Task-based methodology to characterise immersive user experience with multivariate data

Florent Alain Sauveur Robert, Hui-Yin Wu, Lucile Sassatelli, Marco Winckler

Task-based methodology to characterise immersive user experience with multivariate data

Florent Robert*
Université Côte d'Azur, CNRS, I3S, Inria, France

Hui-Yin Wu†
Université Côte d'Azur, Inria, France

Lucile Sassatelli ‡
Université Côte d'Azur, CNRS, I3S, France
Institut Universitaire de France

Marco Winckler§
Université Côte d'Azur, CNRS, I3S, Inria, France

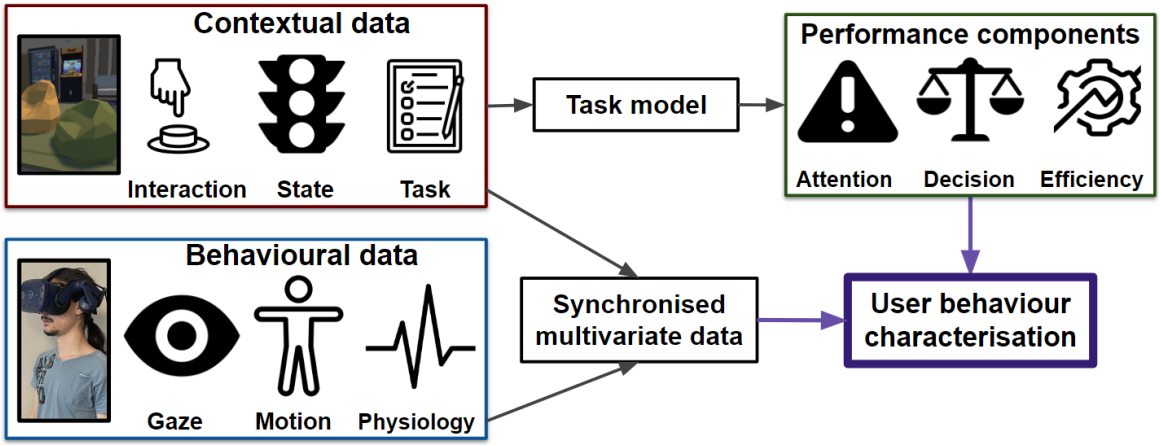

Figure 1: Our task-based methodology comprises three steps : (1) synchronise behavioural data with context data, (2) use task model to define performance baselines, and (3) characterise user behaviour based on their performances using multivariate data.

Abstract

Virtual Reality (VR) technologies enable strong emotions compared to traditional media, stimulating the brain in ways comparable to real-life interactions. This makes VR systems promising for research and applications in training or rehabilitation, to imitate realistic situations. Nonetheless, the evaluation of the user experience in immersive environments is daunting, the richness of the media presents challenges to synchronise context with behavioural metrics in order to provide fine-grained personalised feedback or performance evaluation. The variety of scenarios and interaction modalities multiplies this difficulty of user understanding in face of lifelike training scenarios, complex interactions, and rich context.

We propose a task-based methodology that provides fine-grained descriptions and analyses of the experiential user experience (UX) in VR that (1) aligns low-level tasks (i.e. take an object, go somewhere) with multivariate behaviour metrics: gaze, motion, skin conductance, (2) defines performance components (i.e., attention, decision, and efficiency) with baseline values to evaluate task performance, and (3) characterises task performance with multivariate user behaviour data. To illustrate our approach, we apply the task-based methodology to an existing dataset from a road crossing study in VR. We find that the task-based methodology allows us to better observe the experiential UX by highlighting fine-grained relations between behaviour profiles and task performance, opening pathways to personalised feedback and experiences in future VR applications.

*e-mail: florent.robert@inria.fr
†e-mail: hui-yin.wu@inria.fr
‡e-mail: lucile.sassatelli@univ-cotedazur.fr
§e-mail: marco.winckler@univ-cotedazur.fr

Keywords: virtual reality, user experience, multivariate data, task modeling, behaviour characterisation

Index Terms: Human-centered computing—HCI theory,concepts and models; Human-centered computing—User models; Computing methodologies—Computer graphics—Graphics systems and interfaces—Virtual reality

1 Introduction

In the early rise of virtual reality (VR), research in neuroscience such as that of Alcañiz et al. [1] have shown that interactions in VR stimulate the brain in ways comparable to real-life, physical interactions. This makes VR particularly promising for applications and research in training and rehabilitation, a few recent salient examples including the creation of virtual wheelchair training scenarios [23], sports simulation [41], or to recreate stressful situations for firefighter training [6].

However, when a user misstep occurs when carrying out a task, whether it is overlooking an important indication, forgetting necessary steps, or committing an error resulting in the failure of the task, finding the reason can be extremely difficult: the misstep can be linked to invisible cues such as the person's physical, attentional, or emotional state, triggered by specific scene elements. In the case of training scenarios that involve a sequence of varied, non-trivial tasks, carrying out a precise performance evaluation to gain user understanding becomes even more complex. Yet synthetic 3D environments in VR have this capacity, to allow the designers of a scenario to observe and highlight user behavioural indices in their relevant context. Building a proper understanding of granular user performance in the "lived" experience will bring multiple benefits: on the one hand, to better accompany users in their training and propose personalised feedback according to their performance and interaction profiles; on the other hand, on the system side, design improved affordances and guidance to the observed user behaviours and needs.

We thus propose a task-based methodology to help enrich and characterise the experiential aspect of user experiences (UX) that VR systems offer, by investigating the state of the system in relation to the state of the user [14]. Inspired by performance measures such as GOMS [36] (Goals, Operators, Methods, and Selection rules) as well as task modelling approaches such as CTT [35] and HAMSTERS [28] – popularly used in engineering interactive systems –, our method decomposes high-level VR scenarios and user task into lower-level sub-tasks (or sub-sub-tasks) that can be completed by the user in different ways. The originality of our approach then involves the use of task models to support the analysis of user behaviour on various levels of granularity (high and low-level tasks), aligning task execution in the scene context (e.g., current task, state of the scene, user position) with multivariate data describing the embodied UX (i.e. attention, motion, emotion). It is this alignment of user tasks with multivariate data that allows a better understanding of the performance in relation to behaviours for every low-level task (i.e., take an object, go somewhere) within a scenario. To evaluate a user's performance for each task, we defined three performance components (i.e., efficiency, attention, decision) with set baseline values to determine the success criteria for a given task. Once the tasks and scenarios are settled, our approach allows the comparative analysis of task performance and multivariate data that jointly characterise the immersive UX.

To illustrate the advantages, we apply the task-based methodology on the CREATTIVE3D multimodal dataset of user behaviour in virtual reality from our previous user study [32], which is publicly available [1]. This dataset fulfils two important criteria for our approach: (i) it includes multivariate data providing varied metrics to characterise the experiential UX (ii) these metrics can be spatio-temporally aligned with low-level tasks and scene context. The dataset contains a substantial amount of sensor data collected from 40 participants crossing virtual roads in VR, throughout six scenarios (with a varying amount of interactions and stress-inducing elements) and under four conditions (a combination of normal vision and simulated low vision, with real physical walking and simulated walking with a joystick), making a total of 24 scenarios. The multivariate data comes from integrated and external sensors (e.g., attention, motion, emotion) and system logs on contextual information (i.e., environment state, user interactions in the environment).

The proposed methodology introduces three primary advances for the analysis of VR training scenarios :

- Definition of user performances following three components (i.e. efficiency, attention, decision) and baseline values for each component to observe the performance missteps on a low-level tasks
- Characterisation of behaviour using multivariate data based on the performance missteps
- Creation of user profiles based on observed link between behaviour and performance missteps, to improve personalised feedback and VR experiences

The remainder of the paper is organised as follows. We first present in Section 2 the related work on the analysis of UX in VR to position our work. In Section 4, we detail the task-model method for fine-grained VR experience analysis, and show how we can achieve a fine-grained characterisation of user behaviour based on task performance. Then in Section 5 we show the results of our analysis on the variation in metrics for different performance components (i.e., attention, decision, and efficiency), which leads to individualised behavioural profiles in relation to the performance missteps. Finally, we discuss these findings, their implications, and present the next steps of this work in Sections 6 and 7.

[1] https://zenodo.org/doi/10.5281/zenodo.8269108

2 Related work

In this section we review the use of VR technologies to better understand users experiences, underlining those that apply this understanding for improving and personalising training experiences. Specifically, we look at (1) cognitive science approaches to Affective user understanding in VR with multivariate data (i.e. attention, motion, emotion), and how they are used to synthesise personalised training experiences based on observed user performance, (2) task modelling in human-machine interaction and robotics to represent human actions and task performance in a hierarchical and structured way, which enable a focus on the fine-grained analysis of user behaviour and task performance in a scenario.

2.1 Cognitive approaches to user understanding in VR

Measures of perception and emotion from the cognitive sciences have been popularly adopted to validate observations of user behaviour in VR applications, providing insight into the influence of the content on the affective user experience, surveyed by Luong et al. [25]. These approaches can involve explicit tools such as questionnaires or implicit data captured from sensors such as for motion or physiology. A number of individual studies have strongly inspired our work. As early as 1999, motion and specifically posture has been a subject of interest [20] to measure the level of body sway as a function of visual motion in VR. Jicol et al. [16] adopted questionnaires for emotional intensity and positivity, and employed structural equation models to find complex correlations between emotion, sense of presence, and agency, such as the observation that sense of agency augmented sense of presence only while mediated by an emotional fear condition. Seinfield et al. [37] measured performance and the sense of embodiment as a function of object placement and modalities of interaction (e.g., hand gestures, joystick, keyboard), finding a positive increase in performance and sense of embodiment for virtual hand paradigms. Using eye tracking and electrodermal activity (a.k.a. skin conductance) for a study of viewing 360°videos with various levels of emotional intensity, Guimard et al. [12] found that videos with highly salient visual content coupled with high user arousal allowed for a higher accuracy when predicting user attention. Keighrey et al. [17] investigate the possibility to evaluate the user's perceived quality of experience for virtual speech and language assessment applications on different interfaces (i.e., tablet, VR, AR) using heart rate and electrodermal activity measurements in addition to post-test questionnaires. These studies brought about new ways to analyse the global user behaviour within a scenario, and measure the behaviour variation between different scenario conditions using modern cognitive science approaches. They do not however conduct fine-grained analysis of user behaviour in relation to the scenario context, such as for scenarios with continuous lifelike training scenarios composed of multiple sub-tasks, to pinpoint the elements at a specific time of the study that trigger a user behaviour. The experimental conditions in such studies are also often difficult to replicate (e.g., framework reproducibility, differences in setup, data availability, population bias), limiting the possible conclusions to the ones reported in the work itself.

An important advantage of VR training frameworks is the capacity to adapt the scenarios to each person's needs. Systems that analyses their relation to the user, and use this understanding for personalizing the UX move towards the domain of experiential user experience. Rule-based performance metrics and baselines such as characteristics of the training scenario [24], completion time [33,38], and error rate [33] have been dominantly used as parameters for such adaptations. The more recent inclusion of user behaviour metrics as parameters for synthesising personalised training scenarios or adapting existing ones is gaining traction. Lang et al. [21] used gaze tracking and noted event baselines such as improper habits in driving scenarios, then used an optimisation approach to generate personalised map layouts for improving driving habits. Chen et al. [5]

designed a framework that analyses locomotion parameters such as pose accuracy and motion speed to generate new target points that train specific stances. Dey et al. [8] explored the use of EEG to estimate task load from alpha peaks and adapt the difficulty level of a target selection task.

2.2 Fine-grained analysis through task models

Task analysis is a cornerstone process for understanding how users perform their tasks and how they achieve their intended goals [9]. Task analysis is often done though direct observation of users interacting with the system. Task models provide an abstract structure for the analyst to organise information gathered during task analysis that can be further detailed where needed [30]. Santoro [35] characterised task models representations of the human activity in a hierarchical structured way (such as a state machine, action tree, or graph) allowing the analysis of task feasibility (formal demonstration that a task can be achieved) and (estimated) user performance. Task models have strong advantages to support task analysis, including coping with complex scenarios (by structuring tasks and sub-tasks), supporting abstract and generalised reasoning about tasks (beyond individual's cases of use), and serving as non-ambiguous documentation of tasks and observations made [9, 39].

Such task models are essential to a fine-grained analysis of user tasks. They bind together a deep understanding and characterisation of the interactive system constraints, the user performance, and the qualitative aspects of the experience. For this purpose, various works have adopted task models for the evaluation on various levels of granularity. KLM-GOMS [36] is a well known method used for traditional media to quantify how much time users take to perform a given task by aggregating the duration of low-level tasks (e.g., point mouse to a target = 1.1 seconds, move hand to keyboard from mouse = 0.4 seconds) that compose more complex interactions. Moving from traditional media to more recent works using KLM-GOMS method, we find Rice et al. [31] who investigate the description of different inputs composing touchscreen and mobile device interactions (e.g., pinch, zoom, tilt). Guerra et al. [11] extended the available interactions and measured interaction time to provide a new GOMS methods to evaluate task performances (e.g., grab an object, using teleportation movement) for augmented and extended reality systems. Zhou et al. [44] built a method focused on quantifying user performances using a virtual hand. These works aim primarily at the description of a scenario purely based on the time spent on each tasks, to measure the efficiency of performance, without considering behavioural performance components such as the user attention or emotion. Whilst tasks models have strong advantages in evaluating the usability and experiential UX of various interactive systems (including applications such as web applications [39], iTV [3], systems of command and control [30]...), few studies address 3D interaction in VR systems.

In robotics, there is a strong interest in performance analysis on fine-grained slicing of a scenario, such as to pinpoint performance flaws. Analysing and identifying when there is a flaw and the cause are crucial to improving the robot and avoiding future failures, corresponding to a similar need in VR applications where adapting and personalising UX are concerned. For example, Lee and Lozano-Pérez [22] focus on the design of interaction graphs that finely segment tasks carried out by a robot, in order to precisely detect when an error occurs and proceed to a failure mode. Kroemer et al. [19] focused on the prediction of phases of manipulation by a robot using a probabilistic model to represent the steps composing a manipulation task. These existing works focus on performance analysis on the system level, without taking into account the performance of a potential user interacting with the system. To adapt these methods to VR training scenarios, a model allowing the joint performance analysis and characterisation of the system and the user who is interacting is required.

The above review highlights the need of experiential UX methodologies for the granular observation of the user performances and behaviours in continuous lifelike VR scenarios, in relation to the scenario context. We propose an approach inspired by human-computer interactions and robotic task models, slicing a scenario in multiple low-level task (i.e., take an object, go somewhere, interact with something) each with precise performance criteria synchronised with multivariate metrics, opening multiple possibilities for granular analysis in VR:

- Definition of user performances according to different components (e.g., efficiency, attention, decision) for multiple types of tasks
- Characterisation of behaviour using multivariate data (e.g., gaze, motion, emotion)
- Discovery of user behaviour profiles in correlation with the performance missteps to explore possibilities of improvement for personalised feedback and experiences

Being able to identify and apply the metrics that are most efficient at characterising user behaviour for a given type of task is very valuable, as it help to define with more accuracy when a user misstep or system limitation happen, which can be translated to a certain form of refinement or user guidance to improve the global experience.

3 Dataset

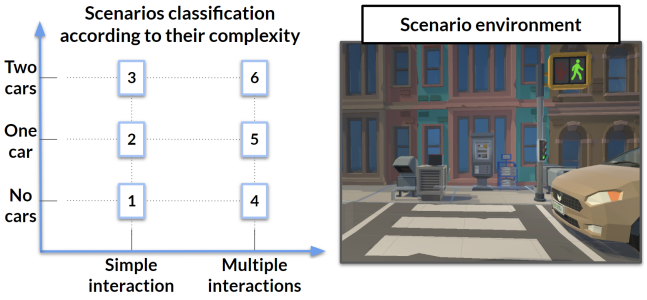

Figure 2: Our study in VR [32] that produced the dataset. In the study, users crossed a virtual road with simulated car traffic, in six scenarios with a variable amount of interactivity and number of cars on the road.

The dataset we chose to exemplify our task-based methodology is the CREATTIVE3D multimodal dataset of user behaviour in virtual reality, publicly available from the VR study we conducted in our previous work [32]. The dataset presents the two characteristics necessary for our methodology: (1) a clear temporal segmentation of tasks and sub-tasks in scenarios, and (2) behaviour and 3D context data containing timestamps for synchronisation. The study involved 40 participants who performed scenarios of around 2 minutes during which they had to achieve multiple tasks including to cross a virtual road with simulated car traffic. This dataset is composed of 6 scenarios with varying interactivity and stress-inducing elements, as shown in Figure 2. Multivariate data was recorded across all scenarios :

- *System scene logs* recording on Unity with the GUsT-3D tool [32] at 10Hz containing time-stamped entries of (1) the current state of the environment (i.e., traffic light colour, cars position, object positions), and (2) the object interactive properties (e.g., the book is on the table, the box is grabbed by the player)
- *System user logs* recording on Unity with the GUsT-3D tool [32] at 10Hz containing time-stamped entries of: (1) the user's

Table 1: Data contained in a dataframe of synchronised behavioural and 3D contextual data, decomposed by tasks. Task objects row gives the name of 3D objects relevant for the current task. Performance metrics row list the variables used to identify the missteps during the task.

Task	# of logs	Task objects	Performance metrics
Get the key	1570	key	efficiency (unixTimestamp)
Open the door with the key	480	key, door	efficiency (unixTimestamp)
Get the trashbag	2243	garbage bag	efficiency (unixTimestamp)
Cross the road safely	2929	car, traffic light, traffic light button	efficiency (unixTimestamp) attention (lookedAtItemName) decision (carHonk, trafficLightColor)
Put the trashbag in the trashcan	642	garbage bag, trashcan	efficiency (unixTimestamp)

visual attention in the environment (i.e., object on gaze focus, objects in the field of view, object in the centre of the vision), (2) the user interactions with the environment (i.e., grab an object, drop an object, press a button, open a door, location in the environment), and (3) the current task they are performing (e.g., take a box, cross the road)

- Motion capture recording with Xsens MVN Awinda suit at 60Hz how users physically move in the environment with 17 sensors position and rotation (x,y,z) for head (1), torso (4: shoulders, hip, and stern), arms and legs (8: upper and lower limb), and feet and hands (4)
- Physiological sensors recording with Shimmer GSR3+ at 15Hz how are users physiologically feeling through electrodermal activity (EDA, a.k.a. skin conductance) and heart rate (HR).
- Gaze and head tracking recording with HTC Vive Pro Eye at 120 Hz of gaze and head movement for left, right, and cyclopean eye (combined gaze vector of both eyes), with the following data entries: gaze vector (x,y,z) , pupil size, eye openness percentage, and data validity mask.

In a pre-processing step, we first synchronised all the data mentioned above using the system user logs as the base timeline. The motion, gaze and physiological data were all recorded in their respective software and data frequencies, with one record per scenario for gaze, one record every six scenarios for motion, and one record per user for the physiological data. The dataset provides synchronised data entries with the system user log data using Unix Timestamp values, with each type of data sliced according to the start and end time of the scenarios, making a total of 24 files per data type for each user, constructing a pool of 960 files per data type. The gaze data is processed to produce point of regard (POR), which were directly calculated on the Unity scene of the experiment through raycasting the gaze vector with the 3D scene mesh. From this step, we converted all data to *csv* format and used the python pandas library to interpolate the data at 125Hz, which allows us to obtain a dataframe composed of synchronised context and multivariate data, for each user and each scenario. Figure 3 provides a sample visualisation of the multivariate data for one user in one scenario: the raw EDA values (in microsiemens μS), POR (x,y,z) (in meters), and inclination of the center of pressure (COP) motion (x,y,z) (in degrees), all synchronised with the context of the experience.

Table 1 highlights the most relevant data contained in a synchronised dataframe and the variables used for our performance baseline (Section 4.2) and the behavior metrics (Section 4.3). Here are some of the meaningful variables contained in the dataframe :

- *unixTimeStamp*: unix timestamp of the data entry,
- *position*: (x,y,z) position of the player in the 3D environment,
- user state variables: *currentTask* (the name of the task the user is performing) and *item* (object the player is holding in VR),

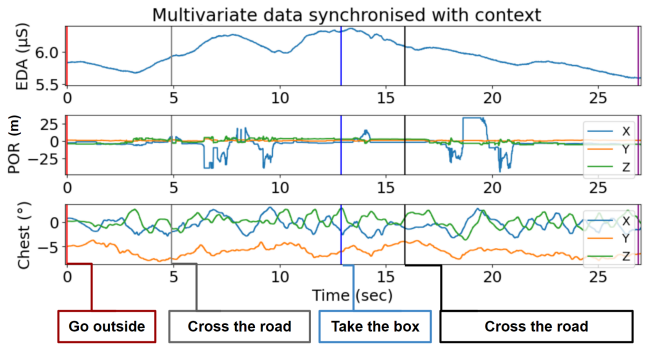

Figure 3: An example of the raw multivariate data (i.e., skin conductance (EDA), point of regard positions with x, y, z axis (POR), chest rotation value) synchronised with the context for one user in a single scenario. During this scenario, the users carried out four tasks: (1) Go outside their house, (2) Cross the road safely, (3) Take the box, (4) Cross the road safely to come back home

- attention data: *lookedAtItemName* (the object the user is looking at) and *inViewItems* (all objects in the user's field of view),
- scene state variables: *trafficLightColor* (current color of the traffic light) and *honk* (boolean indicating if a car is currently honking),
- *EDA*: raw EDA value in microseimens μS,
- point of regard: *PorXYZ* as a (x, y, z) position,
- motion capture sensor positions: *MotionPos* including (x, y, z) positions for each of the 17 motion sensors).

The synchronised dataframes have been made available in the latest version of the dataset repository to ensure the reproducibility of this work.

4 Methodology : Task Model

To tackle the challenges of characterising experiential UX on a granular level for lifelike VR scenarios, we conceived a task-based methodology inspired from existing methodologies in human-machine interaction and robotics to finely observe and characterise user performance using multivariate data. This methodology can be applied widely to studies presenting two characteristics : (1) a clear temporal segmentation and (2) a set of synchronised behavioural and 3D contextual data. The former allows us to build a task model that characterises the sequence and success criteria for each task, and more importantly, detect the missteps, i.e., when a task is incorrectly carried out. The latter then allows us to perform a fine-grained analysis of behavioural metrics in correlation with the 3D context at the moment a misstep occurs.

In this section we first present the task-model definition from high-level scenario to low-level task. Based on the task model, we then define the three baseline components that will be used to identify user performance missteps : efficiency, attention and decision. Finally, we present and justify the selected behavioural metrics derived from the multivariate data used to characterise the user behaviour in relation to performance missteps.

4.1 Task model conception

Our first step is to build our task-based approach inspired from task models [35], describing a scenario as a graph of high to low-level tasks to perform, in a precise order. Each task is composed of precise set of criteria that classifies the execution of the task as a success, or with a misstep.

We constructed our task-model using Hamsters tool[2] [29]. Each task has its own separate task-model with the following elements:

- task description from the high abstraction level – for example "open the door" – to the sequence of motor tasks composing this task – "find the key", "walk to the key", "press the trigger to grab the key", "walk to the door", "interact with the door" – in order to define precisely the ideal sequence of actions the user is expected to perform.

- precise definition of objects included in a task and the interactions expected with them (i.e., touch, grab, look at, place in, place on), which are the criteria to the success of a task.

Our scenarios feature 9 high-level tasks in the following sequence : (1) Open the door, (2) Take the garbage bag, (3) Go outside your home (4) Cross the road safely, (5) Put the garbage bag in the trashcan, (6) Take the box, (7) Cross the road safely, (8) Come back home, (9) Place the box on the table. Each scenario is a subset of tasks from this selection. For example, the simple interaction scenario with one car (Scenario #2 in Figure 2) is composed of the task sequence $3 \rightarrow 4 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 9$. All scenarios are at least composed of one interactive task (e.g., take the box / garbage bag), and one road crossing task.

The Figure 4 illustrates a shorter, plausible scenario composed of only three high-level tasks, with a simplified view of the three task models for each high-level task. Each sub-figure highlights the expected interactions and the sequence to perform them in order to succeed at the task. In the "cross the road safely", the user must, for example, cross the road while the traffic light is green, after checking that the car has stopped. The sequence however can be done in multiple orders creating multiple variation of a same scenario (i.e. State 1: Look at green light, look at the stopped cars, cross the road; State 2: Look at red light, wait, look at green light, look at stopped cars, cross the road). The individual task models are made for all scenario variants. As such, we could for example look only at the road crossing task for one user in a single scenario individually, or we can aggregate all instances of the road crossing task across all users and scenarios in one analysis, and all instances of the opening door task for a different analysis.

The road crossing task is the longest and most complex task, with multivariate metrics to observe and analyse. It also incorporates all of the interactive mechanisms that appear in the other tasks. In the rest of the paper, we therefore use this task as the example to demonstrate our task model approach to UX analysis and performance evaluation.

4.2 Performance baselines

From this task model, we can already begin to observe the different components that could be used to quantify the success at performing

[2] https://www.irit.fr/recherches/ICS/softwares/hamsters/

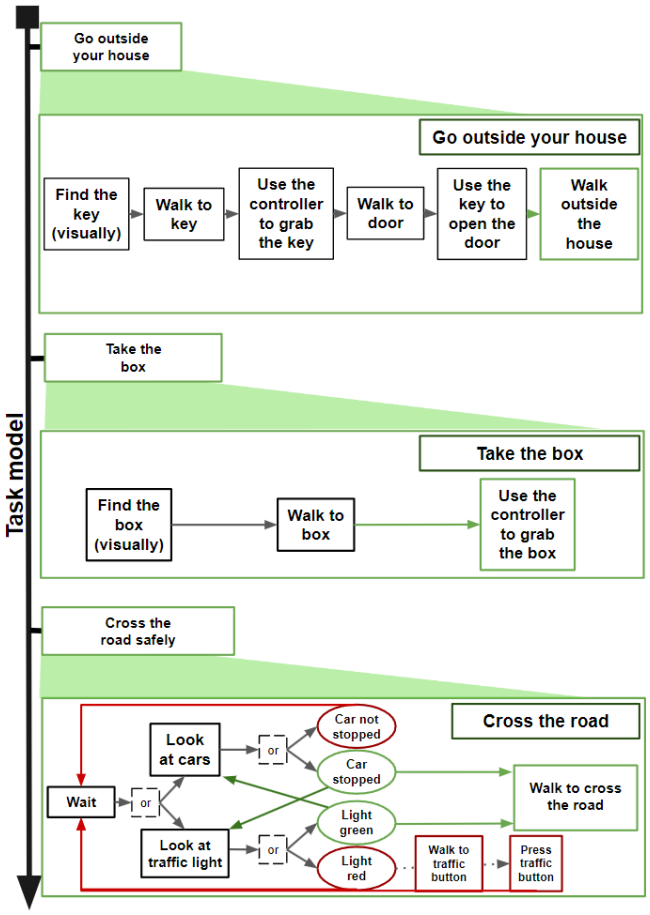

Figure 4: A simplified task model for three high level tasks: (a) "Go outside your house", (b) "Take the box", (c) "Cross the road safely", to properly achieve the task, the user is expected to perform the right interactions at the right time, and in the case of crossing the road, also look at traffic light and cars, and cross when the light is green and the cars are stopped. Rectangles indicate tasks and ovals indicate object states.

the task (e.g., cross in time, cross during the green light, look left and right before crossing, don't get hit by a car). Out of this model, we can define the performances baselines values that an analyst would like to observe and analyse when a user is crossing the road.

We therefore defined three components needed for the successful completion of the "cross the road safely" task : (1) efficiency : the user must finish the task within a given amount of time, (2) attention : the user must pay attention (i.e., look at) certain elements during the task to know when to do an action, and (3) decision : the user must make correct choices in the sequence of executing actions.

First, we define the efficiency component, that is the time limit to complete the task. We calculated our baseline value on the existing human-computer interaction metric GOMS, which is a well known predictive model to quantify how much time it will take to perform a given task. Based on the extension to VR proposed by Guerra et al. [11], we defined times to grab object, drop object and press the traffic light button. For the task of physical walking, we estimated the preferred walking speed of users in VR at 0.9m/s based on Wodarski et al. [40], resulting in the list of actions presented in Table 2 to define the efficiency baseline value of a task.

When starting a scenario, we take into account that users will need a short adaptation time to look around at the environment,

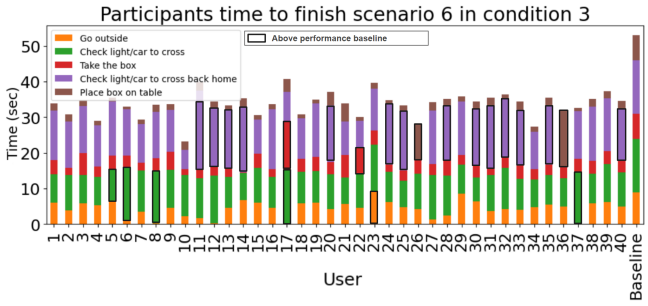

Figure 5: Time taken by all 40 participants (sorted by study completion order) to finish the tasks of the scenario #6 (multiple interactions with two cars). The baseline column on the far right shows the baseline value for each individual task stacked. All the tasks where the user surpasses this baseline are outlined in black.

Table 2: List of actions and the associated expected time to perform them in VR, inspired from GOMS methodology, to define task efficiency baseline value.

Action	Time
Scenario starting delay (S)	2.0 seconds
Audio instruction delay (I)	1.0 seconds
Grab an object (G)	2.6 seconds
Drop an object (D)	2.6 seconds
Press a button (P)	1.45 seconds
Red traffic light duration (T)	8.0 seconds
Walk physically (W)	(Distance / 0.9) seconds

corresponding to the time *S*. For every task, users get a short audio instruction with task instructions, which corresponds to the time *I*. When users want to cross the road, the time they will have to wait for the light to go from red to green corresponds to time *T*.

Figure 5 allows us to visualise the time participants took to finish each task in scenario #6 (multiple interactions with two cars) under one experimental condition. The column *Baseline* in Figure 5 shows the efficiency component baseline defined for every task of the scenario. In this scenario, out of 200 tasks performed, we can see that 24 efficiency missteps occurred (i.e., the execution time surpassed the baseline value), highlighted by the black frames.

This definition process is also done for the two other performance components: attention and decision, which are a more straightforward. The attention baseline defines objects in the scene that the user must look at. The decision baseline designates certain states under which objects must be when the user executes a task. The right-most column in Table 1, summarises the the data entries in the dataframe that are relevant to each performance component for each task. Using these data entries, Table 3 then lists in detail the three performance components for each task the same scenario in Figure 5. The efficiency baseline of each task is defined as the sum of the actions comprising the task listed in Table 2. For example, during the second task "Cross the road safely", to fulfil the efficiency component, users must finish the task within 13.50 seconds (i.e., the user can wait 1 second following the audio instruction to start the task, wait up to 8 seconds for the traffic light to turn to green, and finally walk a distance of 5 meters at a speed of 0.9m/s, making a total expected duration of up to 13.50 seconds to complete the task); to fulfil the attention component, users must look at the cars and traffic light prior to going on the road; finally, they must also cross road while the traffic light is green without being honked by a car that wants to pass during their own green light to fulfil the decision component.

After defining these performance components, we can then group the users in each task into those who fulfilled the performance baselines and successfully executed the task under the baseline value, and those who did not, thus incurring a performance misstep. We can then correlate user behavioural metrics to observed performance missteps, which allow a more fine-grained analysis of the behaviour in relation to the performance.

In order to carry out our task-based analysis, the field *currentTask* in the synchronised dataframe is used to partition a scenario file in task groups corresponding to the rows in Table 1. With this partitioning, we apply our baseline values the relevant variables, to characterise if a user did a misstep for the given task. Again using the example of the "Cross the road safely" task, the missteps for the three performance components are well-defined :

- efficiency : the *unixTimeStamp* value gap between the beginning and the ending of the task is above the baseline value,
- decision : the user crossed the road while *trafficLightColor* value was at "red", or during the road crossing task, the *honk* value has been equal to true at least once,
- attention : the user crossed the road without *lookedAtItemName* value being equal to "car" at least once during the task.

4.3 Behavioural metrics

The last step in our approach is to select the metrics with which we wish to characterise the user performance. It is important to note that multiple possibilities can be imagined from such a dataset for each data type. For example, the emotion data is composed of both electrodermal activity and heart rate information. From motion capture data, gait characteristics, stability, limb coordination etc. can be derived. In this work we limit our analysis to only a subset of these metrics to illustrate our approach, using the following principles to help us in our choice:

1. Only one metric is chosen for each data type (i.e. motion, emotion and attention), since the goal is to exemplify our task-based methodology along with its flexibility to be applied to diverse type of data, and not present an overall analysis of all possible metrics.
2. Avoid metrics that are directly related to the person's absolute physical characteristics, such as height or step size, which can be correlated to additional factors such as gender, which could merit new sets of research questions into other aspects of user experience analysis currently outside of the scope of this work.
3. Prioritise metrics that have already been of interest for existing studies in VR, and are also relevant to the performance baselines in our task model, which highlights the added value of our task-based approach for fine-grained analysis.

Emotion we select skin conductance (or electrodermal activity, EDA) which shows the level of arousal of the user, indirectly indicates the level of stress the user is experiencing in the long term and the intensity of their timely physiological responses to stimuli, represented by skin conductance response (SCR) peaks. Skin conductance (EDA) is often used in the literature ([42], [13]) to evaluate the arousal of the user considering multiple context with more or less stress-inducing elements. This metric is particularly interesting in the context of this dataset which includes scenarios with more or less stress inducing elements that could lead to missteps on user performance and observable variation on user behaviour.

Table 3: The performance baseline values defined for the scenario six of the experiment. When crossing the road in this scenario, the user is expected to look at the car and the traffic light, cross only when the traffic light is green and the cars are stopped, in a total of 13.50 seconds.

Task	Efficiency	Attention	Decision
Go outside the house	S+I+W=2+1+3.5=6.50s	\	\
Cross the road safely	I+W+T=1+4.5+8.0=13.50s	Look at : Cars, Traffic light	Cross : Traffic light green, did not get honked
Take the box	I+W+G = 1+3.4+2.6 =7.0s	\	\
Cross the road back home	I+W+T=1+4.5+8.0=13.50s	Look at : Cars, Traffic light	Cross : Traffic light green, did not get honked
Place the box on the table	I+W+D=1+3.4+2.6=7.0s	\	\

Attention we choose gaze fixation duration (GFD), which has been used to observe visual recognition processes in Chan et al. work [4] or gaze behaviour for sports in Klostermann et al. [18], who have identified a potential link between longer duration of gaze fixations for professional players compared to amateurs. We want to investigate notably if GFD is potentially a metric that could help characterise attention and decision performance missteps, such as being correlated to shorter average fixations.

Motion we choose center of pressure inclination (COPI), adopted in existing studies to measure age and height effect on body balance through center of mass (COM) and center of pressure (COP) inclination [15], or to measure body sway when viewing moving visuals in VR headsets [20]. It is an important metric to measure body muscle stress [43], respiratory mechanics [27], and body balance [10]. The stability of users is a major concern for VR experiences that involve physical walking, making this metric interesting in the context of mobility in VR for our selected task, where there is physical walking involved to cross a road. We therefore investigate the link COPI could have with performance.

Setting out from the synchronised dataframe, we process each data type to calculate the above-mentioned metrics. The skin conductance data (EDA) is analysed with the Python toolkit Neurokit [26], as was done by Guimard et al. [13] in which the phasic component is calculated as the first derivative of the normalised EDA, allowing us to observe short term variations of physiological response in relation to the scene context. For gaze fixation, we used the I-VT algorithm [34] to obtain the list of fixations and their duration. Center of pressure inclination is computed on raw motion data records, using chest and ankle segments to compute the body angle, with a higher value when the body is more inclined and zero meaning the body is 90°vertical.

This three step process: establishing the task model, defining performance components and baseline values, and finally selecting relevant behaviour metrics comprises our methodology which allows us to (1) finely segment tasks into identifiable actions, (2) define which components of performances are evaluated and how they are quantified for each tasks done by the user, in order to classify the performance result (i.e., success or misstep), and (3) analyse correlations to behavioural metrics.

5 Results

In the previous section, we presented the multivariate data synchronisation with context, the task model based on the multivariate data, establishing baselines to define the components on which performance is evaluated, and selection of metrics for characterising user behaviour. Here we take a deep dive into the analysis of user behaviour using multivariate data based on performances, and further investigate on the possibility of identifying profiles of users whose performances are similar on certain observable metrics. We continue using the road crossing task and the three defined performance components and baselines: (1) efficiency: the time to finish the task, (2) attention: the elements looked at during the task and (3) decision: the choices made by the user to finish the task.

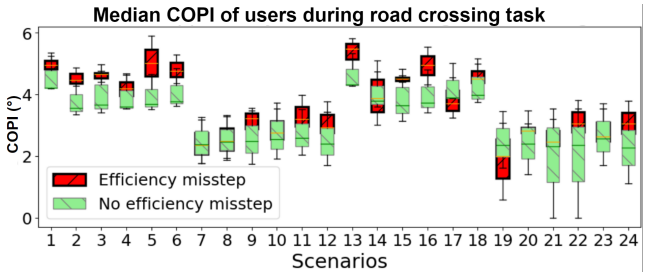

Figure 6: Median COPI of all users for each scenario of the experiment based on the efficiency component of performance. The red boxes represent the COPI values for TM_e, and the green boxes represent the COPI values for non TM_e.

The dataset contains the record of 40 participants who each completed 24 scenarios, for a total of 960 scenarios with road crossing tasks to observe. Using the performance component baselines, we can identify the missteps for each performance component. We then analyse the behavioural metrics selected: electrodermal activity (EDA), center of pressure inclination (COPI), and gaze fixation duration (GFD) in order to characterise the behaviours that are in relation to performance missteps.

The following notation is used:

- TM_a, TM_e, and TM_d denote a task with the occurrence of an attention, efficiency, and decision misstep respectively
- the metrics are abbreviated as COPI, EDA, and GFD for center of pressure inclination, electrodermal activity, and gaze fixation duration respectively

5.1 Behaviour at a scenario level

Out of the 960 scenarios with road crossing done by the 40 participants, 630 missteps of either attention, decision or efficiency were observed. The majority of the missteps were TM_a representing 47.9% (302) of all the misstep occurrences, then 27.1% (172) for TM_e, and 24.9% (156) of TM_d. The average missteps for users was 7.6 TM_a, 4.3 TM_e, and 3.9 TM_d.

For the first part of the analysis, we would like to observe potential trends that could signify a link between one performance component and one behaviour metric. We take the road crossing task data from the 24 scenarios from all users, each classified according to the three tasks performance components, resulting in a dataframe where each row represents the user, scenario, success or misstep associated to each performance component, and the average value of each behavioural metric. To identify if the distribution for each metric is significant between the TM_c and non-TM_c groups for a given component c, we then compute the null-hypothesis significance testing with p-values ([2, 7]). As shown in Table 4, we can find significance ($p < 0.05$) for TM_a+GFD and TM_d+GFD, and strong significance ($p < 0.001$), for the TM_e+COPI, TM_a+COPI, TM_a+EDA, and TM_d+COPI. The median values of each metric are

Table 4: The average values of each metric for tasks with misstep TM_c on a performance component c and those without a misstep. We calculate the p-value significance of each behavioural metric in combination with the performance component. The values in bold indicate significance $p < 0.05$. We can for example observe a strong significance in COPI when comparing the groups with and without TM_e.

	Efficiency			Attention			Decision		
Metrics	COPI	EDA	GFD	COPI	EDA	GFD	COPI	EDA	GFD
TM_c median	3.72°	4.56e-4	153.23ms	3.70°	5.37e-4	153.87ms	3.61°	4.93e-4	153.91
Non TM_c median	3.28°	4.52e-4	148.11ms	3.26°	4.55e-4	148.78ms	3.24°	4.55e-4	147.94
p-value	**4e-8**	0.582	**0.084**	**6e-14**	**4e-7**	**0.010**	**2e-4**	0.159	**0.011**

also shown for the different group and component combinations. We can see an increase of COPI, EDA, and GFD for the different missteps compared to the scenarios without missteps. If we look in detail at Figure 6, we can visualise the trend of the average COPI of three users. The group of tasks with performance missteps has a globally higher median COPI than those without, the exceptions being scenarios 14, 17 and 19.

5.2 Behaviour on user level

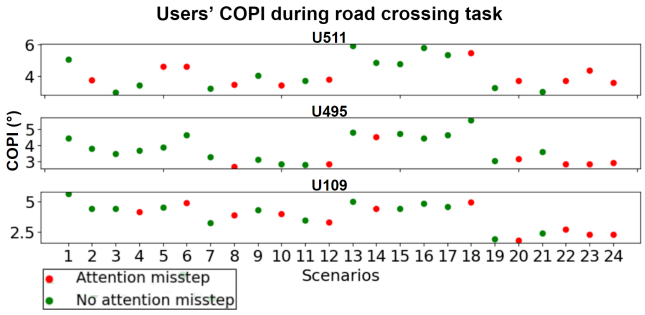

Figure 7: Visualisation of the median COPI value of each scenarios with road crossing for the users U511 (top), U495 (middle), U109 (bottom) with significant negative COPI variation in TM_a tasks. A red dot represent the median COPI value during road crossing task for a scenario done with an attention misstep, a green dot represent the median COPI values during road crossing task for a scenario without attention misstep.

In order to observe on a more fine-grained level if a user presents specific behaviours in relation to their performance, we conducted the same type of analysis for users individually, looking at each road crossing task during the experiment, grouping their scenario data based on success or misstep in a performance component. On these groups of behavioural metrics classified by performance, we computed the p-value in order to identify those users whose behaviour is significantly different for a given metric based on their performance. For this analysis, we kept only users with a p-value below 0.05 and at least 10% of tasks with occurrences of missteps in order to have a sufficient amount of data to compare between the performances groups. The users included this analysis are shown in Table 5.

These results outline unique behavioural profiles. We found at least two people per combination of performance-metric, and can observe that there is rarely one consistent trend of behaviour: significantly higher for one user but the exact opposite for another. Four interesting exceptions exist: a higher EDA for TM_e, a higher GFD for TM_a, and a lower COPI for both TM_e and TM_d. First, the revelation of the relation between the gaze metric (GFD) with attention missteps makes logical sense, as the user does not switch their gaze as frequently to take into account various pieces of information. The relation between emotional metric (EDA) could also be an indicator of the level of stress the user is experiencing. Finally, a lower COPI indicates that the person is more upright, which can be a sign of less engagement or walking slower. The observations can then facilitate the conception of hypotheses to further validate these relations between the performance and behaviour. We can visualise more in Figure 7 the COPI of users U511, U495 and U109 who have a significant difference in value when committing a TM_a, with a decreasing tendency of their COPI of -14% to -18% compared to non TM_a.

6 Discussion

We have presented a task-based methodology to evaluate and characterise experiential UX, taking inspiration from existing work in human-machine interactions and robotic task modelling, which have not yet been fully explored and used for analysis of VR experiences. Our methodology proposes new ways to evaluate the UX by looking at the fine-grained behaviour synchronised with context on three performance components, and finally allowing the characterisation of specific user profiles. By applying the methodology on a selected parameter of performance (i.e., attention, efficiency, decision) with suitably selected behavioural metrics, we can observe whether a metric could potentially be interesting for a better characterisation of the UX in a given context, task, or type of performance measure.

While our analysis and results have mainly been focused on a deep dive into the performance of a single road crossing task, it has already allowed us to observe some user behaviour profiles, and highlight the metrics are potentially more effective for the characterisation of the experience, such as gaze for tasks requiring visual attention, and emotion for task efficiency. We would also like to outline the inclusion of center of pressure inclination (COPI), which is important for tasks that require spatial displacement within a limited amount of time. We believe this is a novel and important contribution for VR that can benefit applications such as training or rehabilitation applications, where the understanding of a user's difficulties and needs when working with a VR system is crucial to improve the both global experience and provide personalisation.

One major strength of the methodology is its flexibility, as it is applicable to various types of studies, allowing the definition of performance parameters according to the task, and the analysis of corresponding behavioural data. In this work, we demonstrated this analysis on a pre-existing dataset, applying new baselines on performance that were not part of the original study design.

The dataset presented in this paper proposes notable advantages in providing synchronised multivariate data with the low-level task context for fine-grained analysis. The public availability of more similar dataset proposing similar features can allow us to flexibly adapt our evaluation methodology to a wide variety of tasks and metrics, which we believe can open new doors for user understanding in media with such rich interaction possibilities as VR. It will also potentially allow validation of hypotheses across different study designs to identify global trends of human behaviour.

However, this evaluation method alone is not sufficient to characterise the complete user experience. It would be better coupled with post-study interviews or qualitative studies. On its own, it is limited to identifying the correlation between behaviour and performance, but we cannot claim causation on whether a performance parameter is the cause of a behavioural metric variation, or vice versa. We

Table 5: Individual users whose values for a given behavioural metric are significantly different ($p < 0.05$) between tasks with and without missteps. *uID* corresponds to the ID of the user, *Misstep* corresponds to the percentage of tasks that have a misstep for a performance component, and *Var* corresponds to the variation (with positive/negative sign) observed median of the metric between TM_c and non TM_c for the given performance component *c*

	Efficiency misstep				Attention misstep				Decision misstep			
	uID	**p-value**	**Misstep**	**Var**	**uID**	**p-value**	**Misstep**	**Var**	**uID**	**p-value**	**Misstep**	**Var**
COPI	851	0.034	54%	-29%	321	0.020	29%	+56%	666	0.001	13%	-43%
	963	0.009	42%	-22%	511	0.010	46%	-14%	877	0.004	25%	-40%
					495	0.031	29%	-15%	828	0.026	38%	-19%
					109	0.029	46%	-18%				
EDA	981	0.002	21%	+55%	748	0.037	62%	-18%	666	0.001	13%	-71%
	361	0.006,	25%	+47%	091	0.047	50%	-35%	981	0.016	13%	-46%
	658	0.009	17%	+52%	213	0.004	21%	+71%	361	0.021	25%	-50%
									446	0.026	29%	+73%
GFD	361	0.032	25%	+27%	666	0.020	42%	+28%	748	0.017	21%	+33%
	940	0.029	17%	-29%	682	0.027	56%	+17%	446	0.033	21%	-17%

used an existing dataset in order to highlight the flexibility of the approach, which doesn't allow us to verify our hypotheses through additional questionnaires. Such conclusions could be drawn, for example, in a follow-up user study designed based on our findings.

Another limitation of this work lies in the sole focus in the results section on the most complex task of the dataset: the road crossing, calling for three different performance parameters. It would be interesting in future work to compare such an analysis to different types of tasks, to observe the variability of relevance of a given behavioural metric depending the type of task. Another potentially interesting analysis would be to analyse multiple behavioural metrics together for a given performance component in order to observe if a broader characterisation of users behaviour could be achieved.

7 Conclusion

In this article, we have introduced a task-based methodology that uses a task model to define performance components and baseline values for the analysis of fine-grained user behaviour based on the context and multivariate data. To demonstrate the advantages of our approach, we used a dataset of VR experiences in order to characterise behaviour in relation to a certain type of performance misstep. Finally, we identified individual profiles of users for whom a behavioural metric is strongly correlated to the performance of the task. This could help us reveal profiles of user behaviour in order to better understand their needs and propose personalised experiences.

The next step of this work calls for a deeper analysis of specific profiles of users across a wider range of tasks and scenarios, such as focusing on users who may have specific needs for VR training and rehabilitation. We open the possibility of investigating complex behaviour characterisation in-context, as well as the possible to push this type of evaluation for the improvement of the experience, to propose better feedback, guidance, and refine experiment systems in VR. Finally, we also emphasise the benefits of open datasets and experimental protocols for the advancement of UX understanding for new and immersive media platforms such as VR.

Acknowledgments

This work has been partially supported by the French National Research Agency though the ANR CREATTIVE3D project ANR-21-CE33-0001 and UCA^{JEDI} Investissements d'Avenir ANR-15-IDEX-0001 (IDEX reference center for extended reality XR^{2}C^2).

References

[1] M. Alcañiz, B. Rey, J. Tembl, and V. Parkhutik. A neuroscience approach to virtual reality experience using transcranial doppler monitoring. *Presence: Teleoperators and Virtual Environments*, 18(2):97–111, 2009.

[2] S. Aseeri and V. Interrante. The influence of avatar representation on interpersonal communication in virtual social environments. *IEEE transactions on visualization and computer graphics*, 27(5):2608–2617, 2021.

[3] R. Bernhaupt, C. Martinie, P. A. Palanque, and G. Wallner. A generic visualization approach supporting task-based evaluation of usability and user experience. In R. Bernhaupt, C. Ardito, and S. Sauer, eds., *Human-Centered Software Engineering - 8th IFIP WG 13.2 International Working Conference, HCSE 2020, Eindhoven, The Netherlands, November 30 - December 2, 2020, Proceedings*, vol. 12481 of *Lecture Notes in Computer Science*, pp. 24–44. Springer, 2020.

[4] S. Chan, H. Zhang, and S. Nanayakkara. Eye movement analysis of human visual recognition processes with camera eye tracker: Higher mean and variance of fixation duration for familiar images. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2023.

[5] H. Chen, Y. Wang, and W. Liang. Vcoach: Enabling personalized boxing training in virtual reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 546–547. IEEE, 2022.

[6] R. M. Clifford, S. Jung, S. Hoermann, M. Billinghurst, and R. W. Lindeman. Creating a Stressful Decision Making Environment for Aerial Firefighter Training in Virtual Reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 181–189. IEEE, Osaka, Japan, Mar. 2019. ISSN: 2642-5254. doi: 10.1109/VR.2019.8797889

[7] T. Delrieu, V. Weistroffer, and J. P. Gazeau. Precise and realistic grasping and manipulation in virtual reality without force feedback. In *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 266–274. IEEE, 2020.

[8] A. Dey, A. Chatburn, and M. Billinghurst. Exploration of an eeg-based cognitively adaptive training system in virtual reality. In *2019 ieee conference on virtual reality and 3d user interfaces (vr)*, pp. 220–226. IEEE, 2019.

[9] D. Diaper and N. Stanton. *The Handbook of Task Analysis for Human-Computer Interaction*. CRC press, USA, 2003.

[10] G. Eklund. General features of vibration-induced effects on balance. *Upsala journal of medical sciences*, 77(2):112–124, 1972.

[11] E. Guerra, B. Kurz, and J. Bräucker. An extension to the keystroke-level model for extended reality interactions. *Mensch und Computer 2022-Workshopband*, 2022.

[12] Q. Guimard, F. Robert, C. Bauce, A. Ducreux, L. Sassatelli, H.-Y. Wu, M. Winckler, and A. Gros. On the link between emotion, attention and content in virtual immersive environments. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2521–2525. IEEE, 2022.

[13] Q. Guimard, F. Robert, C. Bauce, A. Ducreux, L. Sassatelli, H.-Y. Wu, M. Winckler, and A. Gros. Pem360: A dataset of 360 videos with continuous physiological measurements, subjective emotional ratings and motion traces. In *Proceedings of the 13th ACM Multimedia Systems Conference*, pp. 252–258. ACM, Athlone, Ireland, 2022.

[14] M. Hassenzahl and N. Tractinsky. User experience-a research agenda. *Behaviour & information technology*, 25(2):91–97, 2006.

[15] S.-C. Huang, T.-W. Lu, H.-L. Chen, T.-M. Wang, and L.-S. Chou. Age and height effects on the center of mass and center of pressure inclination angles during obstacle-crossing. *Medical engineering & physics*, 30(8):968–975, 2008.

[16] C. Jicol, C. H. Wan, B. Doling, C. H. Illingworth, J. Yoon, C. Headey, C. Lutteroth, M. J. Proulx, K. Petrini, and E. O'Neill. Effects of Emotion and Agency on Presence in Virtual Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, Yokohama Japan, May 2021. doi: 10.1145/3411764.3445588

[17] C. Keighrey, R. Flynn, S. Murray, and N. Murray. A physiology-based qoe comparison of interactive augmented reality, virtual reality and tablet-based applications. *IEEE Transactions on Multimedia*, 23:333–341, 2021. doi: 10.1109/TMM.2020.2982046

[18] A. Klostermann and S. Moeinirad. Fewer fixations of longer duration? expert gaze behavior revisited. *German journal of exercise and sport research*, 50(1):146–161, 2020.

[19] O. Kroemer, H. Van Hoof, G. Neumann, and J. Peters. Learning to predict phases of manipulation tasks as hidden states. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4009–4014. IEEE, 2014.

[20] S. Kuno, T. Kawakita, O. Kawakami, Y. Miyake, and S. Watanabe. Postural adjustment response to depth direction moving patterns produced by virtual reality graphics. *The Japanese journal of physiology*, 49(5):417–424, 1999.

[21] Y. Lang, L. Wei, F. Xu, Y. Zhao, and L.-F. Yu. Synthesizing Personalized Training Programs for Improving Driving Habits via Virtual Reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 297–304. IEEE, Reutlingen, Mar. 2018. doi: 10.1109/VR.2018.8448290

[22] G. Lee, T. Lozano-Pérez, and L. P. Kaelbling. Hierarchical planning for multi-contact non-prehensile manipulation. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 264–271, 2015. doi: 10.1109/IROS.2015.7353384

[23] W. Li, J. Talavera, A. G. Samayoa, J.-M. Lien, and L.-F. Yu. Automatic synthesis of virtual wheelchair training scenarios. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 539–547. IEEE, 2020.

[24] W. Li, J. Talavera, A. G. Samayoa, J.-M. Lien, and L.-F. Yu. Automatic Synthesis of Virtual Wheelchair Training Scenarios. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 539–547, Mar. 2020. ISSN: 2642-5254. doi: 10.1109/VR46266.2020.00075

[25] T. Luong, A. Lecuyer, N. Martin, and F. Argelaguet. A survey on affective and cognitive vr. *IEEE transactions on visualization and computer graphics*, 28(12):5154–5171, 2021.

[26] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, feb 2021. doi: 10.3758/s13428-020-01516-y

[27] M. Mezidi and C. Guérin. Effect of body position and inclination in supine and prone position on respiratory mechanics in acute respiratory distress syndrome. *Intensive Care Medicine*, 45(2):292–294, 2019.

[28] P. Palanque, E. Barboni, C. Martinie, D. Navarre, and M. Winckler. A model-based approach for supporting engineering usability evaluation of interaction techniques. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems*, pp. 21–30, 2011.

[29] P. Palanque and C. Martinie. Designing and assessing interactive systems using task models. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 976–979, 2016.

[30] P. A. Palanque, E. Barboni, C. Martinie, D. Navarre, and M. Winckler. A model-based approach for supporting engineering usability evaluation of interaction techniques. In F. Paternò, K. Luyten, and F. Maurer, eds., *Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing System, EICS 2011, Pisa, Italy, June 13-16, 2011*, pp. 21–30. ACM, 2011. doi: 10.1145/1996461.1996490

[31] A. D. Rice and J. W. Lartigue. Touch-level model (tlm) evolving klm-goms for touchscreen and mobile devices. In *Proceedings of the 2014 ACM Southeast regional conference*, pp. 1–6, 2014.

[32] F. Robert, H.-Y. Wu, L. Sassatelli, S. Ramanoel, A. Gros, and M. Winckler. An integrated framework for understanding multimodal embodied experiences in interactive virtual reality. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, pp. 14–26, 2023.

[33] N. Rossol, I. Cheng, W. F. Bischof, and A. Basu. A framework for adaptive training and games in virtual reality rehabilitation environments. In *proceedings of the 10th international conference on virtual reality continuum and its applications in industry*, pp. 343–346, 2011.

[34] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pp. 71–78, 2000.

[35] C. Santoro. *A task model-based approach for design and evaluation of innovative user interfaces*. Presses univ. de Louvain, 2005.

[36] J. Sauro. Estimating productivity: Composite operators for keystroke level modeling. In *Human-Computer Interaction. New Trends: 13th International Conference, HCI International 2009, San Diego, CA, USA, July 19-24, 2009, Proceedings, Part I 13*, pp. 352–361. Springer, 2009.

[37] S. Seinfeld, T. Feuchtner, J. Pinzek, and J. Müller. *Impact of Information Placement and User Representations in VR on Performance and Embodiment*, vol. PP. IEEE, remote, Sept. 2020. Journal Abbreviation: IEEE transactions on visualization and computer graphics Publication Title: IEEE transactions on visualization and computer graphics. doi: 10.1109/TVCG.2020.3021342

[38] K.-C. Siu, B. J. Best, J. W. Kim, D. Oleynikov, and F. E. Ritter. Adaptive virtual reality training to optimize military medical skills acquisition and retention. *Military medicine*, 181(suppl_5):214–220, 2016.

[39] M. Vigo, C. Santoro, and F. Paternò. The usability of task modeling tools. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 95–99. IEEE, 2017.

[40] P. Wodarski, J. Jurkojć, J. Polechoński, A. Bieniek, M. Chrzan, R. Michnik, and M. Gzik. Assessment of gait stability and preferred walking speed in virtual reality. *Acta of bioengineering and biomechanics*, 22(1):127–134, 2020.

[41] E. Wu, M. Piekenbrock, T. Nakumura, and H. Koike. SPinPong - Virtual Reality Table Tennis Skill Acquisition using Visual, Haptic and Temporal Cues. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2566–2576, May 2021. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2021.3067761

[42] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar. CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos. *IEEE Transactions on Multimedia*, 0(0):1–1, 2021. doi: 10.1109/TMM.2021.3124080

[43] T. Zander, A. Rohlmann, J. Calisse, and G. Bergmann. Estimation of muscle forces in the lumbar spine during upper-body inclination. *Clinical Biomechanics*, 16:S73–S80, 2001.

[44] X. Zhou, F. Teng, X. Du, J. Li, M. Jin, and C. Xue. H-goms: A model for evaluating a virtual-hand interaction system in virtual environments. *Virtual Real.*, 27(2):497–522, jul 2022. doi: 10.1007/s10055-022-00674-y