A conceptual framework for immersive behavioral data exploration assisted by a deep learned representation of data

Victor Duvivier
Nantes Université,
École Centrale Nantes, CNRS,
LS2N, UMR 6004,
44300 Nantes, France
0009-0003-2994-7409

Matthieu Perreira da Silva
Nantes Université,
École Centrale Nantes, CNRS,
LS2N, UMR 6004,
44300 Nantes, France
0000-0003-3921-5132

Yannick Prié
Nantes Université,
École Centrale Nantes, CNRS,
LS2N, UMR 6004,
44300 Nantes, France
0000-0002-7068-0836

***Abstract*—By allowing to conduct experiments involving ecologically valid tasks within controlled environments, Virtual Reality (VR) offers novel opportunities for studying human behavior. Several modalities can be leveraged, including event logs, motion trajectories, eye-tracking data, or physiological signals. However, analyzing such multimodal data presents considerable challenges due to their inherent complexity, the varied structures they imply, as well as the necessity to use exploratory approaches. There is therefore a need to design visual analytics tools for the exploration of immersive behavioral data without prior knowledge. Our idea is to integrate deep learned computational models —which leverage advanced techniques to extract valuable high-level features from unlabeled data— into exploratory Visual Analytics tools. We introduce a conceptual framework for integrating deep learning models into the Visual Analytics process of immersive behavioral data analysis, also focusing on the services that such systems should provide to the analysts.**

***Index Terms*—Visual Analytics, Immersive Behavior Analysis, Self-supervised Learning, Deep Representation Space**

I. Introduction

New approaches for the study of human behavior are made possible by VR; indeed, it allows setting up rich and complex experiments while maintaining control and repeatability over the scenario, and to compare behaviors recorded from different users in the same virtual environment.

Several works have shown the interest of analyzing immersive data to study human behaviors from different perspectives. For example, we developed a virtual kitchen (figure 1) as a tool that allows psychologists to understand and compare the behavioral patterns exhibited by individuals [1]. The work from Yuan et al. [2] examined the influence of spatial configurations of exhibits on visitors' explicit reactions. A study from Yaremych et al. made use of a virtual buffet to analyze the behavior of parents preparing meals for their children, exploring micro-behaviors and emphasizing the need to explore behaviors at different temporal scales [3].

Moreover, VR makes it possible to easily record behavioral data that was very difficult to collect in reality. This temporal data, referred to hereafter as immersive behavioral data, comprises several modalities, such as trajectories of limbs, event logs, eye-tracking measures, or physiological signals.

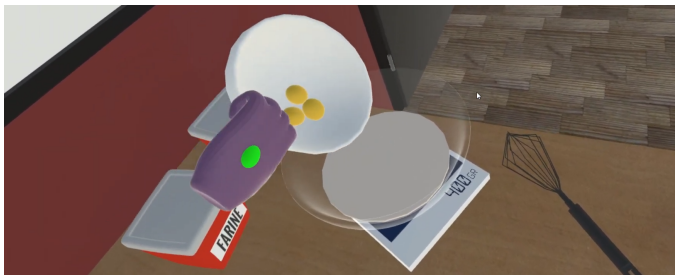

Fig. 1. A participant of a virtual kitchen pours egg yolks into a preparation. Analyzing this action and others may be of use to an analyst trying to assess issues with cognitive functions, *e.g.*, planning.

There are at least three challenges related to analyzing such data. First, the recording of behaviors with (relatively) high sample rate generates a large amount of data that is difficult for humans to handle (even with a single modality). Visualization tools can help when the analyst needs to deal with few participants' data, but the combination of a high sample rate and large number of participants makes the visual analysis of such data very challenging. Second, the plurality of modalities, their different structures, and the complex relationships between them, make the analysis even more challenging. This is amplified by the difference in the density of information provided by each modality. For instance, while the event log modality offers valuable qualitative insights, it is considerably sparser than the spatio-temporal modality. Third, as such type of behavioral data is new, and should allow uncovering new qualitative and quantitative insights on human behavior, it has to be analyzed in exploratory ways.

Exploratory data analysis falls within the scope of Visual Analytics [4], [5], which seeks to make the best use of human capabilities (*e.g*, visual pattern recognition, intuition from domain knowledge) and those of computers (*e.g*. massive data management, model building) to build up new knowledge.

With respect to our third challenge, it seems that visual analytics tools are needed to help analysts make sense of immersive behavioral data while overcoming the other two challenges. The iterative process of knowledge construction on exploratory tasks has been widely studied [4], [5]. Notably, Brehmer and Munzner proposed a multi-level typology describing the different interactions taking place between the analyst and a visual analytics system [6].

Deep learning has proven to be a suitable approach for the automated building of computational models from both massive and multimodal data, be it in a supervised [7]–[10] or unsupervised [7], [11] ways. As our concern is exploratory analysis we cannot provide labeled data to a supervised machine learning model. It is then our proposal to try and leverage deep self-supervised learning to develop computational models of immersive behavioral data that could be used in visual analytics systems.

In this paper, we first introduce our general proposal to incorporate a deep learned representation of immersive behavioral data into visual analytics systems, so that analysts can make use of such models to carry out their exploratory tasks. Our second contribution is a conceptual framework that describes in more details the behavioral data, the generated latent space, and the services provided to an analyst using such a deep learning based visual analytics system.

II. Our conceptual framework

The framework we propose encompasses the concepts described in the framework proposed by Sacha et al. [4], our main modification being that we introduce a deep learning based latent representation of data as an intermediary between the data and the interpretable models (Fig. 2)

This representation space is based on the processing of the entire dataset to generate a model of the data, leveraging deep learning capacities to produce a rich, yet not directly interpretable, representation. This is why such a deep learned representation must be translated into various interpretable models that can then be presented to analysts, depending on their needs.

Figure 2 details several of the components of the proposed enriched visual analytics process: the general structure of immersive behavioral data ("Data" component), the generated latent space ("Latent Space" component) and how it is generated (gray arrow from "Data" to "Latent Space"), as well as the services that should be provided to an analyst using such a system.

A. Structure of immersive behavioral data

We define a trace as the set of all immersive behavioral data recorded during a session. It is composed of two distinct types of data: metadata and temporal data from various modalities.

While metadata provides contextual information about an immersive session (*e.g.*, age of participants), *temporal data* includes information such as the spatial trajectory of limbs, event logs or physiological signals, *etc*. Each piece of temporal data recorded is associated with a timestamp. Temporal data

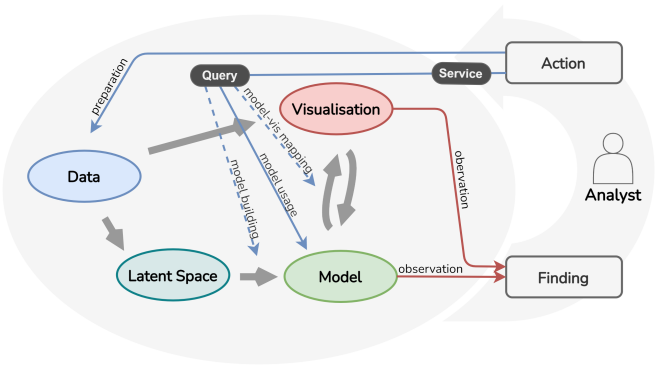

Fig. 2. The visual analytics conceptual model of Sacha et al [4] completed with a latent space, and some interactions between the analyst and the system (blue arrows) or their construction process (blue dashed arrows). The analyst generates findings by observing the models and visualizations (red arrows).

can be associated to one of the three modalities presented in figure 3 . Spatial data usually correspond to *position* and *rotation* of a user's limb. Event logs are collected at a given timestamp and are composed of multiple attributes. *Actor* is the entity that is responsible for the realization of the event (*e.g.* a user, or the system if the event is self-generated). *Verb* is a lexical item that denotes an action or a process (*e.g.*, "grab"). *Object* is the entity upon which the action is performed. For example, an object associated to the verb "grab", could be the virtual object (*e.g.*, "Butter"). *Properties* correspond to information that complete the description of the event (*e.g.*, "with: leftHand"). An *Event Type* is a generic way to describe event logs, for instance : "Event Logs that are linked with the Butter object", or "Events Logs that are emitted by the system".

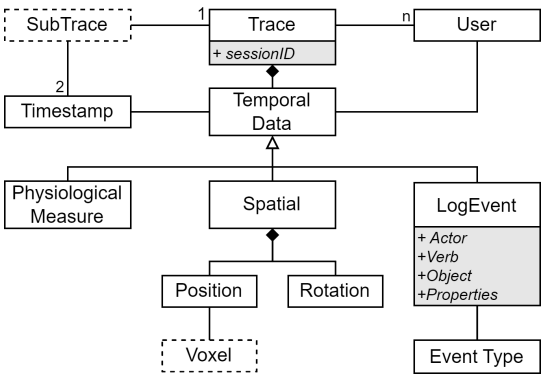

Fig. 3. A possible structure of immersive behavioral data which can be used to build a deep learning model.

All the data structures on which the deep learning model is built correspond to a vocabulary that is shared with the analyst. Yet, analysts always have a broader understanding of immersive behavior, which allows them to understand and manipulate data in ways a deep learning algorithm was not designed to address (dashed rectangles in figure 3). That is why various structures can be derived from the data, that may be of help to the analyst, yet unknown to the deep learning

algorithm. For instance, the algorithm may have been trained to handle spatial coordinates, but an analyst could have an intuitive understanding of situated 3D regions in the immersive environment (*i.e. voxels*, see figure 3). Another example could be that of a sub-trace, defined as the set of all the immersive behavioral data recorded between two timestamps during an immersive session. Such notion is flexible enough to cover the analysis of gestures (few seconds), tasks (few minutes) or even complete activities (up to hours).

B. Latent space

Deep Learning techniques enable the extraction of high-level features from complex data, including immersive behavioral data, without labels. Despite particularities inherent to the choice of the architecture, deep learning models yield vectorized representations of individual data elements. The latent space is the aggregation of these vectorized representations.

The latent space we are considering in our framework is the result of the training of a self-supervised algorithm on pseudo-tasks related to immersive behavioral data. Training on pseudo-tasks makes it possible to address the lack of labeled data by inferring higher-level knowledge about the data's inherent structure. For instance, a pseudo-task could be to reorder a sequence that has been previously shuffled. The definition of the pseudo-tasks is the main constraint that affects the structure of the produced latent space. Therefore, the behavioral assumptions used to define the pseudo-tasks have an influence on the structure of the latent representation space that will be built.

Here are a few examples of pseudo tasks used for self-supervised pre-training of deep learning models to improve their final performance on predictive supervised tasks. *Temporal coherence* considers the behavior of a user to be consistent over short periods of time [12]–[14]. *User coherence* considers that the behavior of a user has some similarity over time. This hypothesis is used especially in the field of user identification [15]–[17]. *Spatial independence* considers that behaviors are unrelated to the locations where they were recorded [12], [18], [19]. On the opposite, *spatial coherence* considers that behaviors are related to the location where they were recorded. This last assumption is supported by multiple visual analytics tools that enable exploring immersive data from environments that are more conducive to interactions [20]–[25]. However, to the best of our knowledge, no deep learning model has used it to constrain the training of their model.

C. Example scenario and related concepts

Besides the "classical" services it can offer to the analyst, a latent space-based visual analytics system can leverage the latent space to facilitate the exploration of the dataset.

In the remainder of this section we define a few concepts (namely, service, query, context, and object of interest) describing the process of an analyst making use of the latent space to explore the data to gain insights (see fig. 2).

Let us first remark that our approach is designed for analysts that are not introduced to deep learning concepts. Therefore, we consider the training of the deep learning model and the projection of the data into the latent representation space to have taken place prior to the exploration. This means that the analyst is not allowed to influence the meta-parameters (number of layers, decay of the learning rate *etc.*) of the deep learning model, nor to interact directly with the generated latent representation space.

Scenario. A short scenario describing how Ms. D., an analyst, explores immersive behavioral data may help to introduce our concepts. At the start of her exploration, Ms. D. faces the whole dataset and has no idea where to start. The system offers a few services, and she chooses "explore similar behaviors". This choice is translated by the system into a query which is processed to produce a list of 5 groups of sub-traces deemed similar in regard to behaviors. Each group is presented as a set of trajectories superposed to the map of the virtual environment. The second group G2, which appears to be related to a behaviors happening in a same place seems interesting, but there are too many of them, and Ms. D. cannot really make sense of it.

Therefore, she continues her exploration from all the sub-traces of group 2, and asks to be shown the associated prototypes, *i.e.*, the best representative sub-traces. 7 of them are then presented with the associated heads and hands trajectories, that she can easily look at. One of them is intriguing because one of the hand trajectory reaches the ground. By looking to the associated events, she notices that an object has been dropped before. Looking at the sub-traces associated to the prototype, she sees that many of them also contain a drop event, and it seems that some users tend to unintentionally drop objects around this place in the environment.

But are there any other places where objects would also be dropped? She selects the "show me associated places" service on the prototype sub-traces, precising that she is interested in the "Drop" event type only. She is then shown with a map of the places where the sub-traces corresponding to this criterion occur, that she can further explore, *etc.*

Definitions. Analysts ask the system for *services*, in various contexts, for different objects of interest. The services translate into queries that create and make use of interpretable models. An *interpretable model* is a model of immersive behavioral data that is understandable by the analyst, because it is directly related to understandable raw data. For example, in our scenario, the 5 clusters of similar trajectories is the first model that is interpreted and explored by the analyst. The 7 further subgroups of G2 correspond to a second model, which is presented throughout the associated prototypes. A *latent space-based service* is a service offered by the system that makes use of the latent space to generate an interpretable model, which can either be directly or indirectly (after further transformation) explored by the user. A service can have parameters. For instance, in our scenario, the clustering algorithm used to generate groups of sub-traces sharing similar behavior has a parameter to limit to 5 the number of groups generated.

Two notions can be of use to further specify a service. The *context* is the subset of the data that the analyst is currently

TABLE I
OVERVIEW OF THE KEY SERVICES AND RESULTS ({*object*} REFERS TO A SET OF *objects*).

Intent	Object(s) of interest		Service description	Result
Explore			Show singular behaviors in context	{sub-trace}
Summarize			Cluster behaviors into groups	groups of {sub-trace}
Summarize			Show the most representative behaviors	{sub-trace}
Locate	Sub-trace		Show similar behaviors in context	{sub-trace}
Identify	Sub-trace		Identify timespans best representing this behavior	{timespan}
Summarize	User		Cluster users according to their behavior	{user}
Browse	Voxel		Show behaviors that are specific to this voxel	{sub-trace}
Summarize	Voxel		Cluster voxels according to the behavior they contain	groups of {voxel}
Locate	Event type		Locate behaviors most influenced by this event type	{sub-trace}
Compare	Sub-trace	Sub-trace	Show the similarities between these sub-traces	{timespan} per user
Explore	Sub-trace	Voxel	Show voxels that contain similar behaviors	{voxel}

engaged with. For example, in our scenario, the context of exploration evolves when the analyst focuses on G2, the second group. An *object of interest* is the reference object upon which the service is specified. For instance, in our scenario: the prototype sub-trace is used as a reference object to induce a specific behavior search.

A *query* is the translation of a service into low-level prompts. Among these prompts are those that 1/ translate the service and its parameters into inputs for an algorithm exploiting the latent representation space, 2/ translate the output of such algorithm into an interpretable model, 3/ specify the way under which the resulting model will be presented to the analyst. For example, for the "show prototypes" service in our scenario, a prompt 1/ mobilized a clustering algorithm with G2 context, 2/ translated clusters from the latent space into 7 clusters of sub-traces, and 3/ extracted a prototype for each cluster, so that it could be shown. In figure 2, the first two prompts correspond to the "model building" arrow. The others are either related to model usage or model-vis mapping.

D. Latent space-based services for visual analytics of immersive behavioral data

Let us now describe what services can be offered by building on the intents proposed in [6]. First, *search* refers to the intent of finding interesting elements in the data, for known or unknown target (what is searched) and location (where it is in the data). *Lookup* is then when the analyst knows both the target and the location, *explore* neither the target nor the location, while *browse* corresponds to looking at unknown pieces of data in a known location, and *locate* to finding the location of known data. Second, *query* refers to further analysis of found element(s): *identify* consists in looking at supplementary information, *compare* in comparing various elements, and *summarize* in getting some overview.

A way to discover the services that can be offered over a latent space is to systematically explore these intents and apply them to either *none*, *one* or *several* objects of interest. This is what we propose in Table I, which shows some of the key services we were able to identify. These services provide an overview of high-level interactions that an analyst can perform on immersive behavioral data. The integration of the latent representation space enables new services and exploratory processes. For instance, the "Show similar behaviors in context" service leverages the latent space to find similar behaviors in context. The implicitly learned behavioral similarity is based on intricate cross-modal relationships that would otherwise be impractical to define explicitly. On the other hand, the "Show the similarities between these users" service enables the comparison of complete traces based on all of their immersive behavioral data, a comparison that would be impractical to consider visually due to the amount of information.

III. CURRENT STATE OF OUR IMPLEMENTATION

The immersive behavioral data for our study comes from a virtual kitchen project [1], involving 60 participants in 30-minutes sessions. We recorded headset and controller trajectories at 10 Hz and logged events during the virtual sessions (*e.g.*, `user grabs butter3 at 01"22'`).

To create a latent representation of this data, we adapted the model from Lin et al. [12], modifying it to train on both spatio-temporal and event log modalities. This baseline was selected for its superior performance in action recognition and its multi-head design which provides robustness to the learned latent space. This enabled us to develop a "semantic" latent representation of our immersive behavioral data, and we have began to design and develop a visual analytics tools (latent-space assisted) for data exploration.

IV. CONCLUSION

Our proposals open the way to a whole new class of visual analytics systems dedicated to the exploration of immersive behavior, that will benefit from the representation capabilities offered by deep learning models. More work, however, is needed at various levels to develop and validate the conceptual approach of latent space based visual analytics systems, together with real working systems. Future works on our side include: 1/ finishing the design and the development of our visual analytics system based on the network we have constructed, 2/ consolidate the evaluation methodology of such a system and evaluating it at various levels, and with experts; while also 3/ considering other deep learning architectures, notably transformer architectures that would enable the processing of immersive behaviors of any given duration.

References

[1] R. Malo, "Flexibilité psychologique, psychopathologie et réalité virtuelle : de la mesure subjective à la possibilité de mesure objective des processus transdiagnostiques," PhD in psychology, Nantes Université, Dec. 2023. [Online]. Available: https://theses.hal.science/tel-04587234

[2] L. Yuan. Echoes in the Gallery: A Collaborative Immersive Analytics System for Analyzing Audience Reactions in Virtual Reality Exhibitions. [Online]. Available: https://osf.io/zmyx9

[3] H. E. Yaremych, W. D. Kistler, N. Trivedi, and S. Persky, "Path tortuosity in virtual reality: A novel approach for quantifying behavioral process in a food choice context," *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 7, pp. 486–493, 2019.

[4] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1604–1613, 2014.

[5] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," in *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, S. J. Simoff, M. H. Böhlen, and A. Mazeika, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 76–90. [Online]. Available: https://doi.org/10.1007/978-3-540-71080-6_6

[6] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2376–2385, 2013.

[7] Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue, "Meta-transformer: A unified framework for multimodal learning," *arXiv preprint arXiv:2307.10802*, 2023.

[8] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6959–6968.

[9] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–6.

[10] S. Bianco and P. Napoletano, "Biometric recognition using multimodal physiological signals," *IEEE Access*, vol. 7, pp. 83 581–83 588, 2019.

[11] S. Ganguli, C. V. K. Iyer, and V. Pandey, "Reachability Embeddings: Scalable Self-Supervised Representation Learning from Mobility Trajectories for Multimodal Geospatial Computer Vision," in *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*, Jun. 2022, pp. 44–53, iSSN: 2375-0324.

[12] L. Lin, S. Song, W. Yang, and J. Liu, "Ms2l: Multi-task self-supervised learning for skeleton based action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. ACM, Oct. 2020. [Online]. Available: http://dx.doi.org/10.1145/3394171.3413548

[13] B. Kim, H. J. Chang, J. Kim, and J. Y. Choi, "Global-local motion transformer for unsupervised skeleton-based action learning," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 209–225.

[14] Y.-B. Cheng, X. Chen, J. Chen, P. Wei, D. Zhang, and L. Lin, "Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[15] C. Rack, T. Fernando, M. Yalcin, A. Hotho, and M. E. Latoschik, "Who is alyx? a new behavioral biometric dataset for user identification in xr," *Frontiers in Virtual Reality*, vol. 4, p. 1272234, 2023. [Online]. Available: https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2023.1272234

[16] R. Miller, N. K. Banerjee, and S. Banerjee, "Combining real-world constraints on user behavior with deep neural networks for virtual reality (vr) biometrics," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2022, pp. 409–418.

[17] C. Rack, A. Hotho, and M. E. Latoschik, "Comparison of data encodings and machine learning architectures for user identification on arbitrary motion sequences," in *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 2022, pp. 11–19.

[18] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2017, pp. 601–604.

[19] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9628–9637.

[20] M. Nebeling, M. Speicher, X. Wang, S. Rajaram, B. D. Hall, Z. Xie, A. R. E. Raistrick, M. Aebersold, E. G. Happ, J. Wang, Y. Sun, L. Zhang, L. E. Ramsier, and R. Kulkarni, "Mrat: The mixed reality analytics toolkit," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376330

[21] C. Javerliat, S. Villenave, P. Raimbaud, and G. Lavoué, "Plume: Record, replay, analyze and share user behavior in 6dof xr experiences," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 5, pp. 2087–2097, 2024.

[22] W. Büschel, A. Lehmann, and R. Dachselt, "Miria: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445651

[23] S. Villenave, J. Cabezas, P. Baert, F. Dupont, and G. Lavoué, "Xrecho: a unity plug-in to record and visualize user behavior during xr sessions," in *Proceedings of the 13th ACM Multimedia Systems Conference*, ser. MMSys '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 341–346. [Online]. Available: https://doi.org/10.1145/3524273.3532909

[24] S. Kloiber, V. Settgast, C. Schinko, M. Weinzerl, J. Fritz, T. Schreck, and R. Preiner, "Immersive analysis of user motion in vr applications," *The Visual Computer*, vol. 36, no. 10, pp. 1937–1949, Oct 2020. [Online]. Available: https://doi.org/10.1007/s00371-020-01942-1

[25] 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 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491102.3517550