“All Right, Mr. DeMille, I’m Ready for My Closeup:” Adding Meaning to User Actions from Video for Immersive Analytics

Andrea Batch[*] Niklas Elmqvist[†]

College of Information Studies
University of Maryland, College Park, MD, USA

Abstract

While the use of machine learning and computer vision to classify human behavior has grown into a large, well-established, interdisciplinary area of research, one area that is somewhat overlooked is the intersection of computer vision as a tool for evaluating user behavior in Virtual Reality, particularly in the context of immersive analytics and visualization. We draw on the literature from pattern recognition, computer vision, and machine learning to compose a simple, comparatively resource-cheap pipeline for camera-based extraction of features of professional analyst users and of their sessions in an existing VR visualization system, ImAxes. Our results show high accuracy in predicting self-reported features of the users, even as survey responses about user experience with the immersive interface are somewhat ambiguous in varying based on these features.

Keywords: Visualization, visual analytics, ubiquitous analytics, evaluation, video analytics, machine learning, deep learning.

Index Terms: Human-centered computing—Visualization—Visualization techniques; Human-centered computing—Visualization—Visualization design and evaluation methods

1 Introduction

The full-body interactions common in many virtual reality (VR) interfaces present a rich opportunity for the use of camera and sensor data to inform the human-computer interaction (HCI) researcher’s understanding of participants and their experiences. Research interest at the intersection of machine learning techniques and VR environments is growing with the development of increasingly affordable consumer VR hardware coinciding with the dramatic improvements in prediction accuracy of neural network architectures and machine learning. This interest is evidenced by the formation of conferences such as, for example, the IEEE International Conference on Artificial Intelligence & Virtual Reality (AIVR).

What constitutes an appropriate use of computer vision and machine learning to a VR visualization problem? We argue that a good area of application is in classifying the user and evaluating their experience. Specifically, this application could be used to augment ambiguous findings and improve the scalability of the costly process of qualitative evaluation of user sessions.

Our vision for the use of computer vision in HCI research is to shift away from video input from being an intractable media format, cheap to capture but expensive to analyze in evaluation studies for HCI and visualization, toward the use of video data as a revealed behavior dataset that is time-cost cheap and therefore scalable for the analysis of large user populations. In such a scenario, a visualization or HCI researcher can simply add a video recording setup, or turn on the onboard camera of a test computer and mobile or wearable device, to collect additional information about the user’s physical behavior and actions while participating in the user study. The video footage can then be easily and quickly analyzed using off-the-shelf models, as discussed here, resulting in both a time sequence of actions synchronized to the rest of the study telemetrics, as well as a summary of the actions performed. Actions include both gross motor skills, such as movement of the head, arms, and legs, as well as fine motor skills, such as hands and fingers, as well as even facial expressions. Even biometric information such as pulse, pupil dilation, respiration rate, etc, can be deduced using appropriate models. All these metrics can then be used as complements to task performance data collected in a user study.

Toward that end, we here demonstrate a specific application of computer vision and machine learning for low-cost and low-resource camera-based tracking of human behavior from video footage using a pre-trained neural network and random forest models. We use this prototype ad hoc machine learning pipeline to extract participant behavior from video footage of a user study involving analytical professionals from the U.S. Department of Commerce interacting with data in Virtual Reality. We argue that approaches like this could potentially be fed back into the interactive system to enable it to react to user behavior extracted from live video footage.

2 Related Work

If our goal is to make better use of video data in evaluating user behavior, we should consider work that has already been done in defining and tracking users and their interactions.

2.1 Personas in HCI: A User Classification Technique?

There is good cause to argue that the HCI and visualization communities should be mindful of the context in which people actually use the things we build [8,38]. In conducting more applied or “in the wild” studies, however, we run the risk of encountering the same issues in replication that plague many other domains [21]. One common approach to understanding factors influencing users’ needs for information visualization involves constructing personas–representational archetypes of “typical” users and their daily lives [38]. 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 [38, 76]. Alternative persona-based approaches also exist in which events within the interface, such as mouse activity [3], are used to develop “data-driven personas” [90] for characterizing types of users. Access to large-scale, user-generated datasets and platforms for crowdsourcing experiments such as Amazon Mechanical Turk [45] make the creation of these types of personas more manageable at larger scales. Similarly, data-driven methods in HCI are increasingly used to classify interactions [53,63,68].

2.2 Action Classification: Actions as User Behavior?

The HCI community has scratched the surface of using artificial neural networks (ANNs), including recurrent and convolutional neural networks (RNNs and CNNs), in evaluating user behavior. The

[*] e-mail: ajulca@umd.edu

[†] e-mail: elm@umd.edu

tasks of identifying gesture [61,71,82] and gaze [65,91], classifying user emotion and facial expression [34,63,81], and detecting characteristics of the user, such as gender [83], by constructing and implementing neural network architecture have been lightly explored within HCI-related discourse. The visualization community has also made contributions to the toolkit of methods used in evaluating user video, logs, transcripts, and other qualitative data [18], as well as user gesture analysis [40,42].

In the computer vision (CV) and machine learning (ML) communities, however, there have been more than a decade's worth of literature evaluating observed behavior via methods for action classification [51,66], motion and path prediction [58], eye tracking [46], and gesture detection [64]. While there have been a few position papers [14] and more serious studies [35] advocating for a closer relationship between the HCI and machine intelligence communities, the current body of literature on the subject is surprisingly sparse. If a trained neural network can identify individuals' emotions and expressions [27, 55, 89, 92] and can accurately predict whether a basketball player is good or bad [10], why is there so little work identifying whether a user's feelings about implementations are positive or negative, or their task performance good or bad?

2.3 The Intersections of Visualization and ML

Visualization has contributed to the domains of CV and ML not only through TensorBoard, the visual tool embedded in the wildly popular TensorFlow [1], but also in methods and algorithms [25] and in more recent and nuanced techniques (e.g., dataflow graphs used to show model structure [87]) and systems (e.g., ActiVis [41], a system deployed at Facebook to assist in model training and subset discovery). With that said, there is still much to do in developing techniques that not only aide CV/ML researchers, but also move us closer to what is increasingly referred to as "Explainable AI" (XAI) [12,32] and evoke a sense of trust by the lay person [72].

Conversely, work within the CV and ML communities to augment information visualization and interactive sensemaking is slightly sparser [26]. A notable exception to this claim is that computational models of perception have long been proposed as a means to make inferences about biological vision [69], and on rare occasion have been successfully applied to enhance and optimize information visualization using neural networks [44,70].

Recent work presenting semantic models for visualizing large image datasets [88] on the one hand, and visualization models for classifying semantic datasets [67] on the other, poetically illustrate the mutual benefit to be had in deepening this relationship for a closely related set of topics. In the former, a CNN is used to caption an image, and the caption is then used in determining the layout of the visualization via a model of semantically associated concepts, thus presenting a novel pipeline for handling a visualization problem (graph layout based on conceptual similarity) using CV and natural language processing (NLP) methods. In the latter, an interactive visualization system, ConceptVector, is implemented to support users in building lexicons of related concepts; these lexicons can then be used to improve recommendations by the visualization system, thus presenting a novel system for supporting NLP modeling, addressing the ML issue of semantic lexicon creation. These are both studies that draw from a heavily overlapping combination of topics in visualization and in ML to present solutions relevant to both fields: a ML solution to a visualization problem, and a visualization solution to a ML problem, respectively.

Another case of CV/ML techniques for improving visualization is represented in models trained to infer 3D scatter points of interest [6, 77]. For the general-purpose computer graphics, neural networks have also been used to predict incomplete regions of 2D images [56]. In broader interface design, latency reduction is another area that has been touched upon, but is ripe for further exploration [36].

2.4 Immersive Analytics and Neural Networks

Ubiquitous [24] and immersive analytics [17] are recently-established sub-domains of visual analysis research, with visualization techniques [23] and evaluation methods [5,22] for augmented reality (AR) and VR emerging with growing frequency. This direction for research is further validated by recent work indicating that information recall tasks are improved by working in VR [47].

There are two strong arguments for AR, VR, and mixed reality (MR) implementations as being more well-suited to DNN architectures for evaluation and for contextual-aware and adaptive design relative to traditional implementations. First, these environments are heavily reliant on cameras and sensors, and thus already collect a wealth of input data that can be used for evaluation. Second, the wide gestures and movements made by users in these environments fit well with the fairly saturated domain in CV research of action classification [15,31,51,66]. A model for detecting hand and foot gestures on smartphones [57], for example, could trigger events or inform evaluation for an AR visualization for mobile devices.

The concept and implementation of multimodal interfaces, another form of immersive environment, are not new [75]. Like humans, neural network architecture may be constructed to interpret multimodal stimuli [62, 92]. Semantic interpretation of chemosensory stimuli—taste and smell—is an open problem, but sound is increasingly well-explored. Srivastava et al. [79] use audio data to infer semantic meaning about video data; similarly, in *DeepEar* [50], an artificial neural network is trained to make inferences about the user's surroundings on mobile devices, and Soundnet is trained to identify the context of a video by combining video and audio inputs [4]. Schissler et al. [74] extend the concept of acoustics in audio analysis to present a visual-acoustic display system based on aural signal processing.

Star Trek's "Holodeck" is a popular metaphor for immersive storytelling [16,37]; we argue that it implies not just immersion, but learning and adaptation of the implementation to the user. The metaphor has been conjured up to describe the potential for scientific and engineering use [60] and for education [20]. It is a metaphor that has captured the public eye, and it is not an unreasonable one, given the emergence of immersive analytics and recent evidence of its benefits [17,47]. At present, the best path toward this end—toward creating the adaptive, immersive analytical environment—is in making strides in combining visualization techniques in immersive environments with CV and ML techniques.

2.5 Datasets and Pre-Trained Weights

The networks discussed thus far have largely been semi-supervised: They depend on semantic labels created by users of large-scale platforms (e.g., YouTube's 8M dataset [2]) or by researchers themselves (e.g., the SumMe dataset [33]). A movement toward unsupervised learning has resulted in the creation of architectures, techniques [9], and datasets for identifying "atomic actions" (e.g., the AVA dataset [31]) or simply predicting future frames in unlabeled video data [80].

With that said, training the weights of a deep neural network (DNN)—an ANN featuring a large number of hidden layers, the approach taken in most contemporary CV research—is still typically computationally costly, albeit not as costly as human labeling of interactions by the researcher. For the immediate future, the use of pre-trained networks to create semantic labels for evaluating user interactions is an easy point of entry (e.g., the Kinetics Human Action Video Dataset [15]).The major benefit of this approach is that it does not require the researcher to embark on the challenging and computationally costly journey of constructing and training their own DNN. In other words, it is labor-cheap and easy to implement, not just compared to comprehensive large-scale qualitative evaluation studies, but also compared to the approach of constructing and training one's own CV pipeline from end to end.

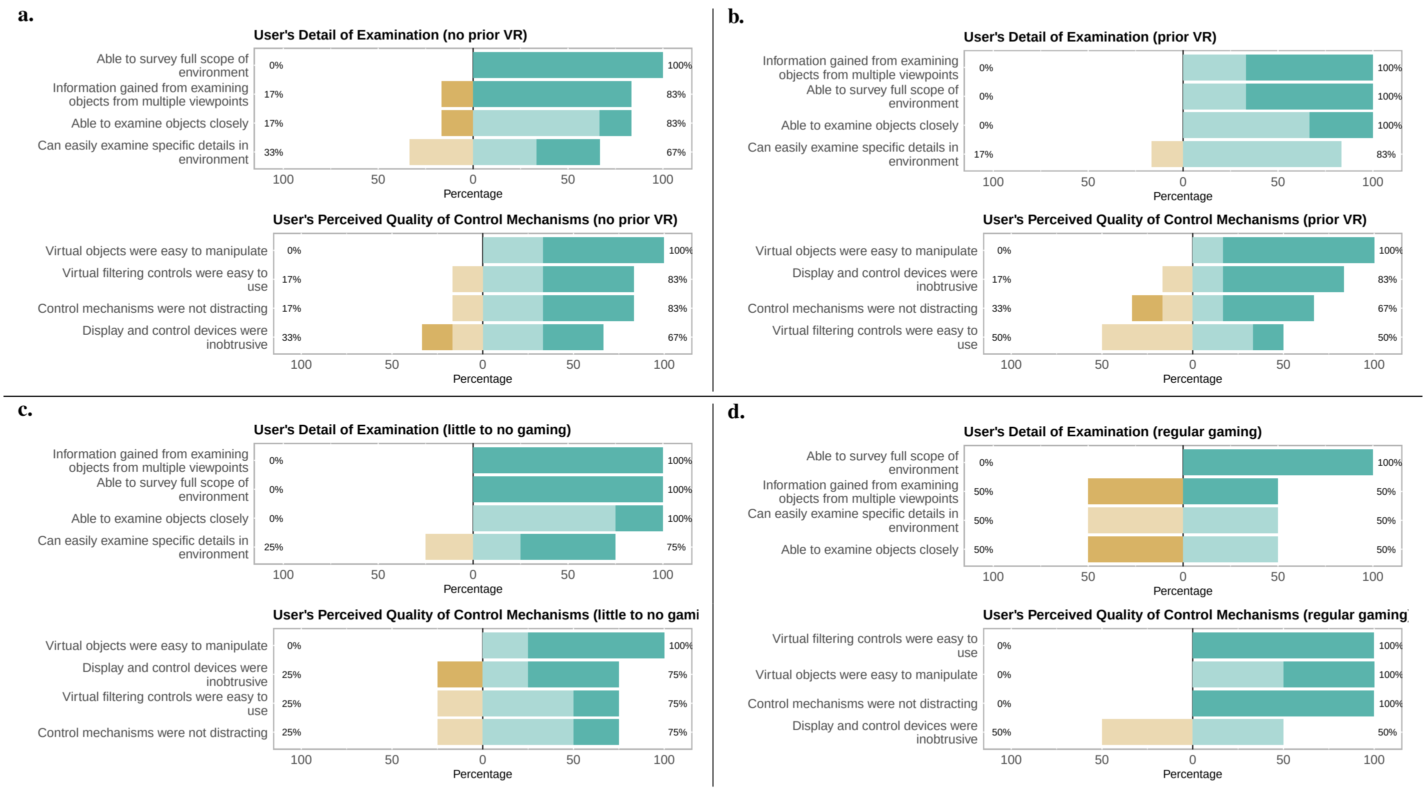

Figure 1: Ratings from users with prior VR use (**a**) versus without prior VR use (**b**), and from users without prior VR use who do not regularly play computer games (**c**) versus users without VR experience who regularly play computer games (**d**).

3 Study Methods

The purpose of our study was to evaluate the use of the immersive analytics environment ImAxes [22] by domain experts, and to extend the implementation.[1] The data was collected from participants employed as data scientists, economic analysts, and economists at the U.S. Bureau of Economic Analysis (BEA), where one of the authors is embedded as an employee.

3.1 Procedure and Experimental Design

The study had two stages. The first was a "formative stage," which included pilot user sessions with a sample drawn from the target population and "in-the-wild" user sessions that were open-ended and self-directed following a brief tutorial. The second was a "summative stage." During the summative stage, user sessions involved an exploration phase during which users were asked to freely explore the data in the environment, and a presentation phase during which users were asked to prepare and then give a presentation of a narrative about features of the data to the researcher.

The mixed-methods study was conducted in a small office of approximately 10×10 feet (3×3 meters). The computing equipment was a personal computer equipped with a Nvidia GeForce GTX 1060 (6GB) GPU, Intel Xeon E5-2620 v3 (2.40GHz) CPU, and 16GB RAM, and running Microsoft Windows 10. The ImAxes application [22] was built using Unity 5.6.5f1, and was installed locally on the aforementioned PC. The rig was equipped with an HTC Vive VR system, including a head-mounted display (HMD) and two base stations.

Two video streams of user interactions were captured using two Raspberry Pi Zeros with 8MP Pi cameras and with MotionEyeOS. The cameras served as motion-activated webcams, and were mounted in different positions in the room. One Raspberry Pi was positioned at chest height in front of the user's starting position, and the other was positioned in a top corner of the room near one of the Vive's base stations. Additional evaluation of video and telemetry data was conducted using a PC equipped with an EVGA GeForce GTX 1080 SC (8GB) GPU, Intel Core i7-7700 CPU (3.60GHz, 4 cores), and 24GB RAM, also running Windows 10.

Immediately after the user sessions, participants were asked to complete a survey with questions related to twenty pre-registered predictions[2] about users' sense of presence, feelings about the control mechanisms, engagement with the tasks being asked of them, and other aspects of their sessions both observed and self-reported [7]. Of those twenty predictions, the following three had weak or mixed results based on evidence from survey responses:

A3.3 Participants will report fatigue from their physical navigation and interaction. *Motivation:* The use of gross body motor controls to navigate in the virtual environment and interact with its objects will yield significant exertion and fatigue on the participants.

A4 Participants will encounter significant navigation and interaction hurdles due to a lack of VR expertise. *Motivation:* Our participant pool has no specific VR training, and will thus be challenged by 3D navigation and interaction concerns.

[1] These findings are drawn from posthoc analysis of data collected as part of a mixed-methods study at IEEE InfoVis 2019 [7]. However, we did not include any results derived from computer vision or machine learning methods in that paper. Thus, the data presented here is all unique to this paper and has not been previously published. However, much of the methods discussed here are by necessity similar to the InfoVis 2019 paper.

[2] https://osf.io/phxr2/

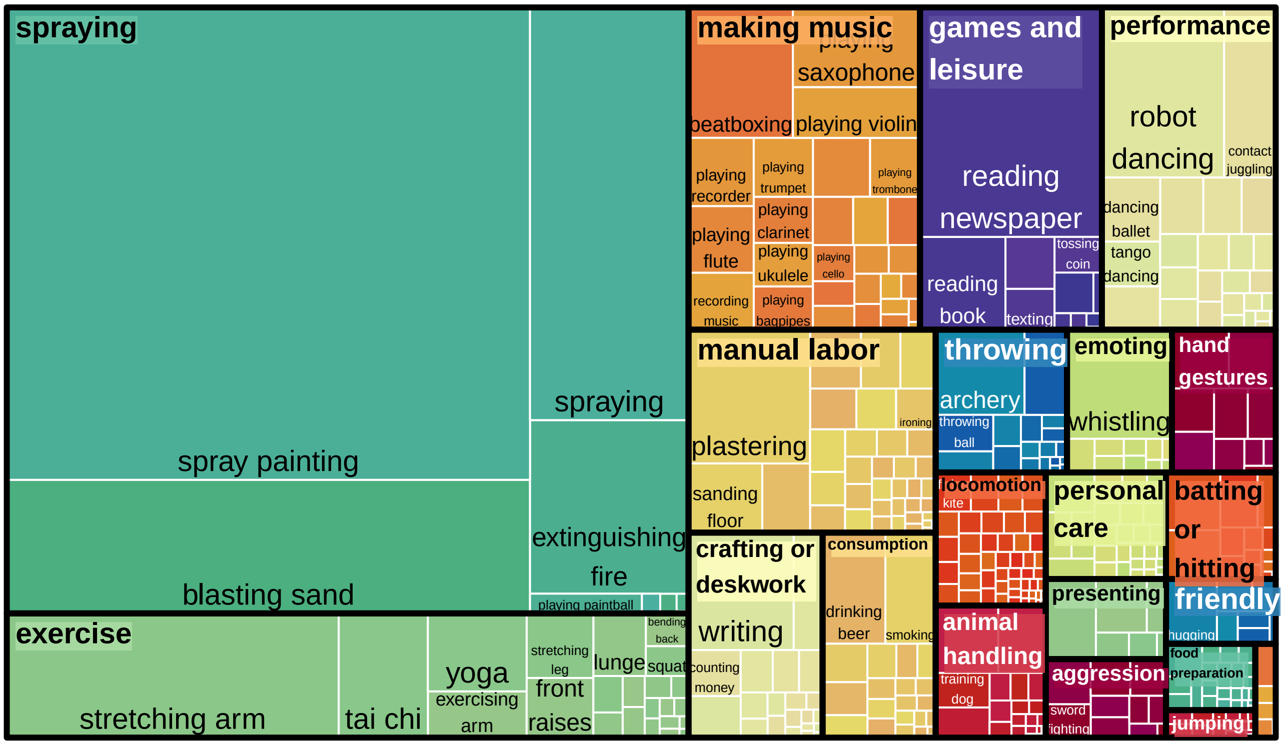

Figure 2: Relative sums of predicted probabilities of semantic labels assigned to all video segments: This may be considered to represent the comparative frequency of each semantic label for all video segments for all users.

A4.1 Participants with 3D computer gaming experience will be less hindered by lack of VR training. *Motivation:* 3D gaming experience will help people interact more efficiently.

We anticipated users reporting fatigue from their navigation and interaction (**A3.3**) without prompting, but this did not occur; our telemetry logs indicated increased user activity as time went on, and we did not have other measures for evaluating user fatigue. Contrary to our prediction, our survey did not indicate that there was a relationship between prior VR usage and user-reported hurdles in navigating the environment (**A4**), and 3D gaming did not have a substantial positive effect (**A4.1**) with respect to the users' perception of the control mechanisms (Figure 1). The results for hypothesis **A4.1** were mixed, with regular gamers reporting that examining objects in detail was more difficult, while also reporting that the control mechanisms were slightly higher quality than participants who did not regularly game. Because "fatigue" was not explicitly measured during the study, we used proxy measures of whether or not the participant experienced nausea and whether or not a given 10-second window was within 2 minutes of deciding to voluntarily stop the session.

The results for these hypotheses were ambiguous or mixed, and as such, presented an opportunity for exploiting video data to fill in some missing pieces. Furthermore, as our user sessions were split into two activities (explore and present), many of our predictions were contingent on an assumption that users would behave in different ways during the exploration stage relative to how they behaved during the presentation stage. We see this as another assumption that can be called into question or confirmed using the technique described in Section 3.2.

3.2 Data Analysis

We used pre-trained weights from the 3D Convolutional Neural Network (3DCNN) applied to the Kinetics Human Action Video Dataset [15] to classify segments of the video captured for users who consented to be video recorded. Our primary rationale for using pre-trained weights is that it is labor-cheap and easy to implement, as noted in Section 2.5; the Kinetics 3DCNN weights, specifically, were chosen for their relative popularity and accessibility. The 3DCNN assigns semantic action class labels to 10-second windows of video data for all video captured of users interacting with the environment. Each window was assigned 400 action class labels, each label with a prediction probability score; the aggregates of these probability scores (Figure 2) may be viewed as a pseudo-frequency for the actions observed during the study.

We checked this method in two ways: a validity check and a value check. As a measure of action class validity, we captured video data using the two cameras set up to record the user from different points of view. We then compared the similarity of the assigned labels by finding the rank-biased overlap (RBO) of each temporal segment [86].

To demonstrate the value added by applying this technique, we used the 3DCNN output as random forest input features to predict features of the user (e.g., how regularly they play computer games or if they engage in athletic or sporting activities that involve moving faster than a running speed), their experience (whether or not they have used VR before), and the type of activity that they were engaging in during the sessions (exploration versus presentation). We then trained a random forest [13] to predict characteristics of the participants relevant to their activities—namely, whether or not they had prior experience in VR, how regularly they play computer games in their free time, and whether or not they engage in any strenuous sporting or athletic activities.

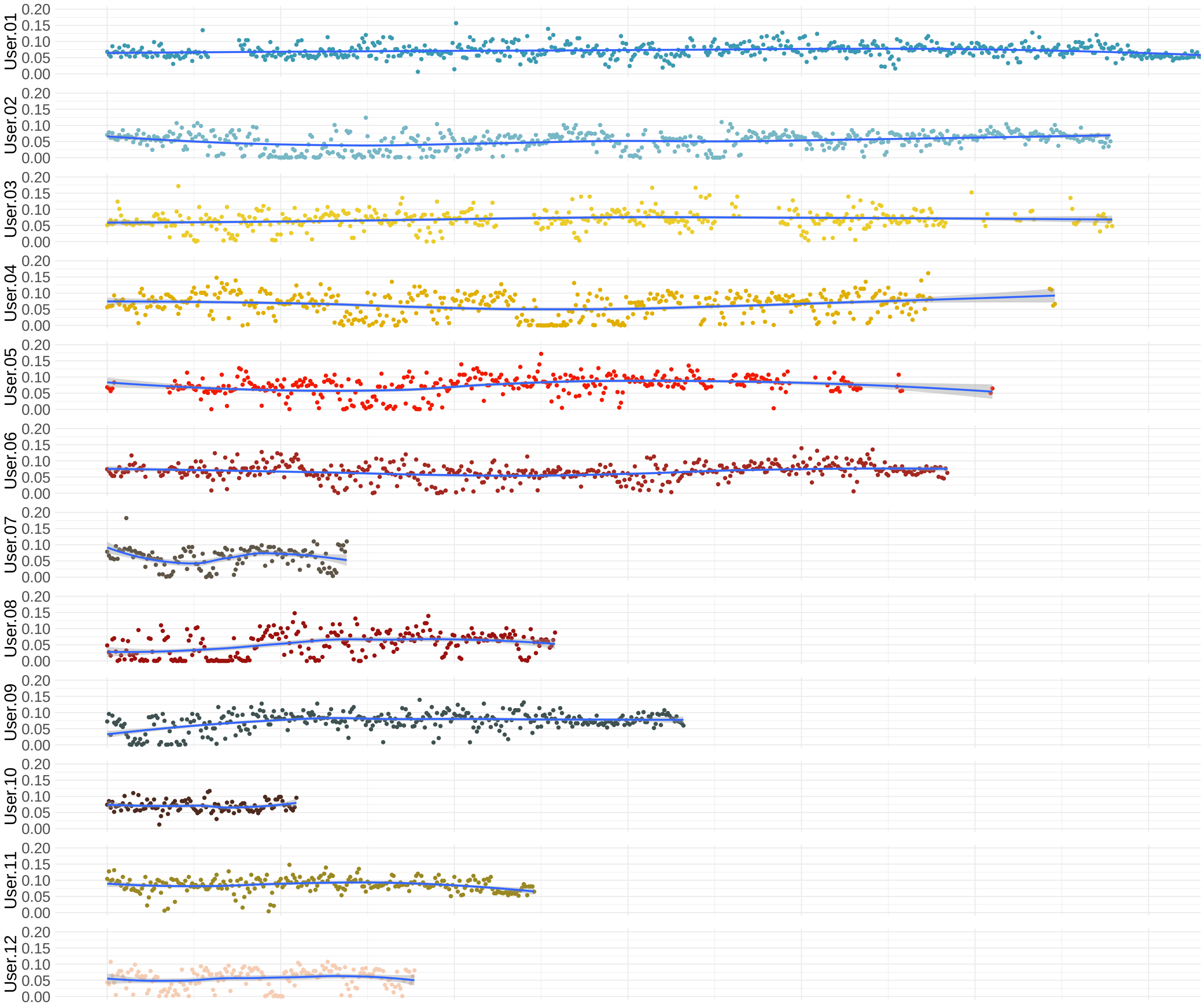

Figure 3: **Probability-weighted rank-biased overlap (RBO) distance scores:** A lower RBO distance score indicates a better match between the two ranked lists of the semantic action labels and probabilities assigned at each ten-second window of time. This can be thought of as indicating that there was always at least a 85% match between the actions assigned to video from two cameras from different angles of the room.

4 Results

The video data was used as an input for the pre-trained Kinetics 3DCNN [15], which predicted semantic labels to ten-second chunks of video (Figure 2). The prediction-probability-weighted ranked semantic labels resulting from the pre-trained 3DCNN yielded strong matches across both cameras (Figure 3), indicating that the network did a reasonably consistent job of predicting user actions. The random forest model was able to predict the participants' engagement in fast-motion sporting activities with **97.35%** accuracy, their likelihood of experiencing some form of VR sickness with **93.63%** accuracy, whether the user had prior VR experience with **92.74%** accuracy, and their gaming habits with **91.93%** accuracy (Table 1). This relatively high prediction accuracy for prior VR experience supports our prediction (**A4**) that users with no VR experience will have a characteristically different experience than users with VR experience. The prediction accuracy for regular gaming activities supports the notion that traditional gaming habits do influence the user's VR experience in an analytical setting, albeit in a more nuanced fashion than we predicted (**A4.1**), given that participant survey responses ran counter to this hypothesis.

To evaluate temporal features of the user session, we applied a model similar to the one used to predict features of the user to a combination of the video data collected during the formative studies and the video collected during the exploration phase of the summative studies. The difference in the models is primarily in the omission of the time from start feature as an independent variable from the model predicting temporal features of the session, as this feature shared a deterministic relationship with the dependent variable. We hoped a proxy measure of user fatigue could be extracted: i.e. whether or not the user was nearly ready to stop what they were doing. Users in the exploration stage of the summative studies and in the formative studies had control over when they ended the activities they were engaged in. Under the assumption that user fatigue plays a role in the user's decision that they are ready to exit the environment or end their current activities and that this fatigue will appear within a two minute window of this decision event, this method was not able to

Model	Out-of-bag error %	Hold-out error %	Cross-validation error %	CM[0]	CM[1]	CM[2]	CM[3]	CM[4]
FMAA*								
ranger	6.43	5.48	6.25					
randomForest	6.41	—	6.65					
CM [No/0]				1131	105			
CM [Yes/1]				75	1629			
Gaming								
ranger	8.91	7.3	7.86					
randomForest	1.83	—	8.07					
CM [Least/0]				648	9	3	28	6
CM [1]				16	722	5	6	38
CM [2]				36	11	258	6	9
CM [3]				33	14	0	508	11
CM [4/Most]				9	3	0	4	557
Prior VR								
ranger	5.0	4.75	6.65					
randomForest	5.41	—	7.26					
CM [No/0]				1456	63			
CM [Yes/1]				88	1333			
VR sickness								
ranger	5.78	4.63	6.37					
randomForest	6.10	—	6.37					
CM [No/0]				1937	44			
CM [Yes/1]				130	916			
Phase								
ranger	28.09	24.25	24.51					
randomForest	24.90	—	23.73					
CM [Explore/0]				801	182			
CM [Present/1]				258	966			
Quit in 2 min								
ranger	10.34	10.16	11.00					
randomForest	11.45	—	11.20					
CM [No/0]				2501	1			
CM [Yes/1]				288	24			

* User participated in fast-moving athletic activities (FMAA) during their leisure time.

Table 1: Prediction errors and confusion matrices for random forest.

detect approaching fatigue. While results yielded **88.8%** accuracy, 92.3% of moments within this 2-minute window were misclassified as being outside of it (i.e., they were classified as not being in proximity to a quit event). These results bolster our negative findings to our prediction (**A3.3**) that users will report fatigue, as we were not able to find such empirical evidence.

Relative to this approach for picking up on user fatigue, the model performed slightly better in predicting whether the users were in the "explore" or "present" phase of the study, or in neither phase (**76.27%** accuracy; 81.7% of users preparing to present were correctly classified, and 73.5% of exploring users were correctly classified, while video of users engaging in neither activity was classified correctly only 43.2% of the time). This indicates that random forest classification of 3DCNN output performs at levels significantly better than chance in predicting whether the user is engaged in types of behaviors entirely novel to the environment and the experiment activities. In general, however, we found that our approach did not perform as well in temporal segmentation of video data as it did in picking up on relevant user traits, such as their prior VR experience.

5 Conclusion and Discussion

In using this technique to evaluate user behavior, we have achieved our goal of implementing a more robust method for testing our hypotheses than what would have been possible without a computer vision approach. By finding that the RBO for cameras capturing user activity from different angles is reasonably low, we have also contributed additional validation for the pre-trained 3DCNN we used to derive our semantic label predictions. These methods are not only robust, but because we are using semantic labels as our random forest input, they produce results (Figures 4 and 5) that convey semantic significance the lay reader: Behavior that looks like beekeeping or slacklining (or doesn't) is important for predicting whether a VR user has had prior VR experience (Figure 4). Gestures that look like extinguishing a fire are a reliable indicator for whether the user is a regular gamer and also for whether or not the user should anticipate experiencing VR sickness (Figure 4).

In a result that is specific to both the implementation and the user study design, actions that look like laughing or throwing an axe are important for predicting whether a user is exploring the data or presenting a narrative to an onlooker. This holds semantic interest in the context of the implementation in particular, because the gesture for destroying an object in ImAxes is to *throw it*, and the Vive controllers may register as looking somewhat like a small axe: It indicates that the destruction of objects may be an important indicator for what kind of activity a user is engaged in. Similarly, it may be that the VR users' speech during their presentations looks like laughter when the face is obscured by a VR headset.

Another goal in this essay is to urge HCI researchers to shift in the direction of setting standards for using models for CV and/or ML in user evaluation and in visual system architecture. We have made the argument that this practice has implications that make it well-suited for VR implementations, but it is also promising in other applications beyond the desktop (e.g., mobile device augmented reality implementations).

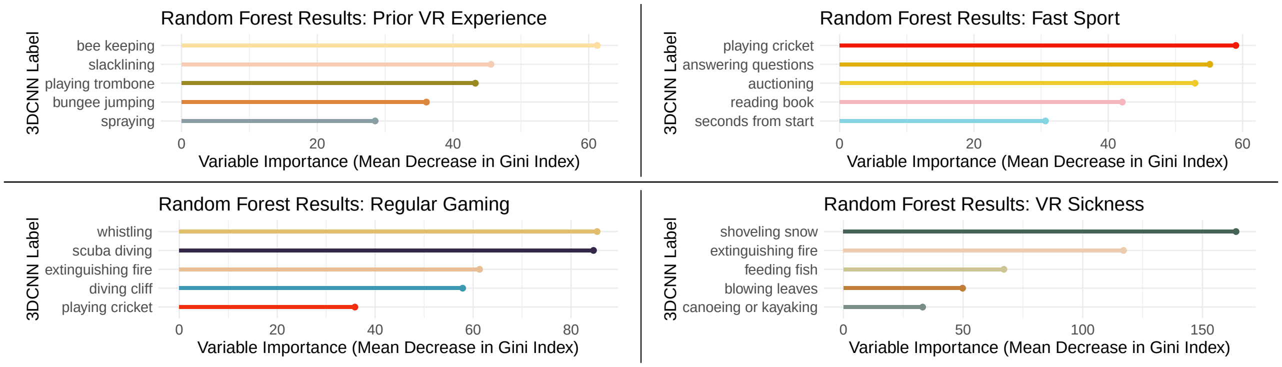

Figure 4: Relative importance of semantic labels for predicting participants' involvement in fast-moving sports, the amount of time spent gaming, prior VR experience, and user's predisposition to VR sickness. These models included time, to improve the fit in the context of how a user's activities may be characterized relative to their time in the session, given features of the user unrelated to time.

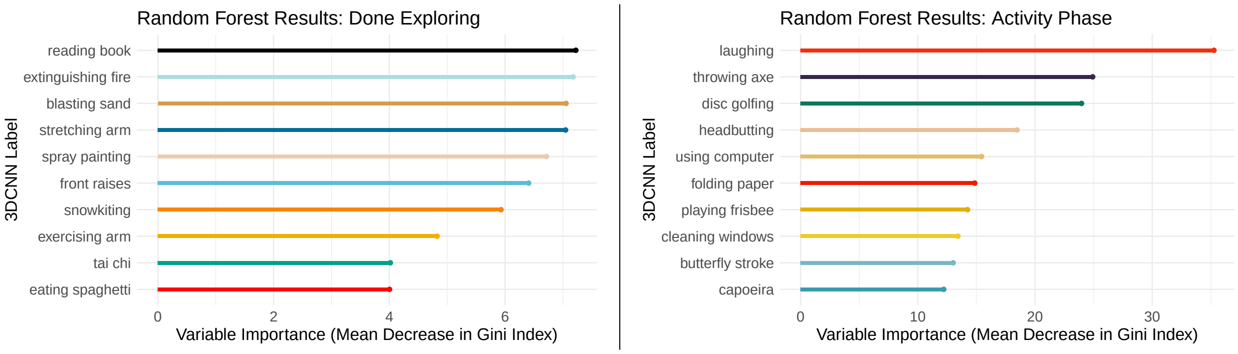

Figure 5: Relative importance of semantic labels for predicting whether participants were engaged in exploration vs. presentation, or if they were within two minutes of quitting. These models *did not* include time, because the dependent variables were directly related to their time in session.

5.1 Limitations

This study exploits computer vision and machine learning techniques to fill in information that is ambiguous and further illuminate patterns observed during user sessions. The analysis in this paper has been conducted as an extension of data collected in work that was accepted at IEEE InfoVis 2019 and publication in TVCG [7]; the scope of this work is limited to a deeper exploration of hypotheses left partly answered by more conventional means. As such, it does not take the next step in analyzing possible interventions to improve user experience; this remains an open area for future research.

5.2 Future Context-Aware, Adaptive Environments

The concept of the context-aware computing environment is nearly as old as HCI itself [73]. A typical manifestation of this model is in the use of sensor and telemetry information for predetermined semantic inference about the user's environment, but the idea of using neural networks to aide in achieving this end is not new [11, 48, 54]. In fact, work implementing ANN architecture and RGB camera input for face detection in the context of HCI emerged no later than the very same year as the concept of context-aware computing [39]. Contemporary work in query prediction [78] combining context-aware information-seeking techniques and neural networks has seen some success toward that end, but it is generally an under-explored area of research. Within the visualization community, the direction is typically in using visualization to assist CV and ML work (e.g., ConceptVector [67]), although this is certainly not universally true [88]. Simple context detection in the CV and ML communities, however, is a problem space that has seen many achievements within the past few years [19, 28, 43, 85].

For an interface to be truly context-aware, we argue that it must be capable of semantically evaluating users' behaviors and environments *in real time*. Hardware capable of doing this on consumer-ready electronics, including mobile devices, is a maturing technology several years old [29]. It is one foundation of AR, which in mobile devices commonly depend on non-visual sensors (e.g., compass, GPS, and accelerometer) due to computational costs and performance, but mobile AR implementations using RGB cameras and CV techniques have been around for decades [59]. Interfaces using specialized sensors and CNN architecture have been capable of real-time gesture interpretation for years [52].

Semantic segmentation of video and images for task evaluation is an old [69] and increasingly optimized [49] area of CV research. Most contemporary architectures for video analysis incorporate long short term memory (LSTM) autoencoders [80]. Generative adversarial networks (GANs), first proposed by Goodfellow et al. [30] and first successfully implemented by Vondrick et al. [84], are also increasingly the state of the art in the use of neural networks for creating imagery.

Training semantic and generative models, however, is still too computationally intensive to do in real-time on consumer electronics. Here, again, the use of pre-trained models shines as an immediate next step—their potential application in user study evaluation may be seen as a stepping stone to embedding the evaluation process into iterations of implementations following initial results. By constructing architecture for change events around subsets of model output values based on evaluation findings, for example, visualization and interface design can take the next steps in being *context-aware*. By constructing architecture to update the trained model (not necessarily in real-time), visualization and interface design can take its first steps toward being *adaptive*.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: A system for large-scale machine learning. In *Proceedings of the USENIX Conference on Operating Systems Design and Implementation*, pp. 265–283. USENIX Association, Berkeley, CA, USA, 2016.

[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[3] A. A. E. Ahmed and I. Traore. A new biometric technology based on mouse dynamics. *IEEE Transactions on Dependable and Secure Computing*, 4(3):165–179, 2007. doi: 10.1109/TDSC.2007.70207

[4] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pp. 892–900, 2016.

[5] B. Bach, R. Sicat, J. Beyer, M. Cordeil, and H. Pfister. The hologram in my hand: How effective is interactive exploration of 3D visualizations in immersive tangible augmented reality? *IEEE Transactions on Visualization and Computer Graphics*, 24(1):457–467, 2018. doi: 10.1109/TVCG.2017.2745941

[6] J. Barhak and A. Fischer. Parameterization and reconstruction from 3D scattered points based on neural network and PDE techniques. *IEEE Transactions on Visualization and Computer Graphics*, 7(1):1–16, 2001. doi: 10.1109/2945.910817

[7] A. Batch, A. Cunningham, M. Cordeil, N. Elmqvist, T. Dwyer, B. H. Thomas, and K. Marriott. There is no spoon: Evaluating performance, space use, and presence with expert domain users in immersive analytics. In *IEEE Transactions on Visualization and Computer Graphics*, to appear, 2020.

[8] A. Batch and N. Elmqvist. The interactive visualization gap in initial exploratory data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):278–287, 2018. doi: 10.1109/TVCG.2017.2743990

[9] 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 (AIVR)*, pp. 1–10, Dec 2018. doi: 10.1109/AIVR.2018.00009

[10] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi. Am I a baller? Basketball performance assessment from first-person videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2196–2204, 2017. doi: 10.1109/ICCV.2017.239

[11] C. Biancalana, F. Gasparetti, A. Micarelli, A. Miola, and G. Sansonetti. Context-aware movie recommendation based on signal processing and machine learning. In *Proceedings of the Challenge on Context-Aware Movie Recommendation*, pp. 5–10. ACM, New York, NY, USA, 2011. doi: 10.1145/2096112.2096114

[12] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, and U. Muller. Explaining how a deep neural network trained with end-to-end learning steers a car. *arXiv preprint arXiv:1704.07911*, 2017.

[13] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324

[14] E. Cambria, G.-B. Huang, L. L. C. Kasun, H. Zhou, C. M. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li, V. C. M. Leung, L. Feng, Y. S. Ong, M. H. Lim, A. Akusok, A. Lendasse, F. Corona, R. Nian, Y. Miche, P. Gastaldo, R. Zunino, S. Decherchi, X. Yang, K. Mao, B. S. Oh, J. Jeon, K. A. Toh, A. B. J. Teoh, J. Kim, H. Yu, Y. Chen, and J. Liu. Extreme learning machines [trends & controversies]. *IEEE Intelligent Systems*, 28(6):30–59, 2013. doi: 10.1109/MIS.2013.140

[15] 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*, pp. 4724–4733, 2017. doi: 10.1109/CVPR.2017.502

[16] M. Cavazza, J.-L. Lugrin, D. Pizzi, and F. Charles. Madame Bovary on the Holodeck: Immersive interactive storytelling. In *Proceedings of the ACM International Conference on Multimedia*, pp. 651–660. ACM, New York, NY, USA, 2007. doi: 10.1145/1291233.1291387

[17] T. Chandler, M. Cordeil, T. Czauderna, T. Dwyer, J. Glowacki, C. Goncu, M. Klapperstueck, K. Klein, F. Schreiber, and E. Wilson. Immersive analytics. In *Proceedings of the International Symposium on Big Data Visual Analytics*, pp. 1–8, Sep 2015. doi: 10.1109/BDVA.2015.7314296

[18] 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*, 100:66 – 80, 2017. doi: 10.1016/j.ijhcs.2016.12.007

[19] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):13–27, 2015. doi: 10.1109/TPAMI.2014.2343217

[20] T. C. Chen, C. F. Chiu, A. Klimenko, and T. K. Shih. Toward a Holodeck like edutainment game using wearable device and motion sensors. In *Proceedings of the International Conference on Ubi-Media Computing*, pp. 242–247, 2015. doi: 10.1109/UMEDIA.2015.7297462

[21] E. Coiera, E. Ammenwerth, A. Georgiou, and F. Magrabi. Does health informatics have a replication crisis? *Journal of the American Medical Informatics Association*, p. ocy028, 2018. doi: 10.1093/jamia/ocy028

[22] M. Cordeil, A. Cunningham, T. Dwyer, B. H. Thomas, and K. Marriott. ImAxes: Immersive axes as embodied affordances for interactive multivariate data visualisation. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 71–83. ACM, New York, NY, USA, 2017. doi: 10.1145/3126594.3126613

[23] M. Cordeil, T. Dwyer, K. Klein, B. Laha, K. Marriott, and B. H. Thomas. Immersive collaborative analysis of network connectivity: CAVE-style or head-mounted display? *IEEE Transactions on Visualization and Computer Graphics*, 23(1):441–450, 2017.

[24] N. Elmqvist and P. Irani. Ubiquitous analytics: Interacting with big data anywhere, anytime. *IEEE Computer*, 46(4):86–89, 2013. doi: 10.1109/mc.2013.147

[25] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews. The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems*, 43(3):411–435, Dec 2014. doi: 10.1007/s10844-014-0304-9

[26] A. Endert, W. Ribarsky, C. Turkay, B. W. Wong, I. Nabney, I. D. Blanco, and F. Rossi. The state of the art in integrating machine learning into visual analytics. *Computer Graphics Forum*, 36(8):458–486, 2017. doi: 10.1111/cgf.13092

[27] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5562–5570, 2016. doi: 10.1109/CVPR.2016.600

[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014. doi: 10.1109/CVPR.2014.81

[29] V. Gokhale, J. Jin, A. Dundar, B. Martini, and E. Culurciello. A 240 G-ops/s mobile coprocessor for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 696–701, 2014. doi: 10.1109/CVPRW.2014.106

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of Advances In Neural Information Processing Systems*, pp. 2672–2680, 2014.

[31] C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. *arXiv preprint arXiv:1705.08421*, 2017.

[32] D. Gunning. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*, 2017.

[33] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *Proceedings of the IEEE European Conference on Computer Vision*, pp. 505–520, 2014. doi: 10.1007/978-3-319-10584-0_33

[34] 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*, pp. 95–100, 2014. doi: 10.1109/HSI.2014.6860455

[35] 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*, pp. 350–354, 2014. doi: 10.1109/ICVRV.2014.6

[36] N. Henze, M. Funk, and A. S. Shirazi. Software-reduced touchscreen latency. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 434–441. ACM, New York, NY, USA, 2016. doi: 10.1145/2935334.2935381

[37] R. Hill, J. Gratch, W. L. Johnson, C. Kyriakakis, C. LaBore, R. Lindheim, S. Marsella, D. Miraglia, B. Moore, J. Morie, J. Rickel, M. Thiébaux, L. Tuch, R. Whitney, J. Douglas, and W. Swartout. Toward the Holodeck: Integrating graphics, sound, character and story. In *Proceedings of the International Conference on Autonomous Agents*, pp. 409–416. ACM, New York, NY, USA, 2001. doi: 10.1145/375735.376390

[38] K. Holtzblatt and H. Beyer. *Contextual Design: Evolved*. Synthesis Lectures on Human-Centered Informatics. Morgan & Claypool Publishers, 2014.

[39] M. Hunke and A. Waibel. Face locating and tracking for human-computer interaction. In *Proceedings of the Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1277–1281, 1994. doi: 10.1109/ACSSC.1994.471664

[40] S. Jang, N. Elmqvist, and K. Ramani. MotionFlow: Visual abstraction and aggregation of sequential patterns in human motion tracking data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):21–30, 2016. doi: 10.1109/TVCG.2015.2468292

[41] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. P. Chau. ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):88–97, 2018. doi: 10.1109/TVCG.2017.2744718

[42] 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 (CGIV)*, pp. 419–428, 2007. doi: 10.1109/CGIV.2007.88

[43] D. Kim, C. Park, J. Oh, S. Lee, and H. Yu. Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the ACM Conference on Recommender Systems*, pp. 233–240. ACM, New York, NY, USA, 2016. doi: 10.1145/2959100.2959165

[44] Y. Kim and A. Varshney. Saliency-guided enhancement for volume visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):925–932, 2006. doi: 10.1109/TVCG.2006.174

[45] 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*, pp. 453–456, 2008. doi: 10.1145/1357054.1357127

[46] 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*, pp. 2176–2184, 2016. doi: 10.1109/CVPR.2016.239

[47] E. Krokos, C. Plaisant, and A. Varshney. Spatial mnemonics using virtual reality. In *Proceedings of the 2018 10th International Conference on Computer and Automation Engineering*, pp. 27–30, 2018.

[48] M. Krstic and M. Bjelica. Context-aware personalized program guide based on neural network. *IEEE Transactions on Consumer Electronics*, 58(4):1301–1306, 2012. doi: 10.1109/TCE.2012.6414999

[49] A. Kundu, V. Vineet, and V. Koltun. Feature space optimization for semantic video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3168–3175, 2016. doi: 10.1109/CVPR.2016.345

[50] N. D. Lane, P. Georgiev, and L. Qendro. DeepEar: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 283–294. ACM, New York, NY, USA, 2015. doi: 10.1145/2750858.2804262

[51] 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*, pp. 1–8, 2008. doi: 10.1109/CVPR.2008.4587756

[52] J. H. Lee, T. Delbruck, M. Pfeiffer, P. K. Park, C.-W. Shin, H. Ryu, and B. C. Kang. Real-time gesture interface based on event-driven processing from stereo silicon retinas. *IEEE Transactions on Neural Networks and Learning Systems*, 25(12):2250–2263, 2014. doi: 10.1109/TNNLS.2014.2308551

[53] H. Li, C. Ye, and A. P. Sample. IDSense: A human object interaction detection system based on passive UHF RFID. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 2555–2564. ACM, New York, NY, USA, 2015. doi: 10.1145/2702123.2702178

[54] T. Lin, C. Wang, and P.-C. Lin. A neural-network-based context-aware handoff algorithm for multimedia computing. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 4(3):17:1–17:23, 2008. doi: 10.1145/1386109.1386110

[55] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1749–1756, 2014. doi: 10.1109/CVPR.2014.226

[56] P. Liu, J. P. Lewis, and T. Rhee. Low-rank matrix completion to reconstruct incomplete rendering images. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2018. doi: 10.1109/TVCG.2017.2722414

[57] Z. Lv, A. Halawani, S. Feng, H. Li, and S. U. Réhman. Multimodal hand and foot gesture interaction for handheld devices. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(1s):10:1–10:19, 2014. doi: 10.1145/2645860

[58] 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*, pp. 4636–4644, 2017. doi: 10.1109/CVPR.2017.493

[59] S. Mann. Humanistic computing: 'WearComp' as a new framework and application for intelligent signal processing. *Proceedings of the IEEE*, 86(11):2123–2151, 1998. doi: 10.1109/5.726784

[60] S. Marks, J. E. Estevez, and A. M. Connor. Towards the Holodeck: Fully immersive virtual reality visualisation of scientific and engineering data. In *Proceedings of the International Conference on Image and Vision Computing New Zealand*, pp. 42–47. ACM, New York, NY, USA, 2014. doi: 10.1145/2683405.2683424

[61] 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*, pp. 132–139, 2015. doi: 10.1109/EUC.2015.25

[62] H. P. Martínez and G. N. Yannakakis. Deep multimodal fusion: Combining discrete events and continuous signals. In *Proceedings of the International Conference on Multimodal Interaction*, pp. 34–41. ACM, New York, NY, USA, 2014. doi: 10.1145/2663204.2663236

[63] 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*, 46(4):916–929, 2016. doi: 10.1109/TCYB.2015.2418092

[64] 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*, pp. 4207–4215, 2016. doi: 10.1109/CVPR.2016.456

[65] S. S. Mukherjee and N. M. Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015. doi: 10.1109/TMM.2015.2482819

[66] 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*, pp. 1–8, 2007. doi: 10.1109/CVPR.2007.383132

[67] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist. ConceptVector: Text visual analytics via interactive lexicon building using word embedding. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):361–370, 2018.

[68] K. Park and G. Lee. FingMag: Finger identification method for smartwatch. In *Extended Abstracts of the ACM Conference on Human*

Factors in Computing Systems, pp. LBW2216:1–LBW2216:6. ACM, New York, NY, USA, 2019. doi: 10.1145/3290607.3312982

[69] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. doi: 10.1109/CVPR.2007.383337

[70] D. Pineo and C. Ware. Data visualization optimization via computational modeling of perception. *IEEE Transactions on Visualization and Computer Graphics*, 18(2):309–320, 2012. doi: 10.1109/TVCG.2011.52

[71] 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*, pp. 639–643, 2015. doi: 10.1109/IADCC.2015.7154785

[72] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, New York, NY, USA, 2016. doi: 10.1145/2939672.2939778

[73] B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *Proceedings of the Workshop on Mobile Computing Systems and Applications*, pp. 85–90, 1994. doi: 10.1109/WMCSA.1994.16

[74] C. Schissler, C. Loftin, and D. Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics*, 24(3):1246–1259, 2018. doi: 10.1109/TVCG.2017.2666150

[75] R. Sharma, V. I. Pavlović, and T. S. Huang. Toward multimodal human-computer interface. *Proceedings of the IEEE*, 86(5):853–869, 1998. doi: 10.1109/5.664275

[76] K. Shilton. This is an intervention: Foregrounding and operationalizing ethics during technology design. In K. D. Pimple, ed., *Emerging Pervasive Information and Communication Technologies: Ethical Challenges, Opportunities and Safeguards*, pp. 177–192. Springer, Dordrecht, 2014.

[77] Z. Shu, S. Xin, X. Xu, L. Liu, and L. Kavan. Detecting 3D points of interest using multiple features and stacked auto-encoder. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2018. doi: 10.1109/TVCG.2018.2848628

[78] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 553–562. ACM, New York, NY, USA, 2015. doi: 10.1145/2806416.2806493

[79] M. Srivastava and A. Agarwal. Classification of emotions from speech using implicit features. In *Proceedings of the International Conference on Industrial and Information Systems*, pp. 1–6, 2014. doi: 10.1109/ICIINFS.2014.7036518

[80] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using LSTMs. In *Proceedings of the International Conference on Machine Learning*, pp. 843–852, 2015.

[81] P. Suja, K. V. Kumar, and S. Tripathi. Dynamic facial emotion recognition from 4D video sequences. In *Proceedings of the International Conference on Contemporary Computing*, pp. 348–353, 2015. doi: 10.1109/IC3.2015.7346705

[82] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 33(5):169:1–169:10, 2014. doi: 10.1145/2629500

[83] 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*, pp. 188–195, 2015. doi: 10.1109/SSCI.2015.37

[84] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Proceedings of Advances In Neural Information Processing Systems*, pp. 613–621, 2016.

[85] T.-H. Vu, A. Osokin, and I. Laptev. Context-aware CNNs for person head detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2893–2901, 2015. doi: 10.1109/ICCV.2015.331

[86] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4):20:1–20:38, Nov. 2010. doi: 10.1145/1852102.1852106

[87] K. Wongsuphasawat, D. Smilkov, J. Wexler, J. Wilson, D. Mané, D. Fritz, D. Krishnan, F. B. Viégas, and M. Wattenberg. Visualizing Dataflow Graphs of deep learning models in TensorFlow. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):1–12, 2018. doi: 10.1109/TVCG.2017.2744878

[88] X. Xie, X. Cai, J. Zhou, N. Cao, and Y. Wu. A semantic-based method for visualizing large image collections. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2018. doi: 10.1109/TVCG.2018.2835485

[89] J. Yi, X. Mao, L. Chen, Y. Xue, and A. Compare. Facial expression recognition considering individual differences in facial structure and texture. *IET Computer Vision*, 8(5):429–440, 2014. doi: 10.1049/iet-cvi.2013.0171

[90] 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*, pp. 5350–5359. ACM, New York, NY, USA, 2016. doi: 10.1145/2858036.2858523

[91] 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*, pp. 4511–4520, 2015. doi: 10.1109/CVPR.2015.7299081

[92] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3438–3446, 2016. doi: 10.1109/CVPR.2016.374