the task. These findings also suggest ideas for optimizing productivity even if we cannot yet define it, such as helping annotators with a high constraint task more easily be able to identify which objects should be annotated to minimize time spent navigating through the task. There is also reason to believe that worker perceptions of time spent can act as an indication of worker well-being and productivity, even if it does not correspond to objective time usage.

Limitations and Future Work

Our study has several limitations. First, our study was conducted in the context of vehicle tracking under several task designs that vary in their complexity. Our study cannot explain how time may be spent in types of annotation tasks, video object tracking annotation tasks that do not involve vehicle tracking, or in vehicle tracking settings that characterize complexity through different means. Second, our study was conducted with 40 full-time annotators based in India that work at a large technology corporation. Our study cannot draw conclusions about annotators who are located elsewhere or are employed in other capacities.

Overall, we cannot say how our findings would change if we included additional levels of constraint or granularity within our study design or considered experience as continuous rather than a binary variable. Lastly, the scope of responses available to our participants when reporting their perceptions of time spent may have been too narrow to determine potential differences between groups. Presenting these questions either in a more open-ended fashion or with a larger scale (e.g., 1 - 100) may have changed our findings.

Future work involves efforts to explore how productivity in data annotation may be defined from the combined lens of label quality and a refined conceptualization of time spent. To further assess the limitations of conventional telemetry data collection, we will conduct a comparative analysis of events through the lens of screen recordings and telemetry data collected during our study. As our understanding of productivity and means of time tracking improve, we will explore how productive behaviors can be modeled and leveraged to aid workers in spending their time well.

Acknowledgments We thank the data annotators who participated in our study for their time and attention. We also thank Jonathan Buck, Koushik Kalyanaraman, Min Bai, and Patrick Haffner for providing feedback on telemetry instrumentation prototype and analysis of telemetry data.

References

Agarwal, A.; and Suryavanshi, S. 2017. Real-time* multiple object tracking (MOT) for autonomous navigation. *Technical report*.

Anjum, S.; Lin, C.; and Gurari, D. 2021. CrowdMOT: Crowdsourcing strategies for tracking multiple objects in videos. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3): 1–25.

Bai, X.; Wang, J.; Simons, D.; and Sapiro, G. 2009. Video snapcut: robust video object cutout using localized classifiers. *ACM Transactions on Graphics (ToG)*, 28(3): 1–11.

Barnett, J. 2023. Scale AI awarded $250M AI contract by Department of Defense. Accessed: 2023 - 08 - 23.

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322.

Bridgelall, R.; and Stubbing, E. 2021. Forecasting the effects of autonomous vehicles on land use. *Technological Forecasting and Social Change*, 163: 120444.

Campbell, D. J. 1988. Task complexity: A review and analysis. *Academy of management review*, 13(1): 40–52.

Chen, Z.; Ding, R.; Chin, T.-W.; and Marculescu, D. 2018. Understanding the impact of label granularity on cnn-based image classification. In *2018 IEEE international conference on data mining workshops (ICDMW)*, 895–904. IEEE.

Das, B.; and Shikdar, A. A. 1999. Participative versus assigned production standard setting in a repetitive industrial task: a strategy for improving worker productivity. *International journal of occupational Safety and Ergonomics*, 5(3): 417–430.

Doroudi, S.; Kamar, E.; Brunskill, E.; and Horvitz, E. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2623–2634.

Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 1013–1022.

Fortune Business Insights. 2021. Market Research Report: Visual Analytics. Accessed: June 9, 2023.

Guillou, H.; Chow, K.; Fritz, T.; and McGrenere, J. 2020. Is your time well spent? reflecting on knowledge work more holistically. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–9.

Hiniker, A.; Hong, S.; Kohno, T.; and Kientz, J. A. 2016. MyTime: designing and evaluating an intervention for smartphone non-use. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 4746–4757.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794.

Joshi, S. 2019. How artificial intelligence is creating jobs in India, not just stealing them — India News - Times of India. . Technical report.

Kiani, K.; Cui, G.; Bunt, A.; McGrenere, J.; and Chilana, P. K. 2019. Beyond" One-Size-Fits-All" Understanding the Diversity in How Software Newcomers Discover and Make Use of Help Resources. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Kim, Y.-H.; Choe, E. K.; Lee, B.; and Seo, J. 2019. Understanding personal productivity: How knowledge workers define, evaluate, and reflect on their productivity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Kim, Y.-H.; Jeon, J. H.; Choe, E. K.; Lee, B.; Kim, K.; and Seo, J. 2016. TimeAware: Leveraging framing effects to enhance personal productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 272–283.

Kim, Y.-H.; Jeon, J. H.; Lee, B.; Choe, E. K.; and Seo, J. 2017. OmniTrack: a flexible self-tracking approach leveraging semi-automated tracking. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(3): 1–28.

Kube, A.; Das, S.; and Fowler, P. J. 2019. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 622–629.

Kuznetsova, A.; Talati, A.; Luo, Y.; Simmons, K.; and Ferrari, V. 2021. Efficient video annotation with visual interpolation and frame selection guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3070–3079.

Lascau, L.; Gould, S. J.; Brumby, D. P.; and Cox, A. L. 2022. Crowdworkers' Temporal Flexibility is Being Traded for the Convenience of Requesters Through 19 'Invisible Mechanisms' Employed by Crowdworking Platforms: A Comparative Analysis Study of Nine Platforms. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–8.

Lascau, L.; Gould, S. J.; Cox, A. L.; Karmannaya, E.; and Brumby, D. P. 2019. Monotasking or multitasking: Designing for crowdworkers' preferences. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–14.

Mann, H. B.; and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Meyer, A. N.; Barr, E. T.; Bird, C.; and Zimmermann, T. 2019. Today was a good day: The daily life of software

developers. *IEEE Transactions on Software Engineering*, 47(5): 863–880.

Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–25.

Northcutt, C. G.; Athalye, A.; and Mueller, J. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*.

Retelny, D.; Robaszkiewicz, S.; To, A.; Lasecki, W. S.; Patel, J.; Rahmati, N.; Doshi, T.; Valentine, M.; and Bernstein, M. S. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 75–85.

Rzeszotarski, J.; Chi, E.; Paritosh, P.; and Dai, P. 2013. Inserting micro-breaks into crowdsourcing workflows. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1, 62–63.

Savage, S.; Chiang, C. W.; Saito, S.; Toxtli, C.; and Bigham, J. 2020. Becoming the super turker: Increasing wages via a strategy from high earning workers. In *Proceedings of The Web Conference 2020*, 1241–1252.

Schubert, R.; Richter, E.; and Wanielik, G. 2008. Comparison and evaluation of advanced motion models for vehicle tracking. In *2008 11th international conference on information fusion*, 1–6. IEEE.

Shikdar, A. A.; and Das, B. 2003. The relationship between worker satisfaction and productivity in a repetitive industrial task. *Applied ergonomics*, 34(6): 603–610.

Siu, K.; Guzdial, M.; and Riedl, M. O. 2017. Evaluating singleplayer and multiplayer in human computation games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*, 1–10.

Veksler, O.; Boykov, Y.; and Mehrani, P. 2010. Superpixels and supervoxels in an energy optimization framework. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V 11*, 211–224. Springer.

Vondrick, C.; Patterson, D.; and Ramanan, D. 2013. Efficiently scaling up crowdsourced video annotation: A set of best practices for high quality, economical video labeling. *International journal of computer vision*, 101: 184–204.

Wang, D.; Prabhat, S.; and Sambasivan, N. 2022. Whose AI Dream? In search of the aspiration in data annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–16.

Wang, J.; Xu, Y.; Shum, H.-Y.; and Cohen, M. F. 2004. Video tooning. In *ACM SIGGRAPH 2004 Papers*, 574–583.

Whittaker, S.; Kalnikaite, V.; Hollis, V.; and Guydish, A. 2016. 'Don't Waste My Time' Use of Time Information Improves Focus. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1729–1738.

Williams, A. C.; Kaur, H.; Mark, G.; Thompson, A. L.; Iqbal, S. T.; and Teevan, J. 2018. Supporting workplace detachment and reattachment with conversational intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.

Williams, A. C.; Mark, G.; Milland, K.; Lank, E.; and Law, E. 2019. The perpetual work life of crowdworkers: How tooling practices increase fragmentation in crowdwork. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–28.

Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; and He, L. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*.

Yang, J.; Redi, J.; Demartini, G.; and Bozzon, A. 2016. Modeling task complexity in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, 249–258.

Yao, R.; Lin, G.; Xia, S.; Zhao, J.; and Zhou, Y. 2020. Video object segmentation and tracking: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4): 1–47.

Yilmaz, A.; Javed, O.; and Shah, M. 2006. Object Tracking: A Survey. *ACM Computing Surveys (CSUR)*, 38(4): 13–es.

Yuen, J.; Russell, B.; Liu, C.; and Torralba, A. 2009. Labelme video: Building a video database with human annotations. In *2009 IEEE 12th International Conference on Computer Vision*, 1451–1458. IEEE.

Zhang, H.; Law, E.; Miller, R.; Gajos, K.; Parkes, D.; and Horvitz, E. 2012. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 217–226.

Zheng, Y. 2015. Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3): 1–41.