Design and Evaluation of Abstract Aggregated Avatars in VR Workspaces

György Persa*, and Ádám B. Csapó†‡

***Abstract*—Avatars are commonly used in digital platforms to provide a visual representation of individual users to each other. Generally, avatar design in the past has focused on achieving visual fidelity and realism of social interactions. In this paper, we broaden the concept of avatars to incorporate displays using an abstract visual language and conveying information on aggregated, interpersonal information from the perspective of the digital platform as a whole. We propose a general design methodology for such aggregated avatars, and also introduce and experimentally evaluate an aggregated avatar which we have developed on the MaxWhere VR platform. Results are promising in that users were able to discern several key states of the avatar and correctly associate them with the correct virtual reality scenarios in a statistically meaningful way.**

***Index Terms*—virtual reality, avatar design, abstract avatars, aggregated avatars**

I. Introduction

The word '*avatar*' originates from Sanskrit, and refers to the meaning of 'embodiment', or a 'divine being made of flesh' [1]. In the rapidly evolving fields of virtual reality and metaverse technology, it evokes a very specific connotation of a human-like visual representation that can convey real-time information on the appearance, activities and even the mental state of a given user [2], [3].

Despite this clear picture of what an avatar is, or should be, it is nevertheless worth noting that avatars can take many shapes and forms, depending on the specific information they are used to represent, the visual and cognitive fidelity of their representation, and even depending on the context in which they are used. At one 'corner' of this multi-dimensional spectrum, an avatar can appear as a simple pictogram, showing a static photo of the user and indicating their presence on the social / virtual platform. At another 'corner' of the spectrum, highly visually realistic avatars are used that digitally re-represent the user's facial and head movements based on real-time tracking. In a third 'corner' of the spectrum, one can consider digital environments that involve many users interacting at once, instead of focusing on one-on-one, or small-group social interactions. In the case of a personalized virtual reality working environment, helping users to either streamline their digital workflows, or to focus on personal productivity and personal growth in an engaging environment would be the most relevant feature. In such scenarios, avatars that provide motivation to users by reflecting back to them key insights on their activities as a kind of meta-cognitive feature of the platform might be most valuable. Such meta-cognitive functionality for an avatar can also in fact be considered in cases where a platform has many users and the aggregate statistics reflecting the way in which the platform is being used will be of more interest than the interaction patterns of any given user. In this scenario, each user might contribute in a small way to the overall behavior of the avatar.

In this paper, we broaden the traditional understanding of avatars to include representations that mirror not just the specifics of a single user to others but also provide a holistic view of the digital environment to the user. This includes the state of the digital platform and the interactions of the user and possibly numerous other users with the platform. By easing the rigid definitions of what an avatar represents and to whom, we propose the possibility of previously unexplored avatar types, which we label as 'abstract avatars' and 'aggregated avatars'. We further posit that these types of avatar hold significant potential in our evolving digital landscape, where human and digital interactions are deepening not only in the short run but also in a more sustained, co-evolutionary manner as often described in the literature on cognitive infocommunications and cognitive aspects of virtual reality [4], [5]. To validate this concept, we describe a possible design methodology for such avatars, develop an example implementation, and perform validation of this implementation to demonstrate the viability of this approach.

The paper is structured as follows. In Section II, we consider different dimensions along which avatars can be qualified, and define novel categories of avatars based on these dimensions, including abstract and aggregated avatars. In Section III, we describe the design steps we have taken to design an aggregated avatar based on emotional features. In Section IV, we describe a framework within which we validate the design of our aggregated avatar through test subjects. Finally, Sections V and VI present details on the experimental design we have used, and the results we have obtained.

II. Conceptualizing novel types of avatars

A. Visual and cognitive fidelity

As discussed in the previous section, the visual representations used in avatars can range from simple, static pictograms to highly accurate visual representations of specific users [2], [6]–[10]. Based on this, it is possible to consider the *visual*

* Doctoral School of Multidisciplinary Engineering Sciences, Széchenyi István University, Győr Hungary
† Corvinus Institute for Advanced Studies & Institute of Data Analytics and Information Systems, Corvinus University of Budapest, Hungary
‡ Hungarian Research Network, Budapest, Hungary
E-mail: persa.gyorgy@sze.hu
E-mail: adambalazs.csapo@uni-corvinus.hu

DOI: 10.36244/ICJ.2024.5.11

fidelity of an avatar, in terms of how it relates to the user it represents in a visual sense.

At the same time, avatars can also be assessed in terms of their behaviors, whether in terms of how naturally they behave in social interactions [11]–[15], or in terms of how well they reflect the key aspects of user interactions, thereby contributing to increased productivity. Based on this, we have defined the term *cognitive fidelity* as follows [16]:

Definition 1: The **cognitive fidelity** of an avatar is a an assessment, both qualitative and potentially quantitative, of how well an avatar reflects the state of an underlying process, viewed from the attendant benefits to the cognitive capabilities of the users to whom it appears.

We note that the term cognitive fidelity has been used in other contexts, for example, to describe how well the use of a virtual tool corresponds to users' actions and possible choices [17], [18]; or in other cases, to describe the ecological validity of an environment from the standpoint of cognitive tasks being carried out inside it [19]. In the above definition, we focus instead on the mental capabilities of the user to whom the avatar is displayed.

As an example, let us consider a flight simulator. In such an application, the designer would have a choice as to whether to represent the co-pilot of the user as a photo-realistic human whose body posture and head movements are fully in accordance with normal social interactions, or instead, to use more abstract representations that can offer useful feedback as to what is going wrong or what aspects of flight control require attention. In some cases, even the feedback that "something is going wrong" can be highly relevant and can improve cognitive performance.

B. Abstract and Aggregated Avatars

When the main focus is on cognitive rather than visual fidelity, questions of information modeling and representation mapping naturally arise.

In terms of information modeling, it is often the case that parameters that are not so directly linked with physical reality (for example, statistics on different interaction types, or other parameters closely related to the application scenario) need to be communicated through the avatar.

In terms of finding a useful representation, the goal is to map the information to be represented onto visual (and perhaps other sensory) channels in a way that is recognizable and intuitively meaningful to users. From the theory of cognitive infocommunication (CogInfoCom) channels, if we consider avatars to be analogous to CogInfoCom messages, then the information mapping types of direct (both low-level and high-level), as well as indirect (including structural, co-stimulation or scenario-based) can be especially relevant [4], [20].

Based on these considerations, we have introduced the following definitions of abstract and aggregated avatars, with the intention of encompassing a broader scope of potential avatar designs [16]. Here, we propose a slightly revised version of this definition so as to focus – in the case of abstract avatars – solely on the language of the avatar, irrespective of the application scenario:

Definition 2: An **abstract avatar** is an avatar representation that relies on an abstract, low-level visual language involving the use of dynamic shapes and colors without reference to anthropomorphic or zoomorphic concepts.

Definition 3: An **aggregated avatar** is an abstract avatar that is used to display the features of an impersonal set of interactions and contextual events in a computational environment.

On the one hand, aggregated avatars are capable of providing users with a mirror of their position within a cooperative process, as seen through the perspective of the digital environment, instead of merely presenting data about other users. Conversely, due to its largely impersonal character, aggregated avatars can also serve to inform users about the comprehensive 'status' of a platform like a workplace or a virtual reality setting. This includes collective, environmental concepts such as the 'vibrancy of the surroundings' or the 'enthusiasm level of participants'. These aspects are somewhat influenced by the conduct of the users involved, yet they cannot easily be broken down into the distinct individual behaviors of each user.

C. Design principles for the development of abstract and aggregated avatars

Based on the dichotomy of information and representation modeling, the design of an abstract or aggregated avatar can be broken into the following steps:

1) Answering the question of what information types to model, including potential range of numerical values. In case the state to be modeled includes (fuzzy) categorical features as well, e.g. fuzzy modeling can be used to convert states into numerical fuzzy membership values.
2) Answering the question of how the resulting variables can be mapped onto an intermediate language suitable for driving discernible and interpretable avatar behaviors. One example of such an intermediate language can be a valence-arousal based emotional model, which is capable of representing distinct emotional states that can be recognized with relative ease by many users.
3) Finally, defining the visual features of the abstract avatar that can be used to drive its behavior, and mapping onto them the states of the intermediate language described in the previous point.

III. Design of an Emotionally Based Aggregated Avatar

We have given design principles for aggregated avatars in previous sections, aiming to enhance cognitive fidelity by finding suitable mappings between cognitive attributes of users and visual features of avatars. In our prior works, we created and implemented an abstract avatar inspired by ethology that can express emotional states [21]. We used this avatar design to present the aggregated space data to users.

A. Emotion displaying agent

The abstract agent that we designed to display emotions has two distinct parts: A colored sphere that can change

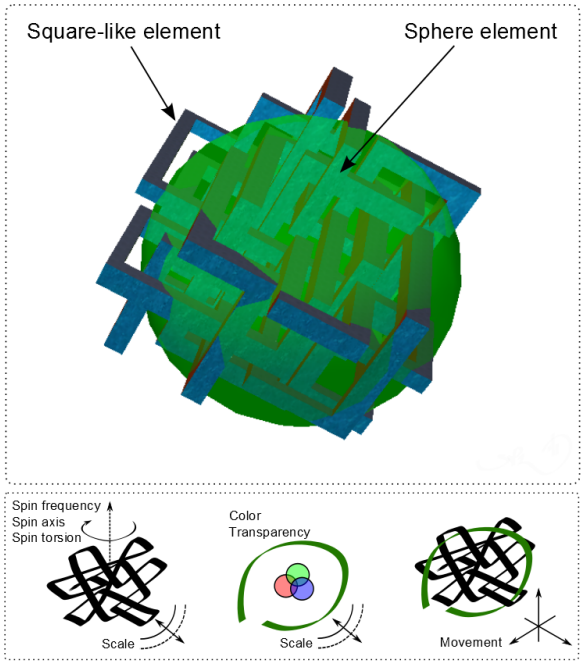

Fig. 1. The appearance and degrees of freedom of the emotional display agent

its transparency and size; and a cube-like structure that has different colors on each side and resembles a maze 1. Degrees of freedom included the position of the component, size of elements separately, color and transparency of the sphere element and rotation speed and angle of the square-like element. This setup provided us with the possibility of evoking associations with emotional states [22].

A possible way to illustrate the aggregated state of a virtual environment is to use the emotion displaying agent as a collective representation of the space. The agent acts as an abstract aggregated avatar that reflects the overall mood of the virtual space through its visual features. This approach requires a mapping function that can translate the aggregated state of the space into the visual parameters of the agent's appearance and behavior.

B. Mapping emotions to agent

Based on the evaluation of expressive features of the Emotion Display Agent, we designed a continuous mapping which can construct the look of the Emotion Display Agent for any combination of emotions. For this purpose we used the well-known Valence-Arousal model of emotions. This psychological framework describes how emotions are represented in two dimensions: valence refers to the positive or negative quality of an emotion, while arousal refers to its intensity or activation level.

The main goal with the mapping was to produce the kinds of appearances in the case of different valence-arousal value compbinations that have already been associated with corresponding emotion (happiness, sadness, fear, etc.) by users in the previously cited study [22]. Thus, we designed the following rules to achieve this:

- High valence and high arousal drove the avatar into a state where it expressed happiness by growing in size, with the sphere element exceeding the inner maze shape and appearing in a yellowish color while rotating at a relaxed pace
- High arousal with low valence caused the avatar to display anger: the sphere element turned red and shrunk while the maze element exceeded the sphere in size and rotated with high speed and non-linear easing.
- Low valence with low arousal caused the avatar to display a sad expression. It rotated evenly, moved away from the camera and shrunk in every dimension while the maze elements greatly exceeded the size of the sphere. The color looked approximately pale purple in this state
- High valence with low arousal reflected a bored or sleepy state of the avatar. To express this emotion, the maze element rotated at a low speed but unevenly, while the sphere grew big and appeared in a purple-like color
- In its natural state the avatar looked relaxed, with a green colored sphere slightly exceeding the size of the maze element. Rotation and easing were adjusted to normal pace in this case.

The top-right corner of Figure 2 shows several examples of how arousal and valence were mapped onto the avatar parameters.

C. Mapping aggregated interactions to emotions

As described in [5], modern infocommunication platforms are evolving to encompass a wide variety of novel interfaces, including virtual reality interfaces, AI-driven interfaces and distributed Web 3.0 applications. In this context, new kinds of aggregated information types are emerging which, when communicated to users in a way that represents the system as a whole, can provide intuitive feedback on parameters such as how active, how overloaded, or how quiet the system is – whatever the case may be.

In the context of virtual reality, parameters such as number of users, activity level of users in terms of moving around in the space, exploring localized subsets of the space, or interacting with others in the space could be of interest.

The specific mapping functions that are used will naturally depend on the aggregated data and the goal of the system in communicating it to users. In Section IV-B, we present a detailed example of one specific application we have developed to test the aggregated avatar concept.

IV. A Test Framework for Evaluating Abstract and Aggregated Avatars

To evaluate the effectiveness of the abstract aggregated avatar, we developed a test framework in the MaxWhere VR application. The main objective was to obtain quantitative data

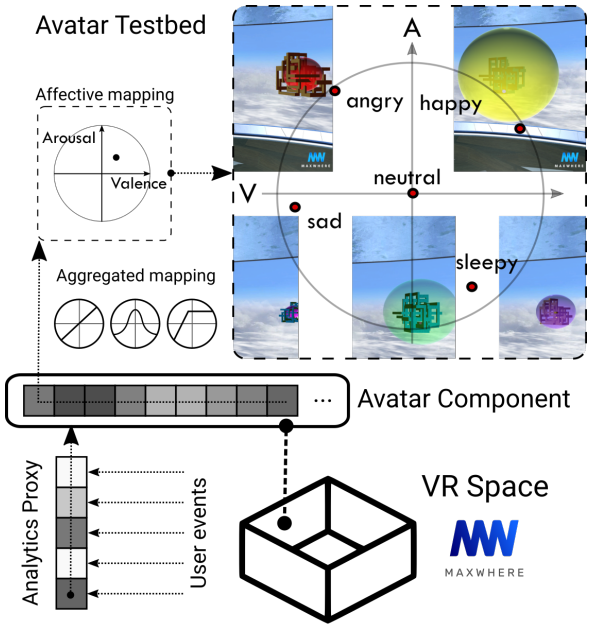

Fig. 2. Architectural diagram of testbed framework used for investigating aggregated avatar in MaxWhere VR.

on how well the participants could understand and differentiate the aggregated state of the virtual space. The basic idea was to design a test in which we could present various user interaction scenarios in the virtual environment and also display the state of the abstract aggregated avatar. We could then test whether a participant could identify the relationship between a user interaction and the corresponding changes in the visual appearance of the avatar by either swapping or keeping aligned the interaction videos and the avatar recordings, and repeating this idea with several different kinds of interactions. The overall structure of the framework is shown in Figure 2.

A. MaxWhere VR

MaxWhere VR is an innovative virtual reality platform that enables users to create and experience immersive virtual environments for various applications. MaxWhere VR offers a versatile and customizable framework that can suit different needs and goals. Users can access a rich library of 3D scenes or import their own content to design their own virtual worlds with custom functionalities. It can be used for presenting, showcasing, collaborating or research purposes as well.

Unlike other platforms, MaxWhere VR offers a unique feature called the "Where Object Model" (WOM), which is similar to a Document Object Model (DOM) used in web programming, but for 3D spaces. The WOM allows users to access and manipulate the properties and behaviors of 3D objects using Javascript code. Users can also create and load components, which are reusable pieces of code that can enhance the functionality and interactivity of the 3D spaces. With MaxWhere, users can design and program their own 3D spaces in a flexible and dynamic way, using powerful programming interfaces.

The platform also provides users with so-called "smartboards", which are essentially 2D display panels that can be used to show web content in a customizable size, position and orientation within the 3D space.

Due to its ability to customize the environment MaxWhere VR is convenient for designing scientific experiments in virtual spaces. Adding custom functionality to virtual spaces allows researchers to perform an experiment directly in the virtual space or collect data from the application for further analysis. The business logic of the experiment can be implemented as a component and can be easily added to any MaxWhere space. Many examples of such environments can be found in the literature (see e.g. [20], [23]–[25]).

B. Analytics Proxy and Testbed Component

MaxWhere provides several proxies to aid component developers accessing higher level information about 3D objects or application state. Proxies are modules which use the MaxWhere engine WOM API to implement functionalities and encapsulate them into a dedicated interface. Interfaces defined this way can be made available for MaxWhere component developers by exposing functions to the WOM API thus making them available for any MaxWhere user. Using this structural possibility we developed a Space Analytics MaxWhere Proxy with the aim of reporting about the user interaction in the space conveniently. The proxy provides basic information about the virtual environment, such as the sum axis aligned bounding box (AABB) of the space calculated from each visual object located in the space. Furthermore event listeners can be registered via the interface which reports about the following events:

- User (camera) movement in the space
- User changes between 3D and menu of the space
- User discovers a 3D object by orbiting around it
- User interact with a smartboard in the space with mouse cursor

The reporting methods for these events can be adjusted on the proxy. Users can specify the reporting interval if the continuous event sending is not desirable. Also, the behavior for unchanged states can be modified as the user can choose whether the idle system should report the identical events or not.

Using the generic Space Analytics Proxy we implemented a custom Testbed MaxWhere component for creating statistics from the space data. For the calculations we used a customizable time window (default is 40 seconds) in which each type of data is aggregated. We construct the following properties from the gathered information:

- Ratio of discovered area of the space relative to the total AABB. Covered area is calculated by creating an AABB from camera positions
- User mobility: Ratio of time the user spent moving
- Ratio of time the user spent interacting on a smartboard. Calculated from mouse cursor moves while 2D content is displayed
- Ratio of time the user spent in a smartboard, on the menu or in 3D

- Ratio of time the user spent orbiting around a 3D object

The produced statistics are then converted into the intermediate interpretation of Arousal and Valence values. For this calculation we defined the processing functions for each parameter of the aggregated Space Analytics data. We used three different transition functions: One for a high effect with a quick slope, one which increases the effect of a parameter as it nears the average value and one for a small contribution with stretched out slope. Using these transfer functions we tailored the Arousal and Valence values so that:

- High valence was associated with the statistical parameter values all being around average, which means that the user interactions are not one-sided and the space capabilities are all used for some degree.
- High Arousal values are directly but not linearly correlated with user movements and orbiting around objects, or with the discovered area increasing. Switching between 3D space and 2D menu / interaction on smartboards only had a limited effect of this value.

The calculated affective values were then used to drive the look of the aggregated avatar inside the MaxWhere space. The avatar – as summarized in section III-A - is able to express emotions via changes in its visual parameters. To map affective values derived from spatial events to the adjustable attributes of the abstract avatar we use the mappings presented in section III-B. The data flow behind the framework is presented on Figure 2.

Note that we have also implemented utilities for recording and replaying user interactions and avatar states. When recording, the avatar is temporary hidden and the user can interact with the space normally. Each interaction received on the Analytics Proxy is recorded and written into a file. The proxy is configured to report every state of the space periodically, thus a continuous sequence of states is stored as a result. Using the replay function, it is then possible to read these files at a later time and to drive the aggregated avatar with the recorded space state values. During replay the space is hidden and only the avatar is visible. Hiding the avatar or the space in these utilities was important for creating sample sequences and videos for further analysis.

V. Experimental design

In the following section we describe the details of the experiment we performed to validate the feasibility of our aggregated abstract avatar.

A. Test videos

We used the record features of the MaxWhere testbed component to create several pairs of test videos to evaluate the performance of the avatar. Each pair of videos included a user interaction video and an avatar behavior video. The user interaction videos captured only the scene and the interactions without the avatar, while the reactions of the avatar to these interactions were recorded from the saved log file using the replay function. During replay, we recorded the corresponding video showing only the expressions of the avatar. To facilitate comparison, we selected three different aspects of user behavior as dimensions for choosing typical interaction patterns for the videos. The three dimensions were as follows.

- *Discovery*: Spatial behavior when the user moves around the virtual space and discover multiple 3D objects. It consists of the space analytics data of camera movements and orbiting.
- *Manipulation*: Describes the scenario when the user interacts with the 3D objects using the Editor capabilities of MaxWhere. It includes a little bit of orbiting from the space analytics data and mostly specified by menu transitions and cursor movements. During this scenario the user opens and closes the menu multiple times, inputs values on the user interface or moving around 3D object or resize them in the space with dragging gesture of the cursor.
- *2D Operation*: Characterized by smartboard interactions. Using the space analytics data of entering or leaving a smartboard and cursor moments while a smartboard is selected for usage. This is the case of browsing the web or working on a content in smartboards.

We recorded 8 pairs of videos in total along these dimensions. For each interaction types we made two videos where only that behavior was shown (six videos in total). We also made one video where the user showed no behavior and one video where the user showed all the behaviors. The original videos were about one minute long, but we sped them up to 30 seconds for the experiment.

A frame from a video of a user interaction and the corresponding avatar behavior can be seen on Figure 3.

B. Survey

We conducted an experiment to test how well participants could match user interaction videos with avatar behavior videos. We have chosen the aggregated state of the virtual space to be the *"usefulness"* of user interactions. This means that the "mood" of the affective avatar was driven by the user interaction types and based on the quality and the quantity of interactions the avatar displayed a combination of "happy", "frustrated" or "angry" state.

We created 8 video pairs in total, each consisting of a user interaction video and a corresponding avatar behavior video as described in the previous section. We then randomly selected 2 video pairs for each task and presented them to the participants in a mixed order: first a user interaction video, then another user interaction video, then an avatar behavior video, and then another avatar behavior video. The participants had to decide which of the two avatar behavior videos matched the first user interaction video they saw. In other words, they had to choose between the combinations of (1-3; 2-4) and (1-4; 2-3). This task is presented to the participants four times with different video pairs each time.

The complete experiment consisted of the following steps:

- Language selection (we prepared an English and a Hungarian version)
- First we introduced the details of the research and explained the tasks ahead

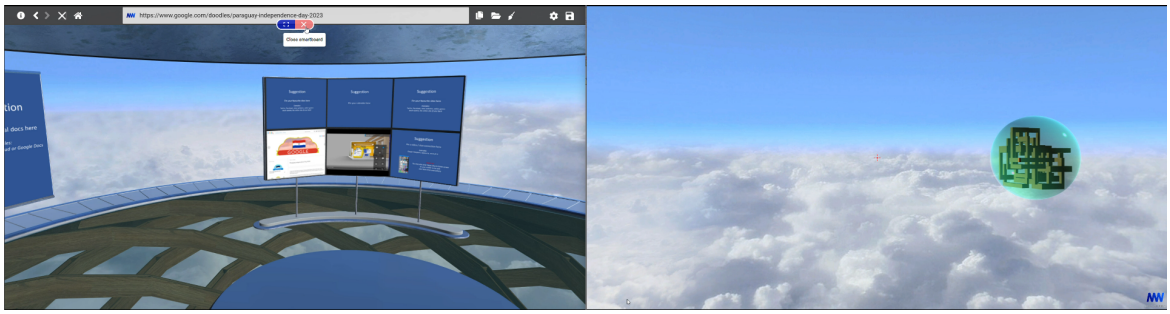

Fig. 3. Example of user interaction video (on the left) with the corresponding avatar video (on the right) showing a 2D operation scenario.

- Collection of informed consent from the participants for anonymized data collection
- Presentation of the 4 matching tasks

The briefing text was the following:

"*This study explores the expressive capabilities of an abstract-shaped avatar.*
The avatar aims to convey a summary (aggregated state) of events occurring in the virtual space. Since the avatar has an abstract shape, it can use only abstract methods (based on emotional associative techniques) to express its state. This test evaluates how well the avatar's expression is understood.
In the following sections, we will present you with 4 tasks, each consisting of 4 videos. The first 2 videos in each task will show user interactions in the virtual space, while the third and fourth video will show the avatar's response to user behaviors from one of the user videos. Your goal will be to pair the first two (user) videos with the second two (avatar) videos.
We used MaxWhere VR application to create the virtual environment for this experiment. This application allows users to access or upload media contents through interactive boards in the 3D space, also known as "smartboards". The user interaction patterns in this system can be classified into three types:

- navigation: moving in the space
- 2D interaction: viewing content on smartboards
- manipulation: moving, resizing smartboards

We recorded user interaction videos in this study that include these patterns. These patterns affect the "mood" of the abstract avatar, which changes the displayed overall state of the environment to combination of these values: "bored", "excited", "happy" or "frustrated". In the following tasks, the overall state will be the "usefulness" *of user interactions, which means how much the user takes advantage of every function available in the space. For instance,*

- a 'pointless wandering' pattern will trigger a frustrated mood,
- a task pattern will trigger a combination of bored mood and excitement based on how repetitive is the task,
- and a balanced use of all interaction types will trigger a happy mood.

The recordings are about 30 seconds long and each are accelerated to the same degree."

We also present a brief video to the participants after the introduction, which demonstrates the typical user interactions in MaxWhere VR that are described in the briefing text. This is essential for ensuring that they can recognize and differentiate the actions on user interaction videos with as much confidence as a more experienced MaxWhere user.

C. Automated data collection

The survey used an automatic recording system to store each response in a spreadsheet that was linked to the survey. The recording also triggered several custom routines that analyzed the answer and produced some statistical metrics. To do this, we created a custom scripting interface that connected the survey with the data collector modules that were integrated into the survey. The custom routines generated the survey for the next participant. This way, the next version of the test had different order and combinations of answers and videos.

The survey for the experiment is created in Google Forms linked to a Google spreadsheet to store the results. Custom routines were implemented in Google Apps Script framework to process the data. The script runs automatically after each Form submission, triggered by the framework settings.

D. Custom routines

Custom routines performed upon survey completion includes randomization of the next test and evaluation of the current one.

Video randomization is performed using the following steps in the script.

- Read out the URLs of video pairs from a spreadsheet page. Each video pair (user interaction and avatar video) is associated with an ID.
- Associate groups to the video pairs according to which user interaction type they contain
- For the first task, select the Idle video and pair it with one of the one-dimensional videos (Discovery, Manipulation, Operation2D).

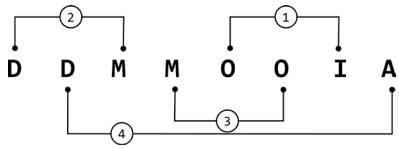

Fig. 4. Example run of the randomization routine. Numbers depicts the order of operation. Letters are the user interaction types (D - discovery, M - manipulation, O - 2D operation, I - idle, A - all)

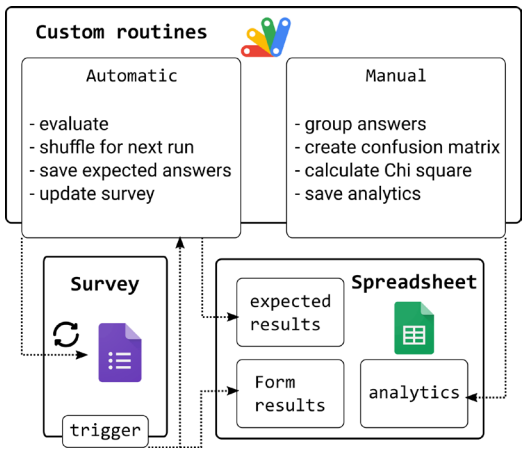

Fig. 5. Data flow of survey, connected spreadsheet and custom routines. Automatic routines run when a survey is completed. It evaluates current answers and shuffle pairs for a new run. Manual routines run on demand and create analytics.

- For second and third task, pair the least used one-dimensional videos randomly. Least used videos are determined by counting how many times a video is already used for pairing in previous tasks. (thus, it is actually forced for the second task but not for the third)
- Pair the All dimensional video with the remaining video for the last (fourth) task.
- Randomize order within pairings to alter display order of videos in the survey.

The outcome of random pairing for the tasks of the next survey is then registered in a spreadsheet. It represents the expected answers for the next submission. An example run of the randomization routine is demonstrated on Figure 4.

The form submission trigger runs the script after the connected spreadsheet has been updated with the latest responses. This allows the script to automatically check the accuracy of the matching tasks. The script retrieves the correct answers from the sheet based on the latest response.

The overall functionality of the evaluation and randomization script can be seen on Figure 5.

VI. Results

In this section we summarize the statistical analysis of the survey responses. The main objective of the analysis is to determine how well the users can match the avatar behaviors with the corresponding user interactions.

The survey has been distributed only for a smaller group at first for fine-tuning. Based on the preliminary findings described below, we modified the survey and recruited a larger group of participants through social media and technical forums. We analyzed the final responses using the custom manual methods explained in the previous section.

A. Preliminary results

For the test video pairs, we initially had a slightly different setup. In the first version of the experiment we combined 2 different interaction types as well instead of using only one, zero or all dimensions for creating test videos. The randomization algorithm also did less instructed shuffling, which made the video match tasks more challenging.

We recorded answers from 12 people (5 male, 7 female) and found that none of the interaction types produced acceptable recognisability. We also observed that the randomized task generator created complex combinations that increased the difficulty level for even the most skilled MaxWhere users.

B. Revised experiments

Based on the preliminary results we simplified the experiment tasks by reducing the number of user interaction types depicted in the test videos. This made the videos more distinguishable as the participants did not see tasks with combinations of the same user interaction type.

We also modified the shuffle algorithm in order to fix one half of the random pairings in half of the tasks. The first task has always presented the zero dimensional case and the last task had all the user interaction types for one of the video pairs. The revised video pairing is described in steps in section V-D. The new routine solved the distribution problems as well, because each video pair is used once for each participant.

The explanatory text at the beginning of the survey has also been modified in order to provide a better understanding of the upcoming tasks. We added information about MaxWhere VR in general and introduced the typical user interaction types performed in the user videos of the tasks in a bullet point list and on a short video. We emphasised that the aggregated state of the space is the ”usefulness” of user interactions in this experiment and listed several examples for mappings between the user interaction and displayed emotions of the avatar. For example, we described that a happy avatar means that the user interaction was useful, while a frustrated or bored avatar means that it was not. These sentences helped the participants to understand the situation presented on user interaction videos and emotional states shown on avatar videos. We considered this to be reasonable for our experiment, realizing that it would have been difficult to expect users to associate interactions with the avatar if they had not been attuned to the kinds of interactions that exist in the first place, or the general purpose of the avatar.

C. Statistical analysis

We recorded responses from 31 participants (8 female, 22 male, 1 undefined). Most of them were middle-aged with average age between 30 and 41. Based on the media channels we distributed the survey, we could ascertain that participants

TESTS	Discover	Manipulate	Operate2D	Idle	All	MATCHES	Discover	Manipulate	Operate2D	Idle	All
Discover	0	20	23	8	11	Discover	0	14	12	7	7
Manipulate	20	0	19	14	9	Manipulate	14	0	12	11	7
Operate2D	23	19	0	9	11	Operate2D	12	12	0	7	7
Idle	8	14	9	0	0	Idle	7	11	7	0	0
All	11	9	11	0	0	All	7	7	7	0	0
EXPECTED	Discover	Manipulate	Operate2D	Idle	All	OBSERVED	Discover	Manipulate	Operate2D	Idle	All
Discover	31	10	11.5	4	5.5	Discover	40	6	11	1	4
Manipulate	10	31	9.5	7	4.5	Manipulate	6	44	7	3	2
Operate2D	11.5	9.5	31	4.5	5.5	Operate2D	11	7	38	2	4
Idle	4	7	4.5	15.5	1.00E-10	Idle	1	3	2	25	0
All	5.5	4.5	5.5	1.00E-10	15.5	All	4	2	4	0	21

Fig. 6. Tables used for extracting input data for Chi Square test. Tests table consists of participation of interaction types in tests. Matches table shows the successful matches of the user interaction types. Expected and Observed tables are calculated for Chi Square test using the basic assumption that user interaction types can be recognized only as good as 50%

Confusion / Occurences	Discover	Manipulate	Operate2D	Idle	All
Discover		0.7	0.522	0.875	0.636
Manipulate	20		0.632	0.786	0.778
Operate2D	23	19		0.778	0.636
Idle	8	14	9		0
All	11	9	11	0	

Fig. 7. Confusion and occurrence matrix of the matching results. The matrix has two parts: the lower left part shows the total number of appearances of each user interaction type combination, and the upper right part shows the rate of successful matches for each combination. Darkness of background color of cells indicates the significance of values

CHIDIST	Discover	Manipulate	Operate2D	Idle	All
p values	0.141611638	0.02256998518	0.3981787884	0.0193337095	0.3849561179

Fig. 8. P values of Chi Square test. The result shows significant difference for Idle and Manipulate state and unconfirmed difference for Discover and Operate2D.

had adequate computer skills and at least half of them were familiar with 3D video games.

We performed various statistical tests and analyses to assess how well our avatar can convey the aggregated state of the space during different types of user interaction. To obtain meaningful statistics, we first grouped the data by dimensions of user interaction type. We counted the successful matches from the tasks for every participant and registered a successful recognition of the aggregated state for each successful match. Since the matching tasks involved two different dimensions each time, we recorded success or failure for both of the involved interaction types.

We constructed two basic tables from the extracted data: The total number of participations and number of successful matches for each user interaction dimension. Both tables were derived from the spreadsheet that contained the pairing information about each test run. Using the pairing information alone we could construct the participation matrix. The automated evaluation script registered the success for each task of each test run, which was used to construct the correct matches matrix. Results of extracted tables can be seen on Figure 6.

Based on the participation and correct match tables we constructed a confusion matrix. This matrix shows how well the participants matched the items correctly in each task. It helps us to evaluate how well the expressive ability of our avatar performs for different types of user interactions. We obtained the accuracy by dividing the number of correct matches by the number of total appearances. Figure 7 shows the confusion matrix we generated. The Idle and Discover scenario had the highest type combination score. This makes sense because the avatar shows anger in one case and boredom in the other.

For our statistical test we used Chi square test to see how the aggregated avatar performed. A Chi square test is a statistical method that can be used to test the association between pairs of categorical variables. The test compares the observed frequencies of each category with the expected frequencies under the null hypothesis of no association. The p-value of the test is the probability of obtaining a test statistic as extreme or more extreme than the observed one, under the null hypothesis. A small p-value indicates that there is strong evidence to reject the null hypothesis and conclude that there is an association between the variables – in our case, that users peformed better than randomly in the case of at least one category.

As a baseline, or null hypothesis we used the assumption that the avatar does not help in recognition of the aggregated state of the virtual environment. This means that we suggested that the participants choose randomly during the matching

tasks giving us 50% success rate for each user interaction types.

Therefore, we set the expected table for Chi square test reflecting this statement. Each value in the matrix has been calculated by taking the half of the total occurrences of the given type combination. We used the matrix diagonal to store the summed result for a given dimension. For example, the total expected value for Discover user interaction type is the sum of each test combination containing Discovery divided by two, which is 62 / 2. Note, that we also simplified the case of zero values in the expected matrix to keep the calculation straightforward. We replaced zeros (which would cause a divide error during the upcoming calculations) with 1E-10 values. The expected table can be seen in lower left side of Figure 6.

We obtained the observed values for the Chi square test from the correct match matrix. In the diagonal we use the number of observed successful matches for a given interaction type which is calculated by the sum of correct matches for a given dimension in each combination it participates. For non-diagonal cells we used the number of *incorrect* choices of a given interaction type combination as the expected values describes the choice of the given value *despite of* the correct answer. We calculated these values by subtracting the correct match value for each combination from the total number of occurrences. For example, see the first two cell of the Observed matrix on Figure 6. The observed total of successful matches of Discover dimension is 14 + 12 + 7 + 7 = 40. The total number of times when participants mistakenly chose Manipulate instead of Discover when these two dimensions were combined is 20 - 14 = 6

Using the composed matrices we could calculate the P values of the Chi square test for each user interaction type as shown on Figure 8. We used 5% for the statistical threshold to reject the null hypothesis.

Based on the results we can state that Idle and Manipulate state show extreme deviation from the values picked by chance (Manipulate 2.2%, Idle 1.9%), thus the effect of our avatar for choosing the correct matching were statistically significant. With other words it is extremely improbable that the results are unrelated to the avatar setup. All and Operate types showed no real deviation from the randomly selected answers thus they could not produce evidence for the usefulness of the avatar. Discover type however shows high probability for existing effect of the avatar without being considered as significant result (14% chance for no connection).

VII. Conclusions

In this paper, we broadened the scope of the avatar concept to include multi-user avatars using an abstract visual language, referred to as aggregated avatars. We argued that such avatars could be useful in multi-user platforms such as VR-based working environments. We proposed a design methodology and developed a reference implementation of the aggregated avatar concept on the MaxWhere VR platform.

Based on our statistical analyses we can conclude that two major types of user interactions (idle state, and rearrangement of spatial content) could be very well recognized by users with the use of our abstract aggregated avatar. A third type of interaction, referring to users moving around in the space, were also likely to be recognized by users at a higher success rate than chance. Other dimensions provided a promising but statistically not relevant result – although in the case of the "All" state, this could have been due to the fact that this state included some examples of all other interactions. Based on these results, we conclude that the proposed avatar design could be applicable to some contexts, and is worthy of further study and refinement.

VIII. Acknowledgement

Project no. C1015653 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the KDP-2020 funding scheme.

This research is contributing to project no. 2021-1.1.4-GYORSÍTÓSÁV-2022-00081 that has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2021-1.1.4-GYORSÍTÓSÁV funding scheme.

The research presented in this paper was also supported by the Hungarian Research Network (HUN-REN), and was partly carried out within the HUN-REN Cognitive Mapping of Decision Support Systems research group.

References

[1] L. de Wildt, T. H. Apperley, J. Clemens, R. Fordyce, and S. Mukherjee, "(Re-) Orienting the Video Game Avatar," *Games and Culture*, vol. 15, no. 8, pp. 962–981, 2020. DOI: 10.1177/155541201985889.

[2] M. E. Latoschik, D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch, "The Effect of Avatar Realism in Immersive Social Virtual Realities," in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, pp. 1–10, 2017. DOI: 10.1145/3139131.3139156.

[3] F. Sibilla and T. Mancini, "I Am (Not) My Avatar: A Review of the User-Avatar Relationships in Massively Multiplayer Online Worlds," *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, vol. 12, no. 3, 2018. DOI: 10.5817/CP2018-3-4.

[4] P. Baranyi, A. Csapo, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)*. Springer, 2015. DOI: 10.1007/978-3-319-19608-4.

[5] I. Horváth, Á. B. Csapó, B. Berki, A. Sudár, and P. Baranyi, "Definition, Background and Research Perspectives Behind 'Cognitive Aspects of Virtual Reality'(cVR)," *Infocommunications Journal*, no. SP, pp. 9–14, 2023. DOI: 10.36244/ICJ.2023.SI-IODCR.2.

[6] D. Aneja, D. McDuff, and S. Shah, "A High-Fidelity Open Embodied Avatar with Lip Syncing and Expression Capabilities," in *2019 International Conference on Multimodal Interaction*, pp. 69–73, 2019. DOI: 10.1145/3340555.3353744.

[7] D. Roth, J.-L. Lugrin, D. Galakhov, A. Hofmann, G. Bente, M. E. Latoschik, and A. Fuhrmann, "Avatar Realism and Social Interaction Quality in Virtual Reality," in *2016 IEEE Virtual Reality (VR)*, pp. 277– 278, IEEE, 2016. DOI: 10.1109/VR.2016.7504761.

[8] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Detailed Human Avatars from Monocular Video," in *2018 International Conference on 3D Vision (3DV)*, pp. 98–109, IEEE, 2018. DOI: 10.1109/3DV.2018.00022.

[9] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li, "Avatar Digitization from a Single Image for Real-Time Rendering," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 6, pp. 1–14, 2017. DOI: 10.1145/3130800.31310887.

[10] J. L. Ponton, V. Ceballos Inza, L. Acosta, A. Rios, E. Monclús, and N. Pelechano, "Fitted Avatars: Automatic Skeleton Adjustment for Self- Avatars in Virtual Reality," *Virtual Reality*, vol. 27, pp. 1–20, 07 2023. DOI: 10.1007/s10055-023-00821-z.

[11] C. Pedica and H. Högni Vilhjálmsson, "Spontaneous Avatar Behavior for Human Territoriality," *Applied Artificial Intelligence*, vol. 24, no. 6, pp. 575–593, 2010.DOI: 10.1080/08839514.2010.492165.

[12] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, "Cross-Cultural Differences in Recognizing Affect from Body Posture," *Interacting with computers*, vol. 18, no. 6, pp. 1371–1389, 2006. DOI: 10.1016/j.intcom.2006.04.003.

[13] M. Fabri, D. J. Moore, and D. J. Hobbs, "The Emotional Avatar: Non-Verbal Communication between Inhabitants of Collaborative Virtual Environments," in *Gesture-Based Communication in Human-Computer Interaction: International GestureWorkshop, GW'99 Gif-sur-Yvette, France, March 17-19, 1999 Proceedings*, pp. 269–273, Springer, 1999. DOI: 10.1007/3-540-46616-9_24.

[14] J. Yang, R. T. Marler, H. Kim, J. Arora, and K. Abdel-Malek, "Multi-Objective Optimization for Upper Body Posture Prediction," in *10th AIAA/ISSMO multidisciplinary analysis and optimization conference*, p. 4506, 2004. DOI: 10.2514/6.2004-4506.

[15] R. Rivu, D. Roth, F. Alt, and Y. Abdelrahman, "The Influence of Avatar Personalization on Emotions in VR," *Multimodal Technologies and Interaction*, vol. 7, p. 38, 03 2023. DOI: 10.3390/mti7040038.

[16] G. Persa and Á. B. Csapó, "A Fuzzy Driven Framework for the Cognitive Modeling and Quantitative Analysis of Aggregated Avatars," in *CINTI-MACRo 2022*, pp. 000 143–000 150, IEEE, 2022. DOI: 10.1109/CINTI-MACRo57952.2022.10029574.

[17] P. S. Moyer, G. Salkind, and J. J. Bolyard, "Virtual Manipulatives used by K-8 Teachers for Mathematics Instruction: The Influence of Mathematical, Cognitive, and Pedagogical Fidelity," *Contemporary Issues in Technology and Teacher Education*, vol. 8, no. 3, pp. 202–218, 2008.

[18] B. Bos, "Virtual Math Objects with Pedagogical, Mathematical, and Cognitive Fidelity," *Computers in Human Behavior*, vol. 25, no. 2, pp. 521–528, 2009. DOI: 10.1016/j.chb.2008.11.002.

[19] D. Lansel, T. Kergoat, S. Bardin, and L. Deruy, "Using Virtual Reality Domain to Set Up Sports Simulation as Part of Rehabilitation," 2021.

[20] Á. B. Csapó, I. Horvath, P. Galambos, and P. Baranyi, "VRasa Medium of Communication: From Memory Palaces to Comprehensive Memory Management," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 000 389–000 394, IEEE, 2018. DOI: 10.1109/CogInfoCom.2018.8639896.

[21] G. Persa, A. Csapo, and P. Baranyi, "CogInfoCom Systems from an Interaction Perspective – A Pilot Application for EtoCom," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 16, no. 2, pp. 297–304, 2012. DOI: 10.20965/jaciii.2012.p0297.

[22] B. Korcsok, V. Konok, G. Persa, T. Faragó, M. Niitsuma, Á. Miklósi, P. Korondi, P. Baranyi, and M. Gácsi, "Biologically Inspired Emotional Expressions for Artificial Agents," *Frontiers in psychology*, vol. 9, p. 1191, 2018. DOI: 10.3389/fpsyg.2018.01191.

[23] A. Sudár and Á. B. Csapó, "Descriptive Markers for the Cognitive Profiling of Desktop 3D Spaces," *Electronics*, vol. 12, no. 2, p. 448, 2023. DOI: 10.3390/electronics12020448.

[24] B. Berki, "Does Effective Use of MaxWhere VR Relate to the Individual Spatial Memory and Mental Rotation Skills?," *Acta Polytechnica Hungarica*, vol. 16, no. 6, pp. 41–53, 2019. DOI: 10.12700/APH.16.6.2019.6.4.

[25] G. Stankov, B. Nagy, *et al.*, "Eye Tracking Based Usability Evaluation of the MaxWhere Virtual Space in a Search Task," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 469–474, IEEE, 2019. DOI: 10.1109/CogInfoCom47531.2019.9089922.

György Persa This author obtained his MSc degree in 2009 from Pázmány Péter Catholic University in Budapest.
Following his graduation, he conducted scientific research at the Institute for Computer Science and Control of the Hungarian Academy of Sciences.
In 2020, he began his doctoral studies at Széchenyi István University in Győr. His primary research focus is on design and implementation in the areas of virtual reality, 3D applications, avatars, human-machine interfaces and cognitive computing.

Ádám B. Csapó obtained his PhD degree at the Budapest University of Technology and Economics in 2014. From 2016 to 2022, he has worked as an associate professor at the Széchenyi István University in Győr; and, since 2022, he continued his academic journey as as an associate professor at Óbuda University, Budapest, Hungary. Since September 2023, he has held the position of associate professor at Corvinus University of Budapest. Dr. Csapó's research focuses on cognitive infocommunication channels in virtual collaboration environments, i.e. enabling users to communicate with each other and with their spatial surroundings in novel and effective ways. At the same time, he has been involved in the development of assistive technology for the visually impaired, as well as in the development of a commercial VR platform. Dr. Csapó has over 50 publications, including 1 co-authored book and 20 journal papers, and has actively participated in the organization of numerous international conferences and special issues.