An Image-Based Encoding to Record and Track Immersive VR Sessions

Bruno Fanini[1,2(✉)] and Luigi Cinque[2]

[1] CNR ITABC (Institute for Technologies Applied to Cultural Heritage), Area della Ricerca Roma 1, Via Salaria Km. 29,300, 00015 Rome, Italy
[2] Department of Computer Science, Sapienza University, Via Salaria 113, Rome, Italy
{fanini,cinque}@di.uniroma1.it

Abstract. Capturing and tracking immersive VR sessions performed through HMDs in public spaces, may offer valuable insights on users propensities and spatial affordances. Large collected records can be exploited to analyze or fine-tune locomotion models for time-constrained experiences. The transmission or streaming of such data over the web to analysts or professionals in distance learning field although, can be challenging due to network bandwidth or involve computationally intensive decoding routines. This work investigates compact encoding models to volumetrically absorb user states and propensities during running VR sessions, using image-based encoding approaches. We focus on quantization methods and data layouts to smoothly record immersive sessions and how they compare to standard approaches in terms of storage and spatio-temporal accuracy. Qualitative and quantitative results obtained from public exhibits are presented in order to validate the encoding model.

Keywords: Immersive VR · Visual analytics · Encoding models · Saliency

1 Introduction

Recent consumer head-mounted displays (HMDs) are often employed in public or shared spaces to create engaging and educational immersive experiences with high sense of presence. During a limited amount of time, several users (casual visitors, students, etc.) explore and perform interactions within the 3D virtual environment. An in-depth investigation of spatial affordances, user propensities and saliency for a given immersive virtual environment (IVE) can be useful to improve the overall immersive experience. Capturing, recording and visualizing entire VR sessions can in fact provide valuable insights. For instance within IVEs for education, the ongoing data collection as students interact with the 3D space can be useful for learning assessment [18]. In distance learning technologies and education, such data may support the remote validation of IVEs consumed by students through an HMD and supports the remote analysis of their behaviors and interactions performed in the 3D space. Furthermore, improving or enhancing an IVE often present huge efforts in terms of 3D modeling

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11620, pp. 299–310, 2019.
https://doi.org/10.1007/978-3-030-24296-1_25

and/or optimization tasks: unexpected elements may capture users' attention, depending on locomotion models used and spatial/temporal constraints. From the analyst perspective, in order to understand user sessions from quantitative and qualitative perspectives, collecting per-user fine-grained data is generally a strong requirement. When directly mapped to the original 3D scene, recorded data (e.g. user locomotion) may provide easier interpretation for spatial analysis, when supported by usable UIs. Furthermore, remotely located analysts may be interested into inspection, visualization and analysis of user sessions while the VR application (desktop or WebVR[1] application) is still running. The transmission or streaming of collected data over the web to analysts, teachers or other professionals although, can be challenging due to network bandwidth or involve computationally intensive decoding routines. The whole history of a single user session may contain huge amount of information (locomotion, gaze, orientation, etc.) thus potentially preventing interactive data transport online.

Contribution - Within this work, we investigate an image-based encoding model that offers flexible runtime accessories to volumetrically absorb user states of VR sessions over time. Captured spatial attributes are encoded into compact image-based layouts providing (a) lightweight transmission over the web; (b) minimal computation for encoding/decoding routines; (c) direct manipulation on GPU and (d) offline 2D image processing. We apply the encoding model to previously recorded locomotion data (CSV) during a public event and evaluate obtained results in terms of accuracy, also performing comparisons with raw binary formats and different bit-depths for encoded image.

2 Related Work

Visual Analytics [23] is a proven approach to directly perceive patterns and extract knowledge from massive and dynamic information streams, offering analysts, teachers or other professionals tools to detect the expected and discover the unexpected. Interactive installations in public or shared spaces allow to collect large amounts of data related to users sessions and visually analyze them [1]. Data mining methods are often applied to tourist activities discovering landmark preferences from photo mappings [13] while classical clustering methods [12] can be used to offline analyze spatial behaviors. VR locomotion plays a central role [5] within 3D virtual environments users are free to explore. Understanding users spatial preferences may support or validate the effectiveness of interaction model used for temporally constrained sessions (e.g. public exhibits, shared spaces, virtual classrooms, etc.).

Saliency in VR. There is recently a growing interest in researching virtual environment saliency for immersive VR, although most of these works focus on 360 or omnidirectional visual content from fixed viewpoints. [19] analyzes for instance how people explore virtual environments from a fixed viewpoint (360) collecting and visualizing gaze data. [21] investigates a model to obtain fixation locations and maps from head direction, offering a good approximation when eye

[1] https://webvr.info/.

tracking data is not available. Regarding visual attention analysis [14] introduced an efficient metric and visualization method for similarity measures between a director's cut and users scan-paths using color-coded maps.

Image-based encoding can be used to transport spatial data in a compact manner (e.g. coordinates, vectors, etc.) through common 2D images. Regarding image formats for the web, PNG format [22], offers a network-friendly, patent-free, lossless compression scheme that is cross-platform. Previous research employed PNG format for instance as externalized mesh container [7,15] to efficiently transmit geometry data over the network and decoded by client browsers. Previous encoding models in literature also investigated "geometry images" [10] as 2D arrays used to quantize and store spatial information ($< x, y, z >$) as RGB values.

3 Encoding Model

This section describes image-based encoding and layout to capture and store immersive VR sessions. We highlight advantages in terms of model compactness and its lightweight transport over networked scenarios towards remote professionals. Within a previous research [8] we describe user state s as a collection of state attributes (s_a, where a is the attribute we are dealing with). For instance s_p represent user HMD location inside the virtual scene ($< x, y, z >$), s_o represent HMD orientation (for instance as quaternion $< x, y, z, w >$), etc. Each casual visitor performs different interactions within a limited amount of time during the public exhibit, thus we define the session operator S:

$$S_a(u, t) \rightarrow s_a \tag{1}$$

where u represents the user ID and s_a the state attribute at given time $t \in \mathbb{R}$. Regarding immersive user sessions, we are particularly interested in tracking specific state attributes over time, like location (s_p), HMD view direction (s_d), focus (s_f) or more advanced data. For instance, $S_p(u, t)$ captures user u locomotion or $S_f(u, t)$ can be exploited for tracking visual attention.

3.1 Session Volumes

Session volumes (or *Quantized User Session Volume* - QUSV [8]) are AABB[2] accessories deployed to observe a portion of the virtual space, "absorbing" immersive sessions. The role of a single volume V in a virtual space is to track user states within its volumetric extents and record them as session signals: the model allows to encode them as stream of RGB(A)[3] values and to store them as common images. In order to map spatial attributes (e.g. s_p, s_f) into RGB values, a volume V is uniformly subdivided into several voxels: each attribute is thus *quantized* depending on QUSV extents (Fig. 1).

[2] Axis-Aligned Bounding Box.
[3] Red, Green, Blue, Alpha channels.

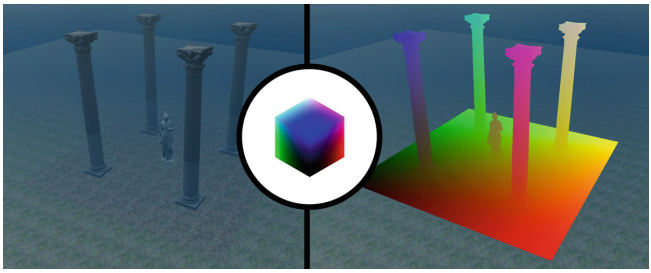

Fig. 1. A session volume arranged in a sample 3D scene (color-coded).

Considering a normalized 3D location ($< x, y, z >$) inside the volume, a direct mapping into a $< R, G, B >$ color using 8 bit per-channel storage allows V to address 256^3 (16.777.216) different voxels. The extents of V have obviously a huge impact on each voxel size, and thus on location quantization. The flexibility of such volumes is to observe limited spatial extents with the purpose of capturing and encoding specific user behaviors. The model allows to deploy multiple volumes in the same virtual environment, each observing different extents and encoding state attributes, offering great scalability and customization.

3.2 Image-Based Encoding and Layout

For a single user u, we can define the encoding of $S_a(u, t)$ for spatial attributes within a given volume as a 1-dimensional stream of RGB(A) values. Given a fixed temporal step increment, each pixel of the 1D stream thus encodes a specific voxel within V boundaries at given time t. V location and extents are indeed crucial for the client application (analyst), in order to decode the signal and retrieve spatial values. The stream can be stored as an image, and common 2D image processing can be optionally employed to perform further data manipulation. This includes for instance: channel isolation (e.g. blue to analyze height variation), temporal compression through resampling methods, common image compression algorithms, and much more. The proposed data layout also allows fast encoding/decoding and direct manipulation on GPU. For instance automatic interpolation of spatial attributes is obtained by common texture filters. From a storage perspective, data compression is furthermore facilitated by PNG format, due to likely continuous RGB data on neighboring pixels [15] (e.g. locomotion data).

Multiple 1D signals can be stacked (see Fig. 2, B) to create image atlases called *QSA* [8]: the volume is thus able to encode the whole $S_a(u, t)$ for a given state attribute a for all recorded users in a single, compact image. Such session atlases allow interactive manipulation on GPU and offline 2D image processing. For instance a QSA tracking a spatial attribute, can be shrinked along the x-axis (time) using nearest-neighbor algorithm for a coherent (but lossy) data reduction (see Fig. 2, C). Such basic 2D operations can be employed to maintain

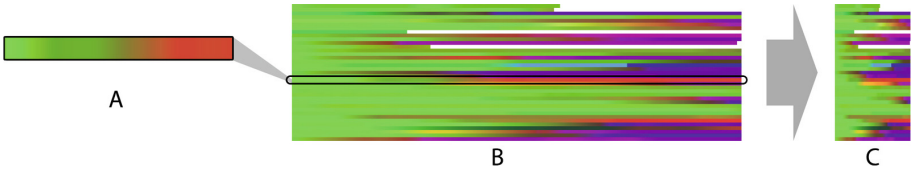

Fig. 2. QSA layout.

an approximation of original record (fetched texels are interpolated on the GPU) or compress old datasets.

3.3 Saliency

Spatial attributes (user location, focus, etc.) recorded during the whole session can be exploited also to compute salient voxels in a given V, providing *volumetric* data on users propensities. Persistence (over time) and other contributing factors (see for instance [1] regarding focus) allow observed voxels in V to be ranked, offering analysts dataset to discover user behaviors in specific parts of the scene. Such data can be encoded as well, to provide a lightweight data transport over the network towards remote analysts.

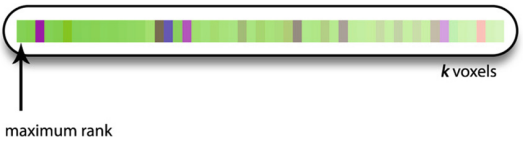

Fig. 3. A sample saliency table with length = k.

The approach of saliency tables (σ) is to maintain a running record of *VOI*s (voxels of interest) inside a session volume, keeping them sorted depending on a rank value (see Fig. 3). The same color-coding method described in Sect. 3.2 is employed to produce a single image, extended with alpha channel to encode rank. The algorithm is based on [16] with an hash-table to compute frequency counts while keeping a constant memory footprint (k, the size of table) and a running record approach. This results in an evolving "fingerprint" of the tracked spatial attribute (e.g. user location) for V. Once the saliency table is encoded as image, the sorted layout allows GPU routines to partially evaluate it by discarding right-most values (lower rank). Within unconstrained (or partially constrained) VR locomotion models, σ_p can be useful for analysts to study explorative behaviors, selectively improve 3D scene detail or support creation of locomotion graphs [11]. Furthermore, there are several advantages of encoded saliency tables:

– they are ultra-compact (see results in Sect. 5), suitable for networked contexts and remote analytics

– sorted layout for σ allows rightmost pixel to be discarded still maintaining good overall approximation
– multiple σ (e.g. different spatial attributes) can be stacked vertically (like QSA layout in Sect. 3.2) to transport several spatial propensities for a volume, using a single lightweight image.

4 Architecture Overview

We describe the general architecture of our framework, employed within public exhibits or events involving large amounts of casual visitors. We describe three main components: (1) VR application; (2) session encoding node and (3) visual inspection node (for analysts) - see Fig. 4. Depending on public space conditions (e.g. absence of internet connection, etc.) some of these components may physically rely on the same machine - for instance the VR application and encoding node.

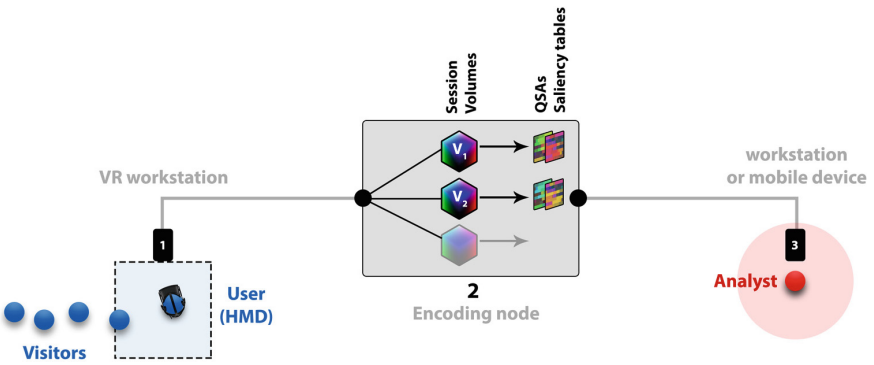

Fig. 4. Architecture of proposed framework.

Casual visitors of the public expo use the VR workstation (1) to explore or interact with a 3D scene: a single session begins when user wear the HMD and user state is being anonymously recorded. The user state is sent to the encoding node (2) including multiple state attributes, for instance: timestamp; virtual location (3D scene coordinates); HMD orientation and view direction; focus (3D coordinates); physical location (local 3D coordinates inside tracked area); etc.

Incoming user states are received by (2) and encoded (see Sect. 3) into multiple QSA and saliency tables as lightweight images. Such operation uses minimal computation for each volume $(V_1, V_2, ...V_k)$, specifically $O(k \cdot h)$ where k is the number of session volumes deployed into the 3D scene and h the number of different state attributes to track $(a_1, a_2, ...a_h)$. Generally h is quite small, since we are interested in a few state attributes (e.g. location, orientation, focus), while number of volumes (k) is highly dependent on analysis requirements and 3D scene

complexity (a single volume can be sufficient to understand user propensities in a specific portion of the virtual environment).

The third component (3) is the inspection node, where a visual analyst can remotely inspect produced datasets (lightweight images) using a desktop application (or web-application). Through the definitions of $(V_1, V_2, ...V_k)$ including extents and positions, the application is able to *decode* loaded QSAs and saliency tables for visual analysis.

An implementation of the architecture has been already realized as part of an open-source project[4] [2,4,17] leveraging on node.js ecosystem [6,20] HTML5 websockets and open-source WebGL libraries for the responsive Front-End. Next section will present encoding results interactively inspected and visualized through the implemented web-based architecture.

5 Results

We present in this section different results obtained by applying proposed encoding model to an existing CSV dataset, recorded during a previous public event ("TourismA" 2018 in Florence, Italy). We focused on locomotion data ($S_p(u,t)$), comparing obtained output to common storage models (ASCII, binary) in terms of accuracy, quantization and data storage. We also compare different formats and compression for QSAs and saliency tables generated images.

For original CSV dataset (ASCII), a timestep of 0.1 s was used to record user states: from previous experiences, this is generally a good interval to capture spatial attributes (like location in the virtual environment) during the VR session. For encoded QSA (see Sect. 3.2) this means each pixel of the image atlas represent $\frac{1}{10}$ s (10 contiguous pixels on x-axis represent one second). The 3D scene we considered is an old virtual environment - employed by CNR ITABC in several past projects [3,9] - due to its extents and size.

In order to test and compare meaningful sessions, we first filtered original locomotion data CSV, using a few acceptance policies suitable for this specific virtual environment:

- *Session duration*: discard very short sessions (< 20 s)
- *Session radius*: compute a bounding sphere for each user locomotion data, if radius $r > 5.0$ m, accept the session
- *Variance*: compute user location variance (s^2) for each dimension. If $max(s_x^2, s_y^2, s_z^2) > 5$, accept the session.

From filtered CSV dataset, we performed different tests to assess storage and accuracy, using standard binary encoding and image-based encoding (presented model). We used a session volume V_1 with extents (50.28 m $\times$ 76.46 m $\times$ 10.0 m): due to voxel quantization, this results in a location error for each dimension of $\epsilon_x = \pm 9$ cm, $\epsilon_y = \pm 14.9$ cm and $\epsilon_z = \pm 1.9$ cm. A second volume V_2 (25.27 m $\times$ 12.85 m $\times$ 8.28 m) was deployed to capture finer users locomotion on the temple

[4] http://osiris.itabc.cnr.it/scenebaker/index.php/projects/aton/.

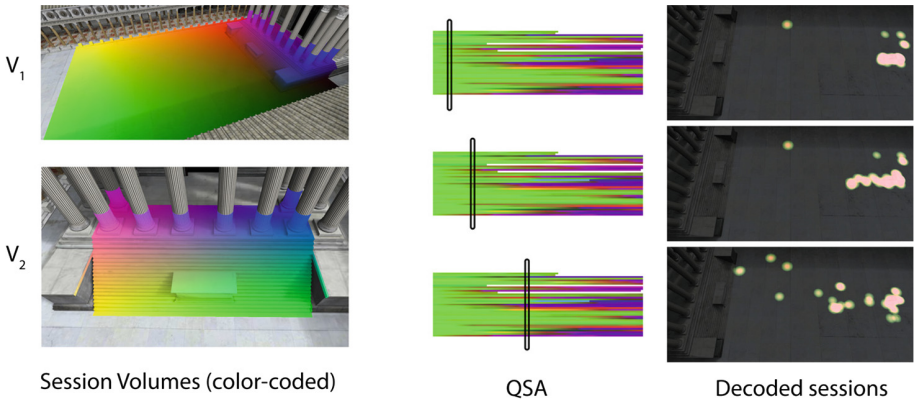

Fig. 5. The two volumes V_1 and V_2 (left, color-coded). Encoded QSA for V_1 and real-time decoding of HMD locomotion data on the GPU (WebGL) through a temporal slider.

stairs with quantization error of $\epsilon_x = \pm 4.9\,\text{cm}$, $\epsilon_y = \pm 2.5\,\text{cm}$ and $\epsilon_z = \pm 1.6\,\text{cm}$ (Fig. 5).

We compare a part of original recorded locomotion dataset size (CSV, 2,86 Mb) with uncompressed binary (double precision and single byte) and QSA (PNG format, bit depth 8, lossless). As shown in Fig. 6, uncompressed binary encoding (double and single byte precision) offer smaller storage, as expected. Proposed encoding model applied to locomotion data using lossless PNG (bit-depth 8) results in very compact QSA (20,1 Kb) storing all users sessions as described. We also evaluate different bit depths for QSA with increasing error during HMD location decoding. From a macroscopic visual inspection of user sessions, 6 bit-depth still maintains a good approximation compared to original 8-bit encoded HMD location, while 4 bit-depth shows major quantization errors

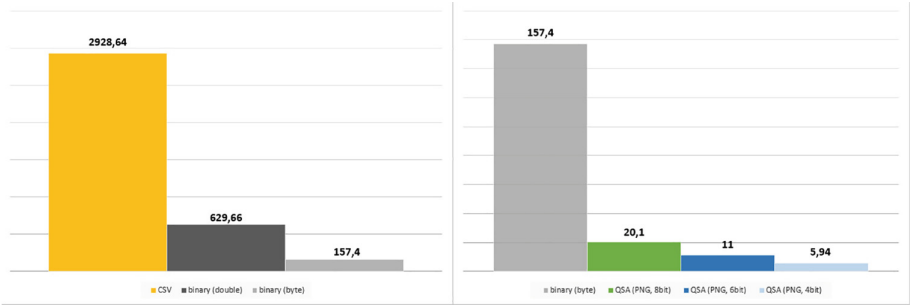

Fig. 6. Left: storage comparison (Kb) between original CSV dataset and binary (double and byte precision). Right: comparison between binary (byte) and proposed image-based encoding (QSA) using different bit depths (8, 6 and 4)

(see Fig. 7). Saliency tables for locomotion data (σ_p) were also evaluated: in this case, the adopted policy to rank voxels was persistence over time. Obtained 1D encoded image (lossless PNG format, bit-depth 8) using a table size of 1024 voxels, resulted in 2, 86 Kb, thus providing a very comfortable data transport over the network.

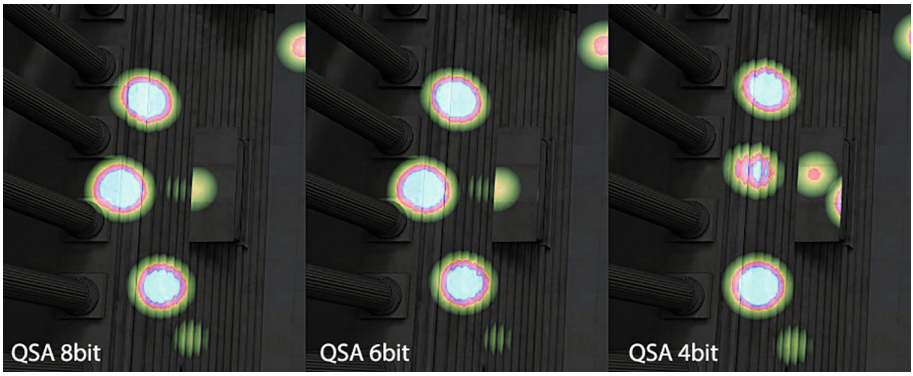

Fig. 7. Comparison between QSAs decoded at runtime using different PNG bit depths (8, 6 and 4)

Figure 8 shows interactive evaluation of σ_p on the GPU through the WebGL Front-End: due to the ranked approach, most salient locations (2%, top left) are related to user starting location as expected, progressively increasing on stairs and entrance of main temple (50% and 80%). The radius (in this case 20 m) also allows to visually understand which parts of the model potentially require

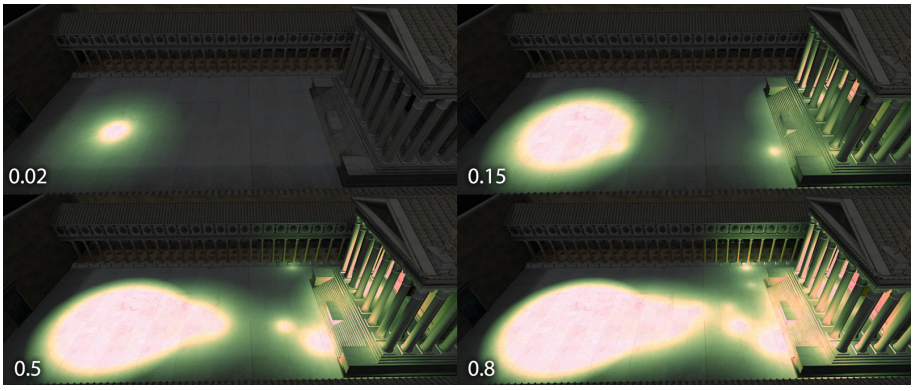

Fig. 8. Interactive evaluation and decoding of saliency table (location persistence over time) on the GPU, using radius of 20 m.

higher detail, supporting selective 3D scene improvements (geometry, texturing, etc.) for immersive VR exploration.

6 Conclusions and Future Developments

We proposed and described an image-based encoding model to volumetrically capture VR sessions during public exhibits or online experiences, producing compact and lightweight data layouts for remote and interactive visual inspection. Within networked scenarios, the model can be particularly useful for analysts, teachers and other professionals to support the remote analysis of spatial affordances and behaviors of users (casual visitors, students, etc.) consuming immersive virtual environments (IVEs) through HMDs.

Session Volumes allow to absorb user attributes and propensities, arranged and deployed at runtime. They are able to encode spatial attributes into special image atlases (*QSA*) allowing to record HMD attributes of several users over time into compact images. Due to the simple RGB mapping and layout, encoding/decoding routines use small computational resources, allowing direct manipulation on GPU and offline intervention through common 2D image processing. We show advantages of proposed solution in terms of scalability and customization, observing only specific portions of the 3D scene and how volume extents impact on accuracy and voxel quantization.

The model has accuracy limitations when encoding HMD spatial attributes due to *voxelization* of the session volume (quantization) introducing controlled error on location decoding (see Sect. 3.1). Such spatial quantization highly depends on bit-depth used for QSA image format (see Fig. 7) and volume extents. On the other hand this issue can be solved or mitigated by arranging multiple smaller session volumes for fine-grained capture, resulting in more accurate decoding. Regarding duration, texture size may limit recording: for instance using a 0.1 s capture interval and a 4096 image width for QSA, the maximum duration we are able to encode for each user is around 6 min. This can be easily solved by using paging techniques [8] thus addressing prolonged immersive sessions. We already employed the model to capture other spatial attributes like focus ($S_f(u,t)$) to visually understand users attention over time and to explore volumetric saliency in a 3D scene. We foresee advantages of encoding model also for the observation of physical motions inside the tracked area and its usage during the session, for instance HMD local motions or VR controllers. Consumer 6DOF headsets allow to track areas of a few meters: for instance a 3 m $\times$ 3 m physical space with session volume attached would result in small quantization errors ($\pm$5 mm) to record spatial attributes.

Acknowledgements. Authors would like to thank D. Ferdani (CNR ITABC) for providing parts of old 3D assets used during Keys2Rome project (http://keys2rome.eu/) and all people involved in VR installation during "TourismA 2018" event.

References

1. Agus, M., Marton, F., Bettio, F., Gobbetti, E.: Interactive 3d exploration of a virtual sculpture collection: an analysis of user behavior in museum setting. In: Proceedings of the 13th Eurographics Workshop on Graphics and Cultural Heritage (2016)
2. Antal, A., et al.: A complete workflow from the data collection on the field to the deployment of a virtual museum: the case of virtual sarmizegetusa (2016)
3. Antonaci, A., Pagano, A.: Technology enhanced visit to museums. a case study: keys to Rome. In: Proceedings of the INTED 2015, Madrid, Spain, pp. 2–4 (2015)
4. Barsanti, S.G., Fanini, B., et al.: The WINCKELMANN300 project: dissemination of culture with virtual reality at the Capitoline Museum in Rome. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **42**(2) (2018)
5. Boletsis, C.: The new era of virtual reality locomotion: a systematic literature review of techniques and a proposed typology. Multimodal Technol. Interact. **1**(4), 24 (2017)
6. Cantelon, M., Harter, M., Holowaychuk, T., Rajlich, N.: Node.js in Action. Manning Greenwich (2014)
7. Dworak, D., Pietruszka, M.: Fast encoding of huge 3D data sets in lossless PNG format. In: Zgrzywa, A., Choroś, K., Siemiński, A. (eds.) New Research in Multimedia and Internet Systems. AISC, vol. 314, pp. 15–24. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-10383-9_2
8. Fanini, B., Cinque, L.: Encoding VR sessions: image-based techniques to record and inspect immersive experiences. In: 3rd Digital Heritage International Congress (DH2018) held Jointly with 24th International Conference on Virtual Systems & Multimedia (VSMM 2018). IEEE (2019, in press)
9. Fanini, B., d'Annibale, E., Demetrescu, E., Ferdani, D., Pagano, A.: Engaging and shared gesture-based interaction for museums the case study of K2R international expo in Rome. In: 2015 Digital Heritage, vol. 1, pp. 263–270. IEEE (2015)
10. Gu, X., Gortler, S.J., Hoppe, H.: Geometry images. ACM Trans. Graph. (TOG) **21**(3), 355–361 (2002)
11. Habgood, M.J., Moore, D., Wilson, D., Alapont, S.: Rapid, continuous movement between nodes as an accessible virtual reality locomotion technique. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 371–378. IEEE (2018)
12. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognit. Lett. **31**(8), 651–666 (2010)
13. Jankowski, P., Andrienko, N., Andrienko, G., Kisilevich, S.: Discovering landmark preferences and movement patterns from photo postings. Trans. GIS **14**(6), 833–852 (2010)
14. Knorr, S., Ozcinar, C., Fearghail, C.O., Smolic, A.: Director's cut-a combined dataset for visual A ention analysis in cinematic VR content (2018)
15. Limper, M., et al.: Fast, progressive loading of binary-encoded declarative-3d web content. IEEE Comput. Graph. Appl. **33**(5), 26–36 (2013)
16. Manku, G.S., Motwani, R.: Approximate frequency counts over data streams. In: VLDB 2002: Proceedings of the 28th International Conference on Very Large Databases, pp. 346–357. Elsevier (2002)
17. Meghini, C., Scopigno, R., Richards, J., Fanini, B., et al.: ARIADNE: a research infrastructure for archaeology. J. Comput. Cult. Herit. (JOCCH) **10**(3), 18 (2017)

18. Shute, V., Rahimi, S., Emihovich, B.: Assessment for learning in immersive environments. In: Liu, D., Dede, C., Huang, R., Richards, J. (eds.) Virtual, Augmented, and Mixed Realities in Education. SCI, pp. 71–87. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-5490-7_5
19. Sitzmann, V., et al.: Saliency in VR: how do people explore virtual environments? IEEE Trans. Vis. Comput. Graph. **24**(4), 1633–1642 (2018)
20. Tilkov, S., Vinoski, S.: Node.js: using JavaScript to build high-performance network programs. IEEE Internet Comput. **14**(6), 80–83 (2010)
21. Upenik, E., Ebrahimi, T.: A simple method to obtain visual attention data in head mounted virtual reality. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 73–78. IEEE (2017)
22. Wiggins, R.H., Davidson, H.C., Harnsberger, H.R., Lauman, J.R., Goede, P.A.: Image file formats: past, present, and future. Radiographics **21**(3), 789–798 (2001)
23. Wong, P.C., Thomas, J.: Visual analytics. IEEE Comput. Graph. Appl. **5**, 20–21 (2004)