Understanding the Role of Video and Data in User Interfaces for Annotation of Activities in Sensor Data

ANONYMOUS, Anonymized, Anonymized

The purpose of this paper is to understand the role of data and video in user interfaces for annotation of events for supervised learning of sensor-based activity recognizers. Over the past few years, many studies have reported on sensor-based activity recognition systems, which depend on supervised learning of event recognizers. Supervised learning requires a human annotator to mark events in sensor data to create a labeled training dataset HOLE: inconsistency between dataset and data set - we use both throughout the paper. Video is often collected along with the sensor data to help the human annotator recognize events in the sensor data stream in HOLE: maybe change 'in' to 'using' for better flow a user interface, which shows the synchronized video and data. The video is not used as an input for activity recognition. A problem with this approach is that it requires collecting both video and data. The video is only used to aid the person who annotates events in the data and collecting video can be slow, costly, difficult, or impossible. In order to better understand the role of video in event annotation interfaces, we conducted two HOLE: need to add info on the new study studies with a total of 116 participants using 3 annotation tasks involving a single camera and a single inertial measurement unit (IMU). We found that after annotating both video and data together, novice annotators working with data alone, compared to novice annotators using video alone, were more efficient, identified event boundaries more accurately, and identified event type HOLE: should 'type' be plural here to match 'boundaries'? somewhat less accurately. Our results suggest that collecting video for all of the training data may not be necessary−or even desirable.

CCS Concepts: • **Human-centered computing** → **Ubiquitous computing**; **Ubiquitous and mobile computing theory, concepts and paradigms**; *User interface toolkits*; • **Computing methodologies** → **Learning settings**; **Learning settings**;

Additional Key Words and Phrases: Interactive learning; Toolkit; Annotation; Computational devices; Machine learning; Internet of things

ACM Reference Format:
Anonymous. 2018. Understanding the Role of Video and Data in User Interfaces for Annotation of Activities in Sensor Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 0, 0, Article 0 (February 2018), 18 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Over the past few years, many studies have appeared involving supervised learning of recognizers for sensor-based activity recognition. In this context, supervised learning is a type of machine learning in which a person annotates or labels events in data. The labeled data are used as input to a machine learning algorithm, which learns to recognize the events in unlabeled data.

In activity recognition tasks, video is often collected for the sole purpose of aiding a person doing the original event annotation. This is commonly done using an interface like the one shown in Figure 1. The video is not used as input to the learning algorithm and is not used in the final recognition system. In this kind of interface, the video and data are synchronized and the data is annotated using a tool which shows the video stream together with a time series plot of the data streams. The view of the video and the data time series are often synchronized so that clicking a point in the data time series advances the video to that point and vice versa. Also, video playback advances the data stream to match video playback.

We call this process the "data plus video activity annotation" process, or "D+V activity annotation". HOLE: should we update the following info for 2019? We found that at least 21% of the articles (37 of 178) published in the 2017 volume of the ACM Journal on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)

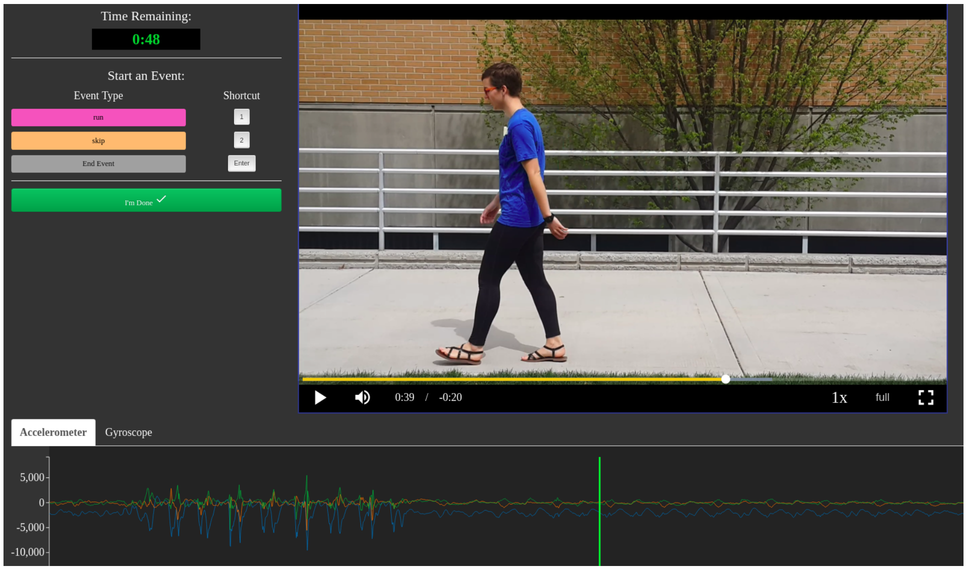


Fig. 1. Screenshot of an example D+V annotation interface. Data and video streams are synchronized and simultaneously shown to the user. The interface shown here, called VAT, could optionally turn off one of these modalities.

include some combination of annotation, sensor data and/or video in the context of activity recognition. 9 articles , or 5% of the articles published in 2017 in IMWUT, included D+V activity annotation with both video and data.

A problem with the D+V activity annotation process is that it requires video. To collect this video, the activity is performed in front of a camera while wearing or using the data collection sensor. The video can be difficult to obtain in some situations due to privacy, logistical, or storage constraints. For example, if the data involves free living in a home, it would be invasive to collect video of a person using the bathroom or having a sensitive conversation. Activities that involve moving are difficult as the camera must follow the activity or the activity must be repeated in a small area. If the activity occurs in a remote location, it can be difficult to deploy power and cameras needed to record the activity. This could be a significant problem for activities in which a person, or animal, wears the sensor for a long period of time or moves in unpredictable ways.

The purpose of this paper is to examine how annotators, who have no experience finding events in IMU data series, use video and data in activity annotation. Understanding how novice annotators use video and data may provide insight into how much video needs to be collected to assist in annotation and when video is of greatest value. In addition, this may inform the design of future interfaces for D+V activity annotation.

We study D+V annotation in the context of a constrained form of the process which serves as a kind of lower bound on the difficulty of D+V annotation. In this constrained form of the problem, the user annotates the event boundaries and the event types in data streams from a single IMU, with a single video stream, and 3 event types.

Author's address: Anonymous, Anonymized, Anonymized, Anonymized, Anonymized, 12345, Anonymized, anon@test.org.

2474-9567/2018/2-ART0 $15.00
https://doi.org/0000001.0000001

This constrained problem is a lower bound in the sense that, if novices cannot succeed at this task without video, then they are unlikely to succeed in more complex scenarios involving more sensors, cameras, or event types. Constraining the problem allows us to carry out detailed controlled studies on the interface and to establish whether or not there are any situations in which video is not needed.

HOLE: need to add info for the new study. Are we completely overriding the old one or just adding? We report the results of 2 studies involving 60 participants in the first study and 56 different participants in the second study. The results of the first study informed the design of the second study. In each study, we presented participants with interfaces involving data alone, video alone, or both data and video together. Figure 1 shows both the data and video elements of an event annotation interface which is similar to those used in other tools [10, 16, 27]. The video interface is the video frame at the top, along with the video playback controls, which are just below the video. The data interface is the sensor trace shown at the bottom, along with a bright green line to indicate the current position of the video frame in the sensor data stream.

We asked participants to mark event boundaries and event types using the buttons and the top left of the interface in Figure 1. We measured both efficiency and accuracy. In order to triangulate observed efficiency and accuracy results, we also asked open and closed ended questions about the experience and collected data on interface commands used by the participant. Results were analyzed using relevant quantitative and qualitative methods.HOLE: This last sentence seems a bit vague - do we need to say more than this?

After annotating events in an interface, which shows both data and video together, novice annotators using a data-only interface, when compared to novice users using a video-only interface, were HOLE: do our new results still support this?

(1) more efficient,
(2) more accurate when marking event boundaries, and
(3) somewhat less accurate when marking event types.

These findings are supported by measurements of efficiency and accuracy, responses to closed and open ended questions and interface usage logs.

There are three main categories of people who can benefit from this work: first, those who are building interfaces for the annotation of data and video; second, those who are spending time or money collecting data for annotation; and third, those giving the annotators instructions about the annotation process and thus indirectly, the annotators themselves. In each case, a better understanding of how annotators use video and data can inform video collection plans, annotation interface designs, and instructions for annotators. Changes to video collection, interface design and annotator instructions can lead to more efficient practices for producing the annotated data needed as input on supervised learning algorithms for activity recognition.

2 RELATED WORK

Prior work in sensor based activity recognition demonstrates a need for better understanding of how annotators use video and data in the annotation process.

We group related work into three areas: event annotation tools which have been used in D+V annotation, event recognition applications, and other approaches for generating sensor based activity recognizers. We focus only on the most closely related prior work and do not list all prior projects involving activity recognition. More complete surveys of sensor based activity recognition [4] and video based activity recognition [24] are available.

2.1 Annotation Tools

We describe three tools which have been used for D+V event annotation in the context of unsupervised learning. These tools present a similar kind of interface involving video and data to the person annotating events and each of these tools have been used in D+V event annotation for supervised training of event recognizers.

The ELAN tool was created for linguistic analysis of videos and sound recorded of people speaking [27]. In ELAN, the user sees the video in one window and a time series plot of a waveform recorded by a microphone in another window. Motion capture data streams were added later to support labeling of sign language [6] but other kinds of time-varying data streams are supported. ELAN has been used for event annotation in activity recognition papers [13, 23].

ChronoViz also includes data and video streams but was designed for assigning codes to video data in behavioral science [10]. ChronoViz allows for multiple data streams collected from different sensors. As with ELAN, ChronoViz has been used in activity recognition papers [2, 26].

Finally, ANVIL is a tool originally designed for audiovisual annotation of multi-modal dialog [16] which also shows video and data traces together in a time-synchronized interface. ANVIL has been used in activity recognition projects as well [12, 22].

Understanding how annotators use video and data in tools like these for activity recognition may inform the design of a new set of tools which are better suited to activity annotation for activity recognition.

2.2 Activity Recognition Applications

The breadth and quantity of activity recognition projects reinforce the importance of understanding the user's experience with annotation of activities in sensor data with and without video. Several applications also highlight the potential difficulty of collecting video for event annotation.

Researchers have studied event recognition for recreation such as evaluating the safety of hiking trails [15], classifying Alpine ski turns [11], and coaching for Nordic walking [7]. Collecting video in these situations can be challenging. For example, collecting video to aid in annotation of Alpine ski turns in sensor data required a second skier wearing a helmet mounted camera to follow the first skier wearing the sensor [11]. Collecting video in Nordic walking required the walker to walk back and forth in front of a fixed camera [7].

In the context of work-related tasks, researchers have carried out event recognition for brick laying [12] and food preparation [22]. Capturing video of bricklayers at work involves bringing a camera to a construction job site and recording 30 to 40 minutes of video per worker [12]. Video of food preparation was recorded in an instrumented kitchen installed in a laboratory which included several cameras for capturing this kind of video [22].

Others have studied activity recognition applications for health related applications such as tracking eating [2, 26]. Collecting video of free eating events can involve collecting video of an entire day including personal events that a participant may not want to be video recorded. Finally, other papers consider kinds of screening related to service animations [3] and detection of certain disorders [5].

2.3 Other Approaches

Other approaches to training event recognizers are being studied. These approaches aim to reduce the amount of work needed to generate a recognizer. Our work may contribute to simplifying these approaches because, in each case, annotated events are still needed.

The Legion [20] and Glance [19] tools by Lasecki et al. explore crowd sourcing as a way to quickly annotate events in data. Crowds can quickly analyze events and using multiple annotators per event can increase accuracy of the aggregated results. Our work on understanding the role of video and data may improve interfaces for crowd sourced approaches to D+V event annotation.

The Transact tool [14] and a project Lane et al. [18] explore learning transfer in the context of activity recognition. Rey et al. took a different approach to automatically generalizing a classifier to include the results of a new sensor added after training [25]. In the cases of both learning transfer and adding a sensor, our work can improve the process of labeling events in data for training the original recognizer.

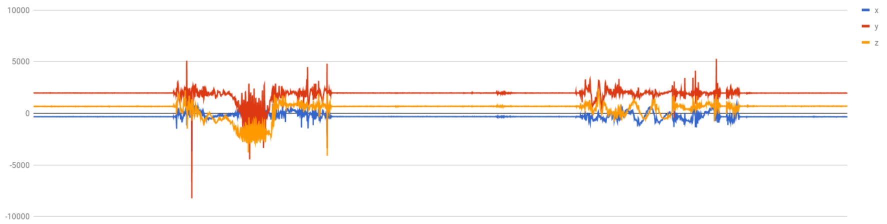


Fig. 2. Sample of data typical for the “pills” dataset; this snippet shows an event where the actor takes a pill, followed by an event where the pill bottle rolls on the table. The flat portions of the graph are when the pill bottle is stationary on a table. The bottle rolling on the table is an example of data in which the bottle moved but a pill was not taken. Annotators need to discriminate between these kinds of motion.

3 DATASETS AND INTERFACES COMMON TO BOTH STUDIES

We conducted two studies HOLE: Are we adding info from the new study or is it replacing the second one completely? if just adding, change to three studies? to explore the effects on efficiency and accuracy for both the data and video interface modalities. Here we give a brief description of datasets and the interface used in both studies. We will then describe each study separately.

3.1 Datasets

Between the two studies HOLE: again, does this need to be changed to three? we ran, we used a total of three different datasets HOLE: change to 5 datasets to include DW and Gait, each of which contained multiple data instances comprising of the raw sensor data synchronized with its corresponding video. All of the sensor data was collected using an IMU that collected six dimensions of accelerometer and gyroscope data.HOLE: <- is this still true for gait and drywall? Need to get info on gait A brief description of datasets and their potential use case are given below.HOLE: need to add potential use cases for gait and dw

The Pills dataset is motivated by a health monitoring system that might remind a person to take a daily medication or notify caregivers whether a person has taken their medication. This data was recorded on an IMU attached to the pill bottle, recording a person periodically performing one of two actions: (1) taking the bottle off a shelf, removing a “pill”[1], taking the “pill” and putting the bottle back on the shelf, or (2) moving the bottle but not taking any pills. The three kinds of events we asked participants to recognize were the two event types listed above and an implicit “nothing” event when nothing is happening.

The sensor streams for this dataset were largely flat except for when an event was occurring. A sample of data from this activity is shown in Figure 2.

The Running dataset modeled the use-case of creating a pedometer: showing video/data of a person alternatively walking, running, and skipping. The data was recorded on an IMU which was attached to the shirt of the pedestrian. The participants were asked to annotate "run" and "skip" events and the default, implicit event was everything else that happened during data collection which was primarily walking.

This dataset contains more variance and noise than the Pills dataset as the subject was always in motion. A sample of data from this activity is shown in Figure 3.

The Cane dataset involves the use of a walking cane as an aid to mobility. HOLE: do we need to note that we used the cane dataset for the third study as well? The IMU was attached to the cane itself. A member of the

[1] The “pills” were pill-sized pieces of candy.

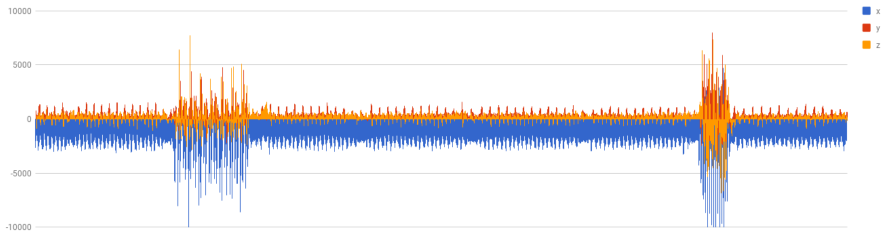


Fig. 3. Sample of data typical for the "run" dataset. The actor walked, skipped, walked, ran, and then continued walking. Although the actor is in constant motion, the run and skip events correspond to higher peaks in the acceleration data.

research team walked with the cane and occasionally dropped the cane. Participants were to annotate "walk" and "drop" events. This dataset was used only as a training task and is not part of any reported results.

3.2 Interface

To measure the effects of both the sensor data and video modalities, we created our own D+V annotation interface we called VAT (Video Annotation Tool), shown in Figure 1. This interface had three modes: one showing the data time series and controls, which we call "*data*", another with the video images and controls, which we call "*video*", and one with both the data time series and the video images, which we call "*both*." All three modes included interface elements for adding, modifying or deleting an event annotation.

The video controls of the interface were modelled after standard YouTube controls, to be as familiar to users as possible. User annotations were overlaid directly onto the sensor data trace (when it is visible) following the common pattern of other D+V annotation tools [1, 8–10, 16, 27]. A green line on the sensor data display was added to mark the current time in the video. When the video was hidden, the standard play/pause controls to navigate in the video interface still worked for the navigation in the data interface.

4 FIRST STUDY: USE OF DATA AND VIDEO INTERFACES IN EVENT ANNOTATION

The first study was designed to measure the effect of using the data only, video only, or both data and video interfaces on annotator efficiency and accuracy.

4.1 Study Design

Each participant annotated events using the *both* interface and one of the *data* or *video* interfaces.

This study design allowed us to compare participants' mean HOLE: do we want "mean" there? performance on an interface with both video and data with their mean performance on the video or data interfaces alone and to compare their performances on the data and video interfaces.

We recruited 60 participants, of whom 47 were male and 13 female. All were undergraduate students. 41 were STEM majors and the others were in other majors. Participants were compensated $15 for their time. To motivate good performance, we told participants that the best three performers would receive an additional $40. We administered a nine question survey at the end of the study. The study was approved by our institutional review board and no adverse events occurred during the study.

We counterbalanced the order in which participants saw information for annotation. The participants were divided into 4 groups as follows:

(1) 15 participants saw only data followed by both data and video,
(2) 15 participants saw both video and data followed by only data,

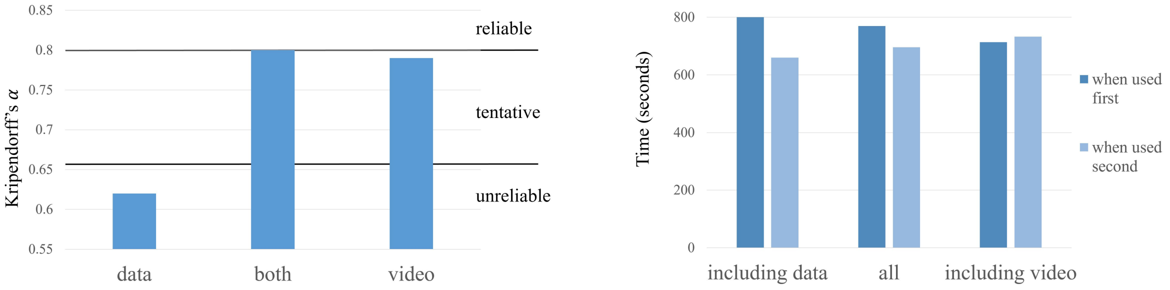


Fig. 4. Mean interrater reliability for all uses of a specific interface, left, and mean annotation time for (i) all tasks, (ii) tasks involving data (with and without video), and (iii)tasks involving video (with and without data), each of these in either the first or second use. See Table 4 and Table 5 for the details of the statistical tests.

(3) 15 participants saw only video followed by both data and video, and
(4) 15 participants saw both data and video followed by only video.

All groups of participants used the interface with both data and video at least once. We did not ask a group of participants to use the data only and video only interfaces because we anticipated that participants would perform about the same with the video only interface compared to the both video and data interface. This reduced the number of participants needed for the study.

Efficiency was measured as the time taken to complete the annotation tasks. Accuracy was measured using Kripendorff's α[17] for the participants' labels in comparison to a gold standard. Kripendorff's α measures inter-rater reliability over a discretization (in time) of the data, it is frequently used as an inter-annotator agreement metric. Higher scores indicate that annotators more often assigned the correct label to data segments.

The gold standard was created by three members of the research team with two members independently annotating event start, end, and type, and the third acting as an arbitrator if the other two disagreed. The same people annotated all events in the gold standard.

4.2 Results

Focusing first on annotation consistency, we examined the annotations generated by participants using two-way ANOVA to look for affects involving the interface modality (users using video only, data only or both) and to check for a learning affect. The interface mode has a significant effect (p-value = 4.23e-06), but neither the learning effect (p-value = 0.279) nor the interaction of modality and learning (p-value = 0.571) have a significant effect. Detailed descriptions of statistical methods used for statistical results presented in this paper, included those used for significance testing, can be found in the appendices.

Since the effect of user interface modality was statistically significant in the ANOVA, we also looked at this affect using various two-tailed t-tests. The left side of Figure 4 compares the interrator reliability and indicates that the use of data alone is significantly inferior to other interface modalities. Annotators were less reliable on the data only interface as compared to the video only (p-value = 0.002) or the interface that offered both video and data (p-value = 4.58e-05).

We also measured the amount of time required to annotate the events in each task and checked for learning affects. The chart on the right side of Figure 4 shows the mean time taken to complete the first and second tasks for all tasks, given interfaces including the data (which are the "both" and "data only" interfaces) and interfaces including the video (which are the "both" and "video only" interfaces). Over all interface types groups HOLE: ?types groups? together, the difference between the first and second tasks is not significant (p-value = 0.156), nor is the difference significant if we consider only interfaces where video is included (p-value 0.802); however, in interfaces where data is included, there is a statistically significant learning affect (p-value = 0.0159, which is less than the Bonferroni corrected level of $p < 0.05/3 = 0.0167$). That is, participants tended to learn to use the data quickly after having seen the data before. This learning effect is statistically significant, but it is not clear from this study exactly what the learning effect means. The design of the second study is intended to provide more clarity about learning effects and sensor data in annotation interfaces.

An examination of responses to post-study interview questions, however, provides some indication for this learning effect. Speaking of using data only, one participant explained that: "Initially the data is very difficult to understand, you just have to make assumptions." It is not difficult to imagine that novice annotators presented only with the sensor data traces shown in Figures 2 and 3 struggled to identify and annotate event boundaries and types.

Seeing both the video and data together helped participants learn to spot patterns in the sensor data stream "The data was almost impossible to use by itself. With the video it was easy to see the correlation between the data and what was happening" (P16). After learning to spot events in the sensor data stream, participants seemed to prefer the data interface: participant P46 said "I like the data, it was nice to see when the movement was going to occur so you could predict. I preferred having both."

4.3 Discussion

The performance and accuracy results along with the participant responses led us to form two hypotheses:

(1) Video seems helpful in allowing one to discover patterns in the data which correspond to activity events. We refer to this as a "Rosetta Stone" effect[2]

(2) Data is useful in moving quickly through the events to find event boundaries. We refer to this as "indexing."

The learning effect for interfaces including data shown in Figure 4 is statistically significant, but the study design does not allow us to see what is producing the learning. Comments from study participants provide some insight about learning from video to recognize events in data but the learning effect and it's causes remain unclear.

5 SECOND STUDY: FURTHER INVESTIGATION OF LEARNING, EFFICIENCY AND ACCURACY

The purpose of the second study was to more carefully evaluate the role of learning from the interface showing both video and data, and to better understand the impact of the data only and video only interfaces on efficiency and accuracy. The second study was also approved by our Institutional Review Board and no adverse events were reported. HOLE: Should we also mention something here about using messier data?

5.1 Methods

In the second study, we measured efficiency by giving participants a fixed amount of time in which to do their work. This reduced the length of the study and allowed us to compare completion rate rather than time to completion. Enforcing a time limit prevented us from seeing decreased efficiency due to fatigue. Future work might evaluate the role of fatigue in the different interface types.

We also split accuracy into two parts: correctly identifying the event type and correctly identifying the bounds (beginning and ending) of an event. We measured the accuracy of event boundary annotations using an asynchronous continuous time model rather than the synchronous discrete time model used in the first study. Krippendorff's α, as used in the first study, involves splitting time into uniform segments and determining if the correct label was applied to each segment. This metric is artificially inflated if the sensor data include long periods

[2] The Rosetta stone is stone slab which contained a single decree written in three languages around 200 BC. The stone was an important part of discovering the meaning of ancient Egyptian hieroglyphs because the same message appeared in multiple languages.

of time in which nothing happens, or long periods during which a single event happens, and the annotators correctly label those long time periods.

We are more interested in how annotators handled the boundaries at the start and end of an event. The accuracy metric in the second study measures only the deviation from the actual event start or end rather than the correctness of the label applied during each uniform discrete time step. Our metric is similar to the metric used in [21].

At the beginning of the study, all participants watched a three-minute training video about the purpose of annotating data for machine learning algorithms in order to create prototypes. HOLE: <- we didn't do this - we only had Austin's training video Immediately after, they watched a three minute video introducing them to the VAT interface. HOLE: <- we did do this

Each participant performed 4 annotation tasks with different interfaces. The datasets and interface types were counterbalanced. In each task, participants used the interface with both video and data first, followed by either the data only or video only interface, and concluding with the third interface type. Table 1 summarizes the tasks, task type, task orders, and interface orders. We also collected interface log data during task completion.

The initial practice task was to familiarize the participant with the interface types and data from this task were not included in any analyses. The first and second tasks involved each interface type as shown and the final task involved the first task dataset on the data only interface.

Participants always first used the interface with both video and data because we were interested in measuring the learning effect created by the interface with both elements. The first study demonstrated a learning effect related to interfaces involving data. The counterbalancing in the second study is intended to capture that effect but to allow us to compare the size of the effect on the data only and video only interfaces.

After the final task, we gave participants a final survey consisting of closed and open ended questions to evaluate when the interfaces with data only, video only, or both data and video were most helpful.

HOLE: Need to update table

Each group in Table 1 contained 14 participants HOLE: probably not true anymore, need to check and fix for a total of 56 HOLE: 73 participants. Participants were recruited from on-campus advertisements. Most were full-time undergraduate students, but 1 was a faculty member and at least one was a graduate student. 75% HOLE: probably 62% were STEM majors. All participants received $15 US for their participation. We told participants that the top 1/3 HOLE: this time we told them the top 3 performers would of performers would receive an additional $40 US, so participants were highly motivated to be as accurate and efficient as possible in that window of time. We told participants that they did not need to find the exact boundary but that a small margin of error (less than 0.5 seconds) was acceptable.

Additionally, we recorded how participants were annotating data. Participants were able to use hot-keys and menus to access various tools in order to annotate the video. We categorized each of these tools into one of three groups. First, data related tools for annotating or navigating the data, like clicking on or scrubbing the data. Second, video related tools HOLE: the other two have examples, should we include one for this or is it just understood?. Third, tools that were not uniquely associated with data or video, like menus and buttons.

5.2 Results

In the results that follow, each type of error is labeled as shown in Table 2. An annotation is correct if the start and end bounds are within the margin of error compared to the gold standard event boundaries and the event type is correct.

5.2.1 Annotation Efficiency and Interface Type. Figure 5 contains two stacked bar charts. The chart on the left summarizes results for the pills dataset and the chart on the right summarizes the running dataset. These results represent annotation on a dataset after annotating events in that data set using the both interface. Each chart is

Table 1. Summary of tasks including the order in which participants used each dataset and interface type. Participants always used the interface with both data and video first in order to measure the impact of learning on the data only and video only interfaces separately as suggested in the first study. Each group contained 14 participants. HOLE: <- need to check if this is still true - not actually sure how many did each one

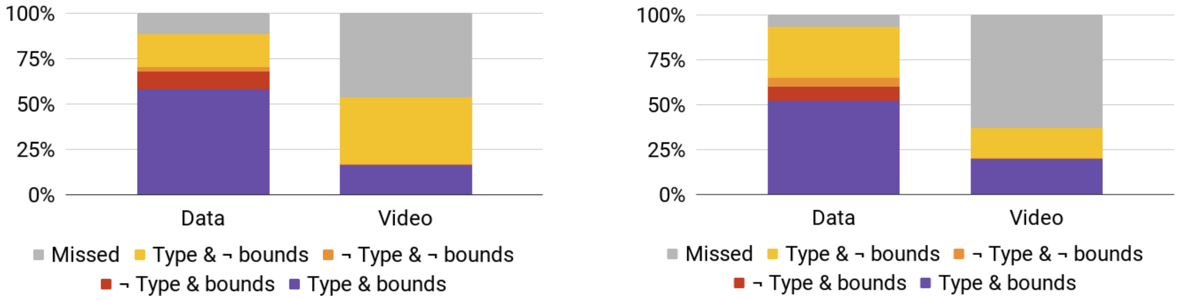
	GROUP A	GROUP B	GROUP C	GROUP D
Practice Task	cane	cane	cane	cane
	both only video only data	both only data only video	both only video only data	both only data only video
First Task	run	run	pills	pills
	both only video only data	both only data only video	both only video only data	both only data only video
Second Task	pills	pills	run	run
	both only video only data	both only data only video	both only video only data	both only data only video
Final Task	run	run	pills	pills
	only data	only data	only data	only data

Table 2. Kinds of annotation errors. The notation "¬ bounds" is intended to be read in the sense of a Boolean formula meaning "not bounds" indicating that the annotator did *not* mark the *bounds* correctly.

Name	Definition
missed	Participant did not add bounds within the event boundaries.
type & ¬ bounds	Type label is correct, but one or more bounds did not fall within our epsilon.
¬ type & ¬ bounds	Type label does not match and one or more bounds did not fall within our epsilon.
¬ type & bounds	Type label type does not match the event type but both bounds are correct.
type & bounds	Type label is correct and both bounds fall within our prescribed epsilon.

normalized to 100% of the events in the dataset, each video had eight events. Purple bars represent events with both type and bounds annotated correctly. Yellow, red and orange bars represent events with bounds, type, or both annotated incorrectly. Grey bars represent events for which the participant made no attempt to annotate the event bounds or type.

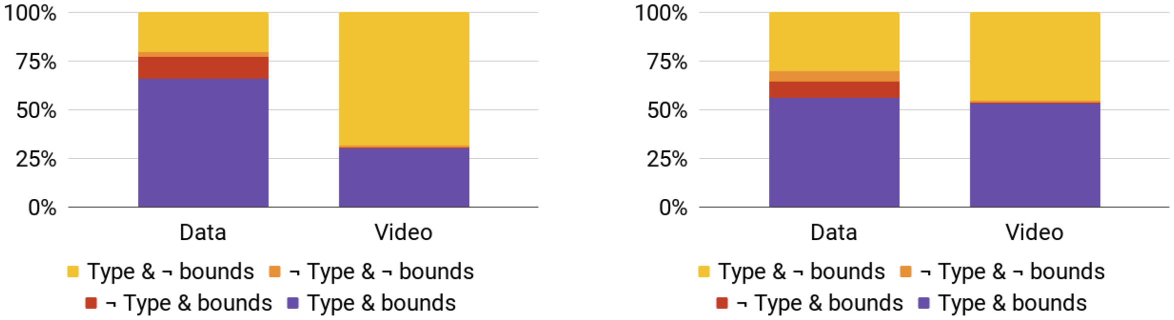
HOLE: need to check if this is still true and add in new analysis for anything different with the messier dataAnnotators were more efficient using the data interface than the video interface for both the pills and the running datasets. Specifically, when time ran out for the task in the study, participants had missed annotating fewer events using the data interface, as shown by the height of the gray bars, than using the video interface (12% vs. 46%, p-value = 4.31e-88). Furthermore, participants annotated more events correctly within the allotted



(a) Pills

(b) Running

Fig. 5. Average annotation efficiency after a training period for both the pills and running datasets given the data and video interfaces and a fixed time limit. Annotators worked more efficiently, attempted to annotate more events and annotated more events correctly using the data interface compared to the video interface. In each graph, events for which the annotator made no attempt at marking the event bounds or type during the allotted time are shown in grey. Events with type and bounds annotated correctly are shown in purple and events annotated incorrectly are shown in red, orange or yellow depending on the kind of error. See Table 6 and Table 7 for the details of the statistical tests.



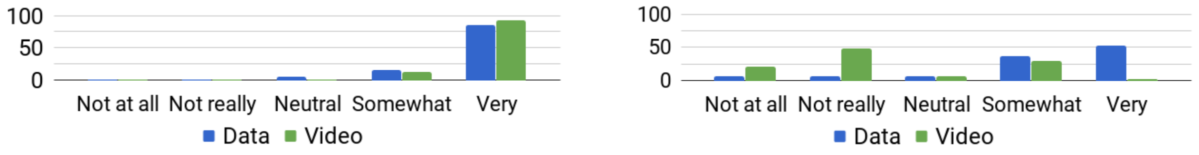
(a) Pills data

(b) Running data

Fig. 6. Annotation results for the pills and running datasets considering *only* the events for which a participant made an attempt at annotating the event. For the pills dataset, participants annotated events correctly more often using the data interface than the video interface. For the running dataset, participants annotated events correctly with about the same frequency. The frequency of annotating either the event type or bounds incorrectly is different across interface type. See Table 6 and Table 7 for the details of the statistical tests, but note that this is the same data and the same statistical tests as for Figure 5, but without the items the participants attempted to annotate.

time using the data interface than the video interface as can be seen by the height of the purple bars (58% vs. 16%, p-value = 7.41e-74). Just as for the previous results, the details of the statistical tests can be found in Appendix B.

5.2.2 Error Frequency and Interface Type. The frequency of errors for the different interface types for each dataset is shown in Figure 6. The stacked bars in Figure 6 represent only the events for which the participant made an attempt at labeling the event bounds or type–or both. Each graph is normalized to the average number



(a) How essential are data and video during your first round of annotation?

(b) How essential are data and video after several rounds of annotation?

Fig. 7. After participants had completed several rounds of annotation with each interface type, we asked them to rate the importance of each interface element for time points in a hypothetical future annotation task. Figure (a) on the left summarizes the responses for the hypothetical first round of annotation and Figure (b) summarizes responses after "several rounds" of annotation. See Table 8 for the details of the statistical tests.

of events attempted by each participant. In terms of Figure 5, the gray bars were omitted and the remaining data normalized by the number of events labeled.

Considering only the events that a participant annotated, participants still labeled events accurately more frequently using the data interface than the video interface for the pills dataset(65.6% vs. 29.8% of the *annotated* data, 58% vs. 16% of the total data; p-value = 7.4e-74). Participants marked the event type incorrectly more often using the data interface but marked the events bounds incorrectly more often using the video interface (14.0% vs. 1.44% of the annotated data,12% 0.7% of the total data; p-value = 7.6e-49). For the running dataset, participants marked events correctly with approximately the same frequency using either the data or the video (55.85% vs. 53.09% of the annotated data, 52% 20% of the total data; p-value = 7.3e-41) but marked event types incorrectly more often using the data interface and marked event bounds incorrectly more often using the video interface.

Marking event types incorrectly more often using the data interface may have happened because participants had an easier time identifying an event when watching the event in a video player than when imagining the event based on a sensor data plot. Participants may have marked event bounds incorrectly more often using the video interface because it may be easier to spot the start of a motion in an abrupt change in an accelerometer data plot than in a video player.

5.2.3 Perceived Importance of Interface Types. After the final annotation task, we asked participants to rate the importance of the data and video interface elements. At this point in the study, each participant had used each interface type twice on two different data sets (except for the data only interface which they had used three times on one of the datasets).

Four questions asked participants to speculate on the value of the data and video interfaces in a future hypothetical annotation task. Figure 7 shows a histogram of the Likert responses to these four questions. Figure 7a indicates that during the first round of a hypothetical future task, participants indicated that both data and video are important (p-value = 0.10). But for a hypothetical scenario after "several rounds" of annotating, participants indicated that data would be more important than video (p-value = 2.0e-18) as shown in Figure 7b.

HOLE: did we get the same sorts of responses this time around? We also asked 11 of the participants 3 open ended questions about video annotation, the importance of the data interface and the importance of the video interface. We identified three themes in participant responses.

First, the video acted as a Rosetta stone to help participants understand the data traces in the data interface. Participant P3 said "The video is essential at first so that you can see what the data is telling you, but once you are able to recognize the pattern of the data it's not essential." Similar statements appeared in 5 other responses.

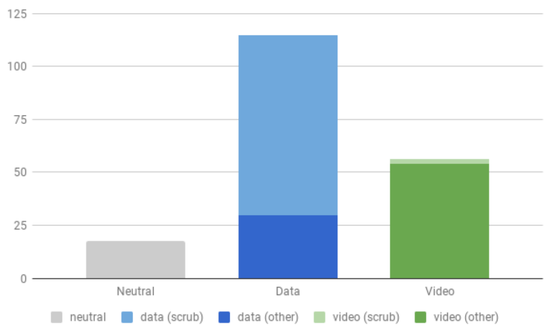


Fig. 8. When both the data and video interfaces are available, participants use data and video controls to different extents. Of the controls used, scrubbing the data control is is the most used. See Table 9 and Table 10 for the details of the statistical tests.

Second, the data interface made it easier to locate event boundaries. Participant P7 said "data was necessary for pinpointing when an action started. Other wise I am left to guessing when something starts and finishes off a crappy video." Similar statements appeared in 5 other responses.

Third, the data interface acted as an index to help participants skip through the video to focus times when events occurred. Participant P8 said "for me data was most beneficial as a means to skip around in the video." Similar statements about using the data for indexing into the video appeared in 5 other responses.

5.2.4 Usage of Each Kind of Interface. During the first and second tasks, we logged all interface events generated by each participant. We divided the events into "data" events which involved the data interface, "video" events which involved only the video interface and "neutral" events which correspond to interface elements which are visible in every interface type.

HOLE: is this still the case? do we have this data? ->We further subdivided interface events into "scrub" and all other events. Scrub events happens when a participant uses the interface to move in the sensor data or video time series.

Figure 8 shows the total usage of each interface type and further separates out the scrub events from the other events. Data events are far more common on average than video events (106.3 vs. 52.1; p-value = 5.1e-07). Participants also used the data scrubbing event far more than the Video scrubbing event (78.6 vs 1.8; p-value 1.6e-10). They also used non-scrubbing video events more frequently than the non-scrubbing data events, but not to a Bonferroni corrected (0.05/2 = 0.025) statistically significant level (p-value = 0.044).

This pattern of tool usage suggests that the data interface was more useful for navigating through time in the data series.

6 DISCUSSION

HOLE: need to update this section based on new resultsVideo is not always essential, or even useful, for D+V annotation tasks for activity recognition. Video was important in the early phases of an annotation project to help annotators learn to spot events in the sensor data stream. But after using an interface with both data and video, novice annotators using an interface with only the sensor data were more efficient than annotators using an interface with only video. Annotators could also identify event boundaries more accurately using only the sensor data compared to using only the video data. Annotators identified event types more accurately with the video only interface than with the data only interface.

HOLE: need to check if this is still the case with new responsesResponses to closed and open ended survey questions in both studies include a preference for using the data only interface *after* learning to recognize events in sensor data. Logs of user interface events show a preference for using the sensor data interface to move through time.

One explanation for this is that novice users learn how to recognize patterns in the data. After that learning has occurred, users may be able to annotate patterns in the data more quickly because the user can see the data vary over time rather than seeing just a single instant of time in a video frame. This allows annotators to annotate more quickly and to identify event boundaries more accurately. However, participants could identify the event type *less* accurately because it is still more intuitive to recognize activity types in a video frame than in sensor data traces.

It should, perhaps, come as no surprise that human annotators can learn to spot patterns in sensor data because machine learning algorithms perform the same task. It could be the case that if a machine can spot a pattern in sensor data, then a person can too–as long as there are not too many sensors.

7 CONCLUSION

Our study has several implications for D+V annotation tasks in the context of supervised learning of sensor-based activity recognizers. First, it may not be essential to collect video of all events for use during annotation. Collecting some video for training is important, but once users have learned to spot patterns, they appear be more accurate and more efficient without the video. This means that projects in which it is unfeasible to collect video paired with *all* sensor data may generate enough video to successfully train annotators.

Second, users may need to start with both video and data together to learn to spot events in the data using the video as reference. This suggests that future tools for D+V annotation might provide an easy way, or an adaptive automatic way, to switch from a mode involving both data and video to a mode involving data only.

Third, it may be useful to more deeply explore the design of "data only" interfaces for annotating events. These interfaces might show examples of the data trace of certain event types for reference or make it easier to zoom in and out of the temporal scale of an event trace.

Fourth, users are somewhat less accurate when labeling event types without video. Investigating machine learning algorithms which are less sensitive to a few mislabeled event types in a larger corpus of labeled data compared to a perfectly labeled smaller corpus may improve the overall accuracy of an activity recognizer.

This study was limited by the simplicity of the data we used in the annotation experiments. While this simplicity was intentionally chosen in order to understand what a novice can do in a best-case simple scenario, it limits the generality of our results. Our results are not likely to generalize to settings involving multiple sensors. For example, it is unlikely that a person can learn to reliably annotate events in the output of 20 acceleration sensors simultaneously recording data.

HOLE: definitely need to update for drywall and gait data - discussion of the messier data and impactsThe study was also limited by the amount of noise in the data we showed participants. It may be more difficult for an annotator to spot events in data with significantly more spurious noise. Data with more noise may lead to different results. Both the pills and the running datasets had very little spurious noise. But datasets involving more noise, such as might be generated by a sensor attached to a pneumatic jack hammer, may produce a different annotation experience.

A APPENDIX 1: STATISTICAL TESTS FROM STUDY 1

We examined the Kripendorff's α[17] scores from annotations in study 1 using two-way ANOVA to look for affects involving the interface modality (users using video only, data only or both) and to check for a learning effect.

Table 3. ANOVA: Interface and Learning Affects in study 1. Items significant at the 0.05 level are shown in bold.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
User Interface Type	2	0.6645	0.3323	13.819	**4.23e-06**
Order	1	0.0285	0.0285	1.184	0.279
User Interface Type: Order	2	0.0271	0.0136	0.564	0.571
Residuals	114	2.7412	0.0240		

Motivated by the ANOVA results we also conducted t-test of Kripendorff's α results. Note that the Data Only vs. Video Only test is not paired:

Table 4. t-test comparisons of Kripendorff's α results. Items significant at the Bonferroni corrected level of 0.05 / 3 are shown in bold. The Video Only vs. Data Only test is a Welch two-sample t-test, the others are paired t-tests. Note that the data labeled "Both" is only from the group that provided the corresponding Data Only or Video Only annotations.

Condition 1	Mean	Condition 2	Mean	t	df	p-value
Data Only	0.62	Both Data and Video	0.80	4.79	29	**4.58e-05**
Video Only	0.77	Both Data and Video	0.79	1.034	29	0.31
Video Only	0.77	Data Only	0.62	3.32	46.13	**0.00175**

We also conducted paired t-test of time our participants required:

Table 5. Pairwise comparisons of the time taken. Items significant at the Bonferroni corrected level of 0.05 / 3 are shown in bold.

Condition 1	Mean	Condition 2	Mean	t	df	p-value
Data Only First	825.77	Data Only Second	659.57	2.56	29	**0.0159**
First	769.61	Second	696.17	1.44	59	0.16
Video Only First	713.46	Video Only Second	732.77	-0.252	29	0.802

B APPENDIX 2: STATISTICAL TESTS FROM STUDY 2

Table 6. Pills dataset: Results which are statistically significant at the Bonferroni corrected level of .05 / 6 = .0083 are bold. This table shows the statistical significance of the data used in Figure 5a as per the Mann Whitney test.

Condition 1	Mean	Condition 2	Mean	Statistic	P-value
Missed (Data)	0.12	Missed (Video)	0.46	23648.0	**4.316e-88**
Type & ¬ Bounds (Data)	0.18	Type & ¬ Bounds (Video)	0.37	53728.0	**5.46e-30**
¬ Type & ¬ Bounds (Data)	0.02	¬ Type & ¬ Bounds (Video)	0.005	84768.0	**1.71e-8**
¬ Type & Bounds (Data)	0.10	¬ Type & Bounds (Video)	0.002	56736.0	**2.78e-46**
Type & Bounds (Data)	0.58	Type & Bounds (Video)	0.16	28128.0	**7.41e-74**
¬ Type (Data)	0.12	¬ Type (Video)	0.007	283040.0	**7.56e-49**

Table 7. Running dataset: Results which are statistically significant at the Bonferroni corrected level of .05 / 6 = .0083 are bold. This table shows the statistical significance of the data used in Figure 5b as per the Mann Whitney test.

Condition 1	Mean	Condition 2	Mean	N-value	P-value
Missed (Data)	0.07	Missed (Video)	0.63	47136.0	**6.70e-132**
Type & ¬ Bounds (Data)	0.28	Type & ¬ Bounds (Video)	17	84704.0	**.000489**
¬ Type & ¬ Bounds (Data)	0.05	¬ Type & ¬ Bounds (Video)	0.002	86080.0	**1.34e-10**
¬ Type & Bounds (Data)	0.08	¬ Type & Bounds (Video)	0.002	82496.0	**1.39e-14**
Type & Bounds (Data)	0.52	Type & Bounds (Video)	0.20	47136.0	**7.28e-41**
¬ Type (Data)	0.13	¬ Type (Video)	0.005	337152.0	**2.60e-23**

Table 8. Combined pills and running dataset: Participants' self evaluation of how essential video and data are. Results which are statistically significant at the Bonferroni corrected level of .05 / 2 = .025 are bold. This table shows the statistical significance of the data used in Figure 7 as per the Mann Whitney test.

Condition 1	Mean	Condition 2	Mean	N-value	P-value
Essential at first (Data)	4.70	Essential at first (Video)	4.80	5541.5	.10
Essential later (Data)	4.13	Essential later (Video)	2.50	2032.0	**2.05e-18**

C APPENDIX 3: HOW OFTEN PARTICIPANTS CHOOSE TO USE WHICH TOOLS

Table 9. Annotation tools: This table shows the statistical significance of the data used in Figures 8 using the Wilcoxon rank-sum test. Results which are statistically significant at the Bonferroni corrected level of .05 / 3 = .017 are bold.

Condition 1	Mean	Condition 2	Mean	Statistic	P-value
Neutral	16.4	Data	106.3	2.0	**1.2e-10**
Neutral	16.4	Video	52.1	32.5	**9.6e-10**
Video	52.1	Data	106.3	170.5	**5.1e-07**

Table 10. Annotation tools: This table shows the statistical significance of the data used in Figures 8 using the Wilcoxon rank-sum test. Results which are statistically significant at the Bonferroni corrected level of .05 / 2 = .025 are bold.

Condition 1	Mean	Condition 2	Mean	Statistic	p-value
Data scrubbing	78.61	Video scrubbing	1.85	0.0	**1.62e-10**
Data no scrubbing	27.72	Video no scrubbing	50.25	488.0	0.044

REFERENCES

[1] Michael Barz, Mohammad Mehdi Moniri, Markus Weber, and Daniel Sonntag. 2016. Multimodal multisensor activity annotation tool. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 17–20.

[2] Abdelkareem Bedri, Apoorva Verlekar, Edison Thomaz, Valerie Avva, and Thad Starner. 2015. Detecting Mastication: A Wearable Approach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, 247–250. https://doi.org/10.1145/2818346.2820767

[3] Ceara Byrne, Jay Zuerndorfer, Larry Freil, Xiaochuang Han, Andrew Sirolly, Scott Cilliland, Thad Starner, and Melody Jackson. 2018. Predicting the Suitability of Service Animals Using Instrumented Dog Toys. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 127 (Jan. 2018), 20 pages. https://doi.org/10.1145/3161184

[4] Liming Chen, Jesse Hoey, Chris D Nugent, Diane J Cook, and Zhiwen Yu. 2012. Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 790–808.

[5] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M. Jones, Agata Rozga, and James M. Rehg. 2017. Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 43 (Sept. 2017), 20 pages. https://doi.org/10.1145/3131902

[6] O. Crasborn, H. Sloetjes, E. Auer, and P. Wittenburg. 2006. Combining video and numeric data in the analysis of sign languages with the ELAN annotation software. In *Proceedings of the 2nd Workshop on the Representation and Processing of Sign languages: Lexicographic matters and didactic scenarios*. 82–87.

[7] Adrian Derungs, Sebastian Soller, Andreas Wesihaupl, Judith Bleuel, Gereon Berschin, and Oliver Amft. 2018. Regression-based, mistake-driven movement skill estimation in Nordic Walking using wearable inertial sensors. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom 2018)*. IEEE, 155–164.

[8] Anind K. Dey, Raffay Hamid, Chris Beckmann, Ian Li, and Daniel Hsu. 2004. A CAPpella: Programming by Demonstration of Context-aware Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 33–40. https://doi.org/10.1145/985692.985697

[9] Alexander Diete, Timo Sztyler, and Heiner Stuckenschmidt. 2017. A smart data annotation tool for multi-sensor activity recognition. In *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*. IEEE, 111–116.

[10] Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D. Hollan. 2011. ChronoViz: A System for Supporting Navigation of Time-coded Data. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. ACM, New York, NY, USA, 299–304. https://doi.org/10.1145/1979742.1979706

[11] Michael Jones, Casey Walker, Zann Anderson, and Lawrence Thatcher. 2016. Automatic Detection of Alpine Ski Turns in Sensor Data. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 856–860. https://doi.org/10.1145/2968219.2968535

[12] Liju Joshua and Koshy Varghese. 2013. Selection of accelerometer location on bricklayers using decision trees. *Computer-Aided Civil and Infrastructure Engineering* 28, 5 (2013), 372–388.

[13] Aftab Khan, James Nicholson, and Thomas Plötz. 2017. Activity Recognition for Quality Assessment of Batting Shots in Cricket Using a Hierarchical Representation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 62 (Sept. 2017), 31 pages. https://doi.org/10.1145/3130927

[14] Md Abdullah Al Hafiz Khan and Nirmalya Roy. 2017. TransAct: Transfer learning enabled activity recognition. In *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*. IEEE, 545–550.

[15] Keunseo Kim, Hengameh Zabihi, Heeyoung Kim, and Uichin Lee. 2017. TrailSense: A Crowdsensing System for Detecting Risky Mountain Trail Segments with Walking Pattern Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 65 (Sept. 2017), 31 pages. https://doi.org/10.1145/3131893

[16] Michael Kipp. 2001. ANVIL - A Generic Annotation Tool for Multimodal Dialogue. (2001).

[17] Klaus Krippendorff. 2012. *Content analysis: An introduction to its methodology*. Sage.

[18] Nicholas D. Lane, Ye Xu, Hong Lu, Shaohan Hu, Tanzeem Choudhury, Andrew T. Campbell, and Feng Zhao. 2011. Enabling Large-scale Human Activity Inference on Smartphones Using Community Similarity Networks (Csn). In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. ACM, New York, NY, USA, 355–364. https://doi.org/10.1145/2030112.2030160

[19] Walter S. Lasecki, Mitchell Gordon, Danai Koutra, Malte F. Jung, Steven P. Dow, and Jeffrey P. Bigham. 2014. Glance: Rapidly Coding Behavioral Video with the Crowd. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 551–562. https://doi.org/10.1145/2642918.2647367

[20] Walter S. Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P. Bigham. 2013. Real-time Crowd Labeling for Deployable Activity Recognition. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1203–1212. https://doi.org/10.1145/2441776.2441912

[21] L. Nguyen-Dinh, A. Calatroni, and G. Troster. 2017. Supporting One-Time Point Annotations for Gesture Recognition. *IEEE Transactions on Pattern Analysis And Machine Intelligence* 39, 11 (Nov. 2017), 2270–2283. https://doi.org/10.1109/TPAMI.2016.2637350

[22] Cuong Pham and Patrick Olivier. 2009. Slice&dice: Recognizing food preparation activities using embedded accelerometers. In *European Conference on Ambient Intelligence*. Springer, 34–43.

[23] Cuong Pham and Tu Minh Phuong. 2013. Real-time fall detection and activity recognition using low-cost wearable sensors. In *International Conference on Computational Science and Its Applications*. Springer, 673–682.

[24] Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and vision computing* 28, 6 (2010), 976–990.

[25] Vitor F Rey and Paul Lukowicz. 2017. Label Propagation: An Unsupervised Similarity Based Method for Integrating New Sensors in Activity Recognition Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 94.

[26] Edison Thomaz, Abdelkareem Bedri, Temiloluwa Prioleau, Irfan Essa, and Gregory D Abowd. 2017. Exploring Symmetric and Asymmetric Bimanual Eating Detection with Inertial Sensors on the Wrist. In *Proceedings of the 1st Workshop on Digital Biomarkers*. ACM, 21–26.

[27] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1556–1559.

Received May 2018