A Clustering Approach to Unveil User Similarities in 6 df Extended Reality Applications

Rossi, Silvia; Viola, Irene; Toni, Laura; Cesar, Pablo

DOI
10.1145/3701734

Publication date
2025

Document Version
Final published version

Published in
ACM Transactions on Multimedia Computing, Communications and Applications

A Clustering Approach to Unveil User Similarities in 6 df Extended Reality Applications

SILVIA ROSSI and IRENE VIOLA, Centrum Wiskunde and Informatica (CWI), Amsterdam, The Netherlands
LAURA TONI, University College London (UCL), London, United Kingdom
PABLO CESAR, Centrum Wiskunde and Informatica (CWI), Amsterdam, The Netherlands and Technische Universiteit (TU) Delft, Delft, The Netherlands

The advent in our daily life of Extended Reality (XR) technologies, such as Virtual and Augmented Reality, has led to the rise of user-centric systems, offering higher level of interaction and presence in virtual environments. In this context, understanding the actual interactivity of users is still an open challenge and a key step to enabling user-centric system. In this work, our goal is to construct an efficient clustering tool for 6 df navigation trajectories by extending the applicability of existing behavioural tool. Specifically, we first compare the navigation in 6 df with its 3 df counterpart, highlighting the main differences and novelties. Then, we investigate new metrics aimed at better modelling behavioural similarities between users in a 6 df system. More concretely, we define and compare 11 similarity metrics which are based on different *distance features* (i.e., user positions in the 3D space, user viewing directions) and *distance measurements* (i.e., Euclidean, Geodesic, angular distance). Our solutions are validated and tested on real navigation paths of users interacting with dynamic volumetric media in both 6 df Virtual Reality and Augmented Reality conditions. Results show that metrics based on both user position and viewing direction better perform in detecting user similarity while navigating in a 6 df system. Such easy-to-use but robust metrics allow us to answer a fundamental question for user-centric systems: 'How do we detect if users look at the same content in 6 df?', opening the gate to new solutions based on users interactivity, such as viewport prediction, live streaming services optimised based on users behaviour but also for user-based quality assessment methods.

CCS Concepts: • **Human-centered computing** → **User studies**; **Mixed / augmented reality**; **Virtual reality**; • **Hardware** → **Analysis and design of emerging devices and systems**; • **Information systems** → **Multimedia streaming**; • **Mathematics of computing** → *Cluster analysis*; *Exploratory data analysis*; Time series analysis; • **Computing methodologies** → *Cluster analysis*;

Additional Key Words and Phrases: Behavioural Analysis, Data Clustering, 6 df, Extended Reality, Virtual Reality, Trajectory analysis, immersive navigation, interaction analysis, clustering tools, similarity metrics

This article is an extended version of work published as Silvia Rossi, Irene Viola, Laura Toni, and Pablo Cesar. 2023. Extending 3 df metrics to model user behaviour similarity in 6 df immersive applications. In *14th Conference on ACM Multimedia Systems*, 39–50.

Part of this work was carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship Programme and in part supported through the European Commission Horizon Europe program under grant 101070109 TRANSMIXR (https://transmixr.eu/).

Authors' Contact Information: Silvia Rossi (corresponding author), Centrum Wiskunde and Informatica (CWI), Amsterdam, The Netherlands; e-mail: s.rossi@cwi.nl; Irene Viola, Centrum Wiskunde and Informatica (CWI), Amsterdam, The Netherlands; e-mail: irene@cwi.nl; Laura Toni, University College London (UCL), London, United Kingdom; e-mail: l.toni@ucl.ac.uk; Pablo Cesar, Centrum Wiskunde and Informatica (CWI), Amsterdam, The Netherlands and Technische Universiteit (TU) Delft, Delft, The Netherlands; e-mail: p.s.cesar@cwi.nl.

ACM 1551-6865/2025/9-ART254
https://doi.org/10.1145/3701734

ACM Reference format:
Silvia Rossi, Irene Viola, Laura Toni, and Pablo Cesar. 2025. A Clustering Approach to Unveil User Similarities in 6 df Extended Reality Applications. *ACM Trans. Multimedia Comput. Commun. Appl.* 21, 9, Article 254 (September 2025), 27 pages.
https://doi.org/10.1145/3701734

1 Introduction

Extended Reality (XR) is transforming the way users interact with media content, surpassing the passive paradigm of traditional video technology, and offering more degrees of presence and interaction in a virtual and immersive environment. This technology is envisioned to lead the next generation virtual worlds [20]. Specifically, the term XR indicates all current immersive technologies, spanning from fully physical to fully virtual world realities. While **Augmented Reality (AR)** combines virtual and real objects on a screen device, **Virtual Reality (VR)** allows users to immerse themselves in an entirely synthetic and virtual experience where they can navigate and interact. Depending on how much a user can move in the 3D space, immersive environments can be classified as 3 or 6 df. In the former scenario, the *de-facto* multimedia content is the *omnidirectional* or *spherical video*, representing an entire 360° environment on a virtual sphere. The viewer is fully immersed in a virtual space where they can navigate and interact thanks to an immersive device—typically a **Head-Mounted Display (HMD)**, which enables to view only a portion of the environment around themself, named *viewport*. The media is displayed from an *inward* position, and the viewer can interact with the content only by changing the viewing direction (i.e., by looking up/down or left/right or tilting the head side to side). In a 6 df system, the user can also change viewing perspective by moving (e.g., walking, jumping) inside the virtual space. The scene is therefore populated by *volumetric objects* (i.e., meshes or point clouds) which are observed from an *outward* position. These extra df bring the virtual experience even closer to reality: A higher level of interactivity makes the user more immersed and present within the virtual environment [9].

Despite their differences, the common denominator of both interactive systems is that the viewer acts as an active decision-maker of the displayed content. This active role defines an *user-centric* era, in which content processing, streaming and rendering need to be tailored to the viewer interaction to remain bandwidth-tolerant whilst meeting quality and latency criteria [37, 57]. Media codecs need to be optimised to maximise the quality experienced by users [17, 47, 59]. Similarly, streaming should be tailored to users interactivity to ensure high-quality content and smooth navigation [31, 46, 51]. From here, the importance to understand, analyse and predict users movements (i.e., *user behaviour*) within an immersive scenario [18, 35, 38, 55]. A better understanding of how the population behave within XR experience has an impact that goes also beyond multimedia system applications, leading to *user clustering/profiling* which is essential for several reasons, from security purposes to medical applications [32]. For example, in the context of authentication, profiling enables secure authentication for specific categories of users or continuous verification based on behavioural analysis, thus increasing security [54]. In medical applications, detecting cluster of similar users allows for personalised healthcare, while identifying outlier behaviour could simplify the detection and treatment of mental disorders [25].

Thanks to the large availability of publicly datasets [21, 28, 34], user navigation in 3 df immersive systems has been deeply investigated, showing the importance of analysing and detecting key behavioural aspects in interactive (user-centric) systems [1, 6, 41, 42, 48]. However, the 6 df counterpart has been overlooked in the literature [2, 19, 50, 60]. Due to the change in the viewing paradigm (from inward to outward) and to more level of interaction in 6 df, current studies in

(a) Point Clouds [23] from a VR experiment [51]. (b) Meshes [30] from an AR experiment [60].

Fig. 1. Volumetric sequences of publicly available navigation trajectory datasets used in our experiments.

3 df cannot be directly applied to 6 df domains [44]. Filling this gap is the main goal of this article by providing new metrics for user behavioural analysis in 6 df. Specifically, we focus on extending the applicability of clustering methods to investigate users similarity (i.e., users sharing common behaviours while interacting with the content) within 6 df environments. Starting from the state-of-the-art clustering algorithm developed in 3 df [36], and the main limitations of the tool when extended to 6 df described in [44], we further extend our previous work presented in [45] by investigating new methodologies for better modelling user similarities and overcoming those limitations. First, we recall the definition of user navigation trajectory in 6 df. Then, we present the exact user similarity metric, which we will be considering as our ground truth. Given its computational complexity, after an exhaustive study, we propose a simpler and yet reliable proxy for it. More concretely, we define and compare 11 similarity metrics which are based on different *distance features* (i.e., user positions in the 3D space, user viewing directions) and *distance measurements* (i.e., Euclidean, Geodesic distance, angular distance). We validate and test our proposed similarity metrics on a publicly available dataset of navigation trajectories collected in a 6 df VR scenario [51] based on four volumetric sequences shown in Figure 1(a). Results highlight that similarity metrics based on more distance features are promising solutions to correctly detect users with similar behaviour while experiencing volumetric content. Finally, we further validate the proposed tool by testing it on navigation trajectories collected in a different setting, a 6 df AR scenario [60] composed by two volumetric sequences shown in Figure 1(b). Similarities among users are detected as well in this new interactive setting, showing that the proposed metrics are general enough to be efficient in multiple interactive systems with 6 df.

In summary, we extend our previous work [45] by including user viewing direction as additional distance feature, and adding three new similarity metrics. We also present a novel use case of behavioural analysis in an AR environment to emphasise the importance of testing the proposed metrics in different XR settings. Thus, the main contributions of this article to the open problem of behavioural analysis in 6 df can be summarised as follows:

—Introduction of the general problem of detecting behavioural similarities in a 6 df system, and definition of novel similarity metrics able to model the user behaviour in this scenario. These are expressed as a function of various distance features and measurements and we divide them into two groups: *single-* and *multi-features metrics*;
—An exhaustive analysis of the different proposed metrics aimed at capturing users behaviour similarity (in terms of displayed content). This analysis based on 6 df VR trajectories reveals

that the position on the floor alone is not always sufficient to characterise the user behaviour and thus the viewing direction cannot be neglected;
— A case study of behavioural analysis in an AR system via a state-of-the-art clustering tool using our proposed similarity metrics. This second example tests our proposed metrics showing their flexibility and validity also in a different XR setting.

The remainder of this article is organised as follows: Section 2 reviews related work on behavioural analysis in 3 df and 6 df systems; the main challenges and the importance of detecting behavioural similarities in 6 df are discussed in Section 3. Our proposed similarity metrics are detailed in Section 4, with experimental setup and validation on real 6 df navigation trajectories collected in VR in Sections 5 and 6. Section 7 demonstrates the applicability of our metrics in a 6 df AR setting. Further discussions, including limitations and future work, are given in Section 8, and final conclusion in Section 9.

2 Related Work

We now describe how user behaviour has been analysed in 3 df systems, showing also the benefit of this type of analysis in user-centric systems. Then, we show which methods have been used for the analysis in 6 df scenarios, highlighting the still outstanding open challenges.

2.1 User Behaviour in 3 df Environment

The user navigation within a 3 df environment has been intensely analysed from different perspectives [37]. Many studies have focused on psychological investigations of user engagement and presence correlated to movements within the spherical content. In [22], a study from a large-scale experiment (511 users and 80 omnidirectional videos) showed a positive correlation between lower interactivity level and higher engagement level (strong focus on few points of interest). A correlation between the perceived sense of presence and the interactivity level was detected in [4], with more random exploratory interactions for less immersed (and hence less engaged) users. Recently, an exploration analysis has also shown benefit on the user of experience by aligning the displayed portion of the content with specific region of interest [3]. This impact has been examined, considering factors such as head motion, sense of presence and discomfort highlighting that innovative editing techniques involving gradual rotation of VR content contribute to improve the overall user experience. To further understand how people observe and explore VR contents, many publicly datasets of navigation trajectories have been made available. Those datasets usually come with statistical analysis aimed at capturing average users behaviour, as a function of average angular speeds under various video segment lengths [10], exploration time [48] or eye fixation distribution [12]. These traditional analysis have been exploited also to investigate the immersive navigation of dynamic scenes characterised by directional sounds [5]. However, no objective metric to properly quantify and characterise user behaviour has been presented in these works.

In [28], the dataset has been analysed through a clustering algorithm presented in [36], specifically built to have in the same cluster users who similarly explore 360° content. This investigation highlighted that movies with few focus of attention lead to higher engagement shown by users with strong similarities and hence collected into few and high-populated clusters. Authors in [40] showed the advantages of employing information theory metrics to study spatio-temporal trajectories, providing a tool to identify and quantify behavioural aspects. This preliminary quantitative analysis not only explored the similarities between users watching the same content but also investigated the similarity of a given user across diverse content. A recent follow-up data analysis using such information theory metrics and across several publicly available VR datasets has also unveiled correlations between users head motion and trajectory predictability [41]. The importance of these

behavioural insights has been proved to be crucial for different VR applications. A critical open problem is the ability to predict users' navigation trajectories within the virtual space. Being able to anticipate viewers' movement is essential to ensure high quality of experience and smooth navigation during the immersive experience. For instance, in a tile-based adaptive streaming scenario, each user receives at high quality only tiles that overlap the predicted displayed portion of the content [39]. This strategy, while effective from both a bandwidth and quality perspective, strongly depends on the performance of the selected prediction algorithm. An erroneous estimate would immediately lead to re-transmissions, and hence, a possible stall or quality reduction effect. Therefore, many new learning models have been proposed to anticipate users' movements [7, 16, 19, 27, 29]. Finally, the analysis and understanding of user navigation in a VR environment have also shown promising results in determining mental health issues (e.g., anxiety, autism spectrum disorder, eating disorders, depression) and to help their treatment [14, 15, 26].

2.2 User Behaviour in 6 df Environment

Extending such behavioural analysis to a 6 df environment is not straightforward due to the change from inward to outward viewing and the addition of translation in 3D space. In the past, user navigation in 6 df scenarios was studied in the context of locomotion and display technology for **Cave Automatic Virtual Environment (CAVE)** environments [33, 53]. A CAVE system is an immersive room on which walls and floor are projected the video content and viewers are free to move inside [11]. For instance, the study in [53] focused on task performance analysis in terms of completion time and correct actions. Authors in [33] compared instead the effect of two different immersive platforms such as CAVE and HMD on the user navigation. More traditional metrics, such as angular distance and linear velocity, alongside completion time, were also used to compare different navigation controllers (i.e., joystick-based vs. head-controlled navigation) in 6 df [8]. Similarly, the navigation with 6 df of users in the form of digital avatars has been also deeply investigated to detect insights into how they behave and move in virtual worlds (e.g., Second Life) [24]. Their focus was into both temporal and spatial dynamics of variations in avatar population and the exploration of spatial distribution, movement patterns and contact interactions among avatars. While the tools and methodologies mentioned above are highly informative to summarise the interaction of users within a 6 df environment, they usually fail to provide other key insights: which users navigate similarly, and which are the dominant interaction behaviour among users.

More recently, subjective quality assessment of both static [2] and dynamic [17, 52] volumetric content has been presented along with general statistical analysis of user movements showing an influence in the navigation given by the perceived content quality, and pointing out a user's preference to visualise volumetric objects from a close and frontal perspective. This last finding was also confirmed in a behavioural navigation analysis conducted in an AR mobile application [60]. Here, viewers movements were analysed in terms of distribution on the floor, viewing angles and relative distance from the content. A behavioural analysis of user navigating in 6 df social VR movie has been also presented in [43]. An investigation on how users are affected by virtual characters and narrative elements of the movie has been conducted through objective metrics, showing a more static behaviour when an interactive task was requested, and more exploratory movements during dialogues. Another exploratory analysis shows user behaviour while displaying volumetric content in a 6 df environment, examining the influence of content features, dynamics, quality and users intrinsic disposition [42]. Specifically, these investigations have been based on both traditional statistical metrics like distribution of viewing position and direction, exploratory velocity and total viewing time, alongside adapted 3 df tools such as information theory metrics [40] and clustering tools [36]. Given the importance of collecting navigation data in 6 df immersive experiences, a novel tool was recently introduced [56].

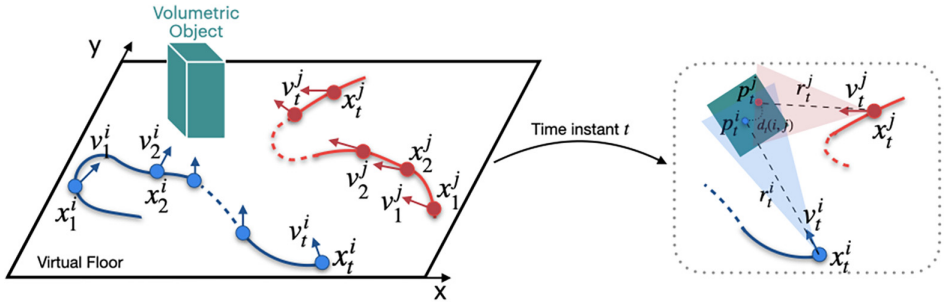


Fig. 2. Example of two 6 df trajectories projected in a 2D domain for user i and j. On the right side, a snapshot at time t: coloured triangles represent viewing frustum per user.

The aforementioned studies are based on conventional metrics, which consider only one user feature at a time, either position on the floor or viewing direction but not together. In this article, we aim to overcome these limitations by proposing a general and efficient tool for detecting similar viewers while experiencing 6 df content.

3 Challenges

Our main goal is to define a new pairwise metric able to capture the (dis)similarity between two 6 df users (in terms of displayed content). This metric needs to be reliable and yet simple to compute. In the following, we first define navigation trajectories in 6 df scenario comparing with its 3 df counterpart and present our definition of similarity among users while navigating in a 6 df environment based on [44]. Then, we show an exact user similarity metric highlighting its limitations, and therefore the need to find a simpler proxy for it. Finally, we emphasise the advantages of having a similarity metric for behavioural analysis via a clique-based clustering approach presented in [36], which detect users who are attending the same portion of an omnidirectional content. This tool relies on a pairwise similarity metric, which is a solid metric in 3 df, but results to be poor in 6 df. Hence, the need to develop a proper metric representative for 6 df system to extend the applicability of such behavioural tool to 6 df scenario.

3.1 User Similarity in 6 df

Following [44, 45], we assume that users behave similarly when they *observe the same portion of volumetric content*. The user behaviour can be identified by the spatio-temporal sequences of their movements within the virtual environment, namely *navigation trajectories*.

In a 3 df scenario, the trajectory of a generic user i can be formally denoted by the sequence of the user's position and the corresponding viewing direction over time: $\{(x_1^i, v_1^i), (x_2^i, v_2^i), .., (x_n^i, v_n^i)\}$ where x_t^i is the user position while v_t^i is the vector representing the viewing direction at timestamp t. In this context, however, users are positioned at the centre of the spherical content; thus, x_t^i is constant and can be neglected. The vector of the viewing direction can be also approximated by p_t^i, which is the centre of the viewport projected on the immersive content (i.e., spherical video), such that the trajectory becomes $\{p_1^i, p_2^i, .., p_n^i\}$ [44]. The viewport centre alone is highly informative of the user behaviour, and it can be used as proxy of viewport overlap among users as shown in [36]. Specifically, the geodesic distance between viewport centres is highly reliable as similarity metric to asses users' similarity, namely a low value indicates high similarity between 3 df users.

In a 6 df setting, however, the added df lead to more challenges in the design of the system and in the representation of user navigation. Figure 2 shows an example of two users

navigating in a 6 df scenario. On the left side, there are navigation trajectories of two users i and j projected on a 2D domain (i.e., floor). Each point x_t represents the spatial coordinates (i.e., [x, y, z]) on the floor, while each associated vector symbolises the viewing direction v_t. The navigation trajectory of a generic user i can be represented as $\{(x_1^i, v_1^i), (x_2^i, v_2^i), .., (x_n^i, v_n^i)\}$ where x_t^i and v_t^i are the user position and viewing direction vector at timestamp t, respectively. However, unlike the 3 df scenario, the users' position changes over time; therefore their distance from the immersive content can also change over time. As a consequence, the viewport centre alone is no longer sufficient to characterise the user behaviour [44]. On the right part of Figure 2, there is indeed a snapshot at a specific time instant t. In more detail, the shaded triangular areas represent the *viewing frustum* per user, which indicates the region within the user viewport, and r_t is the distance between the user and the volumetric content. We have also depicted the viewport centre p_t projected on the displayed volumetric object. Given the two users i and j at time t, in the case of $r_t^i \gg r_t^j$, the user j (very close to the object) is visualising a very focused and detailed part of it; conversely, user i is pointing to the same area but from a much further distance, thus the experienced content is different with less defined details. Despite this difference, the small distance $D_t(i, j)$ between viewport centres p_t^i and p_t^j might suggest a high similarity between the users, which does not reflect the reality in the case of $r_t^i \gg r_t^j$. In this scenario, we cannot rely on the viewport centre alone to characterise the user behaviour; the distance r and the spatial coordinates on the virtual floor x are also needed to define the navigation trajectory for a generic 6 df user i. Thus, an alternative definition of navigation trajectory is given by the following triple over time $\{(x_1^i, p_1^i, r_1^i), (x_2^i, p_2^i, r_2^i), \ldots, (x_n^i, p_n^i, r_n^i)\}$. This information is crucial to define a simple similarity metric among users in this new setting.

3.2 Overlap Ratio as the Ground-Truth Metric

Since we are interested in capturing viewers that are attending similar volumetric content at the same time instance, following the work presented in [45], the straightforward measure that could show this behaviour is the overlap among viewports. Given the two users i and j in Figure 2, we denote their displayed viewport at time t as $\mathcal{S}_t^i$ and $\mathcal{S}_t^j$, respectively, defined as the set of points of the volumetric content falling within their viewing frustum. Then, we denote the overlap set by $\mathcal{S}_t^i \cap \mathcal{S}_t^j$, the portion of points displayed by both users. Equipped with the above notation, we can now introduce a key metric for the analysis: the *overlap ratio* $O(i, j)$. This is defined as the cardinality of the overlap set, normalised by the cardinality of the set containing all points of the volumetric content visualised by both users. More formally, the overlap ratio in a specific time t is

$$O_t(i, j) = \frac{|\mathcal{S}_t^i \cap \mathcal{S}_t^j|}{|\mathcal{S}_t^i \cup \mathcal{S}_t^i|}, \tag{1}$$

where $\mathcal{S}_t^i$ and $\mathcal{S}_t^j$ are the displayed viewport of users i and j, respectively. In particular, a high value of overlap ratio means high similarity between users, and conversely. Even if this metric is exact and a clear indicator of how much similar users are with respect to their displayed content, its evaluation is not trivial as it is intensely time-consuming. For instance, the overlap ratio between two users requires on average 0.8986 seconds per frame on an Intel R machine with CPU E5-4620 at 2.10 GHz. This operation needs to be computed for all the possible combinations of users, leading to a large overhead which does not meet requirements for real-time and scalable applications. A new measure is needed to perform real-time applications. In the rest of the article, we will use this metric as the ground truth of overlap among users and investigate different weights that can approximate viewport overlap between users, and thus being an indication of users' (dis)similarity.

Table 1. Definition of Distance Features and Measurements

	Symbol	Definition
Distance Features	x	User position on the VR floor
	p	Viewport centre projected on the volumetric content
	r	Relative distance between user and volumetric content
	v	Vector of the viewing direction
Distance Measurements	$\mathsf{L}(\cdot,\cdot)$	Difference of relative distance between two users
	$\mathsf{E}(\cdot,\cdot)$	Euclidean distance
	$\mathsf{G}(\cdot,\cdot)$	Geodesic distance
	$\theta(\cdot,\cdot)$	Angle between two vectors

3.3 Clustering as a Tool for Behavioural Analysis

Being able to assess users' similarities in an objective way is crucial to detect users with similar behaviour thorough a clique-based clustering algorithm presented in [36]. This requires indeed a reliable graph where only the nodes identifying similar users (i.e., who are displaying the same portion of the content) are connected. Equipped with such a meaningful graph, the clique-based clustering iteratively finds optimal sub-graphs of all interconnected nodes, ensuring the largest cluster of users who all share a large viewport overlap. Specifically, given a set of users who are experiencing the same content, we can represent their movements in a time-window T as a set of graphs $\{\mathcal{G}_t\}_{t=1}^{T}$. Each unweighted and undirected graph $\mathcal{G}_t = \{\mathcal{V}, \mathcal{E}_t, \mathrm{A}_t\}$ represents behavioural similarities among users at time t, where $\mathcal{V}$ and $\mathcal{E}_t$ denote the node and edge sets of $\mathcal{G}_t$, respectively. Each node in $\mathcal{V}$ corresponds to a user interacting with the content. Each edge in $\mathcal{E}_t$ connects neighbouring nodes defined by the binary adjacency matrix A_t. Assuming that users are connected if they are displaying similar content, we can formally define the adjacency matrix A_t as follows:

$$\mathrm{A}_t(i,j) = \begin{cases} 1, & \text{if } g_t(i,j) \geq G_{th} \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where $g_t(i,j)$ is a similarity metric between user i and j with G_{th} as a threshold value. In [36], this graph construction is based on a pairwise metric specific to 3 df trajectories. On this final graph, the clique-based clustering can be applied to identify clusters of users with similar behaviour.

Identifying a general and reliable metric $g(i,j)$ that approximates behavioural similarities among users who experience the same 6 df content is a key step to extend the applicability of these exiting behavioural tools, and it is the main focus of this article, aimed at formulating various multi-modal metrics, and testing/validating them with real-world data from XR settings.

4 Proposed Metrics

In this section, we present 11 similarity metrics that will be the object of an exhaustive study in the following to understand which one approximates at the best the viewport overlap. Those metrics are expressed as a function of various *distance features* and *measurements* considering either users' position on the floor (x) or users' viewing direction in terms of the viewport centre projected on the volumetric content (p) or viewing vector (v) or a combination of them. We divide the proposed similarity metrics into two groups: *single-feature* and *multi-feature* metrics. For the sake of notation, we omit the temporal parameter t. Table 1 summarises the distance features and measurements that we consider, while our proposed similarity metrics are reported in Table 2.

Table 2. Similarity Metrics: Definitions, Included Distance Features and Measurements, Regulator and Threshold Values

Symbol	Definition	Distance Feature and Metric	Regulator Values	S_{th}
w_1	$k_\alpha^{(\mathsf{E})}(x^i, x^j)$	$\mathsf{E}(x^i, x^j)$	$\alpha = 1$	0.61
w_2	$k_\alpha^{(\mathsf{L})}(r^i, r^j)$	$\mathsf{L}(r^i, r^j)$	$\alpha = 1$	0.78
w_3	$k_\alpha^{(\mathsf{G})}(p^i, p^j)$	$\mathsf{G}(p^i, p^j)$	$\alpha = 1$	0.59
w_4	$k_\alpha^{(\mathsf{E})}(p^i, p^j)$	$\mathsf{E}(p^i, p^j)$	$\alpha = 1$	0.83
w_5	$k_\alpha^{(\theta)}(v^i, v^j)$	$\theta(v^i, v^j)$	$\alpha = 1$	0.76
w_6	$k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot k_\beta^{(\mathsf{L})}(r^i, r^j) \cdot k_\gamma^{(\mathsf{G})}(p^i, p^j)$	$\mathsf{E}(x^i, x^j)$, $\mathsf{L}(r^i, r^j)$, $\mathsf{G}(p^i, p^j)$	$\alpha = 0.5$; $\beta = 0.05$; $\gamma = 0.2$	0.59
w_7	$k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot k_\beta^{(\mathsf{L})}(r^i, r^j) \cdot k_\gamma^{(\mathsf{E})}(p^i, p^j)$	$\mathsf{E}(x^i, x^j)$, $\mathsf{L}(r^i, r^j)$, $\mathsf{E}(p^i, p^j)$	$\alpha = 0.125$; $\beta = 0.05$; $\gamma = 0.2$	0.75
w_8	$k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot k_\beta^{(\mathsf{L})}(r^i, r^j) \cdot k_\gamma^{(\theta)}(v^i, v^j)$	$\mathsf{E}(x^i, x^j)$, $\mathsf{L}(r^i, r^j)$, $\theta(v^i, v^j)$	$\alpha = 0.125$; $\beta = 0.05$; $\gamma = 0.1$	0.75
w_9	$k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot \beta[\eta(r_i) + \eta(r_j)] \cdot k_\gamma^{(\mathsf{G})}(p^i, p^j)$	$\mathsf{E}(x^i, x^j)$, r^i,r^j, $\mathsf{G}(p^i, p^j)$	$\alpha = 0.5$; $\beta = 0.5$; $\gamma = 0.25$	0.69
w_{10}	$k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot \beta[\eta(r_i) + \eta(r_j)] \cdot k_\gamma^{(\mathsf{E})}(p^i, p^j)$	$\mathsf{E}(x^i, x^j)$, r^i, r^j, $\mathsf{E}(p^i, p^j)$	$\alpha = 0.25$; $\beta = 0.5$; $\gamma = 0.5$	0.81
w_{11}	$k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot \beta[\eta(r_i) + \eta(r_j)] \cdot k_\gamma^{(\theta)}(v^i, v^j)$	$\mathsf{E}(x^i, x^j)$, r^i, r^j, $\theta(v^i, v^j)$	$\alpha = 0.5$; $\beta = 0.5$; $\gamma = 0.1$	0.76

4.1 Single-Feature Metrics to Assess Users' Similarity

The first set of similarity metrics is based on one distance feature. We model the similarity functions via radial basis function kernel. Specifically, we consider the following Gaussian kernel [49]:

$$k_\alpha^{(D)}(i, j) = e^{-\alpha D(i,j)}, \tag{3}$$

where $D(i, j)$ is the selected distance between two generic users i and j, while $\alpha > 0$ is a parameter to better regularise the distance. The distance $D(i, j)$ can be evaluated in multiple ways and we consider the distance features and measurements taken into account in [45]. In this work, we also introduce metrics based on the angle between the vector of users viewing direction.

The first two similarity metrics w_1 and w_2 are based only on the location of users in the virtual space with respect to the virtual object or other viewers. The former metric is based on the Euclidean distance $\mathsf{E}(x^i, x^j)$ between user i and j on the virtual floor. Instead, w_2 considers the difference in terms of the relative distance of users to the centroid of the displayed content, $\mathsf{L} = ||r^i - r^j||$. Specifically, we define them as follows:

$$w_1 = e^{-\alpha \mathsf{E}(x^i, x^j)} = k_\alpha^{(\mathsf{E})}(x^i, x^j); \tag{4}$$

$$w_2 = e^{-\alpha ||r^i - r^j||} = k_\alpha^{(\mathsf{L})}(r^i, r^j). \tag{5}$$

The metrics w_3 and w_4 are instead based on the distance between the viewport centres p of user i and user j projected on the volumetric content. To take into account the heterogeneous shape of the volumetric content, the distance in w_3 is measured in terms of the Geodesic distance $\mathsf{G}(p^i, p^j)$ while in w_4 in terms of the Euclidean distance $\mathsf{E}(p^i, p^j)$. More formally, they are defined as

$$w_3 = k_\alpha^{(\mathsf{G})}(p^i, p^j) = e^{-\alpha \mathsf{G}(p^i, p^j)}, \tag{6}$$

$$w_4 = k_\alpha^{(\mathsf{E})}(p^i, p^j) = e^{-\alpha \mathsf{E}(p^i, p^j)}. \tag{7}$$

Finally, the metric w_5 is based on the angular distance $\theta(v^i, v^j)$ between the two vectors of the viewing direction of user i and user j. Specifically, it is defined as follows:

$$w_5 = k_\alpha^{(\theta)}(v^i, v^j) = e^{-\alpha \theta(v^i, v^j)}. \tag{8}$$

4.2 Multi-Feature Metrics to Assess Users' Similarity

As emerged in [44], both user viewing direction and position on the virtual floor are relevant to detect similar behaviour among users. Thus, the last set of proposed similarity metrics considers a combination of distance features and measurements. Appendix A depicts a further analysis of the correlation among these selected distance features and measurements. Despite a general correlation between the selected metrics, this does not result consistent across the different visualised volumetric content. Thus, we consider all of them to propose novel multi-feature similarity metrics and to determine which one best approximates the viewport overlap among users. Specifically, we define w_6, w_7 and w_8 based on the previous similarity metrics w_1 and w_2, but include also the distance of their viewport centres p projected on the volumetric content in terms of Geodesic distance $\mathsf{G}(p^i, p^j)$, Euclidean distance $\mathsf{E}(p^i, p^j)$ and angular distance $\theta(v^i, v^j)$, respectively. More formally, we define w_6 as

$$\begin{aligned} w_6 &= k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot k_\beta^{(\mathsf{L})}(r^i, r^j) \cdot k_\gamma^{(\mathsf{G})}(p^i, p^j) \\ &= e^{-\alpha \mathsf{E}(x^i, x^j)} \cdot e^{-\beta ||r^i - r^j||} \cdot e^{-\gamma \mathsf{G}(p^i, p^j)}; \end{aligned} \tag{9}$$

the second weight is equal to

$$\begin{aligned} w_6 &= k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot k_\beta^{(\mathsf{L})}(r^i, r^j) \cdot k_\gamma^{(\mathsf{E})}(p^i, p^j) \\ &= e^{-\alpha \mathsf{E}(x^i, x^j)} \cdot e^{-\beta ||r^i - r^j||} \cdot e^{-\gamma \mathsf{E}(p^i, p^j)}; \end{aligned} \tag{10}$$

and finally w_7 is equal to

$$\begin{aligned} w_7 &= k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot k_\beta^{(\mathsf{L})}(r^i, r^j) \cdot k_\gamma^{(\theta)}(v^i, v^j) \\ &= e^{-\alpha \mathsf{E}(x^i, x^j)} \cdot e^{-\beta ||r^i - r^j||} \cdot e^{-\gamma \theta(v^i, v^j)}. \end{aligned} \tag{11}$$

For the sake of clarity, β and γ are regulators such as α.

The preliminary analysis presented in [44] has also highlighted a correlation between the viewport overlap of two users and their relative distance from the volumetric content. The closer users are to the volumetric content, the smaller and more detailed is the portion of the displayed content; the farther they are, the bigger but with fewer details becomes the displayed portion. Thus, in the first case, the high overlap between displayed areas of two different users is more difficult. To take into consideration this behaviour, we model the relative distance via a hyperbolic tangent kernel. This function captures the non-linear relationship between user distance and content overlap leading to a better representation of the user interactions. Given the relative distance r_i between the user i and volumetric content, we evaluate it as follows:

$$\eta(r_i) = \tanh(r_i). \tag{12}$$

Thus modelling the relative distance of users with this function, metrics w_9 and w_{10} are based on both user distance in the virtual floor $\mathsf{E}(x^i, x^j)$, and on the volumetric content in terms of Geodesic distance $\mathsf{G}(p^i, p^j)$ and Euclidean distance $\mathsf{E}(p^i, p^j)$, respectively; while w_{11} considers angular distance $\theta(v^i, v^j)$ together with the user distance in the virtual floor $\mathsf{E}(x^i, x^j)$. More formally, we define w_9 as follows:

$$\begin{aligned} w_9 &= k_\alpha^{(\mathsf{E})}(x^i, x^j) \cdot \beta\big[\eta(r^i) + \eta(r^j)\big] \cdot k_\gamma^{(\mathsf{G})}(p^i, p^j) \\ &= e^{-\alpha \mathsf{E}(x^i, x^j)} \cdot \beta\big[\tanh(r_i) + \tanh(r_j)\big] \cdot e^{-\gamma \mathsf{G}(p^i, p^j)}; \end{aligned} \tag{13}$$

while w_{10} is

$$\begin{aligned} w_{10} &= k_{\alpha}^{(\mathsf{E})}(x^i, x^j) \cdot \beta\big[\eta(r^i) + \eta(r^j)\big] \cdot k_{\gamma}^{(\mathsf{E})}(p^i, p^j) \\ &= e^{-\alpha \mathsf{E}(x^i, x^j)} \cdot \beta\big[\tanh(r_i) + \tanh\big(r_j\big)\big] \cdot e^{-\gamma \mathsf{E}(p^i, p^j)}; \end{aligned} \tag{14}$$

and finally, w_{11} is

$$\begin{aligned} w_{11} &= k_{\alpha}^{(\mathsf{E})}(x^i, x^j) \cdot \beta\big[\eta(r^i) + \eta(r^j)\big] \cdot k_{\gamma}^{(\theta)}(v^i, v^j) \\ &= e^{-\alpha \mathsf{E}(x^i, x^j)} \cdot \beta\big[\tanh(r_i) + \tanh\big(r_j\big)\big] \cdot e^{-\gamma \theta(v^i, v^j)}. \end{aligned} \tag{15}$$

5 Experimental Setup

We now validate and test our proposed similarity metrics based on real navigation trajectories collected in a VR setting and selected performance metrics. In the following, we first describe the navigation dataset and how we evaluate the performance of our similarity metrics (Sections 5.1 and 5.2, respectively).

5.1 Dataset and Methodology

Dataset. Existing datasets with user navigation collected while displaying volumetric objects in a 6 df environment are still very limited. In the following, we use the open dataset presented in [51]. This is comprised of navigation trajectories of 26 users participating in a visual quality assessment study in VR. For the study, four dynamic point cloud sequences were employed [23], namely *LongDress* (PC1), *Loot* (PC2), *Red and Black* (PC3), *Soldier* (PC4) (Figure 1(a)). Each sequence was distorted at four different bit rate points with two compression algorithms (i.e., MPEG anchor codec and standard V-PCC). Hidden references were also employed in the test, for a total of 36 stimuli. Similarly to what is shown in Figure 2, a single object of interest was placed in the VR scene, and users were instructed to focus on the volumetric content for the duration of the session and rate its visual quality. Thus, the navigation data adhere to the assumptions listed in Section 3. However, it is important to note that, since this was not a task-free experiment, user navigation may have been influenced by the requirement to rate the content's quality, which we leave to future work for further investigation. Finally, in the experiments conducted in [51], each sequence was 5 seconds long and looped, allowing participants to watch the content as long as they wanted. In the following analysis, to ensure data consistency across all participants, we decided to consider only the first 5 seconds of the collected trajectories, which correspond to the first loop of the volumetric content.

Graph Construction. To implement the graph-based clustering proposed in [36] based on our proposed similarity metrics, we need to construct a binary graph following Equation (2), as described in Section 3.3. To be noted, our proposed similarity metrics are based on distance measurements. As shown in [36], the correlation between overlap and distance is inversely proportional. This means that high values of overlap (and thus, high similarity) correspond to low distance. Therefore, the condition to construct the adjacency metric A_t based on our proposed similarity metrics becomes the following: $w(i, j) \leq S_{th}$ where $w(i, j)$ is one of the similarity metrics in Table 2. S_{th} is a threshold value which identifies similar users and thus, neighbours on the graph. In short, users with a similarity metric below S_{th} are neighbours in the graph. Hence, the first step now is to identify S_{th}. For each proposed similarity metric, we empirically evaluate the **Receiver Operating Characteristic (ROC)** curves based on the navigation trajectories of the entire dataset (i.e., navigation trajectories of both distorted and reference version of the content) described above and select the best value of threshold as originally done in [36]. Specifically, we set the thresholding values such that a good tradeoff between **True-Positive Rate (TPR)** and **False-Positive Rate**

(FPR) is met. As ground truth for the ROC, we assumed that two users are attending the same portion of the content, and thus are classified as similar, if their viewports overlap by at least 75% of their total viewed area as in the original work [39]. The predicted event is instead evaluated using the 11 metrics presented in the previous section, and the corresponding threshold values are selected in order to have TPR equal to 0.75. For the sake of clarity, the ground-truth value of viewport overlap has been set equal to 75% because this ensures per each similarity metric a low probability to have a wrong classification (i.e., FPR below 0.4) without compromising the probability of correctly classifying the similarity event (i.e., TPR) which remains above 0.75. In the last column of Table 2, we provide the selected S_{th} per each similarity metric that will be used in the following. To tune the best set of regulator parameters, we also run an ablation study based on the entire selected dataset of navigation trajectories which is presented in Appendix B. Table 2 also reports all the final values selected after the ablation study and that will be used in the following.

5.2 Performance Evaluation Setup

To test our proposed similarity metrics, we consider three performance metrics: the averaged *overlap ratio* per cluster, the *relevant clustered population*, and the *precision*. The first two are more specific to our navigation trajectory in VR systems, while the last one is a popular index used to evaluate clustering algorithm performance.

Overlap Ratio per Cluster. As defined in Section 3.2, the overlap ratio computes the portion in common of displayed content between two users. Therefore, to compare the performance of our detected clusters with the different similarity metrics, we average the overlap ratio among all the pair of users who are put in the same group. More formally, given a detected cluster C_k, the corresponding overlap ratio O_k is defined as follows:

$$O_k = \frac{1}{n_k} \sum_{\substack{i,j \in C_k \\ i \neq j}} O(i,j), \tag{16}$$

where i and j are two generic users, n_k is the cardinality of elements bellowing to cluster C_k and $O(i,j)$ is the overlap ratio as in Equation (1).

Relevant Clustered Population. The more users are clustered together with high viewport overlap, the more meaningful our clusters become. We consider clusters with more than two elements as relevant. Thus, the relevant clustered population is the ratio of users in these types of clusters.

Precision. In a classification task, this index evaluates the portion of elements that are classified correctly and has values between 0 and 1 [13]. More formally:

$$P = \frac{TP}{TP + FP}, \tag{17}$$

where True Positive (TP) (False Positive (FP)) is the number of viewers classified correctly (incorrectly) together in a cluster. In our case, two users are identified positively if they are in the same cluster and their viewport overlap is actually over the desired value (i.e., 75% of overlap).

6 Results

Equipped with the similarity metrics, the corresponding values of regulators and thresholds given in Table 2, we now conduct our validation study. In this part of the study, we decided to focus only on the analysis of the dataset experienced in VR, Figure 1(a). In particular, we investigate the navigation trajectories experienced with non-distorted content to avoid any bias due to the quality of the content. First, we consider our metrics in a frame-based scenario in which users are clustered in one given frame at a time, then we test our proposed metrics over a time window of duration 1 second.

Table 3. Results in Terms of Average and SD per Each Performance Metric across the Navigation Trajectories Experienced with Not-Distorted Content in the Selected Dataset (Figure 1(a))

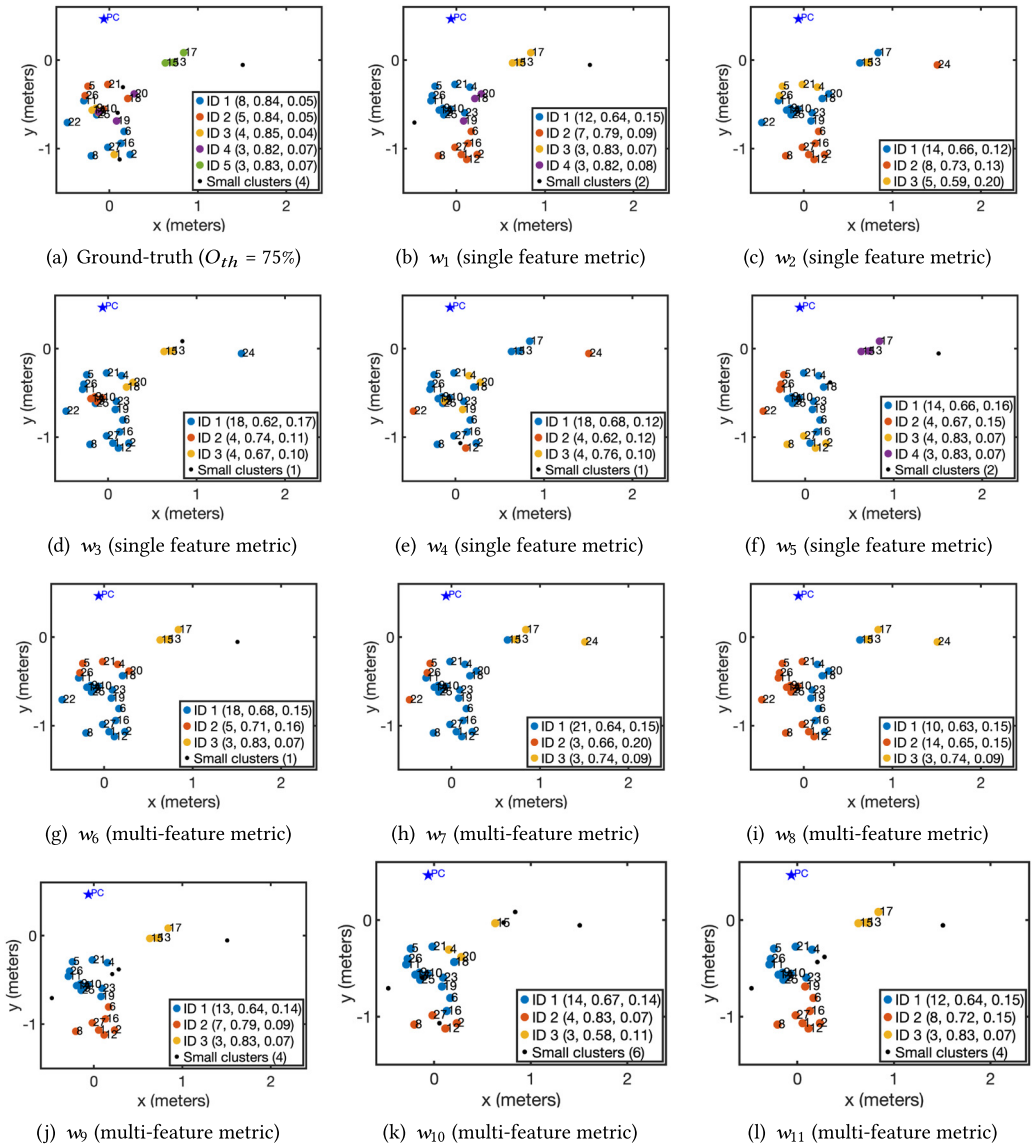
	Metrics	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}
PC1	Overlap	0.69 ± 0.03	0.64 ± 0.04	0.66 ± 0.04	0.68 ± 0.07	0.66 ± 0.04	0.68 ± 0.05	**0.73 ± 0.05**	0.67 ± 0.04	0.68 ± 0.05	0.72 ± 0.05	0.67 ± 0.04
	Relevant Pop.	0.86 ± 0.05	**0.94 ± 0.04**	0.93 ± 0.05	0.87 ± 0.06	0.84 ± 0.05	0.85 ± 0.06	0.81 ± 0.09	0.89 ± 0.04	0.83 ± 0.06	0.82 ± 0.09	0.88 ± 0.07
	Precision	0.42 ± 0.06	0.33 ± 0.05	0.38 ± 0.07	0.31 ± 0.06	0.34 ± 0.04	0.45 ± 0.06	**0.47 ± 0.10**	0.41 ± 0.05	0.43 ± 0.08	0.38 ± 0.08	0.40 ± 0.05
PC2	Overlap	0.56 ± 0.08	0.54 ± 0.08	0.55 ± 0.10	0.55 ± 0.10	0.52 ± 0.08	0.58 ± 0.07	0.56 ± 0.09	0.57 ± 0.07	0.57 ± 0.09	**0.60 ± 0.09**	0.57 ± 0.07
	Relevant Pop.	0.86 ± 0.06	**0.92 ± 0.04**	0.85 ± 0.07	0.91 ± 0.06	0.81 ± 0.08	0.82 ± 0.08	0.80 ± 0.06	0.83 ± 0.06	0.77 ± 0.10	0.69 ± 0.12	0.77 ± 0.07
	Precision	0.43 ± 0.09	0.27 ± 0.07	0.31 ± 0.08	0.27 ± 0.08	0.36 ± 0.08	0.45 ± 0.09	0.39 ± 0.09	0.43 ± 0.09	0.50 ± 0.07	**0.52 ± 0.07**	0.51 ± 0.07
PC3	Overlap	0.64 ± 0.05	0.59 ± 0.06	0.63 ± 0.05	0.68 ± 0.06	0.62 ± 0.06	0.64 ± 0.06	0.65 ± 0.05	0.64 ± 0.05	0.67 ± 0.06	**0.71 ± 0.07**	0.67 ± 0.06
	Relevant Pop.	0.86 ± 0.06	**0.93 ± 0.05**	0.90 ± 0.07	0.84 ± 0.07	0.80 ± 0.06	0.84 ± 0.07	0.80 ± 0.09	0.82 ± 0.05	0.77 ± 0.06	0.59 ± 0.09	0.74 ± 0.07
	Precision	0.47 ± 0.12	0.34 ± 0.09	0.39 ± 0.06	0.38 ± 0.06	0.42 ± 0.08	0.48 ± 0.11	0.49 ± 0.10	0.47 ± 0.10	0.51 ± 0.11	**0.54 ± 0.12**	0.51 ± 0.14
PC4	Overlap	0.58 ± 0.04	0.52 ± 0.05	0.56 ± 0.03	0.59 ± 0.06	0.55 ± 0.06	0.59 ± 0.04	0.60 ± 0.04	0.58 ± 0.04	0.61 ± 0.04	**0.66 ± 0.05**	0.61 ± 0.05
	Relevant Pop.	0.85 ± 0.06	**0.92 ± 0.04**	**0.92 ± 0.06**	0.88 ± 0.08	0.84 ± 0.06	0.85 ± 0.06	0.81 ± 0.07	0.84 ± 0.05	0.83 ± 0.07	0.67 ± 0.09	0.80 ± 0.06
	Precision	0.35 ± 0.07	0.22 ± 0.04	0.31 ± 0.06	0.25 ± 0.07	0.28 ± 0.05	0.37 ± 0.07	0.36 ± 0.07	0.36 ± 0.07	0.40 ± 0.08	**0.42 ± 0.10**	0.39 ± 0.08
All PCs	Overlap	0.62 ± 0.05	0.57 ± 0.06	0.60 ± 0.06	0.62 ± 0.07	0.59 ± 0.06	0.62 ± 0.05	0.64 ± 0.06	0.61 ± 0.05	0.63 ± 0.06	**0.67 ± 0.06**	0.63 ± 0.06
	Relevant Pop.	0.86 ± 0.06	**0.93 ± 0.04**	0.90 ± 0.06	0.88 ± 0.07	0.82 ± 0.06	0.84 ± 0.07	0.81 ± 0.08	0.84 ± 0.05	0.80 ± 0.07	0.69 ± 0.10	0.80 ± 0.07
	Precision	0.42 ± 0.08	0.29 ± 0.06	0.35 ± 0.07	0.30 ± 0.07	0.35 ± 0.06	0.44 ± 0.08	0.43 ± 0.09	0.42 ± 0.08	**0.46 ± 0.08**	**0.46 ± 0.09**	0.45 ± 0.08

Bold represents best performance values per content.

6.1 Frame-Based Analysis

As a first step, we implement a frame-based analysis (i.e., frame-based clustering) on the entire dataset taking into account navigation trajectories experienced only with not-distorted content in the select dataset. In Table 3, we report the average and SD of performance metrics described in Section 5.2 obtained by our proposed similarity metrics per each content. In the last row of the table, we show also the final performance averaged across the volumetric contents. Clusters based on w_2 include the majority of the population within relevant clusters across all the analysed PCs, reaching the maximum value of 0.94 in PC1. However, this comes at the detriment of precision, which falls to values between 0.22 and 0.34. In terms of overlap ratio and precision, the most promising similarity metric is mainly w_{10} followed by w_9 and w_{11}. These outperform the other weights in most of the PCs, ensuring an overlap ratio within the same cluster with values in the range of 0.60 and 0.72 for w_{10}. The only exception is in PC1, where the best performing metric in terms of overlap ratio and precision is w_7, which for the other content cases is always performing worse. Finally, the values of precision are always over 0.38, with an average value above 0.45, for all the three last metrics in the group of multi-feature metrics.

We now visually compare some examples of detected clusters by the different similarity metrics. Figure 3 shows the clusters obtained using the ground-truth metric O to construct the graph (Figure 3(a)), along with the ones based on each proposed similarity metric (Figure 3(b)–(l)) for frame 50 of sequence PC1 (*LongDress*). In particular, each user is represented by a point on the VR floor which is coloured based on the assigned ID cluster, whereas the volumetric content is symbolised by a blue star. Per each relevant cluster (i.e., cluster with more than two users), we provide in the legend the following results: the number of users inside the cluster, and the average and variance of the overlap ratio among all users within the cluster. Finally, we represent the remaining users which are in either single or couple-cluster as black points; the total number of these users is also provided in the legend as 'Small clusters (total number of non-relevant clusters)'. We can notice that our ground truth (Figure 3(a)) generates five main clusters with an average overlap ratio per cluster above 0.82. In particular, cluster ID 1 has the highest number of users (eight), with a high overlap ratio (0.84). Only four users in this case are put in single clusters. The goal is to find a similarity metric that can well approximate these results. In Figure 3(b)–(l), we can notice that our proposed metrics tend to create three main clusters, very populated but with a low overlap ratio. For instance, w_3 and w_7 generate a main big cluster with, respectively, 18 and 21 users, while the corresponding overlap ratio drops drastically to 0.62 and 0.64. Among single-feature metrics, the only exceptions are given by w_1 and w_5, which generate a variable set of



(a) Ground-truth (O_{th} = 75%) (b) w_1 (single feature metric) (c) w_2 (single feature metric)

(d) w_3 (single feature metric) (e) w_4 (single feature metric) (f) w_5 (single feature metric)

(g) w_6 (multi-feature metric) (h) w_7 (multi-feature metric) (i) w_8 (multi-feature metric)

(j) w_9 (multi-feature metric) (k) w_{10} (multi-feature metric) (l) w_{11} (multi-feature metric)

Fig. 3. Cluster results in frame 50 of sequence PC1 (*LongDress*). Each dot represents a user on the virtual floor while the blue star stands for the volumetric content. In the legend in brackets, per each cluster with more than two users are reported: the number of users in the same cluster, averaged pairwise viewport overlap and corresponding variance within the cluster.

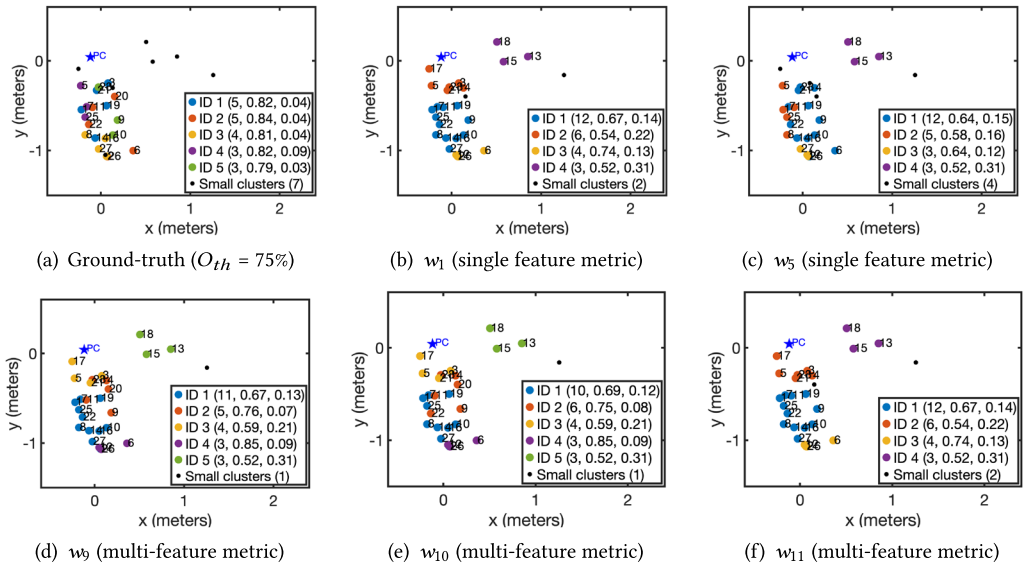
four clusters with consistent values of overlap ratio, over 0.64 and 0.66, respectively. Let us now consider as an example the users 13, 15 and 17, which in the ground-truth case (Figure 3(a)) form their own cluster (i.e., ID 5) with a high overlap ratio (0.83), as well as user 24, who is quite isolated from other users and belongs to a single cluster. Among single-feature metrics (Figure 3(b)–(f)), we can notice that w_2, w_3 and w_4 fail in detecting the group of users 13, 15 and 17 as similar, dividing them instead in different clusters or merging them with existing big clusters. On the contrary, w_1

and w_5 detect this group of participants as similar and assign them to the same cluster. From these observations, we can notice that the viewport centre on the volumetric content, on which w_3 and w_4 are based, is not sufficient to correctly identify similar users. Analogously, considering only the difference in terms of the relative distance between the user and volumetric content, as done in w_2, does not allow the detection of similarity among users. Thus, the most promising metrics in the group of single-feature metrics seem to be w_1 and w_5, which are based on the user position on the virtual floor and the vector of viewing direction, respectively. The second group of metrics in Figure 3(g)–(l) shows clusters based on multi-feature similarity metrics. In this set of metrics, users 13, 15 and 17 are identified within the same cluster only by three metrics, namely w_6, w_9 and w_{11}, which also detect user 24 as a single cluster. On the contrary, the other two metrics w_7 and w_8 create a main cluster with users 17, 24 and 13 while user 15 is assigned to a different main cluster with participants in a different location on the virtual floor; finally, w_{10} assigns most of these users to small clusters. Considering also the analysis shown previously in Table 3, multi-feature metrics appear to be overall better suited to detect similar users than previous single-feature ones. This is expected, as the higher df are given to users, the more challenging the system, and thus detecting users similarities.

Given the above remarks, in the following, we further analyse the selected dataset taking into account only a subset of metrics, namely, w_1, w_5, w_9, w_{10} and w_{11}, based on the best-performing similarity metrics in the previous investigation in terms of precision (w_9, w_{10} and w_{11}). To have a fair comparison, we also keep the most promising among the single-feature metrics, namely w_1 and w_5. In Figure 4, we show a similar visual example of frame-based analysis frame 50 of PC2 (*Loot*). In this case, it is interesting to notice how the ground-truth clusters in Figure 4(a) are very intricate. A total of five main clusters are found, with an overlap ratio consistently over 0.79. A considerable number of participants (seven) is instead put in small clusters. In this case, all the selected similarity metrics fail to detect such a consistent group of participants in terms of overlap ratio. Except for w_9 and w_{10} that create five clusters, the remaining proposed metrics generate four main clusters. However, in all these cases, the overlap ratio drops drastically in a range between 0.52 and 0.76. In this example, the small group of participants located on the right side of the volumetric content, specifically users 13, 15, 18 and 24, are assigned to single clusters in the ground-truth case, as shown in Figure 4(a). On the contrary, users 13, 15 and 18 are assigned to the same cluster with an overlap ratio equal to 0.52 from all the similarity metrics under investigation. We can also observe that, in the case of the ground-truth clustering, users 5, 17, 21 and 23 are all assigned to different clusters, with user 17 being in a small cluster; however, that is not the case in any of the metrics under consideration; in fact, such users are more often not grouped in the same cluster. Among all the metrics, w_5 appears to be the most promising, as it is the only one identifying user 17 as a separate small cluster, and grouping users 26 and 27 in the same cluster, as done by the ground truth. However, the results are still far from adequately matching the ground truth clustering algorithm. To conclude, this example shows also the complexity and critical aspects of addressing the open problem of evaluating the (dis)similarity between 6 df users at each given time of an immersive experience. Our proposed metrics represent a first step in this direction but further investigations are needed to better understand the intricate nature of user navigation in XR systems.

6.2 Trajectory-Based Analysis

We now analyse the performance metrics over time (i.e., trajectory-based analysis). Specifically, we compute clique-based clusters over a time window of 1 second (i.e., chunk) and a time similarity threshold of 0.8 seconds (i.e., users should be similar in the 80% of the chunk length). At each chunk, we evaluate the average overlap ratio per relevant cluster, the average of the relevant population and the precision of detected clusters. In the following, we have as an example, we show in Figure 5



(a) Ground-truth (O_{th} = 75%) (b) w_1 (single feature metric) (c) w_5 (single feature metric)

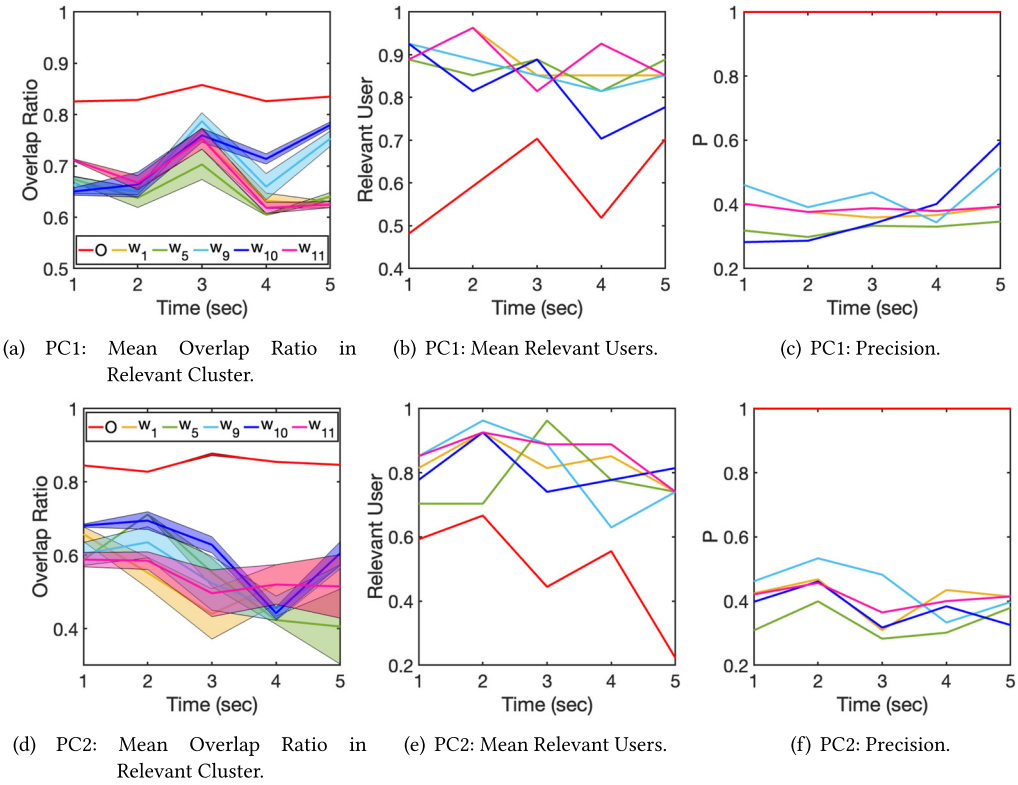
(d) w_9 (multi-feature metric) (e) w_{10} (multi-feature metric) (f) w_{11} (multi-feature metric)

Fig. 4. Cluster results in frame 50 of sequence PC2 (*Loot*). Each dot represents a user on the virtual floor while the blue star stands for the volumetric content. In the legend in brackets, per each cluster with more than two users are reported: the number of users in the same cluster, averaged pairwise viewport overlap and corresponding variance within the cluster.

the performance results as functions of time per each selected similarity metric (w_1, w_5, w_9, w_{10} and w_{11}) per sequence PC1 (*LongDress*) in the first row and PC2 (*Loot*) in the second one. We also add the performance of clusters detected by the ground-truth metric O (i.e., red line); the goal is indeed to find a metric able to perform similarly to our ground truth over time. In PC1, all the similarity metrics reach an average overlap ratio within clusters between 0.6 and 0.75 (Figure 5(a)). Metrics based on single features, such as w_1 and w_5, exhibit lower performance, while others perform quite similarly, with a slight predominance of w_{10}. In the second example (Figure 5(d)), the mean overlap ratio decreases to values between 0.35 and 0.75. However, also in PC2, we observe that clusters based on w_{10} show a slightly better performance. In terms of relevant users, it is worth noting that all the proposed similarity metrics, in both sequences, generate larger clusters compared to the ground-truth metric. The latter considers only 50–70% (0.5–0.7) of the population as relevant in PC1 (Figure 5(b)), and this drops to 0.2 in PC2 (Figure 5(e)). In more detail, the clusters resulting from our proposed metrics consistently put in relevant clusters over 0.7 of the entire population for all the time in both the volumetric sequences. Finally, in terms of precision, the only similarity metric that generated clusters with P consistently over to 0.4 in most of the time of both sequences is w_9. However, it is interesting to notice that the clusters generated based on w_{11} have a constant value of precision over time equal to 0.4. On the contrary, clusters based on w_5 are on average less performing in terms of precision in both PC1 (Figure 5(c)) and PC2 (Figure 5(f)). These investigations show that similarity metrics based on multi-feature, such as w_9 and w_{11}, are more promising for detecting with higher precision similar behaviour while experiencing volumetric content.

From this validation analysis on the VR dataset shown in Figure 1, we can conclude the following:

—Overall, *multi-feature metrics* are more precise in detecting users with similar behaviour (in terms of displayed content) both in a frame- and chunk-based analysis;



(a) PC1: Mean Overlap Ratio in Relevant Cluster.

(b) PC1: Mean Relevant Users.

(c) PC1: Precision.

(d) PC2: Mean Overlap Ratio in Relevant Cluster.

(e) PC2: Mean Relevant Users.

(f) PC2: Precision.

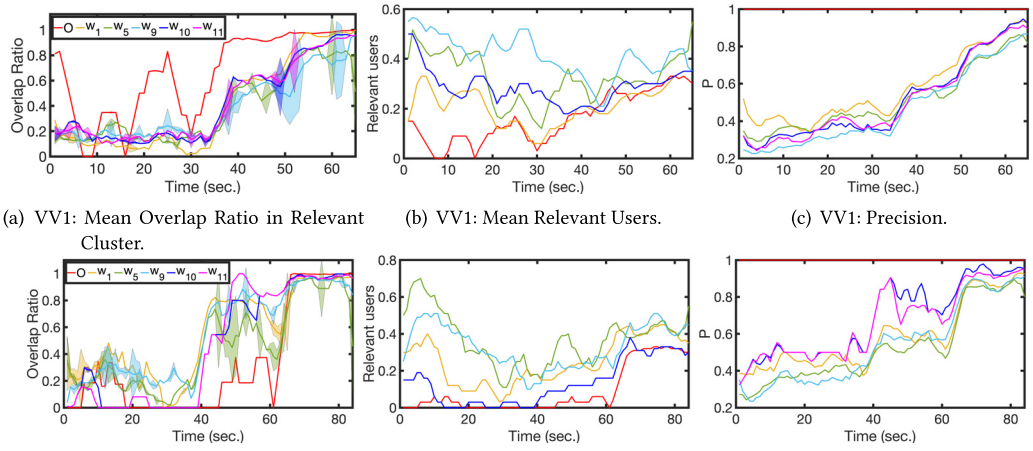
Fig. 5. Clustering over time (chunk = 1 second) results per sequence PC1 (*LongDress*) and PC2 (*Loot*): comparison between ground-truth O and a subset of proposed metrics (w_1, w_5, w_9, w_{10} and w_{11}).

—In particular, in spite of the slightly more complex formulation *multi-feature metrics*, such as w_9, w_{10} and w_{11}, are robust and easy-to-use metrics that ensure a robust and reliable behavioural analysis via clustering tools;
—On the contrary, metrics based only on a single feature (i.e., *single-feature metrics*) are not always sufficient to correctly identify similar users;
—The only exceptions among single-feature metrics are w_1 and w_5 which are based only on the position of the user on the floor and the vector of viewing direction, respectively. Despite their simplicity, these metrics are overall comparable with multi-feature metrics. Hence, they can be used for an easy-to-implement preliminary behavioural analysis.

These outcomes are based on point clouds of human body and trajectories collected in a visual quality assessment study. Thus, it is important to point out that these observations are valid for similar volumetric contents (i.e., human body). The user navigation might be also affected by the task to rate the quality of the content: For instance, participants might have checked only visual impairments rather than freely explore the volumetric content. We leave further analysis across multiple types of content and task-free datasets for future work.

7 Case Study: A Behavioural Analysis in AR Setting

We are now interested in understanding if the insights from the above study could be applied to other human-like datasets. To show also the robustness of our proposed similarity metrics, we



(a) VV1: Mean Overlap Ratio in Relevant Cluster.
(b) VV1: Mean Relevant Users.
(c) VV1: Precision.
(d) VV2: Mean Overlap Ratio in Relevant Cluster.
(e) VV2: Mean Relevant Users.
(f) VV2: Precision.

Fig. 6. Spherical clustering over time (chunk = 1 second) results per sequence VV1 (*Nico*) and VV2 (*Sir Fredrick*): performance comparison between ground-truth, and a subset of proposed metrics (w_1, w_5, w_9, w_{10} and w_{11}).

therefore apply them to a different dataset. In particular, we select the dataset presented in [60]. Authors have collected in a task-free experiment the navigation trajectories of 20 users while displaying volumetric content in an AR scenario. Similarly to the previously analysed dataset presented in Section 5.1, a single object of interest was placed in the scene. Specifically, two dynamic volumetric human body sequences represented as 3D meshes with texture information were used: *Nico* (VV1) and *Sir Frederic* (VV2) in Figure 1(b). In order to conduct our study, both the sequences were kindly made available by Vologrames upon request [30, 58]. The navigation data were collected in a remote scenario through an Android AR application, which allowed users to display the volumetric content from any desired location and portable device (e.g., smartphone) [60]. Participants were also free to display the volumetric content how they most preferred. Thus, the main differences with the previously analysed dataset are the following: a different format of volumetric content (3D mesh instead of point cloud), a different immersive scenario (AR instead of VR application), a different aim of the experiment (task-free instead of a quality assessment) and a heterogeneity of viewing devices (any smartphone device instead of a specific HMD). In particular, the 3D mesh content does not allow for a simple formulation of the overlap ratio as we have described it in Section 3.2. For consistency, we convert the sequences from 3D meshes to point clouds by discarding edge information and only keeping vertices as points; we discuss the inherent challenges to define our ground-truth metric in Section 8.

Similarly to our previous investigations, we now apply to this new scenario the spherical clustering based on the subset of best-performing feature metrics, such as w_1, w_5, w_9, w_{10} and w_{11}. We evaluate clusters in chunks of length 1 second with a time similarity threshold of 0.8 seconds and the threshold values G_{th} reported in Table 2. At each chunk, we compute the average overlap ratio per relevant cluster (i.e., cluster with at least two elements), the average of the relevant population and the precision of the detected clusters. Figure 6 shows these results as a function of time per each selected similarity metric, in particular, the first row refers to VV1 (*Nico*) while the second one to VV2 (*Sir Frederic*). Since viewers were allowed to drop the AR experience at any desired time, in the following we consider only the time window in which 75% of the user population (15 out of 20 viewers) are still in the experiment: 63 and 83 seconds, respectively, for

VV1 and VV2. We observe that both the sequences have an initial moment of adjustment where viewers are displaying different portions of the content. This is detected by clusters based on the overlap ratio (i.e., red line) which do not have a consistent pairwise overlap. For instance, Figure 6(a) shows in the first 40 seconds of the immersive experience for VV1 the average of overlap ratio within the main detected clusters has up and down for the ground-truth and is quite low for all the selected similarity metrics. However, this behaviour stabilises around 40 seconds when the overlap ratio for the ground-truth metric converges to 0.8. Similarly, the performance metric detected by w_1, w_{10} and w_{11} reaches values above 0.6 with a very low variance for both the metrics. On the contrary, the overlap ratio of cluster detected by w_5 and w_9 has a more inconsistent variance over time. Compared to other metrics and to the ground truth, these metrics in terms of relevant users (Figure 6(b)) generate overall bigger clusters. In particular, it considers most of the time half of the population in big clusters, which is quite the opposite behaviour to the ground-truth metric. This metric indeed generates small relevant clusters most of the time; clusters based on w_1, w_{10} and w_{11} follow a very similar trend. In this case, results based on w_1 are the best performing in terms of precision, as shown in Figure 6(c) with values close or above 0.4. A slightly different behaviour is observable for VV2 in the second line of Figure 6. In this volumetric content, users explore the scene more randomly during the first minute of the experience leading to such different behaviours that the ground-truth fails in detecting relevant clusters between 18 and 39 seconds (Figure 6(d)). This divergent user behaviour might be due to the task-free experiment which bring participants to observe different parts of the content. We leave as future work further investigations to understand the impact of structured tasks and not on the user behaviour. However in this case, similarity metrics w_{10} and w_{11} are more precise in reflecting the ground-truth behaviour and thus, detecting viewers with similar behaviour and putting them within the same clusters (Figure 6(f)). Finally, it is worth noticing that all the similarity metrics reach a higher overlap ratio compared to the ground-truth performance (Figure 6(f)). For example, some metrics as w_1 in the first 10 seconds or w_{11} around 50 seconds reach an overlap ratio of 40% and above 80%, respectively, while the ground-truth has very lower values. This shows the ability of our proposed metrics in identifying users with similar displayed viewports, however, raises new questions on the selected ground-truth and how accurately it captures such similarities in different XR conditions.

From the behavioural analysis of this second dataset, which, although composed of only 20 navigation trajectories, was collected in an AR setting, we can conclude the following:

- —Our proposed similarity metrics demonstrate flexibility and generality, proving their suitability for analysing not only navigation trajectories in VR settings but also in AR conditions. This suggests their robustness across different XR environments;
- —Overall, *multi-feature metrics*, in particular w_{10} and w_{11}, are more precise in identifying where there is similarity between users in very eclectic behaviours such as in VV2 (the second line of Figure 6);
- —Interestingly, in this second analysis, our ground-truth similarity metric, designed to capture consistent user similarity, exhibits low performance in terms of overlap ratio. This finding suggests also inherent challenges in accurately capturing user similarity in dynamic AR settings, paving the way for further investigation.

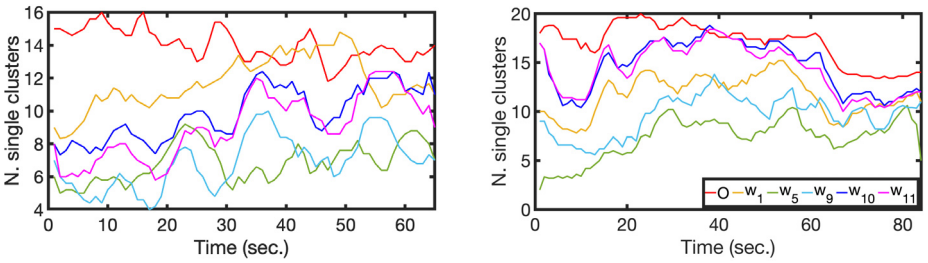
8 Discussion and Future Work

We presented the main challenges of user behavioural analysis in a 6 df system caused by the new settings and the added locomotion functionalities. In particular, our main goal was to extend the applicability of existing behavioural tool, such as clique-based clustering, [36] designed for 3 df scenario to its 6 df counterpart. However, behavioural analysis of 6 df users is not considered

in the literature yet; as such, there is no reference metric available to detect viewers who are displaying the same portion of the content. As first step, we had to define a general ground-truth user similarity metric, namely the *overlap ratio*. To be as general as possible, we established the overlap as the percent of points displayed in common by two users. This is fairly straightforward, albeit time-consuming, to compute for point cloud contents, in which each point is rendered separately. For other types of volumetric contents, determining the overlap ratio is not as simple. Considering the number of vertexes that fall into a given frustum could lead to misleading results when large faces between sparsely distributed vertexes are present. Moreover, the metric requires to render each volumetric video at any given time and for each viewer, making its computation not trivial and intensely time-consuming. To address this challenge and objectively assess users similarity in a simple way, in this article we investigated various similarity metrics aimed at better modelling behavioural similarities between users in a 6 df setting. Specifically, we were interested in modelling similarities among users *observing the same volumetric content*. We defined and compared 11 different metrics based on different *distance features* (i.e., user positions in the 3D space, user viewing directions) and *distance measurements* (i.e., Euclidean, Geodesic, angular distance). More concretely, we considered user information, such as their location in the virtual floor and viewing direction, which is consistently available in immersive systems. Our proposed metrics can be computed in less than 10 milliseconds on average per frame, ensuring their applicability in real-time applications. To test and validate our similarity metrics using a clique-based clustering tool proposed for 3 df scenario, we employed real navigation trajectories collected in a 6 df VR environment [51] (Figure 1(a)). Our extensive analysis showed that overall metrics based on a combination of distance features (*multi-feature metrics*), such as w_9, w_{10} and w_{11}, exhibit encouraging values of overlap ratio and superior precision in detecting users with similar behaviour, whether analysed frame by frame or in chunks of data. On the contrary, metrics based solely on a single feature (referred to as *single-feature metrics*) fall short in consistently identifying similar users accurately. However, exceptions to this trend are found in w_1 and w_5, which leverage user position on the floor and the vector of viewing direction, respectively. Remarkably, despite their simplicity, these metrics perform comparably to multi-feature metrics, making them suitable for a straightforward preliminary behavioural analysis.

To test the flexibility of our proposed metrics, we tested their performance on a different kind of 6 df navigation trajectories [60]. In this second dataset, viewers displayed volumetric content in an AR scenario through smartphones. Therefore, even if users were enabled with the same 6 df locomotion settings, the viewing device and the FoV were different. Despite these differences, our proposed similarity metrics are still good at identifying viewers who are displaying similar content. However, it is also worth mentioning that our ground-truth metric of similarity is very tight in detecting similar users, especially in an AR scenario. As an example, Figure 7 shows the number of single clusters detected over time by the overlap ratio (i.e., red line) and the subset of most performing similarity metrics for both the volumetric sequences of the second analysed dataset. In particular, in VV2 (Figure 7(b)), the clique-based clustering based on the overlap ratio does not detect similar users such that the majority of the population are put in a single cluster. Therefore, further analysis is needed to test if in this scenario a different overlap threshold better model similarity among users. Finally, it is important to point out that these observations are currently only valid for similar volumetric contents (i.e., human body). We leave further analysis across multiple datasets and types of content for future work.

This work opens the gate to further investigations aimed at detecting user behavioural differences in a 6 df experience done in VR and AR settings. These are indeed essential to be exploited in efficient user-centric solutions for XR systems to enable, for example, new modalities of live streaming services optimised for users' profiles but also for user-based quality assessment methods.



(a) VV1 (b) VV2

Fig. 7. Single-user cluster per sequence VV1 (*Nico*) and VV2 (*Sir Fredrick*) obtained via spherical clustering based on overlap ratio, and a subset of proposed similarity metrics (w_1, w_5, w_9, w_{10} and w_{11}).

9 Conclusion

To conclude, this article contributes to advancing the field of behavioural analysis in XR scenarios. By introducing novel similarity metrics tailored to the new physical settings and locomotion functionalities of users in XR environments, we have addressed a critical aspect of user-centric system development. Our behavioural investigation on 6 df navigation trajectories with behavioural tool for 3 df trajectories provided insights into the distinctive features and challenges posed by the former. The proposed 11 similarity metrics, based on various distance features and measurements, were rigorously tested and validated using real navigation paths from both 6 df VR and AR conditions. Our results showed that solutions that consider both user position and viewing direction are promising to correctly detect users with similar behaviour while experiencing volumetric content. Moreover, since these metrics are based on simple operations of data that are typically already known in a multimedia system (i.e., user position in the virtual space and viewing direction), they can be evaluated on average in less than 10 milliseconds. This makes our proposed metrics not only robust but also suitable for real-time applications. Moreover, we have also demonstrated the robustness and versatility of these metrics, which preserve good performance on navigation trajectories collected both in a 6 df VR and AR scenario, showcasing their applicability across diverse XR settings.

References

[1] Avi M. Aizenman, George A. Koulieris, Agostino Gibaldi, Vibhor Sehgal, Dennis M. Levi, and Martin S. Banks. 2023. The statistics of eye movements and binocular disparities during VR gaming: Implications for headset design. *ACM Transactions on Graphics* 42, 1 (Feb 2023), 1–15. DOI: https://doi.org/10.1145/3549529

[2] Evangelos Alexiou, Nanyang Yang, and Touradj Ebrahimi. 2020. PointXR: A toolbox for visualization and subjective evaluation of point clouds in virtual reality. In *International Conference on Quality of Multimedia Experience*. IEEE, 1–6.

[3] Lucas S. Althoff, Myléne C. Q. Farias, Alessandro Rodrigues Silva, and Marcelo M. Carvalho. 2023. Impact of alignment edits on the quality of experience of 360° videos. *IEEE Access* 11, 108475–108492. DOI: https://doi.org/10.1109/ACCESS.2023.3319346

[4] Jesús Bermejo-Berros, and Miguel Angel Gil Martínez. 2021. The relationships between the exploration of virtual space, its presence and entertainment in virtual reality, 360º and 2D. *Virtual Reality* 25, 4 (Dec 2021), 1043–1059. DOI: https://doi.org/10.1007/s10055-021-00510-9

[5] Edurne Bernal-Berdun, Daniel Martin, Sandra Malpica, PedroJ Perez, Diego Gutierrez, Belen Masia, and Ana Serrano. 2023. D-SAV360: A dataset of gaze scanpaths on 360° ambisonic videos. *IEEE Transactions on Visualization and Computer Graphics* 29, 11 (Nov 2023), 4350–4360. DOI: https://doi.org/10.1109/TVCG.2023.3320237 37782595

[6] Alberto Cannavò, Antonio Castiello, F. Gabriele Pratticò, Tatiana Mazali, and Fabrizio Lamberti. 2024. Immersive movies: The effect of point of view on narrative engagement. *AI & Society* 39, 4 (Aug 2024), 1811–1825. DOI: https://doi.org/10.1007/s00146-022-01622-9

[7] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. 2022. Privacy-preserving viewport prediction using federated learning for 360° live video streaming. In *IEEE 24th International Workshop on Multimedia Signal Processing (MMSP).* IEEE, 1–6.

[8] Weiya Chen, Anthony Plancoulaine, Nicolas Férey, Damien Touraine, Julien Nelson, and Patrick Bourdot. 2013. 6df navigation in virtual worlds: Comparison of joystick-based and head-controlled paradigms. In *19th ACM Symposium on Virtual Reality Software and Technology,* 111–114.

[9] Pietro Cipresso, Irene Alice Chicchi Giglioli, Mariano Alcañiz Raya, and Giuseppe Riva. 2018. The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature. *Frontiers in Psychology* 9 (2018), 2086. DOI: https://doi.org/10.3389/fpsyg.2018.02086

[10] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 2017. 360-degree video head movement dataset. In *8th ACM on Multimedia Systems Conference,* 199–204.

[11] Heather Creagh. 2003. Cave automatic virtual environment. In *Electrical Insulation and Electrical Manufacturing and Coil Winding Technology Conference,* 499–504.

[12] Erwan J. David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. A dataset of head and eye movements for 360 videos. In *9th ACM Multimedia Systems Conference,* 432–437.

[13] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (Jun 2006), 861–874. DOI: https://doi.org/10.1016/j.patrec.2005.10.010

[14] Daniel Freeman, Sarah Reeve, Abi Robinson, Anke Ehlers, David Clark, Bernhard Spanlang, and Mel Slater. 2017. Virtual reality in the assessment, understanding, and treatment of mental health disorders. *Psychological Medicine* 47, 14 (Oct 2017), 2393–2400. DOI: https://doi.org/10.1017/S003329171700040X 28325167

[15] Chris N. W. Geraets, Elisabeth C. D. van der Stouwe, Roos Pot-Kolder, and Wim Veling. 2021. Advances in immersive virtual reality interventions for mental disorders–A new reality? *Current Opinion in Psychology* 41 (2021), 40–45.

[16] Quentin Guimard, Lucile Sassatelli, Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. 2024. Deep variational learning for 360° adaptive streaming. *ACM Transactions on Multimedia Computing,Communications and Applications* 20, 9 (Sep 2024), 1–25. DOI: https://doi.org/10.1145/3643031

[17] Jesús Gutiérrez, Gulzhanat Dandyyeva, Matteo Dal Magro, Carlos Cortés, Michele Brizzi, Marco Carli, and Federica Battisti. 2023. Subjective evaluation of dynamic point clouds: Impact of compression and exploration behavior. In *Proceeding of the 31st European Signal Processing Conference (EUSIPCO).* IEEE, 675–679.

[18] Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-aware mobile volumetric video streaming. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, Article 11, 1–13.

[19] Gazi Karam Illahi, Ashutosh Vaishnav, Teemu Kämäräinen, Matti Siekkinen, and Mario Di Francesco. 2023. Learning to predict head pose in remotely-rendered virtual reality. In *14th Conference on ACM Multimedia Systems,* 27–38.

[20] Hupont Torres Isabelle, Charisi Vasiliki, De Prato Giuditta, Pogorzelska Katarzyna, Schade Sven, Kotsev Alexander, Sobolewski Maciej, Duch Brown Nestor, Calza Elisa, Dunker Cesare, et al. 2023. *Next Generation Virtual Worlds: Societal, Technological, Economic and Policy Challenges for the EU* (No. JRC133757). Technical Report. Joint Research Centre.

[21] Yili Jin, Junhua Liu, Fangxin Wang, and Shuguang Cui. 2022. Where are you looking? A large-scale dataset of head and gaze behavior for 360-degree videos and a pilot study. In *30th ACM International Conference on Multimedia,* 1025–1034.

[22] Hanseul Jun, Mark Roman Miller, Fernanda Herrera, Byron Reeves, and Jeremy N. Bailenson. 2022. Stimulus sampling with 360-videos: Examining head movements, arousal, presence, simulator sickness, and preference on a large sample of participants and videos. *IEEE Transactions on Affective Computing* 13, 3 (2022), 1416–1425. DOI: https://doi.org/10.1109/TAFFC.2020.3004617

[23] Maja Krivokuća, Philip A. Chou, and Patrick Savill. 2018. 8i voxelized surface light field (8iVSLF) dataset. In *ISO/IEC JTC1/SC29/WG11 input document m42914,* Ljubljana, Slovenia. Retrieved from https://mpeg-pcc.org/index.php/pcc-content-database/8i-voxelized-surface-light-field-8ivslf-dataset/

[24] Huiguang Liang, RansiNilaksha De Silva, WeiTsang Ooi, and Mehul Motani. 2009. Avatar mobility in user-created networked virtual worlds: measurements, analysis, and implications. *Multimedia Tools and Applications* 45, 1–3 (Oct 2009), 163–190. DOI: https://doi.org/10.1007/s11042-009-0304-x

[25] Robert F. K. Martin, Patrick Leppink-Shands, Matthew Tlachac, Megan DuBois, Christine Conelea, Suma Jacob, Vassilios Morellas, Theodore Morris, and Nikolaos Papanikolopoulos. 2020. The use of immersive environments for the early detection and treatment of neuropsychiatric disorders. *Frontiers in Digital Health* 2 (2020), 576076. DOI: https://doi.org/10.3389/fdgth.2020.576076

[26] Pramit Mazumdar, Giuliano Arru, and Federica Battisti. 2021. Early detection of children with autism spectrum disorder based on visual exploration of images. *Signal Processing: Image Communication* 94 (2021), 116184.

[27] Dario D. R. Morais, Lucas S. Althoff, Ravi Prakash, Marcelo M. Carvalho, and Mylene C. Q. Farias. 2021. A content-based viewport prediction model. *Electronic Imaging* 33 (2021), 1–8.

[28] Afshin Taghavi Nasrabadi, Aliehsan Samiei, Anahita Mahzari, Ryan P. McMahan, Ravi Prakash, Mylene C. Q. Farias, and Marcelo M. Carvalho. 2019. A taxonomy and dataset for 360° videos. In *10th ACM Multimedia Systems Conference,* 273–278.

[29] Afshin Taghavi Nasrabadi, Aliehsan Samiei, and Ravi Prakash. 2020. Viewport prediction for 360 videos: A clustering approach. In *30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video,* 34–39.

[30] Rafael Pagés, Emin Zerman, Konstantinos Amplianitis, Jan Ondřej, and Aljosa Smolic. 2021. Volograms & V-SENSE volumetric video dataset. *ISO/IEC JTC1/SC29/WG07 MPEG2021/m56767* (2021).

[31] Jounsup Park, Philip A. Chou, and Jenq-Neng Hwang. 2019. Rate-utility optimized streaming of volumetric media for augmented reality. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2019), 149–162. DOI: https://doi.org/10.1109/JETCAS.2019.2898622

[32] Ken Pfeuffer, Matthias J. Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. 2019. Behavioural biometrics in VR: Identifying people from body motion and relations in virtual reality. In *CHI Conference on Human Factors in Computing Systems*. ACM, 1–12.

[33] Eric D. Ragan, Siroberto Scerbo, Felipe Bacim, and Doug A. Bowman. 2016. Amplified head rotation in virtual reality and the effects on 3D search, training transfer, and spatial orientation. *IEEE Transactions on Visualization and Computer Graphics* 23, 8 (2016), 1880–1895.

[34] Miguel Fabián Romero Rondón, Lucile Sassatelli, Ramon Aparicio-Pardo, and Frédéric Precioso. 2020. A unified evaluation framework for head motion prediction methods in 360 videos. In *11th ACM Multimedia Systems Conference,* 279–284.

[35] Miguel Fabian Romero Rondon, Lucile Sassatelli, Ramón Aparicio-Pardo, and Frédéric Precioso. 2021. Track: A new method from a re-examination of deep architectures for head motion prediction in 360-degree videos. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 01 (2021), 1–1.

[36] Silvia Rossi, Francesca De Simone, Pascal Frossard, and Laura Toni. 2019. Spherical clustering of users navigating 360 content. In *IEEE International Conference on Acoustics, Speech and Signal Processing,* 4020–4024.

[37] Silvia Rossi, Alan Guedes, and Laura Toni. 2023. Streaming and user behavior in omnidirectional videos. In *Immersive Video Technologies.* Giuseppe Valenzise, Martin Alain, Emin Zerman, and Cagri Ozcinar (Eds.), Academic Press, 49–83.

[38] Silvia Rossi, Cagri Ozcinar, Aljosa Smolic, and Laura Toni. 2020. Do users behave similarly in VR? Investigation of the user influence on the system design. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 2 (May 2020), 1–26. DOI: https://doi.org/10.1145/3381846

[39] Silvia Rossi and Laura Toni. 2017. Navigation-aware adaptive streaming strategies for omnidirectional video. In *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP).* IEEE, 1–6. DOI: https://doi.org/10.1109/MMSP.2017.8122230

[40] Silvia Rossi and Laura Toni. 2020. Understanding user navigation in immersive experience: An information-theoretic analysis. In *12th ACM International Workshop on Immersive Mixed and Virtual Environment Systems,* 19–24.

[41] Silvia Rossi, Laura Toni, and Pablo Cesar. 2023. Correlation between entropy and prediction error in VR head motion trajectories. In *2nd International Workshop on Interactive eXtended Reality,* 29–36.

[42] Silvia Rossi, Irene Viola, and Pablo Cesar. 2022. Behavioural analysis in a 6-df VR system: Influence of content, quality and user disposition. In *1st Workshop on Interactive eXtended Reality,* 3–10.

[43] Silvia Rossi, Irene Viola, Jack Jansen, Shishir Subramanyam, Laura Toni, and Pablo Cesar. 2021. Influence of narrative elements on user behaviour in photorealistic social VR. In *International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE'21),* 1–7.

[44] Silvia Rossi, Irene Viola, Laura Toni, and Pablo Cesar. 2021. A new challenge: Behavioural analysis of 6-DOF user when consuming immersive media. In *IEEE International Conference on Image Processing (ICIP),* 3423–3427.

[45] Silvia Rossi, Irene Viola, Laura Toni, and Pablo Cesar. 2023. Extending 3-df metrics to model user behaviour similarity in 6-df immersive applications. In *14th Conference on ACM Multimedia Systems*, 39–50.

[46] Michael Rudolph, Stefan Schneegass, and Amr Rizk. 2023. RABBIT: Live transcoding of V-PCC point cloud streams. In *14th Conference on ACM Multimedia Systems,* 97–107.

[47] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A. Chou, Robert A.Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, et al. 2019. Emerging MPEG standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2019), 133–148. DOI: https://doi.org/10.1109/JETCAS.2018.2885981

[48] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (Apr 2018), 1633–1642. DOI: https://doi.org/10.1109/TVCG.2018.2793599 29553930

[49] Ljubisa Stankovic, Danilo P. Mandic, Milos Dakovic, Ilia Kisil, Ervin Sejdic, and Anthony G. Constantinides. 2019. Understanding the basis of graph signal processing via an intuitive example-driven approach. *IEEE Signal Processing Magazine* 36, 6 (2019), 133–145. DOI: https://doi.org/10.1109/MSP.2019.2929832

[50] Shishir Subramanyam, Jie Li, Irene Viola, and Pablo Cesar. 2020. Comparing the quality of highly realistic digital humans in 3df and 6df: A volumetric video case study. In *IEEE Conference on Virtual Reality and 3D User Interfaces*. IEEE, 127–136.

[51] Shishir Subramanyam, Irene Viola, Alan Hanjalic, and Pablo Cesar. 2020. User centered adaptive streaming of dynamic point clouds with low complexity tiling. In *28th ACM International Conference on Multimedia*, 3669–3677.

[52] Shishir Subramanyam, Irene Viola, Jack Jansen, Evangelos Alexiou, Alan Hanjalic, and Pablo Cesar. 2022. Subjective QoE evaluation of user-centered adaptive streaming of dynamic point clouds. In *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 1–6.

[53] Colin Swindells, Barry A. Po, Ima Hajshirmohammadi, Brian Corrie, John C. Dill, Brian D. Fisher, and Kellogg S. Booth. 2004. Comparing CAVE, wall, and desktop displays for navigation and wayfinding in complex 3D models. In *IEEE Proceedings Computer Graphics International*. IEEE, 420–427.

[54] Pier Paolo Tricomi, Federica Nenna, Luca Pajola, Mauro Conti, and Luciano Gamberi. 2023. You can't hide behind your headset: User profiling in augmented and virtual reality. *IEEE Access* 11 (2023), 9859–9875.

[55] Jeroen van der Hooft, Tim Wauters, Filip De Turck, Christian Timmerer, and Hermann Hellwagner. 2019. Towards 6-df HTTP adaptive streaming through point cloud compression. In *27th ACM International Conference on Multimedia*, 2405–2413.

[56] Sophie Villenave, Jonathan Cabezas, Patrick Baert, Florent Dupont, and Guillaume Lavoué. 2022. XREcho: A unity plug-in to record and visualize user behavior during XR sessions. In *13th ACM Multimedia Systems Conference*, 341–346.

[57] Irene Viola and Pablo Cesar. 2023. Volumetric video streaming. In *Immersive Video Technologies*. Giuseppe Valenzise, Martin Alain, Emin Zerman, and Cagri Ozcinar (Eds.), Academic Press, 425–443.

[58] Vologrames. 2021. Vologrames Homepage. Retrieved from https://www.vologrames.com/

[59] Mai. Xu, Chen Li, Shanyi Zhang, and PatrickLe Callet. 2020. State-of-the-art in 360° video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing* 14, 1 (2020), 5–26. DOI: https://doi.org/10.1109/JSTSP.2020.2966864

[60] Emin Zerman, Radhika Kulkarni, and Aljosa Smolic. 2021. User behaviour analysis of volumetric video in augmented reality. In *13th International Conference on Quality of Multimedia Experience*. IEEE, 129–132.

Appendices

A Correlation Analysis of Distance Features and Measurements

We now investigate the correlation among the distance features on which we based the multi-feature metrics presented in Section 4.2. Following the preliminary study presented in [44], we consider the following distance features and measurements: the Euclidean distance $\mathsf{E}(x^i, x^j)$ between user i and j on the virtual floor, the relative distance of users to the centroid of the displayed content, $\mathsf{L} = ||r^i - r^j||$, the distance between the viewport centres p of user i and user j projected on the volumetric content both in terms of Geodesic distance $\mathsf{G}(p^i, p^j)$ and Euclidean distance $\mathsf{E}(p^i, p^j)$, and finally the angular distance $\theta(v^i, v^j)$ between the vectors of the viewing direction of user i and user j. To visually explore the relationships between the different distance features, we use both multivariate scatter plots and **Principal Component Analysis (PCA)**. We evaluate the distance features based on the navigation trajectories experienced with non-distorted content of the dataset presented in Section 5.1 and averaged over time. Figure A1 shows a multivariate scatter plot to investigate the pairwise relationships between the different distance features. Specifically, subplots in the diagonal show histograms for the distribution of each variables while the remaining subplots presents a pairwise scatter plot of the analysed metrics. The histograms in the diagonal provide insights on the distribution and variability of each distance feature. For instance, we can notice that the Euclidean distance $\mathsf{E}(x^i, x^j)$ between users on the floor and the Geodesic distance $\mathsf{G}(p^i, p^j)$ between viewport centres projected on the volumetric content cover a quite large range of values while the Euclidean distance $\mathsf{E}(p^i, p^j)$ between viewport centres has a very concentrated distribution around low values. However, the pairwise scatter plots among the different distance features reveal correlations among all of them. In particular, $\mathsf{E}(x^i, x^j)$ and $\theta(v^i, v^j)$ show a clear positive correlation, suggesting that users who are farther apart in terms of Euclidean distance tend also to have a high angular distance between their viewing vectors. On the contrary, the relation

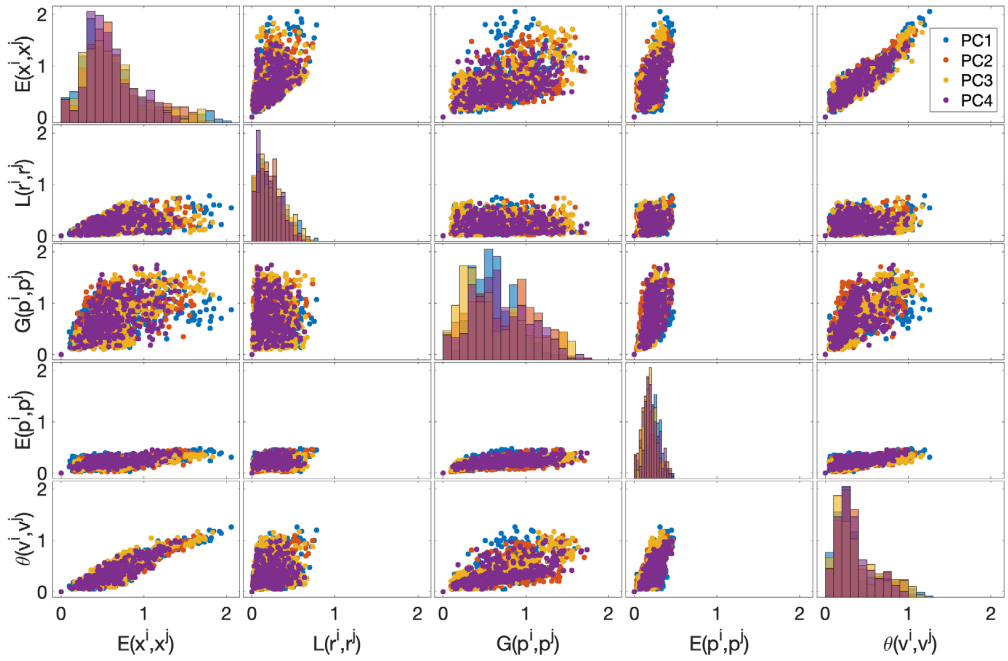
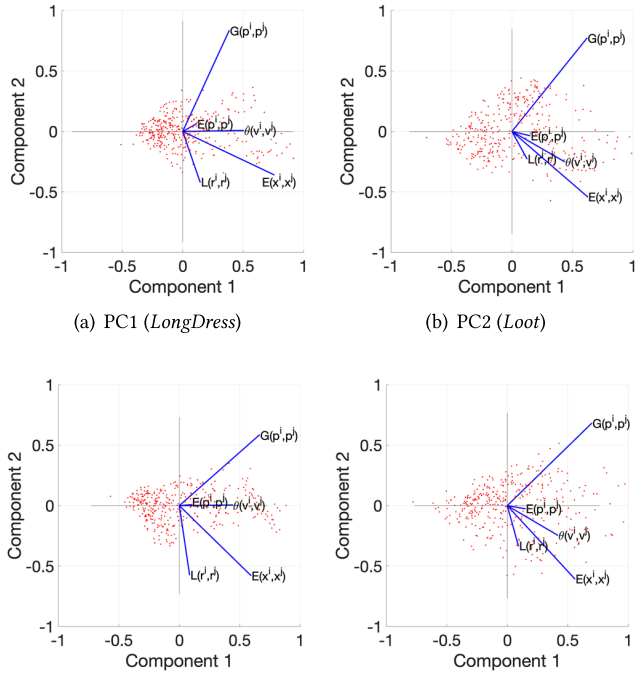


Fig. A1. Multivariate scattering plot of the proposed distance features and measurements per each sequence PC1 (*LongDress*), PC2 (*Loot*), PC3 (*Red and Black*) and PC4 (*Soldier*).



(a) PC1 (*LongDress*) (b) PC2 (*Loot*)

(c) PC3 (*Red and Black*) (d) PC4 (*Soldier*)

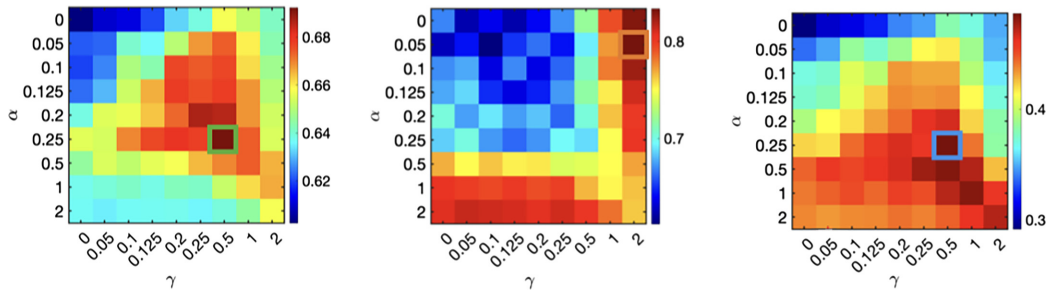
Fig. A2. Correlation plots via principal component analysis of the proposed distance features and measurements per each analysed PC1 (*LongDress*), PC2 (*Loot*), PC3 (*Red and Black*) and PC4 (*Soldier*).

between the Geodesic distance $\mathsf{G}(p^i, p^j)$ and Euclidean distance $\mathsf{E}(x^i, x^j)$ between users' position on the floor is more sparse and a clear correlation cannot be detected. We take a step further in Figure A2 showing correlation plots via PCA per each of the sequence of the analysed dataset. In each subplot red dots represent a transformed data in the principal component space, while the blue vectors indicate the direction and magnitude of the investigated metrics. From this analysis, it is clear that the Geodesic distance $\mathsf{G}(p^i, p^j)$ between viewport centre is not highly correlated with the other metrics, in particular with the Euclidean distance $\mathsf{E}(x^i, x^j)$ between users' position on the floor. Indeed, in all the subplots the corresponding vectors have a right angle indicating no correlation between them; furthermore a negative correlation is shown by the obtuse angle with the relative distance $\mathsf{L} = ||r^i - r^j||$ of users to the displayed content. A positive correlation is instead confirmed among the remaining distance features and metrics. However, there are some differences across the four sequences: the correlation between $\mathsf{E}(p^i, p^j)$ and $\theta(v^i, v^j)$ is very strong in PC1 and PC3 (Figure A2(a) and (c)) but less with $\mathsf{E}(x^i, x^j)$; while in PC2 and PC4 (Figure A2(b) and (d)) the three vectors are more closely aligned, indicating a strong positive correlation among the metrics. Despite a general correlation between the selected distance features and measurements, there is some variability among the sequences that should be deeper investigated in future work.

B Ablation Study

In this section, we present an ablation study to tune the best set of regulator parameters that maximise the performance of each similarity metric. Equipped with the threshold values given in Table 2, we run a frame-based clustering to select the best regulators α, β and σ per each metric. We test their performance based on navigation trajectories collected in the entire dataset of trajectories (i.e., navigation trajectories of both distorted and not-distorted version of the volumetric content) presented in Section 5.1 in terms of the metrics given in Section 5.2 and considering the following range of values $[0, 0.05, 0.1, 0.125, 0.2, 0.25, 0.5, 1, 2]$. For single-feature metrics ($w_1 - w_5$), we notice a very small variance in terms of performance. Thus, we selected $\alpha = 1$ for this set of metrics.

More challenging is the selection parameters for multi-feature metrics ($w_6 - w_{11}$). Each similarity metric depends on three parameters: α, β and γ. To overcome this, we first select three sets of parameters: one group of parameters (set 1) based on the maximum overlap ratio, the second (set 2) on the maximum relevant clustered population and the last group (set 3) as the one reaching the highest precision. As an example, Figure B1 shows the selection of these three sets of parameters for the metric w_{10}. Then, we test these on the trajectories experienced with not-distorted version of the volumetric content to finally select the best set of parameters. Table B1 provides the average of all the performance of the multi-feature similarity metrics obtained by the three selected sets of



(a) Overlap Ratio (b) Relevant population (c) Precision

Fig. B1. Example of parameter selection for w_{10} with $\beta = 0.5$. Values set 1 selected based on max overlap, set 2 max clustered users, set 3 based on precision.

Table B1. Parameter Selections and Their Performance for Multi-Feature Metrics (w_5–w_{11})

		w_6	w_7	w_8	w_9	w_{10}	w_{11}
set 1	$[\alpha, \beta, \gamma]$	[0.2, 0.05, 0.125]	[0.125, 0.125, 0.5]	[0.5, 0.1, 0.05]	[0.25, 0.5, 0.1]	[0.25, 0.5, 0.5]	[0.25, 0.5, 0]
	Overlap Ratio	0.64	0.66	0.63	0.66	**0.69**	0.65
	Relevant Population	0.80	0.76	**0.81**	0.70	0.69	0.69
	Precision	0.43	0.45	0.43	0.46	**0.49**	0.42
set 2	$[\alpha, \beta, \gamma]$	[0.05, 1, 0.05]	[0.05, 2, 0.05]	[0.05, 2, 0.05]	[0.1, 0.5, 2]	[0.05, 0.5, 2]	[2, 0.5, 0.05]
	Overlap Ratio	0.59	0.59	0.59	0.60	**0.64**	0.63
	Relevant Population	0.91	**0.93**	**0.93**	0.88	0.83	0.81
	Precision	0.32	0.30	0.30	0.36	0.35	**0.43**
set 3	$[\alpha, \beta, \gamma]$	[0.5, 0.05, 0.2]	[0.125, 0.05, 0.2]	[0.125, 0.05, 0.1]	[0.5, 0.5, 0.25]	[0.25, 0.5, 0.5]	[0.5, 0.5, 0.1]
	Overlap Ratio	0.64	0.66	0.63	0.65	**0.69**	0.64
	Relevant Population	0.80	0.76	**0.80**	0.77	0.69	0.71
	Precision	0.46	0.46	0.43	0.47	**0.49**	0.45

Bold represents best performance values per metric in each set.

parameters. Since there is no particular configuration that outperforms in terms of overlap ratio, relevant population and precision, we decided to select set 3. This configuration, besides ensuring the highest value of precision, also guarantees acceptable values of overlap ratio and relevant population for all the similarity metrics. For example for w_{10}, selecting values of set 3 means that users are correctly clustered in almost the 50% of the time (precision equal to 0.49); at the same time the 69% of the population is put in clusters with more than the two users (relevant population equal to 0.69) and on average the overlap of viewport between users in the same cluster is consistent (overlap ratio equal to 69%). It should also be noted that in Section 5.1 we assumed that users are classified as similar if their viewports overlap by 75% of their total viewed area. Therefore, we find it acceptable to ensure clusters with on average a consistent viewport overlap ratio of around 70% which is very close to our threshold of similarity, even if the precision values are not very high.

Received 30 January 2024; revised 28 August 2024; accepted 18 October 2024