Delft University of Technology

Extending 3-DoF Metrics to Model User Behaviour Similarity in 6-DoF Immersive Applications

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

DOI
10.1145/3587819.3590976

Publication date
2023

Document Version
Final published version

Published in
MMSys 2023 - Proceedings of the 14th ACM Multimedia Systems Conference

Citation (APA)
Rossi, S., Viola, I., Toni, L., & Cesar, P. (2023). Extending 3-DoF Metrics to Model User Behaviour Similarity in 6-DoF Immersive Applications. In *MMSys 2023 - Proceedings of the 14th ACM Multimedia Systems Conference* (pp. 39-50). (MMSys 2023 - Proceedings of the 14th ACM Multimedia Systems Conference). Association for Computing Machinery (ACM). https://doi.org/10.1145/3587819.3590976

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Extending 3-DoF metrics to model user behaviour similarity in 6-DoF immersive applications

Silvia Rossi
Centrum Wiskunde & Informatica (CWI)
Amsterdam, the Netherlands
s.rossi@cwi.nl

Irene Viola
Centrum Wiskunde & Informatica (CWI)
Amsterdam, the Netherlands
irene@cwi.nl

Laura Toni
University College London (UCL)
London, UK
l.toni@ucl.ac.uk

Pablo Cesar
CWI, TU Delft
Amsterdam, the Netherlands
Delft, the Netherlands
p.s.cesar@cwi.nl

ABSTRACT

Immersive reality technologies, such as Virtual and Augmented Reality, have ushered a new era of user-centric systems, in which every aspect of the coding–delivery–rendering chain is tailored to the interaction of the users. Understanding the actual interactivity and behaviour of the users is still an open challenge and a key step to enabling such a user-centric system. Our main goal is to extend the applicability of existing behavioural methodologies for studying user navigation in the case of 6 Degree-of-Freedom (DoF). Specifically, we first compare the navigation in 6-DoF with its 3-DoF counterpart highlighting the main differences and novelties. Then, we define new metrics aimed at better modelling behavioural similarities between users in a 6-DoF system. We validate and test our solutions on real navigation paths of users interacting with dynamic volumetric media in 6-DoF Virtual Reality conditions. Our results show that metrics that consider both user position and viewing direction better perform in detecting user similarity while navigating in a 6-DoF system. Having easy-to-use but robust metrics that underpin multiple tools and answer the question "how do we detect if two users look at the same content?" open the gate to new solutions for a user-centric system.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Virtual reality**; • **Information systems** → *Multimedia streaming*.

KEYWORDS

Point Cloud, User Behavioural Analysis, Data Clustering, 6-DoF, Immersive Reality, Virtual Reality, Trajectory analysis

ACM Reference Format:
Silvia Rossi, Irene Viola, Laura Toni, and Pablo Cesar. 2023. Extending 3-DoF metrics to model user behaviour similarity in 6-DoF immersive applications. In *Proceedings of the 14th ACM Multimedia Systems Conference (MMSys '23), June 7–10, 2023, Vancouver, BC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3587819.3590976

1 INTRODUCTION

Immersive reality technology has revolutionised how users engage and interact with media content, going beyond the passive paradigm of traditional video technology, and offering more degrees of presence and interaction in a virtual environment. Depending on how much a user can move in the 3D space, immersive environments can be classified as 3- or 6-Degrees-of-Freedom (DoF). In a 3-DoF 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-DoF 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 degrees of freedom bring the virtual experience even closer to reality: a higher level of interactivity makes the user more immersed and present within the virtual environment [4].

Despite their differences, the common denominator of both interactive systems is the viewer as an active decision-maker of the displayed content. This active role defines the *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 [28, 45]. Media codecs need to be optimised such that the quality experienced by the user is maximised [34, 46]. Similarly, streaming should be tailored to users' interactivity to ensure high-quality content and smooth navigation, while remaining bandwidth-tolerant [13, 22, 39]. From here, the importance to understand, analyse and predict users' movements

This work has been carried out with the support of ERCIM 'Alain Bensoussan' Fellowship Programme.

This work is licensed under a Creative Commons Attribution International 4.0 License.

MMSys '23, June 7–10, 2023, Vancouver, BC, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0148-1/23/06.
https://doi.org/10.1145/3587819.3590976

(*i.e.*, *user behaviour*) within an immersive scenario [12, 26, 29, 43]. A better understanding of how the population behave when experiencing immersive reality has an impact that goes beyond system applications, leading to user similarities, *i.e.*, *user clustering/profiling* [23], which is essential for several purposes: from secure authentication [42] to medical application [17].

Thanks to the large availability of public datasets [14, 20, 25], user navigation in 3-DoF immersive systems has been deeply investigated [30, 35], showing the importance of analysing and detecting key behavioural aspects in interactive (user-centric) systems. However, the 6-DoF counterpart has been scarcely considered in the literature [1, 38, 47]. Due to the change in the viewing paradigm (from inward to outward) and to more level of interaction in 6-DoF, current studies in 3-DoF cannot be directly applied to 6-DoF domains [33]. Filling this gap is the main goal of this paper by providing new metrics for user analysis in 6-DoF.

In this work, we focus on extending the applicability of clustering methods to investigate users similarity (*i.e.*, users sharing common behaviours while interacting with the content) to 6-DoF environments. Specifically, clustering techniques usually rely on pairwise similarity metrics, with similarity being in this case in terms of 6-DoF interaction. To the best of our knowledge, such metric has not been proposed yet in 6-DoF context. Starting from the state-of-the-art clustering algorithm developed in 3-DoF [27], and the main limitations of the tool when extended to 6-DoF described in [33], we investigate new methodologies for better modelling user similarities and overcoming those limitations. First, we recall the definition of user navigation trajectory in 6-DoF. 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 8 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). We validate and test our proposed similarity metrics on a publicly available dataset of navigation trajectories collected in a 6-DoF Virtual Reality (VR) scenario [39]. Results have shown that similarity metrics based on more distance features are promising solutions to correctly detect users with similar behaviour while experiencing volumetric content.

Our work contributes to the overall open problem of behavioural analysis in a 6-DoF system with the following main contributions:

- presenting the general problem of detecting behavioural similarities in a 6-DoF system, and introducing 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' trajectory similarity (in terms of distance on the plane or from the object) and the ability to approximate the ground truth. This analysis based on 6-DoF VR trajectories reveals that the position on the floor alone is not sufficient to characterise the user behaviour and that the viewing direction cannot be neglected.

The remainder of this article is organised as follows: related work on user behavioural analysis in both 3-DoF and 6-DoF systems are reported in Section 2. The main challenges of detecting behavioural similarities in a 6-DoF system and the importance of having a tool that approximates such similarities are described in Section 3. Our proposed similarity metrics are described in Section 4; while Section 5 and Section 6 present experimental setup and validation of our proposed metrics on real navigation trajectories collected in a 6-DoF VR setting, respectively. Further discussion and final conclusion are summarised in Section 7.

2 RELATED WORK

We now describe how user behaviour has been analysed in 3-DoF 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-DoF scenarios, highlighting the still outstanding open challenges.

2.1 User Behaviour in 3-DoF environment

The user navigation within a 3-DoF environment has been intensely analysed from many perspectives. Many studies have focused on psychological investigations of user engagement and presence correlated to movements within the spherical content. In [15], 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). Similarly, a correlation between the perceived sense of presence and the interactivity level was detected in [2], with more random exploratory interactions for less immersed (and hence less engaged) users. However, no objective metric to properly quantify and characterise user behaviour has been presented in these works.

To further understand how people observe and explore 360° contents, many public 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 maximum and average angular speeds under various video segment lengths [5], exploration time [35] or eye fixation distribution [7]. A deeper analysis was presented in [20] where the dataset has been analysed through a clustering algorithm presented in [27], specifically built to have in the same cluster users who similarly explore 360° content. However, behavioural analysis based on such clustering tool mainly provides a general idea of similarity among viewers without offering however a quantitative metric. To overcome such limitation, authors in [30], showed the benefit of studying spatio-temporal trajectories by information theory metrics, and thus the possibility of identifying and quantifying behavioural aspects. Key outcomes from this quantitative analysis were the study of similarities between users when watching the same content, but also the similarity of a given user when watching diverse content. The importance of these behavioural insights has been then exploited in different VR applications. For instance, authors in [21] proposed a scalable prediction algorithm for user navigation, which considered previous navigation patterns while in [19] a hybrid approach has been presented based on both dominant user behaviour (detected via a clustering approach) and the video content. Recently, authors in [11] showed that behavioural uncertainty could lead to different navigation in the future even if previously presented similarity; thus, a deep variational learning framework to predict multiple plausible head trajectories was presented. Moreover, in order to extend publicly available navigation datasets, realistic synthetic head rotation data were also generated using a deep learning algorithm given similar data distribution over time [37]. Finally, the analysis and understanding of user navigation in a VR environment have shown promising results also in determining the mental health issues of subjects (*e.g.*, anxiety, autism spectrum disorder, eating disorders, depression) and their treatment [9, 10, 18].

2.2 User Behaviour in 6-DoF environment

Extending such behavioural analysis to a 6-DoF environment is not straightforward, due to the change in the viewing paradigm (from inward to outward) and the addition of translation in 3D space. In the past, user navigation in 6-DoF scenarios was studied in the context of locomotion and display technology for CAVE environments [24, 41]. A Cave Automatic Virtual Environment (CAVE) system is an immersive room on which walls and floor are projected the video content and viewers are free to move inside [6]. For instance, the study in [41] focused on task performance analysis in terms of completion time and correct actions. Authors in [24] 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-DoF [3]. In detail, the authors showed the superiority of head-controlled techniques, allowing more sense of presence and better control with less discomfort in the navigation. While the tools mentioned above are highly informative to summarise the interaction of users within a 6-DoF environment, they usually fail to provide other key insights: which users navigate similarly, and which are the dominant interaction behaviour among users.

Recently, the focus has been put on subjective quality assessment based on different coding techniques of volumetric content, both static [1] and dynamic [40]. These studies present a statistical analysis of user movements in terms of mean angular velocity, the ratio of frames viewed while in movement, most displayed areas of the content showing an influence in the navigation due to the perceived content quality, and point out a users' preference to visualise the volumetric object from a close and frontal perspective. A behavioural analysis of user navigating in 6-DoF social VR movie has been also presented in [32]. 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. Authors in [31] present an exploratory behavioural analysis of users while displaying volumetric content within a 6-DoF environment focusing on the understanding how the way of navigating is affected by the content and its features, such as dynamics and quality, but also by the intrinsic disposition of the single user. Finally, to encourage the collection of navigation data in 6-DoF immersive experience, a new tool was recently released in [44].

These preliminary 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, suffering from the major shortcomings highlighted before. In this paper, we aim to overcome these limitations by proposing a general and efficient tool for detecting similar viewers while experiencing 6-DoF content.

3 CHALLENGES

In this work, our main goal is to define a new pairwise metric able to capture the (dis)similarity between two 6-DoF users (in terms of displayed content). This metric needs to be reliable and yet simple to compute. In the following, we first present our assumption of similarity among users while navigating in a 6-DoF environment based on [33]. 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 [27], which identified users who are attending the same portion of an omnidirectional content in a 3-DoF system. This clustering technique relies on a pairwise similarity metric, and thus, having a proper metric also for 6-DoF system would extend the applicability of this state-of-the-art tool.

3.1 User Similarity in 6-DoF

Similarly to [33], we are interested in analysing user behaviour, assuming that users interact similarly when they *observe the same 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-DoF scenario, the trajectory of a generic user i can be formally denoted by the sequence of the user's viewing direction over time $\{p^i_1, p^i_2, .., p^i_n\}$ where p^i_t is the centre of the viewport projected on the immersive content (*i.e.*, spherical video) at timestamp t. In fact, the viewport centre alone is highly informative of the user behaviour and can be used as a proxy of viewport overlap among users [27]. In particular, the geodesic distance has been proved as a reliable similarity metric such that a low value indicates high similarity between 3-DoF users.

Differently in a 6-DoF setting, the more degrees of freedom are given to the user, the more challenging becomes the system and the description of user navigation within it. The viewport centre alone is no more sufficient to characterise the user behaviour in a 6-DoF scenario since now the distance between the user and immersive content can change over time due to the added degrees of freedom. Figure 1 shows an example of two users navigating in a 6-DoF system. On the left side of the figure, there are navigation trajectories of two users i and j projected on a 2-D domain (*i.e.*, floor). Each point x_t represents the spatial coordinates (*i.e.*, [x,y,z]) on the floor of viewers while each associated vector symbolises the viewing direction. In the right part of Figure 1, we have instead a snapshot of 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 $r^i_t \gg r^j_t$, the user j (very close to the object) is visualising a very focused and

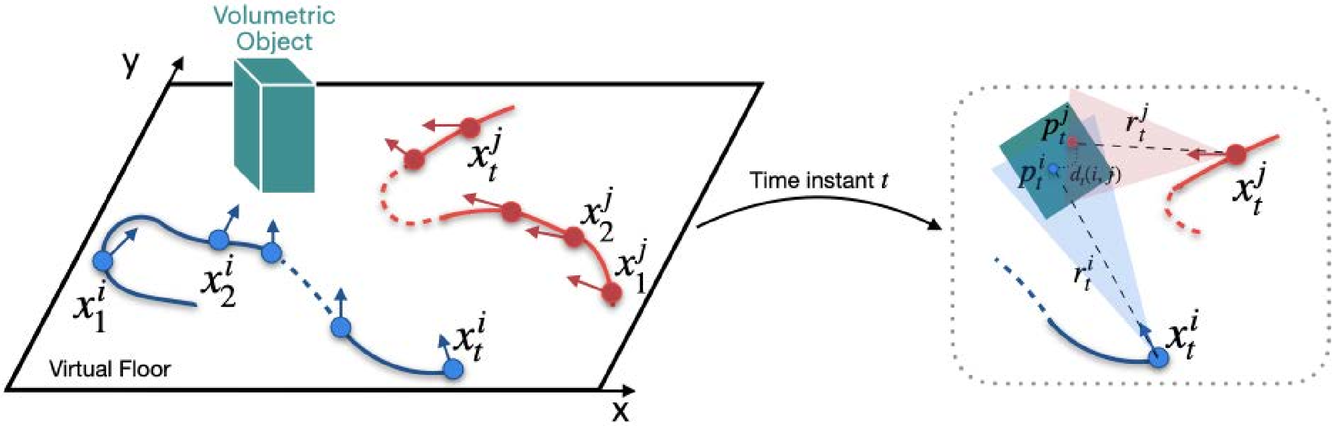


Figure 1: Example of two 6-DoF trajectories projected in a 2D domain for user i and j. On the right side, a snapshot at time t: coloured triangles represent viewing frustum per user.

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 $r_t^i \gg r_t^j$. Thus in this scenario, we cannot rely on the viewport centre only to characterise the user behaviour. The distance r and the spatial coordinates on the virtual floor x are also needed to formally define the navigation trajectory for a generic 6-DoF user i as $\{(x_1^i, p_1^i, r_1^i), (x_2^i, p_2^i, r_2^i), \dots, (x_n^i, p_n^i, r_n^i)\}$ [33]. 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 [33], the straightforward measure that could show this behaviour is the overlap among viewports. Given two users i and j shown in the right part of Figure 1, we denote their displayed viewport 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 of the displayed content, 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 0.8986 seconds per frame on average on an Intel R machine with CPU E5-4620 at 2.10 GHz; the 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 paper, we will use this metric as the ground truth of overlap among users and investigate different weights as a proxy for viewport overlap.

3.3 Clustering as a tool for behavioural analysis

Being able to assess users similarities in an objective way might be crucial for different applications such as behavioural analysis. As shown in [27], a clique-based clustering algorithm is used to detect users with similar behaviour. This requires a reliable graph to be constructed in such a way that only the nodes that identify 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 identifies optimal sub-graphs of all inter-connected nodes, ensuring the identification of the largest cluster of users all sharing a large viewport overlap. In more detail, 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 follow:

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

where $g_t(i,j)$ is a similarity metric between user i and j and G_{th} is a thresholding value. On this final graph, the clique-based clustering algorithm can be applied to identify a set of users all connected (*i.e.*, clique), and therefore with similar behaviour. In [27], this graph construction is based on a pairwise similarity metric specifically for the 3-DoF trajectories.

Identifying a generic and reliable metric $g(i,j)$ that approximates

Table 1: Definition of distance features and measurements.

Symbol	Definition
x	user position on the VR floor
p	viewport center projected on the volumetric content
r	relative distance between user and volumetric content
$\mathsf{L}(\cdot,\cdot)$	difference of relative distance between two users
$\mathsf{E}(\cdot,\cdot)$	Euclidean distance
$\mathsf{G}(\cdot,\cdot)$	Geodesic distance

behavioural similarities among users who experience a 6-DoF content is a key step to enable user behavioural analysis via tools proposed for 3-DoF scenario and the focus on the next section.

4 PROPOSED METRICS

In this section, we present eight similarity metrics and we provide an exhaustive study 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 both. We divide the 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.

4.1 Single-feature metrics to assess users similarity

The first set of similarity metrics is based on one single distance feature. We model the similarity functions via radial basis function kernel. Specifically, we consider the Gaussian kernel [36] defined as follows:

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

where $D(i,j)$ is the distance between two generic users i and j, while $\alpha > 0$ is a parameter to better regularise the distance. This distance can be evaluated in multiple ways and we consider the distance features and measurements taken into account in [33]. Specifically, the first two similarity metrics w_1 and w_2 are based 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 two 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, this 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:

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

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

4.2 Multi-feature metrics to assess users similarity

As emerged in [33], 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. In detail, w_5 and w_6 are 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 $\mathsf{G}(p^i, p^j)$ and Euclidean distance $\mathsf{E}(p^i, p^j)$, respectively. More formally, we define w_5 as:

$$\begin{aligned} w_5 &= k_\alpha^{(\mathsf{E})}(x^i,x^j) \cdot k_\beta^{(\mathsf{L})}(r^i,r^j) \cdot k_\gamma^{(\mathsf{G})}(p^i,p^j) \\ &= e^{-\alpha \mathsf{E}(x^i,x^j)} \cdot e^{-\beta ||r^i - r^j||} \cdot e^{-\gamma \mathsf{G}(p^i,p^j)}; \end{aligned} \tag{8}$$

while the second weight is equal to:

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

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

The preliminary analysis presented in [33] 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. Given the relative distance r_i between the user i and volumetric content, we evaluate it as follows:

$$\eta(r_i) = \tanh(r_i). \tag{10}$$

As previously, metrics w_7 and w_8 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. More formally, we define w_7 as follows:

$$\begin{aligned} w_7 &= k_\alpha^{(\mathsf{E})}(x^i,x^j) \cdot \beta\big[\eta(r^i) + \eta(r^j)\big] \cdot k_\gamma^{(\mathsf{G})}(p^i,p^j) \\ &= e^{-\alpha \mathsf{E}(x^i,x^j)} \cdot \beta\big[\tanh(r_i) + \tanh(r_j)\big] \cdot e^{-\gamma \mathsf{G}(p^i,p^j)}; \end{aligned} \tag{11}$$

while w_8 is:

$$\begin{aligned} w_8 &= k_\alpha^{(\mathsf{E})}(x^i,x^j) \cdot \beta\big[\eta(r^i) + \eta(r^j)\big] \cdot k_\gamma^{(\mathsf{E})}(p^i,p^j) \\ &= e^{-\alpha \mathsf{E}(x^i,x^j)} \cdot \beta\big[\tanh(r_i) + \tanh(r_j)\big] \cdot e^{-\gamma \mathsf{E}(p^i,p^j)}. \end{aligned} \tag{12}$$

5 EXPERIMENTAL SETUP

We now validate the above metrics using a point cloud dataset. We now describe the navigation dataset and how we evaluate the performance of our similarity metrics (Section 5.1 and 5.2, respectively). Then, we run an ablation study to evaluate for each similarity metric the best-performing set of regulators.

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	$k_\alpha^{(\mathrm{E})}(x^i, x^j)$	$\mathrm{E}(x^i, x^j)$	$\alpha = 1$	0.64
w_2	$k_\alpha^{(\mathrm{L})}(r^i, r^j)$	$\mathrm{L}(r^i, r^j)$	$\alpha = 1$	0.80
w_3	$k_\alpha^{(\mathrm{G})}(p^i, p^j)$	$\mathrm{G}(p^i, p^j)$	$\alpha = 1$	0.63
w_4	$k_\alpha^{(\mathrm{E})}(p^i, p^j)$	$\mathrm{E}(p^i, p^j)$	$\alpha = 1$	0.84
w_5	$k_\alpha^{(\mathrm{E})}(x^i, x^j) \cdot k_\beta^{(\mathrm{L})}(r^i, r^j) \cdot k_\gamma^{(\mathrm{G})}(p^i, p^j)$	$\mathrm{E}(x^i, x^j)$, $\mathrm{L}(r^i, r^j)$, $\mathrm{G}(p^i, p^j)$	$\alpha = 0.1$; $\beta = 0.5$; $\gamma = 1$	0.54
w_6	$k_\alpha^{(\mathrm{E})}(x^i, x^j) \cdot k_\beta^{(\mathrm{L})}(r^i, r^j) \cdot k_\gamma^{(\mathrm{E})}(p^i, p^j)$	$\mathrm{E}(x^i, x^j)$, $\mathrm{L}(r^i, r^j)$, $\mathrm{E}(p^i, p^j)$	$\alpha = 0.1$; $\beta = 0.125$; $\gamma = 0.2$	0.87
w_7	$k_\alpha^{(\mathrm{E})}(x^i, x^j) \cdot \beta[\eta(r_i) + \eta(r_j)] \cdot k_\gamma^{(\mathrm{G})}(p^i, p^j)$	$\mathrm{E}(x^i, x^j)$, r^i,r^j, $\mathrm{G}(p^i, p^j)$	$\alpha = 0.25$; $\beta = 0.5$; $\gamma = 0.5$	0.60
w_8	$k_\alpha^{(\mathrm{E})}(x^i, x^j) \cdot \beta[\eta(r_i) + \eta(r_j)] \cdot k_\gamma^{(\mathrm{E})}(p^i, p^j)$	$\mathrm{E}(x^i, x^j)$, r^i, r^j, $\mathrm{E}(p^i, p^j)$	$\alpha = 0.5$; $\beta = 0.5$; $\gamma = 0.5$	0.62

(a) *Long Dress* (PC 1) (b) *Loot* (PC 2) (c) *Red and Black* (PC 3) (d) *Soldier* (PC 4)

Figure 2: Human Body Point Clouds [16] content used in the collection of a public available dataset presented in [39].

5.1 Dataset and Methodology

Dataset. Existing datasets with user navigation collected while displaying volumetric objects in a 6-DoF environment are still very limited. In the following, we use the open dataset presented in [39]. 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 [16], namely *Long dress* (PC1), *Loot* (PC2), *Red and black* (PC3), *Soldier* (PC4) (Figure 2). Each sequence was distorted at four different bit rate points with two compression algorithms: the anchor used for the MPEG call for proposals, and the upcoming MPEG standard V-PCC. Hidden references were additionally employed in the test, for a total of 36 stimuli. Similarly to what is shown in Figure 1, 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. Therefore, the navigation data adheres to the assumptions listed in Section 3.

Graph Construction. To implement the graph-based clustering proposed in [27] 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 [27], 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 proposed in Section 4 and S_{th} a threshold value which identifies similar users and thus, neighbours on the graph. In short, users with a similarity metric below a threshold value S_{th} are neighbours in the graph. Hence, the first step now is to identify S_{th}. Per each proposed similarity metric, we empirically evaluate the Receiver Operating Characteristic (ROC) curves based on the navigation trajectories of the entire dataset above described and select the best value of threshold as originally done in [27]. Specifically, we set the thresholding values such that a good trade-off 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. The predicted event is instead evaluated using the eight 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 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.

5.2 Performance Evaluation Setup

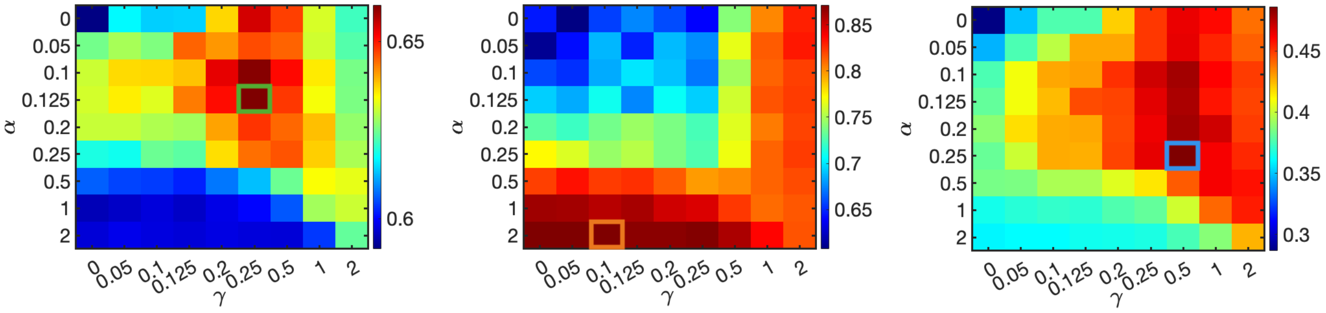
To validate our proposed similarity metrics, we consider three performance metrics: averaged *overlap ratio* per cluster, *relevant clustered population*, and *precision*. The first two are more specific to our navigation trajectory in a VR system, while the latter is a popular index used to evaluate 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 users who are put in the same group. More formally, given a detected cluster C_k is defined as follows:

$$O_k = \frac{1}{n_k} \sum_{\substack{i,j \in C_k \\ i \neq j}} O(i,j) \tag{13}$$

Table 3: Parameter selections and their performance for multi-feature metrics (w_5 - w_8).

		w_5	w_6	w_7	w_8
set 1	$[\alpha, \beta, \gamma]$	[0.12, 0.125, 0.125]	[0.12, 1, 0.25]	[0.125, 0.5, 0.25]	[0.25, 0.5, 0.2]
	Overlap Ratio	0.63	0.64	0.66	**0.69**
	Relevant Population	0.82	0.78	0.69	0.62
	Precision	0.45	0.40	0.47	0.48
set 2	$[\alpha, \beta, \gamma]$	[1, 0.05, 0.05]	[0.5, 0.05, 0.05]	[2, 0.5, 0.1]	[2, 0.5, 0.05]
	Overlap Ratio	0.58	0.59	0.60	0.63
	Relevant Population	**0.91**	0.89	0.87	0.84
	Precision	0.32	0.32	0.36	0.33
set 3	$[\alpha, \beta, \gamma]$	[0.1, 0.5, 1]	[0.1, 0.125, 0.2]	[0.25, 0.5, 0.5]	[0.5, 0.5, 0.5]
	Overlap Ratio	0.63	0.63	0.65	0.66
	Relevant Population	0.83	0.80	0.77	0.74
	Precision	0.45	0.44	**0.49**	0.48



(a) Overlap Ratio

(b) Relevant population

(c) Precision

Figure 3: Example of parameter selection for w_7 with $\beta = 0.5$. Values set 1 selected based on max overlap, set 2 max clustered users, set 3 based on precision.

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

Relevant clustered population: the more users are clustered together with high viewport overlap, the more meaningful are our clusters. Thus, we consider as relevant clustered population the sum of users that have been put in clusters with more than 2 elements.

Precision: in a classification task, this index evaluates the portion of elements that are classified correctly and has values between 0 and 1 [8]. More formally:

$$P = \frac{TP}{TP + FP} \tag{14}$$

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 threshold.

5.3 Ablation Study

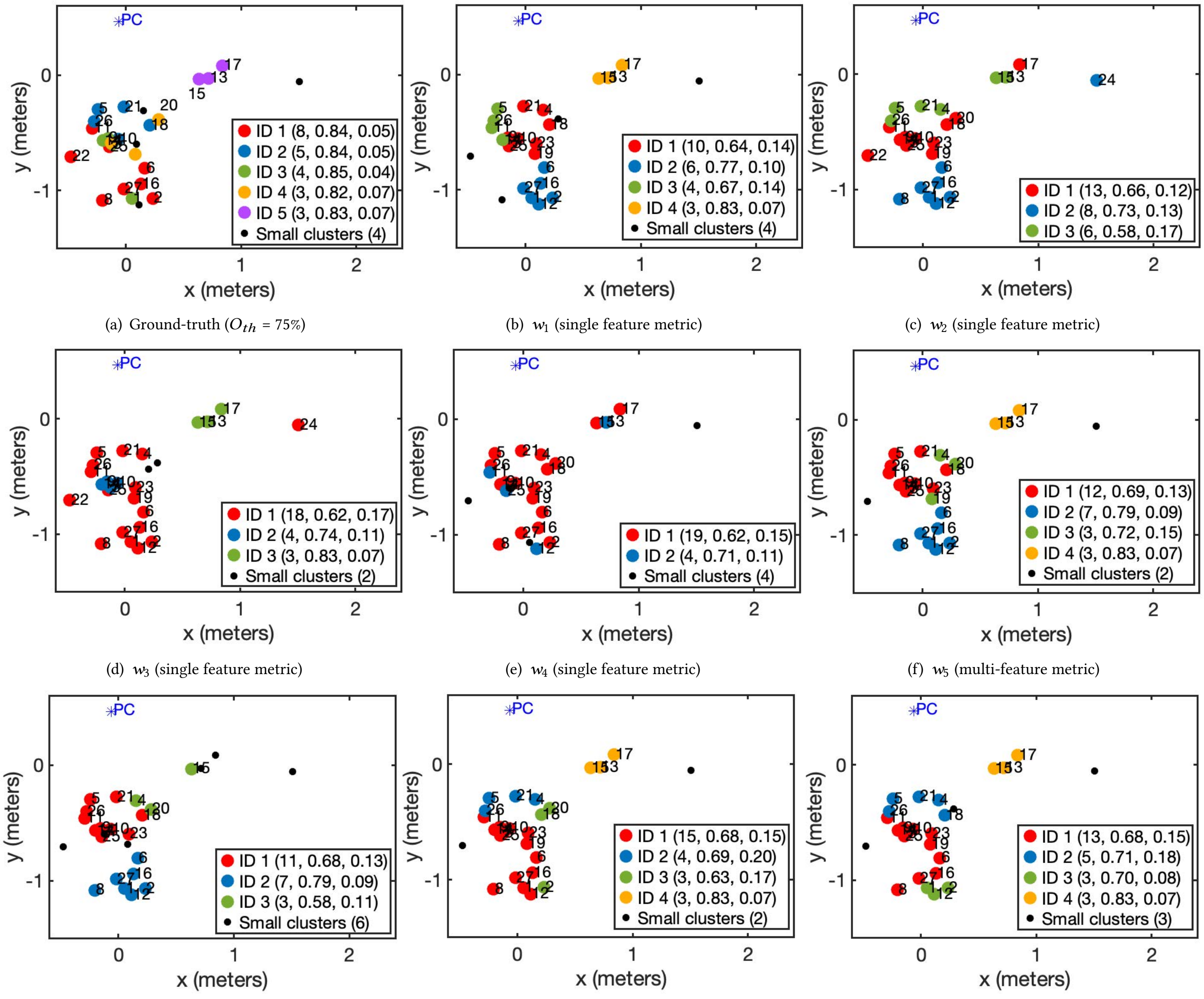
We finally present an ablation study to tune the best set of regulator parameters that maximise the performance of each similarity metric. Equipped with the threshold values given in Table 2, we run a frame-based clustering to select the best regulators α, β and σ. We test their performance in terms of the metrics above described in the following range of values $[0, 0.05, 0.1, 0.125, 0.2, 0.25, 0.5, 1, 2]$ based on navigation trajectories collected in the entire dataset. Finally, we average over time and across content the performance of each cluster obtained by all the similarity metrics.

Single-feature metrics

For single-feature metrics ($w_1 - w_4$), we notice a very small variance in terms of performance. Thus, we selected $\alpha = 1$ for this set.

Multi-feature metrics

More challenging is instead the selection parameters for multi-feature metrics ($w_5 - w_8$). Each similarity metric depends on three parameters: α, β and γ. To overcome this, we first select three sets of parameters taking into account only navigation trajectories for reference content: 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 3 shows the selection of these three sets of parameters for the metric w_7. Then, we test these on all the available trajectories included in the analysed dataset to finally select the best set of parameters. Table 3 provides all the performance of the multi-feature similarity metrics obtained by the three selected sets of 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_7, 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 77% of the population is put in clusters with more than



(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 (multi-feature metric)

(g) w_6 (multi-feature metric)

(h) w_7 (multi-feature metric)

(i) w_8 (multi-feature metric)

Figure 4: 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 2 users are reported: the number of users in the same cluster, averaged pairwise viewport overlap and corresponding variance within the cluster.

the 2 users (relevant population equal to 0.77) and on average the overlap of viewport between users in the same cluster is consistent (overlap ration equal to 65%). These values are similar to the highest value for w_7 of the relevant population and overlap ratio which are 0.87 and 0.66, respectively. Table 2 summarises the values used in the following.

6 RESULTS

Equipped with the similarity metrics, the corresponding values of regulators and thresholds in Table 2, we now conduct our validation study, focusing on analysing navigation trajectories experienced with non-distorted content.

6.1 Frame-Based Analysis

As first step, we implement a frame-based analysis (i.e., frame-based clustering) to visually compare the detected clusters by the different similarity metrics. Figure 4 shows the clusters detected using the ground-truth metric O to construct the graph (Figure 4 (a)) with the ones given based on each proposed similarity metric (Figure 4 (b-i)), for frame 50 of sequence PC1. 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 2 users), we provide in the legend the following results: the number of users inside the cluster, the average and variance of the overlap ratio

Table 4: Results in terms of averaged and standard deviation per each performance metric across the entire dataset.

	Metrics	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
PC1	Overlap Ratio	0.68 ± 0.05	0.65 ± 0.04	0.66 ± 0.04	0.68 ± 0.07	0.70 ± 0.05	0.71 ± 0.05	0.70 ± 0.05	**0.72 ± 0.06**
	Relevant Population	0.85 ± 0.04	**0.94 ± 0.03**	0.92 ± 0.05	0.84 ± 0.08	0.83 ± 0.06	0.83 ± 0.07	0.83 ± 0.06	0.83 ± 0.07
	Precision	0.44 ± 0.06	0.35 ± 0.05	0.39 ± 0.07	0.30 ± 0.06	0.47 ± 0.07	**0.49 ± 0.08**	0.46 ± 0.07	0.44 ± 0.10
PC2	Overlap Ratio	0.57 ± 0.08	0.53 ± 0.09	0.54 ± 0.12	0.54 ± 0.11	0.59 ± 0.08	0.58 ± 0.08	0.59 ± 0.12	**0.60 ± 0.10**
	Relevant Population	0.80 ± 0.07	**0.92 ± 0.06**	0.83 ± 0.07	0.89 ± 0.06	0.80 ± 0.10	0.81 ± 0.07	0.72 ± 0.08	0.73 ± 0.06
	Precision	0.45 ± 0.06	0.28 ± 0.08	0.31 ± 0.08	0.27 ± 0.08	0.47 ± 0.09	0.42 ± 0.08	**0.54 ± 0.08**	**0.54 ± 0.12**
PC3	Overlap Ratio	0.65 ± 0.06	0.60 ± 0.07	0.64 ± 0.05	0.68 ± 0.06	0.65 ± 0.06	0.65 ± 0.06	0.68 ± 0.05	**0.69 ± 0.05**
	Relevant Population	0.82 ± 0.07	**0.93 ± 0.05**	0.88 ± 0.06	0.82 ± 0.08	0.84 ± 0.06	0.81 ± 0.07	0.72 ± 0.07	0.70 ± 0.07
	Precision	0.48 ± 0.11	0.36 ± 0.08	0.39 ± 0.07	0.39 ± 0.06	0.49 ± 0.11	0.49 ± 0.10	0.52 ± 0.08	**0.55 ± 0.08**
PC4	Overlap Ratio	0.60 ± 0.04	0.52 ± 0.06	0.55 ± 0.03	0.59 ± 0.06	0.59 ± 0.04	0.58 ± 0.05	0.61 ± 0.04	**0.66 ± 0.05**
	Relevant Population	0.82 ± 0.07	**0.92 ± 0.05**	0.90 ± 0.08	0.86 ± 0.08	0.83 ± 0.08	0.77 ± 0.07	0.80 ± 0.07	0.71 ± 0.08
	Precision	0.35 ± 0.06	0.22 ± 0.04	0.31 ± 0.06	0.25 ± 0.07	0.38 ± 0.07	0.38 ± 0.09	**0.42 ± 0.06**	**0.42 ± 0.07**
All PCs	Overlap Ratio	0.62 ± 0.06	0.57 ± 0.06	0.60 ± 0.06	0.62 ± 0.07	0.63 ± 0.06	0.63 ± 0.06	0.65 ± 0.06	**0.66 ± 0.06**
	Relevant Population	0.82 ± 0.06	**0.93 ± 0.05**	0.88 ± 0.07	0.85 ± 0.08	0.83 ± 0.07	0.80 ± 0.07	0.77 ± 0.07	0.74 ± 0.07
	Precision	0.43 ± 0.07	0.30 ± 0.06	0.35 ± 0.07	0.30 ± 0.07	0.45 ± 0.09	0.45 ± 0.09	**0.49 ± 0.07**	0.48 ± 0.09

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)".

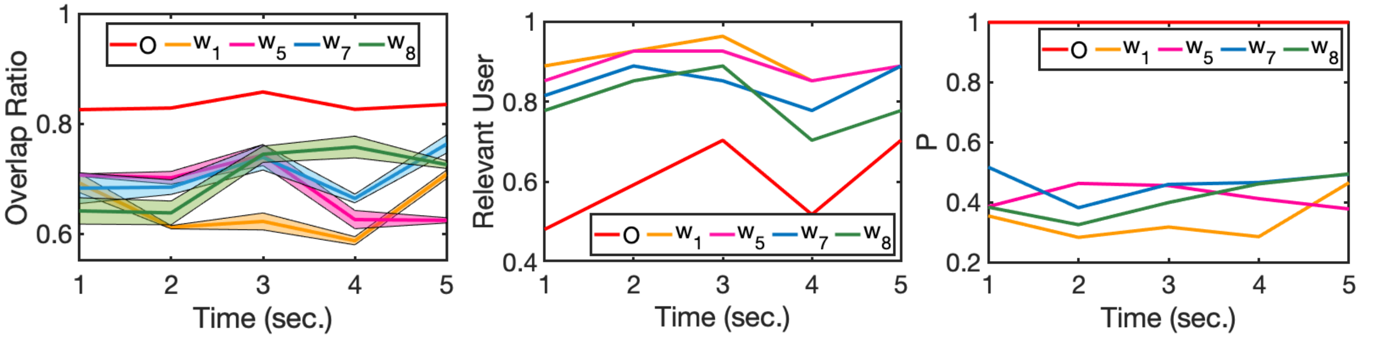
Figure 4 (a) shows the clusters that we consider as our ground truth since they are evaluated considering the overlap ratio O as a similarity metric. In this case, 5 main clusters are detected with an average overlap ratio per cluster above 0.82. In particular, cluster ID 1 has the highest number of users (8) but has a relevant value of overlap ratio (0.84). Only 4 users in this case are put in single clusters. The goal is to find a similarity metric that can detect similar results. We can notice that single feature metrics, Figure 4 (b-e), have the tendency to create very populated clusters but with a low overlap ratio. For instance, w_3 and w_4 generate a main big cluster with 18 and 19 users, respectively, while the corresponding overlap ratio drops drastically to 0.62. The only exception is given by w_1, which generates a variable set of clusters with consistent values of overlap ratio, over 0.64. Let us now consider as an example the users 13, 15 and 17, which in the ground-truth case (Figure 4 (a)) form their own cluster (*i.e.*, ID 5) with a high overlap ratio (0.83), and user 24, who is quite isolated from other users and belongs to a single cluster. We can notice that w_2 and w_4 fail in detecting the group of users 13, 15 and 17 as similar, dividing them instead in different clusters. On the other hand, w_3 detects this similarity but puts user 24 in a relevant cluster (ID 1). 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 metric in this group seems to be w_1, which is based on the user position on the virtual floor.

The last group of Figure 4 (f-i) shows clusters based on multi-feature similarity metrics. In all these settings, a total of four main clusters are detected, except for w_6 which leads to three clusters, as shown in Figure 4 (g). The latter detects the highest number of small clusters (6) while being the only one that does not identify users 13, 15 and 17 within the same cluster. On the contrary, the other three metrics w_5, w_7 and w_8 detect a main cluster and three smaller clusters with a consistent overlap ratio. For instance, the resulting clusters based on w_5 have an overlap ratio always bigger than 0.69 and only two users fall into small clusters. Overall, multi-functional metrics appear to be better suited to detect similar users than previous ones, with the exception of w_6.

In Table 4, we extend our per-frame clustering analysis to the entire dataset: we show the average and standard deviation of performance metrics described in Section 5.2 obtained by our proposed metrics. Clusters based on w_2 group in relevant clusters the majority of the population in all the analysed PCs (reaching the maximum value of 0.94 in PC1) to the detriment of precision, which falls to values between 0.22 and 0.35. As shown also in the previous investigation, the most promising similarity metrics in terms of precision and overlap ratio are both w_7 and w_8 followed by w_5. These outperform the other weights in all PCs, ensuring an overlap ratio within the same cluster with values in the range of 0.59 and 0.70 for w_7, 0.60 and 0.72 for w_8. Similarly, the values of precision are always over 0.42 for both w_7 and w_8. The only exception is in PC1, where the best performing metric in terms of precision is w_6, which for the other contents cases is always the worst performing metric.

6.2 Trajectory-Based analysis

Given the above remarks, we now analyse the performance metrics over time, taking into account only w_1, w_5, w_7, and w_8. Indeed, we decide to select the best-performing similarity metrics in the previous investigation (w_5, w_7 and w_8). To have a fair comparison, we also keep the most promising among the single-feature metrics, w_1. We compute clique-based clusters over a time window of $1s$ (*i.e.*, chunk) and a time similarity threshold of $0.8s$. At each chunk, we evaluate the average overlap ratio per relevant cluster, the average of the relevant population and the precision of detected clusters. As an example, we show in Figure 5 the performance results per sequence PC1 (*Longdress*) as functions of time per each similarity metric. In Figure 5, 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



(a) Mean Overlap Ratio in Relevant Cluster. (b) Mean Relevant Users. (c) Precision.

Figure 5: Clustering over time (chunk = 1 sec.) results per sequence PC1 (*Longdress*): comparison between ground-truth O and a subset of proposed metrics (w_1, w_5, w_7 and w_8).

over time. All the similarity metrics reach an average overlap ratio within clusters between 0.6 and 0.75 (Figure 5 (a)). However, clusters based on w_1 have lower performance, while other metrics are performing quite similarly, although with a slight predominance of w_7. In terms of relevant users (Figure 5 (b)), it is worth noting that all the proposed similarity metrics generate bigger clusters than the ground-truth metric, which considers only half of the population as relevant. In more detail, the clusters resulting from w_1, w_5 and w_8 put in relevant clusters 0.8 of the entire population for all the sequence time. Finally, in terms of precision as highlighted in Figure 5 (c) the only similarity metric that generated clusters with P over to 0.4 in the entire sequence is w_7. These investigations show that similarity metrics based on multi-feature, such as w_7 and w_8, are more promising for detecting users with similar behaviour while experiencing volumetric content.

In summary, from this validation analysis, 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;
- In particular, in spite of the slightly more complex formulation, w_7 and w_8 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 sufficient to correctly identify similar users;
- The only exception among single-feature metrics is w_1 which is based only on the position of the user on the floor. Despite its simplicity, this metric is comparable with multi-feature metrics. Hence, it can be used for an easy-to-implement preliminary behavioural analysis.

However, 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.

7 DISCUSSION AND CONCLUSION

In this paper, we have summarised the main challenges of user behavioural analysis in a 6-DoF system due to the new settings and the added locomotion functionalities. Behavioural analysis of 6-DoF 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. Thus, we considered a general ground-truth user similarity metric, such as *overlap ratio*: the percentage 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 rendering each volumetric video at any given time and for each viewer, making its computation not trivial and intensely time-consuming. To overcome this issue and to assess users' similarity in a simple and objective way, we formulated and investigated several similarity metrics considering different distance features and measurements. We were interested in modelling similarities among users *observing the same volumetric content*. In detail, we investigated different features or combinations of them which consider users' location in the virtual space and their viewing direction. We validated and tested our similarity metrics via a clique-based clustering tool proposed for 3-DoF scenario on real navigation trajectory collected in a 6-DoF VR environment. Therefore, in this article we advanced the state-of-the-art, proposing novel similarity metrics taking into account the new physical settings and locomotion functionalities given to users. 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 a hundredth of a second. This makes our proposed metrics suitable for real-time applications. In future work, we will further test the robustness and versatility of these metrics on 6-DoF navigation trajectories collected in a different virtual scenario, for example in Augmented Reality (AR) applications [47].

REFERENCES

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

[2] 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* (2021), 1–17.

[3] Weiya Chen, Anthony Plancoulaine, Nicolas Férey, Damien Touraine, Julien Nelson, and Patrick Bourdot. 2013. 6DoF navigation in virtual worlds: comparison of joystick-based and head-controlled paradigms. In *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology.*

[4] 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.

[5] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 2017. 360-degree video head movement dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference.*

[6] Heather Creagh. 2003. Cave Automatic Virtual Environment. In *Proceedings: Electrical Insulation and Electrical Manufacturing and Coil Winding Technology Conference.* 499–504. https://doi.org/10.1109/EICEMC.2003.1247937

[7] 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 *Proceedings of the 9th ACM Multimedia Systems Conference.* 432–437.

[8] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[9] 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 (2017), 2393–2400.

[10] Chris NW Geraets, Elisabeth CD 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* (2021).

[11] Quentin Guimard, Lucile Sassatelli, Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. 2022. Deep variational learning for multiple trajectory prediction of 360° head movements. In *Proceedings of the 13th ACM Multimedia Systems Conference.* 12–26.

[12] 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.* 1–13.

[13] Jonathan Harth, Alexandra Hofmann, Mike Karst, David Kempf, Annelie Ostertag, Isabell Przemus, and Bernhard Schaefermeyer. 2018. Different types of users, different types of immersion: A user study of interaction design and immersion in consumer virtual reality. *IEEE Consumer Electronics Magazine* 7, 4 (2018), 36–43.

[14] 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 *Proceedings of the 30th ACM International Conference on Multimedia.* 1025–1034.

[15] Hanseul Jun, Mark Roman Miller, Fernanda Herrera, Byron Reeves, and Jeremy N Bailenson. 2020. 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* (2020), 1–1. https://doi.org/10.1109/TAFFC.2020.3004617

[16] M Krivokuća, PA Chou, and P Savill. 2018. 8i voxelized surface light field (8iVSLF) dataset. In *ISO/IEC JTC1/SC29/WG11 MPEG, input document m42914.*

[17] Robert FK Martin, Patrick Leppink-Shands, Matthew Tlachac, Megan DuBois, Christine Conelea, Suma Jacob, Vassilios Morellas, Theodore Morris, and Nikolaos Papanikolopoulos. 2021. The Use of Immersive Environments for the Early Detection and Treatment of Neuropsychiatric Disorders. *Frontiers in Digital Health* 2 (2021), 40. 10.3389/fdgth.2020.576076.

[18] 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.

[19] Dario D. R. Morais, Lucas S. Althoff, Ravi Prakash, Marcelo M. Carvalho, and Mylene C.Q. Farias. 2021. A Content-Based Viewport Prediction Model. In *Electronic Imaging*, Vol. 2021.

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

[21] A. T. Nasrabadi, A. Samiei, and R. Prakash. 2020. Viewport prediction for 360 videos: a clustering approach. In *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video.* 34–39.

[22] 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.

[23] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–12.

[24] 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.

[25] 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 *Proceedings of the 11th ACM Multimedia Systems Conference.* 279–284.

[26] 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.

[27] 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.*

[28] Silvia Rossi, Alan Guedes, and Laura Toni. 2022. Streaming and User Behaviour in Omnidirectional Videos. *Immersive Video Technologies* (2022), 49.

[29] 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* (2020).

[30] Silvia Rossi and Laura Toni. 2020. Understanding user navigation in immersive experience: an information-theoretic analysis. In *Proceedings of the 12th ACM International Workshop on Immersive Mixed and Virtual Environment Systems.*

[31] Silvia Rossi, Irene Viola, and Pablo Cesar. 2022. Behavioural Analysis in a 6-DoF VR System: Influence of Content, Quality and User Disposition. In *Proceedings of the 1st Workshop on Interactive eXtended Reality.* 3–10.

[32] 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 *Proceedings of the International Workshop on Immersive Mixed and Virtual Environment Systems.* ACM, 1–7.

[33] Silvia Rossi, Irene Viola, Laura Toni, and Pablo Cesar. 2021. A New Challenge: Behavioural Analysis Of 6-DOF User When Consuming Immersive Media. In *2021 IEEE International Conference on Image Processing.* IEEE, 3423–3427.

[34] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A Chou, Robert A Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, et al. 2018. Emerging MPEG standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2018), 133–148.

[35] 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?. In *IEEE Transactions on Visualization and Computer Graphics.*

[36] 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 [lecture notes]. *IEEE Signal Processing Magazine* 36, 6 (2019), 133–145.

[37] Jakob Struye, Filip Lemic, and Jeroen Famaey. 2022. Generating Realistic Synthetic Head Rotation Data for Extended Reality using Deep Learning. In *Proceedings of the 1st Workshop on Interactive eXtended Reality.* 19–28.

[38] Shishir Subramanyam, Jie Li, Irene Viola, and Pablo Cesar. 2020. Comparing the Quality of Highly Realistic Digital Humans in 3DoF and 6DoF: A Volumetric Video Case Study. In *IEEE Conference on Virtual Reality and 3D User Interfaces.*

[39] Shishir Subramanyam, Irene Viola, Alan Hanjalic, and Pablo Cesar. 2020. User Centered Adaptive Streaming of Dynamic Point Clouds with Low Complexity Tiling. In *Proceedings of the 28th ACM International Conference on Multimedia.* 3669–3677.

[40] 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.*

[41] 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.*

[42] Pier Paolo Tricomi, Federica Nenna, Luca Pajola, Mauro Conti, and Luciano Gamberini. 2022. You Can't Hide Behind Your Headset: User Profiling in Augmented and Virtual Reality. *arXiv preprint arXiv:2209.10849* (2022).

[43] Jeroen van der Hooft, Tim Wauters, Filip De Turck, Christian Timmerer, and Hermann Hellwagner. 2019. Towards 6-DoF HTTP adaptive streaming through point cloud compression. In *Proceedings of the 27th ACM International Conference on Multimedia.*

[44] 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 *Proceedings of the 13th ACM Multimedia Systems Conference.*

341–346.

[45] Irene Viola and Pablo Cesar. 2022. Volumetric video streaming. *Immersive Video Technologies* (2022), 425.

[46] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le 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.

[47] 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.