uxSense: Supporting User Experience Analysis with Visualization and Computer Vision

Andrea Batch, Yipeng Ji, Mingming Fan, Jian Zhao, and Niklas Elmqvist, *Senior Member, IEEE*

Abstract—Analyzing user behavior from usability evaluation can be a challenging and time-consuming task, especially as the number of participants and the scale and complexity of the evaluation grows. We propose UXSENSE, a visual analytics system using machine learning methods to extract user behavior from audio and video recordings as parallel time-stamped data streams. Our implementation draws on pattern recognition, computer vision, natural language processing, and machine learning to extract user sentiment, actions, posture, spoken words, and other features from such recordings. These streams are visualized as parallel timelines in a web-based front-end, enabling the researcher to search, filter, and annotate data across time and space. We present the results of a user study involving professional UX researchers evaluating user data using uxSense. In fact, we used uxSense itself to evaluate their sessions.

Index Terms—Visualization, visual analytics, evaluation, video analytics, machine learning, deep learning, computer vision.

1 INTRODUCTION

We live in the age of the disappearing computer [1] where most any gadget now involves computing technology and where the internet increasingly becomes integrated into everyday life. As these devices and technologies reach an ever-expanding audience, the need for compelling user experiences increases. This yields a growing need for usability, user research, and user experience (UX) professionals who can design, evaluate, and revise these interfaces. Unfortunately, UX evaluation is a costly and time-consuming activity that does not scale with increased demand.

The bottleneck is most often the actual analysis of information, where a UX or usability professional often must review hours of video and audio recordings of different individuals who are interacting with a specific piece of technology, such as a smartphone, dishwasher, or website. Further, a UX professional usually needs to attend to multiple behavioral signals (e.g., emotion, facial expression, body movement, and actions) simultaneously at a fast pace and combine them to make thorough and precise assessment of usability problems. These signals are observed or inferred from the video and/or audio recordings, some of which may not always be salient to quick manual checking. The above challenges make such video review a tedious and time-consuming task that often involves stepping forward the video a single frame at a time, repeatedly rewatching a critical segment, or manually coding the mood, actions, or body language of the participants.

We propose a method for extracting multi-modal features of human behavior from video and audio footage using machine learning (ML) to support UX and usability professionals in their analysis of user session data using interactive visualization. Existing off-the-shelf tools for reviewing usability videos (e.g., UserTesting[1] and FullStory[2]) either just offer video playback functions or support extracting a few basic features (e.g., sentiment), lacking a holistic view of user interactions happening in the session. They also do not provide a unified and interactive visual representation of multiple time-based metrics to support in-depth analysis. While some systems (e.g., VA^2 [2] and CoUX [3]) concurrently visualize several sources of data in usability studies, they are still limited in their data features (i.e., eye-tracking, think-aloud, and interaction events) and are not flexible in integrating other time-based metrics.

Our prototype system—UXSENSE—validates the above method by providing a web-based interface to a computational backend that asynchronously runs a range of *filters* to extract different types of data from one or several time-synced video streams uploaded by the user (Figure 1) as *data streams*. Filters in the uxSense system are designed as plugins that can be added or removed at run-time, each responsible for extracting one or more types of data, such as spoken language, gaze direction, arm gestures, head orientation, and even facial expressions. Data streams recovered from each filter are shown in the visual interface as parallel time tracks, similar to the track-centric approach with a horizontal timeline used by video editing software. Finally, the tool provides robust annotation features where the user can select events as well as intervals across the timelines and add notes for future analysis, collaborative iteration and review, and report generation.

To prove the feasibility of our method, we present results from an expert review involving five professional UX designers working at several large tech companies on the U.S. West Coast. We asked these participants to use uxSense to analyze a think-aloud usability session performed on the Tableau desktop visualization tool. To further showcase the utility of uxSense, we used the system itself to evaluate data streams recorded during these sessions. Our findings show promise for our vision of ML to facilitate usability and UX testing, as well as for our uxSense prototype implementation. While one of our participants wanted to know more about the accuracy of the models before she would trust it, the general sentiment among them was that the ideal use of such preprocessing filters could significantly ease their own daily work processes, or at least

- *Andrea Batch and Niklas Elmqvist are with University of Maryland, College Park, MD, United States. E-mail: ajulca@umd.edu, elm@umd.edu*
- *Yipeng Ji and Jain Zhao are with University of Waterloo, ON, Canada. E-mail: y43ji@edu.uwaterloo.ca, jianzhao@edu.uwaterloo.ca*
- *Mingming Fan is with the Hong Kong University of Science and Technology (Guangzhou) and Hong Kong University of Science and Technology. E-mail: mingmingfan@ust.hk*

Manuscript received XXX XX, 2021; revised XXX XX, 2021.

1. https://www.usertesting.com/

2. https://www.fullstory.com/

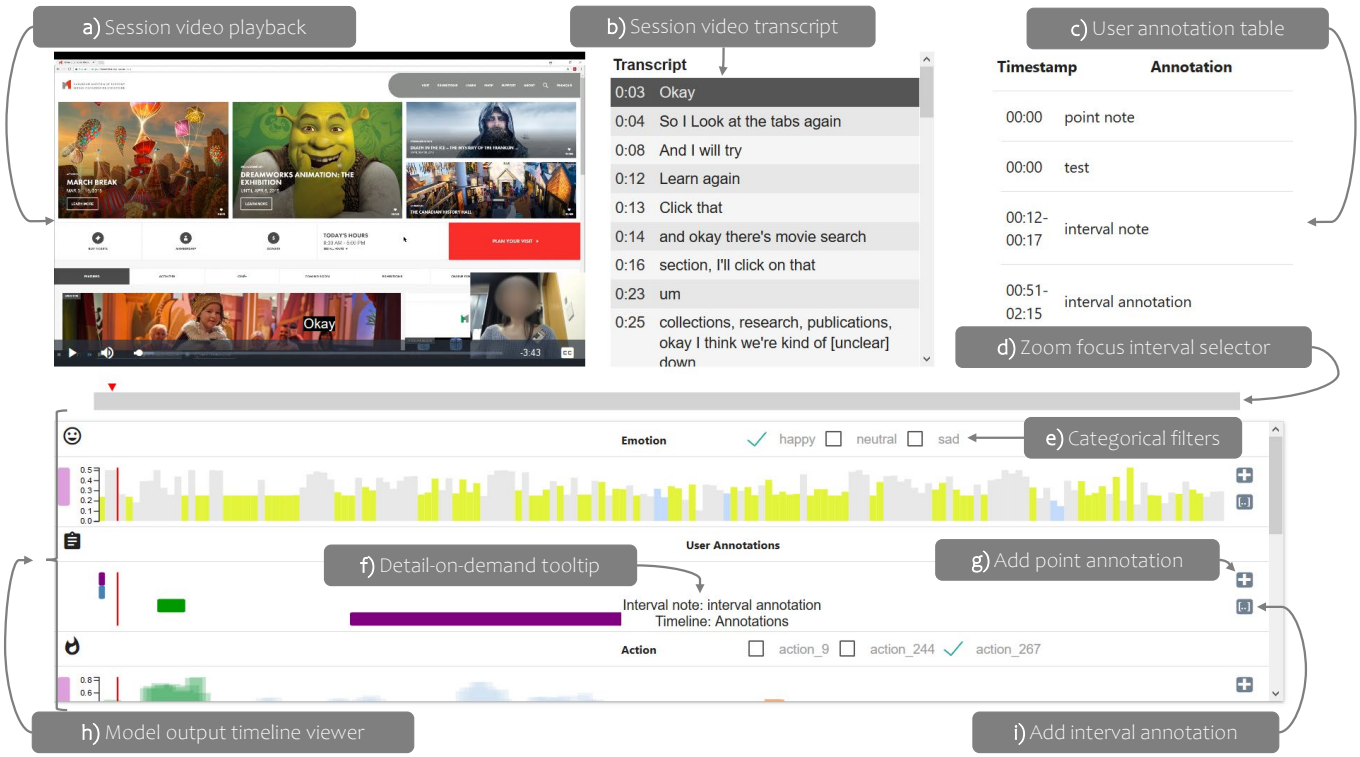


Fig. 1: **Schematic overview.** Analysis interface in the UXSENSE web-based client (view reflects tutorial video). **a)** *Video playback*: View user session video, with or without captions. **b)** *Session transcript*: View timestamped transcript of speech from video, and navigate video by clicking on line of text. **c)** *User annotation Table*: View the text and timestamp of all annotations made by the user. **d)** *Zoom focus*: Select, zoom, and pan whole extent of the video. Red arrow marker indicates current video time, while brushed region shows zoom extent in context of video duration. **e)** *Categorical filters*: When selected, non-selected elements of the view are shown with low opacity. **f)** *Details-on-demand*: Mouseover to get details of observation in model output at given time. **g)** *Point annotation* and **i)** *Interval annotation*: Add an annotation corresponding to given timeline for either the video's current time (**g**) or the brushed interval range (**i**) with (**d**). **h)** *Model output timeline viewer*: The timeline and user annotations are described in Section 3.4.

help them identify points of interest. We close the paper by briefly speculating how future iterations of uxSense could also be used to evaluate mobile and ubiquitous analytics applications, where the participants walk around in physical space rather than being restricted to a workspace with a screen, mouse, and keyboard.

In summary, the primary contributions of this paper include:

1) A design framework for supporting UX research with interactive visual analytics of temporal UX metrics by incorporating off-the-shelf ML modeling of user video and audio;
2) A system—UXSENSE—designed around our framework's design requirements; and
3) Results from a user study evaluating the uxSense prototype.

2 BACKGROUND

In this section, we review the relevant work including visualization and human-computer interaction (HCI) evaluation, ML for extracting user behavior, and visual analytics to support UX studies.

2.1 Data-driven Evaluation in HCI and Visualization

In the domain of HCI and data visualization, some approaches to understanding factors influencing users' experiences and needs for systems involve constructing personas—representational archetypes of "typical" users and their daily lives [4]. This generally involves qualitative and ethnographic methods in which the researcher tracks, records, and interprets the users' daily activities in collaboration with the participant, reaching a shared understanding of the user's thought processes through interview and activity [4], [5]. Alternative approaches exist in which events within the interface, such as mouse activity [6], are used to develop "data-driven personas" [7] for specific types of users. Platforms for crowdsourcing experiments such as Amazon Mechanical Turk [8] make the creation of these types of personas more manageable at larger scales. In these evaluation methods, video and audio recordings tend to be a ubiquitous form of data to capture and analyze.

Video and audio recordings are nuanced data formats that are cheap to capture but expensive to analyze. As a result, HCI and visualization research communities have already begun to shift toward the use of video and audio inputs as a revealed behavior dataset that is time-cost cheap and therefore scalable for the analysis of large user populations. Systems for visualizing and analyzing visual and semantic features of cinematic films in the context of film studies exist. One example is VIAN [9], which represents information about average frame color to the user, who can then manually segment the video with semantic annotations. Kurzhals et al. [10] introduce a system that uses the text of movie scripts to assign semantic labels to frames, which is graphically represented to the user along with motion and other visual frame information in an interactive dashboard that affords user annotation. Pavel et al. [11] present a system for automatically segmenting and summarizing

lecture recordings and append them with crowdsourced transcripts. QuickCut [12] allows for fast video editing and annotation to quickly transcribe, semantically match, and cut together audio annotations corresponding to timestamped clip segments. Leake et al. [13] create a system for automatically generating audio-video slideshows using text and imagery from written articles.

Manual video annotation, such as using the ELAN tool [14], has long been in existence, but recent years have also seen automatic methods that could be used for scalable evaluation based on video. EgoScanning [15] processes first-person (egocentric) videos to detect important passages and adapts playback speed accordingly. VidCrit [16] compiles spoken, textual, and gestural feedback during video production into a visual interface for navigating annotations. Finally, commercial services such as UserTesting and `Frame.io`[3] are based on video annotation and sensing, and while the former is focused on product evaluation and acceptance, it still does not provide sophisticated visual analytics tools to study this data. Compared to these efforts, uxSense provides an extensible list of AI/ML/CV filters that extract various features (e.g., user actions, emotions, speech rate, etc.) from the audio and video recordings, thus providing a holistic view of user behavior.

Finally, in EduSense [17], automatic video annotation is applied at scale to classroom sensing, allowing an array of commodity cameras to capture previously latent metrics about learning environments. We tend to think of the uxSense system as having a similar scope: it captures many previously intangible metrics about user experience from video recordings into a platform that enables UX researcher to explore and analyze this data.

2.2 Characterizing Users with Machine Learning

The HCI community has been harnessing artificial neural networks (ANNs), including recurrent and convolutional neural networks (RNNs and CNNs), for evaluating user behavior. Examples include discovering speech patterns [18], identifying gesture [19]–[21] and gaze [22], [23], classifying user emotion and facial expression [24]–[26], and detecting characteristics of the user, such as gender [27],[4] by constructing and implementing neural network architecture. The visualization community has also made contributions to the toolkit of methods used in evaluating user video, logs, transcripts, and other qualitative data [28], as well as user gesture analysis [29].

In the ML community there has been more than a decade's worth of research exploring methods for action classification [30], [31], motion and path prediction [32], eye tracking [33], and gesture detection [34]. While there have been a few position papers [35] and more serious studies [36] advocating for a closer relationship between the HCI and machine intelligence communities, the current body of literature on the subject is still somewhat sparse.

Our work fills in these gaps by applying suitable ML techniques to the analysis of video and audio recordings in the domain of UX evaluation in HCI and presenting the results interactively for in-depth sensemaking. Specifically, uxSense leverages the available body of research in speech analysis, computer vision, and machine intelligence for characterizing audio and video inputs into different types of information, i.e., data streams. Each of the data streams is generated by independent functional units that we call filters, and then displayed with multiple coordinated timeline visualizations.

3. https://www.frame.io/

4. We note that this approach, like many similar projects conceived with little thought to their sociotechnical impact, are a highly questionable practice.

2.3 Visual Analytics to Facilitate UX Evaluation

As ML continues demonstrating its potential, qualitative researchers are becoming increasingly interested in adopting ML into their analysis flows. However, challenges quickly surface. First, although traditional classification and clustering ML methods are helpful for generating additional labels to inform analysis, these labels alone are often not sufficient for addressing HCI research problems. Instead, HCI researchers need to leverage their skills to make sense of the ML-generated labels to gain a deeper and more nuanced understanding of the data. Second, many ML methods require a significant amount of data to optimize parameters and thus have limited accuracy when dealing with small-scale yet semantically-rich human-behavior data. This has inspired new methods, such as visualization, for better integrating ML into qualitative workflows.

One line of research is to support qualitative coding, which is a powerful yet labor-intensive method. Felix et al. designed a visual data analysis tool that integrates unsupervised learning methods to provide suggestions to help researchers progressively code a large corpus of texts [37]. Another challenge that qualitative researchers often face is to resolve conflicts among researchers when analyzing qualitative data. Drouhard et al. designed a tool, Aeonium, that identifies potential conflicts in codes created by different coders using ML and highlights the conflicts to facilitate coders to spot their disagreements and resolve conflicts efficiently [38].

Another line of research is to support the analysis of user interaction data to uncover users' intentions and reasoning processes. Both low-level user inputs (e.g., mouse clicks, drags, key presses [39], [40]) and high-level graphical structures of user interactions [41] are captured and visualized to help researchers make sense of their analytic activity. Moreover, eye-tracking data have also been visualized to help researchers analyze user interactions and even predict user intent [42], [43]. Furthermore, researchers have investigated user-generated annotations and developed visual interfaces to uncover hidden sensemaking patterns [44], [45]. In addition to using proxy data (e.g., mouse events, eye-tracking data) and manual provenance (e.g., user-generated annotations), researchers have recently begun to investigate think-aloud data, which are generated by asking users to verbalize their thought processes while working on a task, to better understand their hidden thinking process. Think-aloud data have been used to understand analysts' reasoning processes [46], [47] as well as users' interactions [48]. VA^2 visualizes think-aloud, interaction, and eye movement data to facilitate the analysis of multiple concurrent evaluation results [2]. However, VA^2 only supports these three data streams, in raw forms, with a specific visualization design, which lacks a more holistic view of the user behaviors that can be characterized by other data or derived features. Recently, ML models have been employed to predict usability problems of think-aloud sessions (e.g., based on users' speech, verbalization, and scrolling patterns), and the ML predictions are further visualized in timelines to support individual [18] and collaborative [3] UX evaluation. In addition, analytical technologies can detect user moods and facial expressions to facilitate UX evaluation [49]–[52].

Inspired by prior work, we extend this line of research by considering a wider range of modalities of data extracted from video and audio footage that are indicative of users' experiences (speech rate, transcripts, gaze direction, facial expressions, semantic actions). These time-stamped data streams can then be combined into a flexible and extensible timeline visualization panel to enable detailed analysis of user behaviors using visual analytics.

TABLE 1: List of the current filters implemented in uxSense.

Name	Filter Description	Data Stream	Model	Justification
VisTA	Speech transcription	Text, speech rate	[18], [53]	Transcripts are standard for usability analysis; user speech tends to slow down when users encounter a usability issue.
CoUX	Audio	Pitch	[3]	User speech tends to change pitch when users encounter a usability issue.
VideoPose3D	Track the user's position	Continuous vectors	[54]	User position reflects embodiment and intention for MR.
E-Divisive	Joint angle-based intervals	Frame intervals	[55], [56]	User pose reflects embodiment and intention for MR.
Kinetics-I3D	Semantic action classification	Action probabilities	[57], [58]	User actions reflect embodiment and intention for MR.
face-classification	Real-time emotion classification	Emotion probabilities	[59]	User mood often reflects their experience using a tool.

3 Extracting Behavior from Video

Our framework is based on visualizing time-stamped feature streams extracted from audio and video using entirely automatic methods from computer vision, machine learning, and signal processing. This semantic information is presented to the UX researcher in a visual timeline interface that supports user session video evaluation, manual annotation, and report generation. The purpose for our framework is to enable an analyst to triangulate multiple data sources with their own knowledge and experience to understand UX issues in a more accurate and efficient manner. We discuss our design constraints, data model, and applications below.

3.1 Method: Requirements Analysis

User experience is a data-driven discipline based on quantifiable and measurable data [60]. While roles in UX are ill-defined, we can distinguish between two main types: *UX researchers*, who take a proactive role in understanding *what* users want, and *UX designers*, who are more concerned about *how* to implement the product. Our goal is to design a visual analytics tool that serves both roles.

In curating our design requirements and data features below, we draw on both industrial and academic resources documenting UX practices as well as the needs of UX designers and researchers:

- **Industry UX:** Albert and Tullis [60] detail the quantitative metrics that UX research and design needs for effective analysis. Nunnally and Farkas [61] also discuss some of the more qualitative metrics involved. Sauro and Lewis [62] review the formal statistical and mathematical methods commonly used
- **Academic UX:** Section 2.3 gives an overview of current academic tools for UX R&D. In deriving our requirements, we draw upon our past work on supporting think-aloud sessions [18], [53], usability testing [63], and UX research [64].

While we would argue that the requirements listed below are the central one to quantitative and qualitative UX analysis, there are many additional requirements and criteria in the UX discipline. Exhaustively listing these is beyond the scope of this paper.

3.2 Design Requirements

Supporting UX research means keeping UX research workflows central. We propose the following interrelated requirements, each reflecting an important step in the UX research workflows:

- **R1 - Semantic key point detection:** UX research workflows frequently involve identifying semantically significant or pivotal moments in user sessions upon reviewing session data. As such, the UX system should reduce the time cost of identifying important moments in user sessions.
- **R2 - User-defined segment classification:** Identifying key points (R1) is often followed by classifying or tagging segments of the user session based on recurring or novel patterns in the data. The system should support constructing qualitative classification frameworks for session analysis.
- **R3 - Annotation:** Once segments have been identified (R1) and tagged (R2), they will be annotated by the analyst.
- **R4 - Summary Report Generation:** Finally, these identified (R1), tagged (R2), and annotated (R3) data streams should be summarized by the system as reports and figures.

3.3 Data Features

We model the following features from user data:

- **F1 - Multiple concurrent streams:** UX experiments often include multiple metrics measured over time, such as head position, physical location, activity level, speech, mood, etc.
- **F2 - Facial expressions:** The ability to track facial expression may yield an insight into the user's emotional state.
- **F3 - Speech:** User session video typically includes audio that captures dialogue between the researcher and the participant, as well as think-aloud or pair analytics procedures that involve the participant verbally externalizing their sensemaking process. An evaluation system should take advantage of this information by generating a transcript from speech, visually representing features of the audio signal, and using it in computer vision models for predicting activity.

3.4 Data Model and Filters

Due to the computational time costs of predicting semantic features of video data, we anticipate a data model consisting of uploaded video files as input, with server-side computation processing occurring asynchronously over a brief period of time before model output is accessible to the client. However, the design ideal would be in minimizing the latency between video input to model output. For this reason, we have opted to use real-time implementations (e.g., using a real-time emotion classification framework [59]) where doing so would not heavily compromise the accuracy of our output. Table 1 provides a cross-reference of the models (or "filters") involved in our present pipeline.

Model output from all filters is represented as data streams in synchronous timelines that can be flexibly manipulated. These data streams can be linked to annotations in what may be called a "human-in-the-loop" stage of the pipeline. Following this human-in-the-loop evaluation, the final output of the pipeline is generated as report figures. The transcript also appears again at the final stage of the process as part of the micro-document output.

3.5 Practical Considerations

In addition to the above design requirements, it is also possible to combine our sensing mechanisms with data from clickstreams, IR

motion tracking, and sociometric badges. The parallel timelines in uxSense can accommodate any form of data.

The facial emotion classification network, which relies on finer features of the video subject's face, is calculated concurrently and asynchronously over fixed-width intervals using a rolling window. Because of this reliance on finer features, we use 30 FPS video input for emotion classification. However, this does not greatly affect model performance, since we have chosen a real-time model [59] for deriving the emotion labels.

Concurrently, the audio from the video input is used to derive transcripts, speech rate, and pitch [18]. It can be used for captions.

Finally, more specialized filters can be added as new plugins depending on context. For example, it might be useful to track the 2D mouse pointer for desktop applications, whereas a Virtual Reality application could benefit from capturing the user's physical location, pose, and limb movements in 3D (see Section 8.1).

4 The uxSense System

uxSense is a client/server system that implements our framework for extracting user behavior from video (Section 3) with a computational and storage backend and a web-based interactive frontend. In broad terms, uxSense supports analysis of both qualitative and quantitative user experience through a variety of temporal visualization techniques for continuous (time series) and discrete (event) timelines representing *feature extraction filters* (or just filters)—the products of our models (Table 1)—to highlight segments of interest. Because many video analytics algorithms require significant processing to complete, uxSense is based on an asynchronous steering workflow where the user can shut down the interface while the video is processed on the server. Results are streamed back to the client as it is made available. The tool provides an interactive visual analysis interface for viewing multiple structured data streams, annotating them, and generating reports.

4.1 Overall Workflow

The main uxSense workflow is entirely asynchronous; any step of the following (all conducted using the web-based interface) process can be visited at any point (Figure 5):

(1) **Upload** one (or more) video or audio recordings to the server.
(2) **Start** asynchronous computation of the available list of filters.
(3) **Monitor** computation progress in real time, or
(4) **Close down** the interface while computation continues.
(5) **Analyze** the data as it becomes available.
(6) **Generate** reports from data analysis.

Each step is associated with a specific project. The backend constantly runs computations while there are still recordings to process and filters that have not been executed. Uploading new footage will schedule new computation for the unprocessed video. Results are streamed dynamically to the analysis interface.

4.2 Feature Extraction Filters

The basic building block of uxSense is the *feature extraction filter*, an algorithm that is capable of processing sequential video data v_i and generating a corresponding sequence of extracted features d_i (or *data streams*), e.g. $f_{filter} : (v_1, \ldots, v_t) \rightarrow (d_1, \ldots, d_t)$. Video frames consist of both imagery and audio, and extracted features are derived from any of these (or both). For example, a typical feature extraction filter may track the position of a person in the frame over time, the direction of their gaze, and their perceived fatigue. Some features are continuous, such as the user's head direction, whereas others are discrete, such as time intervals when a person is pointing or forming another gesture. In addition to the sequential frame data, filters also maintain summary and aggregate data relevant to the tracked feature, such as the user's cumulative movement, their activity level, individual gestures, etc.

Our prototype uxSense implementation provides an initial library of feature extraction filters (see Table 1). Many practical computer vision models track multiple features at once, such as the user's pose and a semantic label overlaid on top of the video playback. However, the uxSense design philosophy is to provide feature extraction filters such that the user can rearrange, hide, and reveal streams based on what they deem most important, relevant, or revealing for their analysis. This facilitates a more semantically meaningful configuration of which filters to include based on the research question and data being evaluated.

4.3 Analysis Interface

The uxSense analysis interface provides a mechanism for viewing and comparing feature data streams for one or multiple video recordings in parallel and time-synchronized *tracks* akin to video editing software. Figure 1 shows a schematic overview of this interface. The three main elements of the analysis interface are the footage view, the text view, and the timelines (or tracks). A common *time indicator* on the *track* governs which frame of the current footage is being viewed. The *footage view* shows that frame from the currently selected recording. The *text view* displays timestamped textual information about the video from the transcript and from the user's own annotations, and can be clicked to navigate to different points of interest throughout the video.

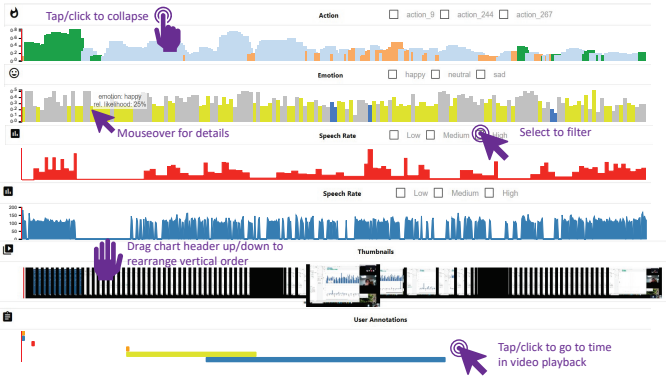


Fig. 2: **Timeline View.** This view acts as both a video scrubber to select current time and an interactive representation of the filter output. Mousing over an element shows text details of the timeline at the current frame. Filtering using the checkboxes on the header highlights all observations meeting the filter criteria by making all other observations semi-transparent. Clicking on the timeline navigates the video to the selected timestamp.

Each of the *multiple concurrent data streams* (F1) are visualized depending on data type (from top to bottom in Figure 2):

- **Action Predictions:** A plot of discrete events (using hues) based on action labels assigned to video segments over time, with prediction probability represented with rectangle height. Detecting actions support *semantic key point detection* (R1).
- **Emotion Prediction**: *Facial expression* (F2) emotion classification labels and prediction probabilities [59] are presented, also for supporting *semantic key point detection* (R1).

- **Speech Rate**: *Speech* (F3) is calculated using the word frequency over fixed time intervals using speech-to-text model output [18]; we argue that this, too, supports a level of *semantic key point identification* (R1) for the user session.
- **Pitch**: Audio signal pitch (see Section 4.5).
- **Thumbnails**: Thumbnails with the moused-over video timestamp frame shown in relief larger than the others, which are dynamically repositioned using Cartesian fisheye distortion.
- **Annotations**: In support of a combination of *semantic key point detection* (R1), *user-defined segment classification* (R2), and *annotation* (R3), the user's own annotations are visualized as a step function, with step colors signifying the annotation's timeline, and mouseover details showing the annotation and name of the corresponding timeline.

Beyond the time-marker based track view of each feature data stream, the user can zoom in on a point of interest by brushing the focus interval selection (Figure 3). Once zoomed, the timelines can be dragged to pan through the video and all of the timelines. To support *annotation* (R3)—and *user-defined segment classification* (R2) by way of *annotation* (R3)—the brushed interval of the video can be annotated as a range, or the user can opt to annotate a single point of the video as they code user behavior.

Fig. 3: **Focus-brushing.** This feature selects an interval of the video, zooms all timelines, and allows the user to drag either the timelines or the selected rectangle to pan through the data.

4.4 Micro-Report Generation with uxSense

The final destination of uxSense—the concluding step in our workflow (Section 4.1)—is to generate figures that can be used to report on user behavior in an academic paper, internal memo, or industry whitepaper. The report generation functionality of uxSense creates a small vector graphic document for each individual annotation created by the user that we have named "annotlettes" (Figure 4) that link the relevant timeline, transcript, and user annotation for the notes the user has created in the analysis. The annotlette feature was added after we completed our user study based on participant feedback and our observations (see Section 7) evaluated through the lens of our design requirements; it directly links *annotation* (R3) with *summary report generation* (R4).

- **Timeline Chunk:** A zoomed view of the annotated timeline (Figure 4d) is used to represent the data stream segment; and
- **Transcript Snippet:** A static version of the transcript (Figure 4b) during the selected time period; and
- **Annotation:** The user's annotation text, with the metadata associated with it formatted as a header (Figure 4a&c).

4.5 Implementation Notes

We implemented uxSense as a Node.js[5] application with several of the server-side components implemented in Python and R that are spawned from the Node.js process. Individual filters used

5. https://nodejs.org/

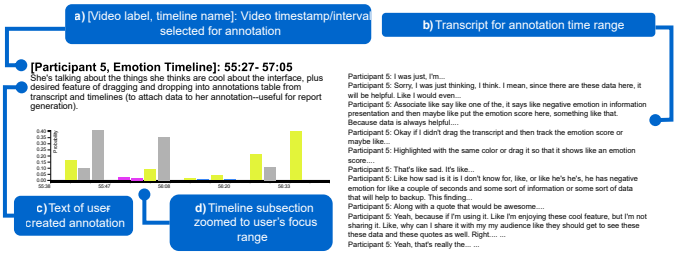


Fig. 4: **Annotlette example.** An annotlette generated by uxSense during our own evaluation of user sessions in which UX professionals used uxSense. **a)** Metadata regarding the annotation. **b)** Transcript from the period selected for annotation. **c)** The annotation text created by the user. **d)** A zoomed view of the timeline associated with the annotation for the selected period.

a combination of freely available model implementations; see Table 1 for details. For audio analysis, we used Praat[6] to extract audio features. We also use the Google Cloud Speech-to-Text API[7] to transcribe audio to text. The web client was implemented in JavaScript, HTML5, and CSS. We used HTML5 video and canvas for video playback, and D3.js [65] for the visualization components. Transcription is implemented using videojs-transcript.[8]

Since some models provide multiple features in the same computation, we used a server-side caching scheme where different filters that rely on the same model look up previously computed data instead of rerunning the same analysis from scratch. For example, a head-tracking filter that uses a full-body 3D tracking model to merely extract the user's gaze direction would store the recovered full 3D skeleton of the user in a local file. If another filter was introduced that relied on the same model (e.g., determining the user's position in 3D space), that filter could merely look up the previously stored data instead of rerunning the same model.

Our framework is Open Source and can be accessed on GitHub at https://github.com/DreaJulca/uxsense. Furthermore, as we are keen to make this technology available to other researchers working in this area, we also distribute prebuilt software packages on the uxSense GitHub website to facilitate dissemination.

5 Expert UX Designer Review

Our evaluation involved several professional UX researchers from large tech companies in the United States. The study process was not only a means for evaluating our participants' responses, but was also an opportunity for us to use uxSense itself to analyze our users' evaluation of user session footage using uxSense. Figure 5 shows the study workflow. Due to the limited availability of our participants, we chose to have participants use uxSense for an in-depth case study rather than conduct a comparative analysis between uxSense and alternative tools or baselines.

Prior to running actual tasks, we piloted the study with a single HCI master's student who was not familiar with the system. This pilot session enabled us to identify and troubleshoot issues with the software, as well as refine the testing protocol.

6. http://www.praat.org/
7. https://cloud.google.com/speech-to-text
8. https://github.com/walsh9/videojs-transcript

Fig. 5: **Evaluation pipeline.** Workflow in the UXSENSE prototype system for extracting user behavior from video footage using deep learning to support in-depth and advanced analysis of participant performance in user studies. We use uxSense to evaluate user sessions in which professional UX researchers use uxSense to evaluate a sample Tableau user session. **a)** sample user session with commercial visual analytics tool (Tableau Public); **b)** sample user session video and audio streams; **c)** server compute of sample user session data: Video pose-estimate-based temporal segmentation, video emotion and action classification, speech detection and audio signal processing; **d)** evaluation user sessions with professional UX researchers using uxSense interface with sample session video, model output; **e)** UX researcher user session video and audio streams; **f)** server compute of video and audio models using professional UX researcher session data; **g)** authors' evaluation of professional UX researcher user sessions using uxSense.

5.1 Participants

All participants were women who held the job title "UX Designer;" one was a "senior UX designer." Two participants had 5 years of professional experience, two had 2-3 years of professional experience, and one had 0.5 years of professional experience (in addition to relevant graduate school experience). Two participants held Ph.D.s in Information Science, one held a Ph.D. in HCI, one participant held a Master's degree in Product Innovation, and one participant held a Bachelor's degree in Media and Advertising.

When asked about their typical evaluation process, participants responded with the following descriptions:

- *"I normally video/audio record the session, take notes during the session and sometime collect data about the tasks we ask people to do during the session. User sessions are typically semi-structured with some tasks and open feedback. I normally code the data based on themes in spreadsheets and also create video clips that demonstrate the key themes."*
- *"There are different UX research methods and they are conducted differently. The most common ones are usability test, interview, survey. In usability test, I mark the success of each task, quantify some of the useful actions, such as user errors, user habits, user preferences and quotes. I write down their pain points and extract themes from there."*
- *"[I]nterview, survey, observation, usability testing."*
- *"[B]ehavioral and perception data [...] or self-reported response about their workflow, interview, survey, diary study."*

Participants reported frequently using the following categories of tools for evaluating user data:

- **Video conferencing**: Hangouts, Zoom, and GoToMeeting.
- **Spreadsheets**: Microsoft Excel and Google Sheets.
- **Programming languages and extensions** for working with quantitative data and for NLP: R, Python, and internal tools.
- **Survey tools**: Validately, Qualtrics, and QuickTimer.

5.2 Apparatus

uxSense was hosted on an institutional server running Linux (Ubuntu version 16.04.1) with an Intel Xeon CPU E5-1650 v3 (3.50GHz). Because our system's backend pipeline involves a processing time longer than real-time and depends on GPU support for efficient model output, the model output data accessed by our users was pre-generated prior to the beginning of their session on a laptop running Windows version 10.0.18362 with an Intel Core i7-8750H CPU (2.2GHz) and a Nvidia GeForce GTX 1060 GPU (8GB RAM). Users were able to access the system remotely.

User interactions with the system and video playback activities were asynchronously posted to the server. Session video and audio activity was recorded using Zoom's Enterprise cloud service, which also generated transcripts of the user sessions that were evaluated both outside and within uxSense. Video of the users' faces and audio of their voices during the session was recorded using their own webcams and microphones. The model output for the video and data collected during the session was generated using the same laptop that was used to pre-generate data for the user sessions.

5.3 Tasks and Procedures

Participant activity involved a user session 60-75 minutes long followed by a Google Forms survey that they were asked to fill out at their earliest convenience. At the end of the session, participants were paid US$50/hour for their time, with the post-session survey being compensated for in advance as a half-hour's worth of work.

Participants were asked to perform an open-ended exploration of the uxSense interface features in a pre-activity training phase. Once they were done with their exploration, the researcher informed the participant of any features not discovered during exploration, and answered their questions about the system. During the second stage of the user session, the participants were tasked with identifying problems with the user's experience in an 11.67 minute-long video of a novice Tableau user exploring a dataset they hadn't seen before. They were told that they would need to give a brief summary of their observations at the end of the session. After this activity ended, participants were paid via Venmo and emailed a link to the post-session survey. All participants who completed the session also completed the post-session survey.

Participants were asked to think aloud during all stages of the session. After the session activities, they were asked about their

general thoughts, and they summarized their experience and offered suggestions on how the interface may be modified to improve the user experience. Finally, after the session, they were asked to complete a post-test survey with both Likert scale questions and open-ended text response questions. All participants completed this survey less than 1 week after their session.

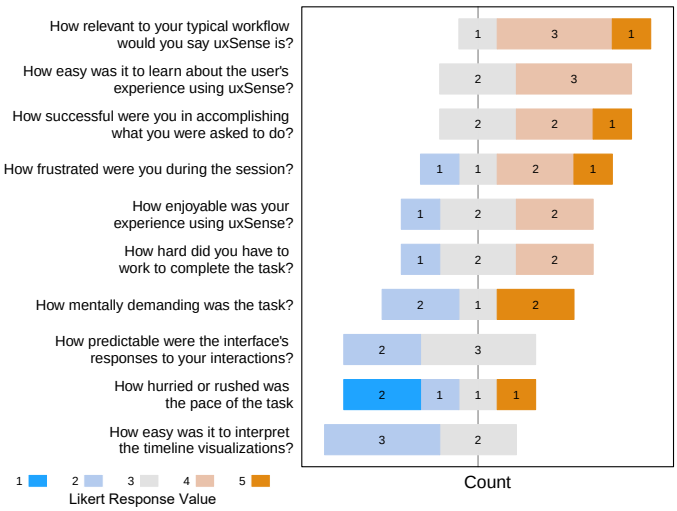


Fig. 6: **Likert responses** to user experience survey (1 = strongly disagree, 5 = strongly agree).

6 Expert Review Results

We originally recruited six UX professionals to participate in our study. However, the third session failed due to incompatibility problems with the participant's browser, as well as unexpected latency due to the geographical distance between the server and the client coupled with large file size and poor compression. For this reason, that participant was unable to complete their session. We revised the software based on these experiences and successfully conducted the study with the five participants reported below (numbered 1-6, omitting participant 3).

6.1 Think-Aloud Transcripts

We asked participants to follow a think-aloud protocol during their session. We used Zoom to automatically transcribe participant utterances, analyzed them, and report our findings below.

In general participants felt that taking notes and marking their corresponding time at the same is demanding. However, the multiple timeline interface can relatively reduce some effort. For example, participants used the annotation timeline to help them keep track of and organize the notes that they could refer to during their analysis. Also, they used the emotion timeline to better understand the user's behavior; for example, P4 commented that *"I found that those positive emotions are related to excitement that he experienced when he [the pictured user] found something really interesting."* Although participants found the user's pitch information was indicative of excitement, they felt that the current visualization of pitch information did not help them quickly spot the moments of unusual pitches. Moreover, participants felt that having access to multiple timelines during their first pass of analysis was overwhelming, because there was too much information to attend to and they wanted to focus on the video. Instead, P2 felt the timelines might *"be more helpful in the second round of analysis."* Lastly, participants hoped to be able to configure the layout of different panels, so that they could temporarily hide panels that are non-essential to their analysis at hand. They also requested a synchronized video transcript view with the timelines.

6.2 User Experience Survey

Expert responses to the Likert-scale user experience survey questions described are shown in Figure 6. Participants also gave open-ended responses to survey questions; their responses are reported in Table 2. Participants generally found uxSense to be relevant to their typical workflow and allowed them to easily learn about the user's experience, sometimes revealing points of interest in the video that they may have otherwise missed. On the other hand, they also found that it sometimes responded in unexpected ways and made them feel frustrated, and that the model output timelines could be difficult to interpret.

TABLE 2: Summary of open-ended survey responses.

Feature	Summary of Participant Responses in Open-Ended Survey Question
Video	• Video was reported as the primary focus for all sessions.
Transcript	• Transcript should be exportable. • Two users found transcript to be most important (after video).
Annotations	• Three participants recommended that the annotations table be exportable. • Time constraints inhibited use of the annotation feature; a typical evaluation of 10 minutes of user session video takes >30 minutes.
Timelines (All)	• Difficult to divert attention to timelines on first watchthrough of video. Two participants said that they would look at it more on a second pass of video. • Timeline brushing to focus and zoom almost went overlooked, and panning with a zoomed interval was confusing at first.
Timelines (Action & Emotion)	• Emotion and Action timelines were deemed most relevant to the evaluation workflow by three participants. • One participant reported a lack of trust in the action and emotion model output, while two participants reported a sense of trust in the model output. Two participants said they could see clear links between emotion labels assigned by the model and their observations in the video and transcript.
Timelines (Pitch & Speechrate)	• One participant found pitch timeline informative when used in conjunction with emotion timeline, but general feedback was that speech rate and pitch were marginally redundant and not relevant to their workflow.
Timelines (Annotations)	• Annotation timeline was deemed helpful to workflow once understood, but was described as confusing at first.
General/ Multiple-Feature Feedback	• Two participants would have liked a stronger link between transcript and annotations, either via interaction for embedding transcript lines in annotations table or vice-versa. • Explicit labeling of interface features and data variable descriptions was suggested by all participants. • In light of video being focus of sessions, the increased cognitive load and information overload of viewing many features at once was commented upon by two participants. • One participant suggested "[adding] a free-form note taking section for researcher to take initial notes and maybe allow them to organize them when do more post analyses."

Their feedback in response to our open-ended survey questions (Table 2) generally suggested design changes that were more in line with universal design guidelines in layout and interaction (e.g., mindfulness of information overload, improved visual feedback to user content creation, and better descriptive labeling of features). Video playback speed control (0.5x speed, 1.5x speed, 2x speed), 10-second skip-forward/skip-back buttons, and video view resizing were the most commonly requested features, as the video was reported to be the central focus for the participants. Most participants also strongly desired a transcript or annotation table export feature, and several participants saw value in visually linking the annotations with the transcripts and/or the selected point or interval on the timelines by embedding two or all three in either the transcript or the annotations table. It was the combination of these final two points in their feedback that motivated us in our design of the post-study implementation of the "annotlettes" feature shown in Sections 4.4 and 6.4—a surprise outcome of our study.

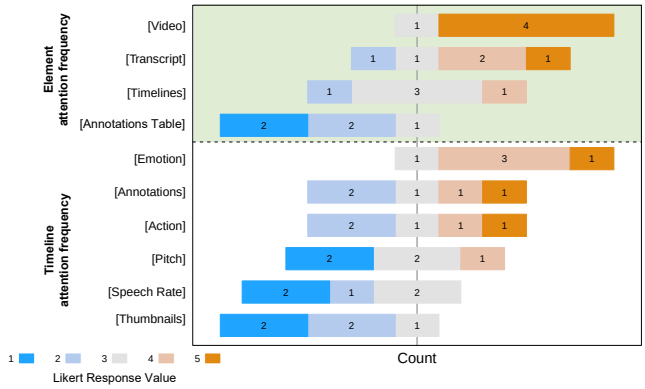


Fig. 7: **Likert responses to time-attention questions.** Users were asked to rank time spent on each feature relative to the others (1 = much less time, 5 = much more time).

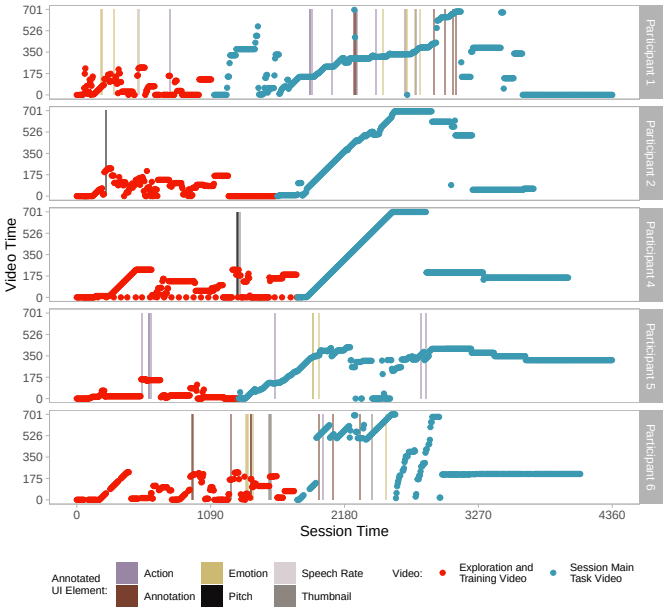


Fig. 8: **Activity logs.** Participant activity in video playbacks.

6.3 Time Use: Observed and Self-Reported

We analyzed event logs, survey responses, and qualitative feedback from user sessions to create a descriptive flow report of their interactions using uxSense. Given the participants' heavy use of the video playback feature, we studied their navigation of the video in detail (Figure 8). All participants spent a similar share of their time in undirected exploration of features of the interface (shown in red) relative to time spent in the UX evaluation stage of the session. However, their navigation of the video shows very different viewing patterns during evaluation. Participants 2 and 4 opted to watch the video the whole way through in a largely linear way before looping back to a small number of points of interest that they spent longer stretches of time on. These participants (2 and 4) also scarcely use the annotation feature during the training video, and do not use it at all during the main task. Participants 1 and 6 start off by skipping around the video in a way that generally progresses forward before looping back and taking a second pass that involves frequent short-range backtracking, and then skip to a small handful of moments of interest. These participants (1 and 6) also made the heaviest use of the annotation feature, particularly after their initial skipping around but before their final pass. Participant 5 performs a lot of short-range backtracking on her first pass, then has a period of frequent skipping back and forth between several intervals in the video that she found most informative before spending longer spans of time examining key points in the video. Participant 5 makes moderate use of the annotation feature throughout both videos.

Participants were asked to assess where their own attention fell most frequently during their user study as part of the post-session survey (Figure 7). As we saw in Section 6.2, nearly all participants reported spending most of their time examining the video playback, and most reported that the transcript was nearly as central to their evaluation process using uxSense.

In order to add depth to our inferences about the participants' use of time during the sessions, we asked them about their own assessment of how and why they distributed their attention during the session. Participants' self-reported attention time use was recorded on a Likert scale (Figure 7).

6.4 uxSense in Three Vignettes

Based on feedback gathered during the user sessions, the authors extended uxSense to produce annotlettes (as described in Section 4.4). After the sessions in which expert UX researchers used uxSense to evaluate a sample user session, we used uxSense to evaluate our sessions with them, and then generated annotlettes with our own annotations. uxSense is intended for more detailed reporting than can be contained in a single section, so we demonstrate this feature in three representative vignettes from our evaluation. While we do not claim that these vignettes are exhaustive, we do think they are representative of participant experiences during our study.

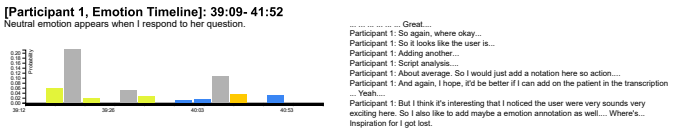


Fig. 9: Researcher made an annotation with the observation the appearance of a high-confidence prediction of a neutral facial expression when they responded to the participant's question.

WHILE exploring the video of Participant 1's user session (video duration: 1 hour and 28 minutes) in uxSense, the researcher notices that there were a small number of high-probability "neutral" facial expression predictions from the model output that appear as peaks in the emotion timeline (the emotion timeline was directly used by all participants during their user sessions). The researcher watches the video from start to until she reaches the first of these peaks, after which point she skips forward directly to the next peak (navigating to a potential point of interest in the video using the emotions timeline was an interaction performed by all participants). Using uxSense's focus+context interval selection feature, she adds an interval note (Figure 9) highlighting what she observes about the first high-confidence 'neutral' label (all participants added at least one interval annotation during their session). Then,

navigating to the next high-probability neutral expression in an entirely different phase of the study, she notices that there is another question-answering dialogue between the researcher and the participant about the system, and adds an annotation for that time interval of video as well (Figure 10). After generating uxSense annotlettes, she makes the observation that high-probability neutral expressions tend to follow a low-probability happy expression label (P1 and P6 made interval annotations that explicitly linked patterns in the emotion labels with user behavior). Using this pattern observation, she is able to quickly identify other instances of question-answering during the session.

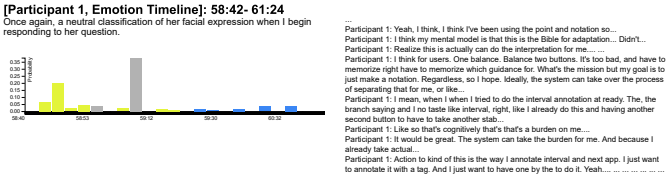


Fig. 10: The researcher makes another annotation with the observation of an association between researcher question-answering and the appearance of a high-probability neutral expression.

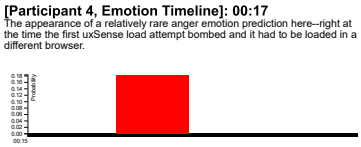


Fig. 11: The first appearance of an "angry" emotion classification label; here it is associated with the inconvenience of an unanticipated system error that had yet to be worked out. There is no associated transcription because there was no dialogue at this specific point in time.

AFTER working her way through user session videos until she reaches Participant 4's session (video duration: 1 hour and 0 minutes), the researcher notices the appearance of the rare "angry" emotion label classification twice (all participants made verbal or annotated observations about the emotions timeline). Because the anger emotion label has not appeared in her data up to this point, she navigates to these times in the video immediately upon observing them to review the participant's behavior. She uses the point annotation feature to make a note of what happens during the session associated with the emotion timeline (P1, P5, and P6 made annotations pointing out events in the video and the emotion at the time of the event; in contrast with interval annotations, point annotations tended to appear more frequently and be slightly more detailed than interval annotations). The bright red emotion indicator appeared when the user ran into a compatibility issue with the browser she was using, and the system failed to load properly (Figure 11). The researcher seeks out another appearance of the uncommon emotion, again in the same participant's session; this time, it corresponds to an equally uncommon evaluation of a user session: Disapproval toward the user's analytical process (all participants had at least one point during their sessions in which their attention was drawn to labels specifically because they appeared less frequently in the timeline, i.e., to outliers). She creates another point note describing her observation (Figure 12).

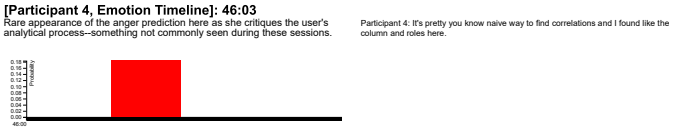


Fig. 12: Another rare "angry" emotion classification label, again within the same participant's timeline. This time it is associated with the equally rare criticism of the user's analytical choice.

MOVING on with her evaluation of the user sessions to Participant 6's session (video duration: 1 hour and 1 minute), the researcher sees that the appearance of action_290 is something of an outlier in the actions timeline. Going directly to that point in the timeline reveals that the participant has said "I wish I could see this more clearly, because right now it's really small" in reference to the video playback viewer. The researcher makes a note of it (P1, P5, and P6 made annotations based on the semantic action label timeline) without looking too closely at the transcript (Figure 13). This particular action is uncommon in this participant's action timeline, so she navigates to its next appearance. Once again, she sees that the participant is expressing her frustration with the small size of the video playback viewer, and she makes a note of it (Figure 14). Satisfied for the moment, she generates annotlettes for the annotations she has made thus far, and inserts them as figures in her report. Upon inspection of the annotlettes she has created, the researcher is now able to see the semantically-loaded action labels that she tagged during her first pass at evaluating her video dataset. She finds that action_290 is "shaking head;" she amends her annotation with this new information. She also sees upon closer review that the actions timeline picked out something that she may have otherwise missed: In the interval selected in the first annotation (Figure 13), the participant speaks quietly enough that the transcript failed to capture her complaint about the video playback viewer size. By using the semantic label visualization, she was able to identify a pattern that would have been missed in an analysis of the transcript alone (P1 and P5 used the actions timeline to support identifying behavior that they did not immediately notice in the transcript).

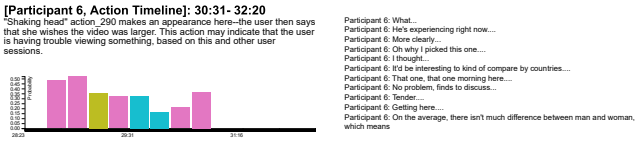


Fig. 13: The researcher observes the appearance of action_290 during a moment in the session in which the participant is expressing frustration at being unable to resize the video playback viewer, but the transcript has not captured her sentiments.

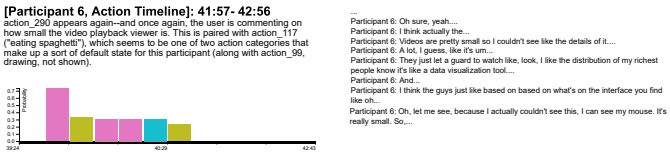


Fig. 14: The researcher observes another appearance of action_290 capturing the participant's complaint about video size, the same design concern featured in Figure 13.

7 Proposed System Redesign

Based on our observations and UX expert feedback, we identified some critical changes to the uxSense interface.

A1 **Annotlettes and data export:** Several users suggested that features that were either generated by the system's backend models or added by the UX researcher, such as the transcript and researcher annotations, should be exportable. These requests, coupled with our review of users interactions with their own annotations, resulted in our development of annotlettes (in addition to traditional data export utilities, which we have also added). As noted in Section 6.2, the annotlettes feature was one of our post-study additions to uxSense. Participants desired a stronger visual link between their annotations, the video, and the data, which annotlettes provide.

A2 **Video playback features:** All participants wished to be able to speed up, slow down, or resize their view of the video playback. For all participants, review of the video was central to their analysis, and there were times during which the screen space taken up by other features of uxSense was unnecessary, even if they wound up using a selection of those features at various times throughout their sessions. A redesigned system would allow users to hide some or all of the non-video features of the view, in addition to the changes they requested directly.

A3 **Affordance signifiers:** Some participants found features of the system to not be implied by the interface design. In response, we identify several signifiers to uxSense's to make important and easily-overlooked affordances visible:

- **Icons** identifying several non-intuitive interactions with the interface, such as the interval selection feature on the zoom focus, and drag directional arrows on the timelines;
- **Tooltips** indicating what the controls in the interface do;
- **Pre-selected interval range on initialization** that highlights the fact that intervals can be selected; and
- **A single annotation button** that is clearly labeled.

A4 **Session comparison:** The patterns identified in or across user studies are often based on recurring themes or novel outliers that occur amongst, or in contrast across, multiple users. The UX system should support qualitative pattern identification across multiple user sessions by allowing the researcher to compare them within and across studies. Furthermore, since researchers may be looking for specific patterns, it seemed appropriate to allow them to use their annotations to tag specific segments of video with codes or labels that they create for comparing only those segments across different sessions. While this was a design consideration that we discussed prior to our user sessions, it was ultimately not included prior to our study out of respect for the time constraints of our participants and the amount of their valuable time that comparing multiple UX sessions would demand.

8 Discussion

The use of visualization for analyzing video recording to evaluate human behavior presents opportunities for progress in designing environments that take better advantage of features of the user, their surroundings, or their session, and respond accordingly. However, there are ethical considerations and limitations in using video data.

While our system may have faced a few design and performance issues during the expert user study, the results were largely affirming of the relevance of, interest in, and frankly, need for ML-supported HCI. Equally importantly, our participants directly and indirectly highlighted improvements that we have implemented in our redesign of the system described in Section 7.

Perhaps our favorite result from the study was in finding a way to not only use the system to evaluate itself, but in using that self-evaluation to extend uxSense so that it could aid in writing narrative vignettes that highlighted important moments for further design revisions. This represents, to us, the unlimited potential for growth and improvement in ML-HCI-VIS evaluation systems.

8.1 Beyond Desktop UX

While we so far have explored the use of uxSense for desktop computer applications, it is clear that the same ideas can be applied to other forms of computing, such as immersive 3D and mobile computing. In fact, one of our original motivations for uxSense was to support 3D immersive analytics evaluation. This would, for example, enable extending tools such as ReLive [66] to include automatically derived metrics. It would require adding the following features to the data model:

F4 - Physical navigation: User physical location over time.
F5 - Limb tracking: The ability to track both fine (fingers and hands) and gross (arms, legs, torso, head) motor interaction.

Several of our filters (Table 1) already support such data, including for detecting 3D pose [54], 3D joint angles [55], and actions [56]. However, we leave such extensions to future work.

8.2 Limitations

There are several limitations with the technical approach we propose in this paper. Even state-of-the-art computer vision and machine learning algorithms are still far from perfect in accurately detecting the features that uxSense visualizes, and thus there is significant noise in the process. The fact that our participants mostly relied on the raw video footage rather than the extracted feature timelines during our expert review may be an indication that the noise was too overwhelming to rely upon. In a sense, this could be seen as a negative result: perhaps automated or even semi-automated UX analysis is doomed to failure?

We offer several answers to this question. On the human side of the equation, it is possible that this bias in favor of raw video is merely a habitual effect arising from our participants' everyday practices as professional UX designers. It could also be due to the limited amount of time that UX evaluators were exposed to the tool, as suggested by Fan et al. who observed similar behaviors of UX evaluators in the evaluation of their UX video analytics tool [18]. Thus, we suspect that this effect will probably decrease as UX evaluators embrace this kind of automated tool in their daily workflow over time. On the technology side, the noise in our

feature streams will likely reduce over time as computer vision and machine learning technologies improve. Furthermore, the problem can be somewhat mitigated by offering several complementary data streams that serve to provide an accurate overview despite inaccuracies in individual streams. However, there is nothing to prevent a user from importing manually annotated data into uxSense as a complement. Furthermore, uxSense also gives the user access to the original video and audio data, thus allowing them to verify (and correct) any automatic annotation in the interface.

The other side of this coin is that automatic metrics may lead to overreliance, anchoring effects, and even a decay in analytical reasoning. Furthermore, the lack of transparency for many of the computer vision models may certainly impact the user's trust in the system. Involving the human in the loop and triangulating multiple metrics may mitigate these dangers; however, it remains a cautionary tale when employing automatic methods. Alternatively, recent research showed that when and how automatic metrics are revealed to UX evaluators can affect their trust and performance [67]. Thus, another possible direction is to investigate human-AI collaborative approaches.

In terms of limitations of our research methodology, our evaluation was done with a very small set of professional UX researchers from tech companies. We only involved a single sample user study session video for the evaluation activity. Furthermore, the qualitative nature of our study provides few quantitative measurements on performance and the lack of a baseline makes comparison to other tools impossible. We made these choices with the intention of collecting actionable feedback for improving uxSense further rather than summatively assessing its utility. However, we acknowledge that our single evaluation is not generally representative of the many potential variations of products and end-users, and that further studies are needed in the future.

Part of our evaluation methodology is based on using uxSense itself to evaluate uxSense. There is obviously a risk in this practice because missing features in uxSense would lead to these aspects of the system not being evaluated; an intrinsic "blind spot" in our methodology. We have mitigated this risk by triangulating the findings with our own experience as visualization and UX researchers, akin to how we envision UX designers will use uxSense in practice.

Finally, in choosing the commercial data analysis and visualization tool Tableau as the platform of choice for our expert review, we restricted the usability study session used as a dataset to a mature software package with a polished user interface. This limited the scope and scale of usability issues that our UX designer participants could find in the user study data. In the future, it would be interesting to perform a follow-up study involving a more experimental and early-stage user interface, or for a more academic user study conducted in a controlled laboratory setting.

8.3 Ethical Considerations of Video

In today's surveillance society, an increasing portion of our lives takes place inside a camera viewfinder. It seems as if every week uncovers some new horror of how digital video can be abused to threaten the privacy, security, and even safety of people just trying to live their lives. Thus, it could well be argued that a system such as uxSense is misguided in that it builds on fundamentally problematic ideas about recording human participants, and could even facilitate future abuses in the same vein. Indeed, while our current prototype does not include a facial recognition component, we could easily see the utility of having such a filter for when a system is deployed in the field and the researchers want to track how specific people use the system over time.

Rather than merely disavow any future such events as beyond our control, let us here acknowledge that this is a possible outcome. We ourselves commit to safeguarding our own approach and prototype so that the recordings are not distributed or used for other purposes than for what participants gave informed consent to.

We also note that researchers already collect plenty of video recordings of their participants, much of which is only lightly analyzed and then archived. These videos will obviously identify the participants. Due to the prohibitive size of video, we suspect that much of this data resides on unprotected network or external drives. However, a fundamental feature of our approach is to process high-bandwidth video data to extract key data streams from the footage. These extracted data streams are refined and precise; the position of a person in three dimensions, their fatigue level, or the direction of their gaze. After all relevant data has been extracted, the original video can be deleted, thereby saving storage space and eliminating identifiable likenesses of participants. Thus, it could be argued that our approach may actually improve privacy, as it allows researchers to safely discard video while retaining deidentified data.

9 Conclusion and Future Work

We have proposed a visual analytics tool called uxSense for visualizing multiple timelines of data streams extracted from video and audio recordings of usability sessions. We then present results from an expert review where UX professionals used the tool to understand a usability session conducted using Tableau. Finally, we used uxSense itself to analyze these expert review sessions.

Our prototype only includes a small set of filters to demonstrate our concept, and we can easily see the need for many additional filters. Fortunately, all feature extraction filters in uxSense are Open Source modules, and we anticipate generously extending this library of filters to include other Open Source modules. Furthermore, it is common for UX professionals to analyze multiple sessions from different users to determine common issues. uxSense is currently restricted to evaluating a single user's session in a post-hoc manner. In the future, UX professionals may need to analyze user experience across multiple users, perhaps even real-time.

In fact, many interactive systems now consist of multiple, connected, and collaborating devices, such as in smart rooms, for groups of smartphones working together, or for an individual user's menagerie of personal devices. By taking advantage of this networked system of many devices, we can extend uxSense and develop new systems like it, and step a little closer to HCI's dream of context-aware and ubiquitous computing [1].

Acknowledgments

This work was partially supported by the U.S. National Science Foundation grant IIS-1908605 and the Natural Sciences and Engineering Research Council of Canada through the Discovery Grant program. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] M. Weiser, "The computer for the 21st Century," *Scientific American*, vol. 265, no. 3, pp. 94–104, Sep. 1991. [Online]. Available: https://doi.org/10.1038/scientificamerican0991-94

[2] T. Blascheck, M. John, K. Kurzhals, S. Koch, and T. Ertl, "VA2: a visual analytics approach for evaluating visual analytics applications," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 61–70, 2016. [Online]. Available: https://doi.org/10.1109/TVCG.2015.2467871

[3] E. J. Soure, E. Kuang, M. Fan, and J. Zhao, "CoUX: Collaborative visual analysis of think-aloud usability test videos for digital interfaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 643–653, 2021. [Online]. Available: https://doi.org/10.1109/TVCG.2021.3114822

[4] K. Holtzblatt and H. Beyer, *Contextual Design: Evolved*, ser. Synthesis Lectures on Human-Centered Informatics. San Rafael, CA, USA: Morgan & Claypool Publishers, 2014.

[5] K. Shilton, "This is an intervention: Foregrounding and operationalizing ethics during technology design," in *Emerging Pervasive Information and Communication Technologies: Ethical Challenges, Opportunities and Safeguards*. Dordrecht, The Netherlands: Springer, 2014, pp. 177–192. [Online]. Available: https://doi.org/10.1007/978-94-007-6833-8_9

[6] A. A. E. Ahmed and I. Traore, "A new biometric technology based on mouse dynamics," *IEEE Transactions on Dependable and Secure Computing*, vol. 4, no. 3, pp. 165–179, 2007. [Online]. Available: https://doi.org/10.1109/TDSC.2007.70207

[7] X. Zhang, H.-F. Brown, and A. Shankar, "Data-driven personas: Constructing archetypal users with clickstreams and user telemetry," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2016, pp. 5350–5359. [Online]. Available: https://doi.org/10.1109/10.1145/2858036.2858523

[8] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2008, pp. 453–456. [Online]. Available: https://doi.org/10.1145/1357054.1357127

[9] G. Halter, R. Ballester-Ripoll, B. Flueckiger, and R. Pajarola, "VIAN: A visual annotation tool for film analysis," *Computer Graphics Forum*, vol. 38, no. 3, pp. 119–129, 2019. [Online]. Available: https://doi.org/10.1111/cgf.13676

[10] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, and D. Weiskopf, "Visual movie analytics," *IEEE Transactions on Multimedia*, vol. 18, no. 11, pp. 2149–2160, Nov. 2016. [Online]. Available: https://doi.org/10.1109/TMM.2016.2614184

[11] A. Pavel, C. Reed, B. Hartmann, and M. Agrawala, "Video digests: A browsable, skimmable format for informational lecture videos," in *Proceedings of the ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2014, pp. 573–582. [Online]. Available: https://doi.org/10.1145/2642918.2647400

[12] A. Truong, F. Berthouzoz, W. Li, and M. Agrawala, "QuickCut: An interactive tool for editing narrated video," in *Proceedings of the ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2016, pp. 497–507. [Online]. Available: https://doi.org/10.1145/2984511.2984569

[13] M. Leake, H. V. Shin, J. O. Kim, and M. Agrawala, "Generating audio-visual slideshows from text articles using word concreteness," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2020, pp. 1–11. [Online]. Available: https://doi.org/10.1145/3313831.3376519

[14] H. Sloetjes and P. Wittenburg, "Annotation by category: ELAN and ISO DCR," in *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association, 2008.

[15] K. Higuchi, R. Yonetani, and Y. Sato, "EgoScanning: Quickly scanning first-person videos with egocentric elastic timelines," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2017, pp. 6536–6546. [Online]. Available: https://doi.org/10.1145/3025453.3025821

[16] A. Pavel, D. B. Goldman, B. Hartmann, and M. Agrawala, "VidCrit: Video-based asynchronous video review," in *Proceedings of the ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2016, pp. 517–528. [Online]. Available: https://doi.org/10.1145/2984511.2984552

[17] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal, "EduSense: Practical classroom sensing at scale," *Proceedings of the ACM Conference on Interactive, Mobile, Wearable, and Ubiquitous Technologies*, vol. 3, no. 3, pp. 71:1–71:26, 2019. [Online]. Available: https://doi.org/10.1145/3351229

[18] M. Fan, K. Wu, J. Zhao, Y. Li, W. Wei, and K. N. Truong, "VisTA: Integrating machine intelligence with visualization to support the investigation of think-aloud sessions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 343–352, 2020. [Online]. Available: https://doi.org/10.1109/TVCG.2019.2934797

[19] G. Marqués and K. Basterretxea, "Efficient algorithms for accelerometer-based wearable hand gesture recognition systems," in *Proceedings of the IEEE International Conference on Embedded and Ubiquitous Computing*. Piscataway, NJ, USA: IEEE, 2015, pp. 132–139. [Online]. Available: https://doi.org/10.1109/EUC.2015.25

[20] Ramakant, N.-e.-K. Shaik, and L. Veerapalli, "Sign language recognition through fusion of 5DT data glove and camera based information," in *Proceedings of the IEEE International Advance Computing Conference*. Piscataway, NJ, USA: IEEE, 2015, pp. 639–643. [Online]. Available: https://doi.org/10.1109/IADCC.2015.7154785

[21] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics*, vol. 33, no. 5, pp. 169:1–169:10, 2014. [Online]. Available: https://doi.org/10.1145/2629500

[22] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015. [Online]. Available: https://doi.org/10.1109/TMM.2015.2482819

[23] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2015, pp. 4511–4520. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7299081

[24] K. Harezlak, P. Kasprowski, and M. Stasch, "Idiosyncratic repeatability of calibration errors during eye tracker calibration," in *Proceedings of the International Conference on Human System Interactions*. Piscataway, NJ, USA: IEEE, 2014, pp. 95–100. [Online]. Available: https://doi.org/10.1109/HSI.2014.6860455

[25] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE Transactions on Cybernetics*, vol. 46, no. 4, pp. 916–929, 2016. [Online]. Available: https://doi.org/10.1109/TCYB.2015.2418092

[26] P. Suja, K. V. P. Kumar, and S. Tripathi, "Dynamic facial emotion recognition from 4D video sequences," in *Proceedings of the International Conference on Contemporary Computing*. Piscataway, NJ, USA: IEEE, 2015, pp. 348–353. [Online]. Available: https://doi.org/10.1109/IC3.2015.7346705

[27] J. van de Wolfshaar, M. F. Karaaba, and M. A. Wiering, "Deep convolutional neural networks and support vector machines for gender recognition," in *Proceedings of the IEEE Symposium Series on Computational Intelligence*. Piscataway, NJ, USA: IEEE, 2015, pp. 188–195. [Online]. Available: https://doi.org/10.1109/SSCI.2015.37

[28] S. Chandrasegaran, S. K. Badam, L. Kisselburgh, K. Peppler, N. Elmqvist, and K. Ramani, "VizScribe: A visual analytics approach to understand designer behavior," *International Journal of Human-Computer Studies*, vol. 100, pp. 66–80, 2017. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2016.12.007

[29] C. Kerdvibulvech and H. Saito, "Vision-based detection of guitar players' fingertips without markers," in *Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation*. Piscataway, NJ, USA: IEEE, 2007, pp. 419–428. [Online]. Available: https://doi.org/10.1109/CGIV.2007.88

[30] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2008, pp. 1–8. [Online]. Available: https://doi.org/10.1109/CVPR.2008.4587756

[31] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2007, pp. 1–8. [Online]. Available: https://doi.org/10.1109/CVPR.2007.383132

[32] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani, "Forecasting interactive dynamics of pedestrians with fictitious play," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2017, pp. 4636–4644. [Online]. Available: https://doi.org/10.1109/CVPR.2017.493

[33] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2016, pp. 2176–2184. [Online]. Available: https://doi.org/10.1109/CVPR.2016.239

[34] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with

recurrent 3D convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2016, pp. 4207–4215. [Online]. Available: https://doi.org/10.1109/CVPR.2016.456

[35] E. Cambria, G. Huang, L. L. C. Kasun, H. Zhou, C. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li, V. C. M. Leung, L. Feng, Y. Ong, M. Lim, A. Akusok, A. Lendasse, F. Corona, R. Nian, Y. Miche, P. Gastaldo, R. Zunino, S. Decherchi, X. Yang, K. Mao, B. Oh, J. Jeon, K. Toh, A. B. J. Teoh, J. Kim, H. Yu, Y. Chen, and J. Liu, "Extreme learning machines," *IEEE Intelligent Systems*, vol. 28, no. 6, pp. 30–59, 2013. [Online]. Available: https://doi.org/10.1109/MIS.2013.140

[36] S. Heng and D. Yunfeng, "Research on cooperative control of human-computer interaction tools with high recognition rate based on neural network," in *Proceedings of the IEEE International Conference on Virtual Reality and Visualization*. Piscataway, NJ, USA: IEEE, 2014, pp. 350–354. [Online]. Available: https://doi.org/10.1109/ICVRV.2014.6

[37] C. Felix, A. Dasgupta, and E. Bertini, "The exploratory labeling assistant: Mixed-initiative label curation with large document collections," in *Proceedings of the ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2018, pp. 153–164. [Online]. Available: https://doi.org/10.1145/3242587.3242596

[38] M. Drouhard, N. Chen, J. Suh, R. Kocielnik, V. Peña-Araya, K. Cen, and C. R. Aragon, "Aeonium: Visual analytics to support collaborative qualitative coding," in *Proceedings of the IEEE Pacific Visualization Symposium*. Piscataway, NJ, USA: IEEE, 2017, pp. 220–229. [Online]. Available: https://doi.org/10.1109/PACIFICVIS.2017.8031598

[39] A. C. Robinson and C. Weaver, "Re-visualization: Interactive visualization of the process of visual analysis," in *Workshop on Visualization, Analytics & Spatial Decision Support at the GIScience conference*. Bern, Switzerland: International Cartographic Association, 2006.

[40] D. Gotz and M. X. Zhou, "Characterizing users' visual analytic activity for insight provenance," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. Piscataway, NJ, USA: IEEE, 2008, pp. 123–130. [Online]. Available: https://doi.org/10.1109/VAST.2008.4677365

[41] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala, "Graphical histories for visualization: Supporting analysis, communication, and evaluation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1189–1196, 2008. [Online]. Available: https://doi.org/10.1109/TVCG.2008.137

[42] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "Visualization of eye tracking data: A taxonomy and survey," *Computer Graphics Forum*, vol. 36, no. 8, pp. 260–284, 2017. [Online]. Available: https://doi.org/10.1111/cgf.13079

[43] N. Silva, T. Schreck, E. Veas, V. Sabol, E. Eggeling, and D. W. Fellner, "Leveraging eye-gaze and time-series features to predict user interests and build a recommendation model for visual analysis," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*. New York, NY, USA: ACM, 2018, pp. 1–9. [Online]. Available: https://doi.org/10.1145/3204493.3204546

[44] J. Zhao, M. Glueck, S. Breslav, F. Chevalier, and A. Khan, "Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 261–270, 2017. [Online]. Available: https://doi.org/10.1109/TVCG.2016.2598543

[45] J. Zhao, M. Glueck, P. Isenberg, F. Chevalier, and A. Khan, "Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 340–350, 2017. [Online]. Available: https://doi.org/10.1109/TVCG.2017.2745279

[46] H. R. Lipford, F. Stukes, W. Dou, M. E. Hawkins, and R. Chang, "Helping users recall their reasoning process," in *IEEE Symposium on Visual Analytics Science and Technology*. Piscataway, NJ, USA: IEEE, 2010, pp. 187–194. [Online]. Available: https://doi.org/10.1109/VAST.2010.5653598

[47] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang, "Recovering reasoning processes from user interactions," *IEEE Computer Graphics and Applications*, vol. 29, no. 3, pp. 52–61, 2009. [Online]. Available: https://doi.org/10.1109/MCG.2009.49

[48] M. Fan, S. Shi, and K. N. Truong, "Practices and challenges of using think-aloud protocols in industry: An international survey," *Journal of Usability Studies*, vol. 15, no. 2, pp. 85–102, 2020.

[49] J. Staiano, M. Menéndez, A. Battocchi, A. De Angeli, and N. Sebe, "UX_Mate: from facial expressions to UX evaluation," in *Proceedings of the Designing Interactive Systems Conference*. New York, NY, USA: ACM, 2012, pp. 741–750. [Online]. Available: https://doi.org/10.1145/2317956.2318068

[50] C. T. Tan, S. Bakkes, and Y. Pisan, "Inferring player experiences using facial expressions analysis," in *Proceedings of the ACM Conference on Interactive Entertainment*. New York, NY, USA: ACM, 2014, pp. 1–8. [Online]. Available: https://doi.org/10.1145/2677758.2677765

[51] K. M. Munim, I. Islam, M. Khatun, M. M. Karim, and M. N. Islam, "Towards developing a tool for UX evaluation using facial expression," in *Proceedings of the International Conference on Electrical Information and Communication Technology*, IEEE. Piscataway, NJ, USA: IEEE, 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1109/EICT.2017.8275227

[52] R. Y. da Silva Franco, R. Santos do Amor Divino Lima, M. Paixão, C. G. Resque dos Santos, and B. Serique Meiguins, "UXmood – A sentiment analysis and information visualization tool to support the evaluation of usability and user experience," *Information*, vol. 10, no. 12, p. 366, 2019. [Online]. Available: https://doi.org/10.3390/info10120366

[53] M. Fan, Y. Li, and K. N. Truong, "Automatic detection of usability problem encounters in think-aloud sessions," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 2, may 2020. [Online]. Available: https://doi.org/10.1145/3385732

[54] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2019, pp. 7753–7762. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00794

[55] A. Batch, K. Lee, H. T. Maddali, and N. Elmqvist, "Gesture and action discovery for evaluating virtual environments with semi-supervised segmentation of telemetry records," in *Proceedings of the IEEE International Conference on Artificial Intelligence and Virtual Reality*. Piscataway, NJ, USA: IEEE, 2018. [Online]. Available: https://doi.org/10.1109/AIVR.2018.00009

[56] D. S. Matteson and N. A. James, "A nonparametric approach for multiple change point analysis of multivariate data," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 334–345, 2014. [Online]. Available: https://www.jstor.org/stable/24247158

[57] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ, USA: IEEE, 2017, pp. 4724–4733. [Online]. Available: https://doi.org/10.1109/CVPR.2017.502

[58] A. Batch and N. Elmqvist, ""All right, Mr. DeMille, I'm ready for my closeup": Adding meaning to user actions from video for immersive analytics," in *Proceedings of the Machine Learning from User Interactions (MLUI) for Visualization and Analytics Workshop at IEEE VIS*, 2019.

[59] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *Computing Research Repository (CoRR)*, vol. abs/1710.07557, 2017.

[60] B. Albert and T. Tullis, *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, 2nd ed. Burlington, MA, USA: Morgan Kaufmann Publishers, 2013.

[61] B. Nunnally and D. Farkas, *UX Research: Practical Techniques for Designing Better Products*. Sebastopol, CA, USA: O'Reilly Media, 2016.

[62] J. Sauro and J. R. Lewis, *Quantifying the User Experience: Practical Statistics for User Research*, 2nd ed. Burlington, MA, USA: Morgan Kaufmann Publishers, 2016.

[63] M. Fan, J. Lin, C. Chung, and K. N. Truong, "Concurrent think-aloud verbalizations and usability problems," *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 5, pp. 1–35, 2019. [Online]. Available: https://doi.org/10.1145/3325281

[64] M. Fan, Q. Zhao, and V. Tibdewal, "Older adults' think-aloud verbalizations and speech features for identifying user experience problems," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2021, pp. 358:1–358:13. [Online]. Available: https://doi.org/10.1145/3411764.3445680

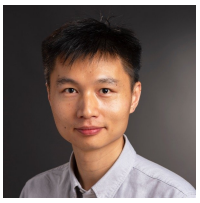
[65] M. Bostock, V. Ogievetsky, and J. Heer, "D^3: Data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec 2011. [Online]. Available: https://doi.org/10.1109/TVCG.2011.185

[66] S. Hubenschmid, J. Wieland, D. I. Fink, A. Batch, J. Zagermann, N. Elmqvist, and H. Reiterer, "ReLive: Bridging in-situ and ex-situ visual analytics for analyzing mixed reality user studies," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2022, pp. 24:1–24:20. [Online]. Available: https://doi.org/10.1145/3491102.3517550

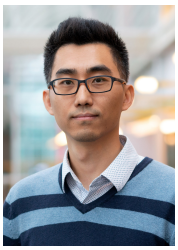
[67] M. Fan, X. Yang, T. Yu, Q. V. Liao, and J. Zhao, "Human-AI Collaboration for UX Evaluation: Effects of Explanation and Synchronization," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW1, apr 2022. [Online]. Available: https://doi.org/10.1145/3512943

Andrea Batch received the Ph.D. degree in information studies in 2022 from the University of Maryland in College Park, MD, USA, and her Masters degree in economics in 2011 from the University at Albany, SUNY in Albany, NY, USA. She is an Economist at the U.S. Bureau of Economic Analysis. At the time of the work involved in this publication, she was also a student member of the Human-Computer Interaction Laboratory (HCIL) at UMD.

Yipeng (Penny) Ji received the bachelor's degree in computer science in 2020 from the University of Waterloo in Waterloo, ON, Canada. She is currently an AI Research Engineer in the Advanced AI Lab at LG Electronics in Toronto, ON, Canada.



Mingming Fan received the Ph.D. degree in computer science in 2019 from the University of Toronto in Toronto, ON, Canada. He is currently an assistant professor at The Hong Kong University of Science and Technology (Guangzhou) and The Hong Kong University of Science and Technology, where he directs the APEX (Accessible & Pervasive User Experience) group. His research interests include human-computer interaction, computational user experience (UX), aging and accessibility, and information visualization.



Jian Zhao received the Ph.D. degree in computer science in 2015 from the University of Toronto in Toronto, ON, Canada. He is currently an assistant professor in the Cheriton School of Computer Science at the University of Waterloo in Waterloo, ON, Canada, where he directs the WatVis (Waterloo Visualization) research group. He is also a member of the Waterloo Artificial Intelligence Institute (WAII). His research interests include information visualization, human-computer interaction, and data science.

Niklas Elmqvist received the Ph.D. degree in 2006 from Chalmers University of Technology in Göteborg, Sweden. He is a professor in the College of Information Studies, University of Maryland, College Park in College Park, MD, USA. He is also a member of the Institute for Advanced Computer Studies (UMIACS) and formerly the director of the Human-Computer Interaction Laboratory (HCIL) at UMD. He is a senior member of the IEEE and the IEEE Computer Society.