Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

EDVAM: a 3D eye-tracking dataset for visual attention modeling in a virtual museum*#

Yunzhan ZHOU†1, Tian FENG†‡2, Shihui SHUAI3, Xiangdong LI4, Lingyun SUN5, Henry Been-Lirn DUH2

[1]*Department of Computer Science, Durham University, Durham DH1 3LE, UK*

[2]*Department of Computer Science and Information Technology, La Trobe University, VIC 3086, Australia*

[3]*Alibaba Group, Hangzhou 311121, China*

[4]*Department of Digital Media, Zhejiang University, Hangzhou 310027, China*

[5]*International Design Institute, Zhejiang University, Hangzhou 310058, China*

†E-mail: yunzhan.zhou@durham.ac.uk; t.feng@zju.edu.cn

Received July 3, 2020; Revision accepted Feb. 15, 2021; Crosschecked Nov. 2, 2021

Abstract: Predicting visual attention facilitates an adaptive virtual museum environment and provides a context-aware and interactive user experience. Explorations toward development of a visual attention mechanism using eye-tracking data have so far been limited to 2D cases, and researchers are yet to approach this topic in a 3D virtual environment and from a spatiotemporal perspective. We present the first 3D Eye-tracking Dataset for Visual Attention modeling in a virtual Museum, known as the EDVAM. In addition, a deep learning model is devised and tested with the EDVAM to predict a user's subsequent visual attention from previous eye movements. This work provides a reference for visual attention modeling and context-aware interaction in the context of virtual museums.

Key words: Visual attention; Virtual museums; Eye-tracking datasets; Gaze detection; Deep learning
https://doi.org/10.1631/FITEE.2000318 **CLC number:** TP391

1 Introduction

Supported by head-mounted displays (HMDs), virtual museums can represent real-world cultural and historical exhibits through virtual reality (VR) techniques, and offer an immersive and satisfactory user experience. Researchers have proposed various interaction methods to enhance user experience with an improved sense of presence using haptic feedback (Azmandian et al., 2016; Hirota and Tagawa, 2016; de Jesus Oliveira et al., 2016; Lopes et al., 2017), hand tracking (LaViola, 2015; Davis et al., 2016; Hirota and Tagawa, 2016), or motion tracking (Suma et al., 2015; Nielsen et al., 2016). In comparison, context-aware interaction enables more significant improvement on user experience in virtual museums, but it requires acquisition of user behaviors and the corresponding adaptations.

Visual attention is a useful type of user behaviors in VR. Therefore, research on its mechanism and prediction becomes meaningful in the area of context-aware interaction. Current research on the visual attention mechanism focuses on visual saliency detection in images (Cerf et al., 2008; Judd et al., 2009; Jian et al., 2011; Lang et al., 2012; Mathe

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61802341), the National Science and Technology Innovation 2030 Major Project of the Ministry of Science and Technology of China (No. 2018AAA0100703), the Research Innovation Plan of the Ministry of Education of China, and the Provincial Key Research and Development Plan of Zhejiang Province, China (No. 2019C03137)

A preliminary version was presented at the 17th International Conference on Virtual-Reality Continuum and Its Applications in Industry, November 14–16, 2019, Australia

ORCID: Yunzhan ZHOU, https://orcid.org/0000-0003-1676-0015; Tian FENG, https://orcid.org/0000-0001-9691-3266

and Sminchisescu, 2012; Zhao and Koch, 2012; Xu et al., 2015; Zhu et al., 2015; Kruthiventi et al., 2017) and videos (Engelke et al., 2010; Riche et al., 2013; Fang et al., 2016; Fu et al., 2017), but is limited to 2D cases. In addition, prediction of "when" and "what" users will notice in a 3D virtual environment, such as a virtual museum, remains unclear. We have also observed that the related datasets were not labeled regarding timestamps, and could hardly represent sequential behaviors or support context-aware interaction in a 3D virtual environment.

In this study, we present a 3D Eye-tracking Dataset for Visual Attention modeling in a virtual Museum, named EDVAM. The EDVAM includes 9 604 480 visual attention records from users during their navigation. We divide these records into two subsets: the raw subset holds the captured eye movement sequences and the practical subset comprises the processed samples. To build the EDVAM, we use a novel approach to achieve gaze-based 3D interaction, which enables user interaction with virtual objects and acquires visual attention records from real-time eye movements. To our knowledge, the EDVAM is the first 3D eye-tracking dataset in a virtual environment. To illustrate its potential contribution to research on visual attention, we devise a deep learning model to predict a user's visual attention in the next moment from previous records. Trained on our dataset, this model provides a benchmark and an approach to context-aware interaction (e.g., displaying interfaces on the next region of space that a user would view).

We summarize the contributions of this study as follows:

1. Construct the first 3D eye-tracking dataset in a virtual museum, with a focus on visual attention modeling;

2. Design a deep learning model to predict user visual attention in the next moment.

This paper is an extension of the work originally presented in Zhou et al. (2019).

2 Related work

Our study relates to topics concerning virtual museums, user experiences in VR, visual attention, and eye-tracking datasets. Selected studies are discussed and compared to ours.

2.1 Virtual museums

Museums present artwork and exhibits to the public and are learning hubs that provide rich interaction experiences, which are now usually integrated with information technologies. For example, a virtual museum, augmented by personal digital assistants (PDAs), provides an intuitive artwork information guide, with which participants retrieve knowledge related to geographic locations (Hou HT et al., 2014). Augmented reality (AR) enables a conventional museum to support direct interactions with exhibits and their augmented images, promoting engagement with content about cultural heritage (Ciolfi et al., 2015). A projector-based virtual museum builds a large-scale museum with a 120° field of view (FoV) for an extraordinary immersive environment (Carrozzino and Bergamasco, 2010; Koskenranta et al., 2013). Recent advances in HMDs represent the entire real world in VR with finely reproduced artworks and an enhanced sense of immersion (Beer, 2015; Barbieri et al., 2018).

2.2 User experiences in VR

A sense of presence can improve the user experience in VR and enable users to feel like they are in the real world. This requires the mechanism of human perception and its implementation in VR.

Recent studies have employed haptics to solve the problem as mentioned above. Azmandian et al. (2016) proposed a method to warp a virtual environment to match a physical device's location in the user's surrounding for haptic feedbacks. Lopes et al. (2017) used electrical muscle stimulation to provide haptic feedbacks. Hand tracking and motion tracking have also received attention in the field. Hirota and Tagawa (2016) implemented a hand-tracking method using manipulation with a deformable hand, and Davis et al. (2016) accomplished a similar task using 3D gesture recognition. Motion-tracking methods focus on user movements with improved ease and accuracy (Suma et al., 2015; Nielsen et al., 2016).

These studies enhanced the sense of presence in VR for a better user experience. However, they paid less attention to user needs, on which this study focuses for a user-adapted interactive experience in a virtual museum.

2.3 Visual attention

Mechanisms of visual attention can be categorized into two types: bottom-up and top-down (Connor et al., 2004). The bottom-up mechanism depends on raw sensory input and rapid and involuntary attention shifts to salient visual features of potential importance. For instance, salient stimuli popping up in the surroundings could attract users (Itti, 2000). Previous bottom-up models focused on saliency detection (Itti et al., 1998; Shokoufandeh et al., 1999; Kadir and Brady, 2001; Hou XD and Zhang, 2007). Itti et al. (1998) proposed a classical model whose architecture mimics the properties of primate early vision, combining multi-scale image features into a single topographical saliency map. The following models detect salient regions from the perspective of multiple scales (Shokoufandeh et al., 1999; Kadir and Brady, 2001). Recently, Hou XD and Zhang (2007) proposed a fast method of constructing saliency maps based on the image's spectral residual, outperforming Itti et al. (1998)'s method.

In contrast, the top-down mechanism concentrates on long-term human cognitive strategies and bias attention toward particular objects in a specific situation (e.g., colored spots when hungry, sudden movements when afraid of predators) (Connor et al., 2004). Cerf et al. (2008) presented a significant combined model of face detection and low-level saliency. Some later models detect salient regions using classifiers on an eye-tracking dataset (Judd et al., 2009; Zhao and Koch, 2012).

These 2D-oriented methods do not perfectly suit 3D tasks, in which saliency detection relates to the temporal aspect of the visual image and becomes more complicated. Understanding temporal visual attention requires collection of sequential eye-tracking data. Specifically, context-aware interaction would be possible if the user's visual attention in the next moment could be predicted based on the analysis of collected eye-tracking data.

2.4 Eye-tracking datasets

An eye-tracking dataset maps collected data (i.e., features) to targets (i.e., labels), that are feature-label pairs for research on saliency models. Its related research has become active in recent years.

The existing eye-tracking datasets can be summarized into image datasets and video datasets. The majority include lightly compressed data, whereas only a few are made up of uncompressed data (Winkler and Subramanian, 2013). No existing image dataset comprises more than 40 participants (Bruce and Tsotsos, 2006; Ehinger et al., 2009; Liu and Heynderickx, 2009; Ramanathan et al., 2010; Kootstra et al., 2011) and no video dataset includes more than 55 participants (Itti, 2004; Carmi and Itti, 2006; Alers et al., 2012; Hadizadeh et al., 2012). In comparison, the EDVAM involves the largest group of 63 participants.

Existing datasets, including the recent ones with a 360° feature (Lo et al., 2017; Rai et al., 2017; David et al., 2018; Sitzmann et al., 2018), are still limited to 2D cases. Regardless of the number of 3D objects being viewed, the recorded eye-tracking data would be mapped onto a 2D plane, remaining at the image level and regarded as a 2D projection of 3D objects. In addition, participants are not allowed to freely move, observe, or gaze at objects from different angles and positions during the creation of these 2D datasets. To fill these gaps, we propose the first 3D eye-tracking dataset, including real-time visual attention records in a virtual museum.

3 EDVAM

To build the EDVAM, we collected the data in the context of a virtual museum supported by HMDs (Fig. 1). We assumed that the virtual museum was holding an exhibition focusing on an antique ceramic bowl placed at the center, together with other related exhibits near the surrounding walls. We designed several user interfaces (UIs) involving texts, images, and video clips to describe the ceramic bowl. A user wearing a VR headset navigated inside the virtual museum and freely viewed exhibits during a visit. At the same time, we recorded the real-time eye movements of the user and the exhibits viewed. For details about the virtual museum, we suggest that interested readers refer to Sun et al. (2018).

We divided the collected data into two subsets. The raw subset included the captured sequences of eye movements with 44 attributes as their features. The practical subset comprised 145 370 items sampled from the raw subset. Each item was derived from a fixed-length eye movement sequence and was given an extra label compared to the one in the raw subset. We formatted

both in CSV files to ensure data accessibility and compatibility. The dataset is publicly available (https://github.com/YunzhanZHOU/EDVAM).

3.1 Participants and devices

Sixty-three participants studying at a university in East China were recruited for data collection, including 26 female and 37 male (age: mean=23.44, standard deviation=1.81). We paid each participant $8 and required them to list any eye-related disabilities before the task. We also informed them that the task would not involve either violent or sexual content, and they would be able to exit whenever they felt nauseated.

The devices for data collection included an Oculus Rift DK2 VR headset for the participant's access to the virtual museum, a joystick for the participant's navigation, and a monitor displaying the real-time video streaming from the participant's view. To enable eye tracking on the VR headset, we attached a gadget (i.e., a Pupil Labs monocular add-on cup with an infrared (IR) mirror, IR LEDs, and an HD camera) to its left-hand display. In particular, the HD camera tracked the participant's point-of-gaze (PoG) with a tracking accuracy of less than 1°.

During the task, the VR headset and the eye-tracking gadget recorded the participant's eye movements and activities in the virtual museum. The navigation stage lasted for 3–5 min without any time constraint. The entire task finished within 10 min, including the preparation stage.

Fig. 1 Virtual museum used in this study (from top to bottom: overview, local-view, and top-view)

3.2 Gaze-based 3D interaction in VR

To enable gaze-based 3D interaction with virtual objects, we introduced a novel approach that maps 2D PoG positions to the corresponding 3D positions in VR, which contributed to obtaining real-time eye movements for recording visual attention. Fig. 2 illustrates the approach.

Our approach took as input an image that describes the user's eye movement captured by the eye-tracking gadget at a sampling rate of 30 Hz. The image included a set of eye-movement parameters $[t, C, P]$, where t denotes the elapsed time since the system's last restart, C denotes the confidence in $[0, 1]$ (i.e., 1 equals 100% confidence) of a comprehensive analysis at the level of image processing, and P denotes the PoG data at time t. These parameters were calibrated and matched to the VR plane α to recognize the corresponding 2D PoG position $[t, C, (x_\alpha, y_\alpha)_{t,C}]$ on α with C confidence at time t. However, this was insufficient for our goal of gaze-based 3D interaction in VR, so we conducted spatial mapping to obtain the exact 3D PoG position $[t, C, (x_s, y_s, z_s)_{t,C}]$ being observed in the VR space s with C confidence at time t.

In our approach, the spatial mapping step was based on the ray-object intersection of the ray casting algorithm (Roth, 1982), which was the methodological basis for 3D modeling and 2D image rendering, as shown in Fig. 3. Given the camera's position O, we first obtained the 3D coordinates of the 2D PoG position $A(x_\alpha, y_\alpha)$ in the VR world space s, denoted by $A'(x_s, y_s, z_s)$, using Unity's method Camera.ViewportToWorldPoint (Unity Technologies, 2019). This step was necessary

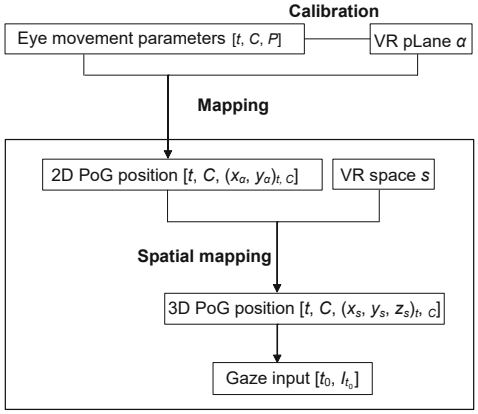

Fig. 2 Pipeline of gaze-based 3D interaction

because the viewport space and the VR world space shared different coordinate systems, and the clipping volume captured by the camera changed in a real-time manner. A ray was then cast from O and through A' to find its first intersection with an object in the VR world space, which was regarded as the 3D PoG position $B(x_s, y_s, z_s)$.

The approach's last step triggered a gaze input $[t_0, I_{t_0}]$ from a series of 3D PoG positions:

$$[t_i, C_i, (x_s, y_s, z_s)_i],\ t_i \in (0, d],\ C_i > C_0, \quad (1)$$

where t_i denotes the timestamp, C_i represents the confidence, $(x_s, y_s, z_s)_i$ denotes the 3D coordinates in the VR world space s (the subscript i denotes the number of each position in the series of 3D PoG positions), d denotes a duration upper-bound of 2 s, and C_0 denotes a confidence threshold of 0.3, subject to

$$(x_s, y_s, z_s)_i \in U, \quad (2)$$

and an operation I_t at time t in an interaction area U. Fig. 4 illustrates this step from the user's perspective. In our virtual museum, the user can interact with objects via the gaze cursor by keeping the eyes on an interaction area for 2 s.

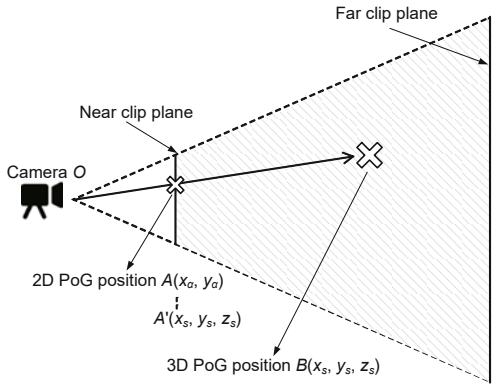

Fig. 3 Mapping a 2D PoG position to the corresponding 3D PoG position in the VR space

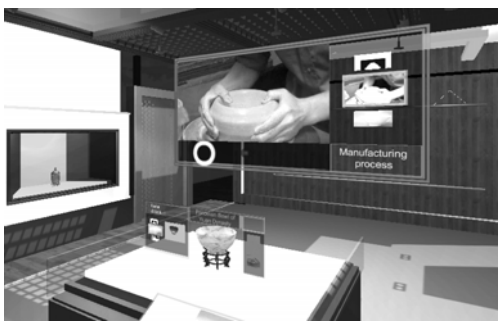

Fig. 4 Gaze cursor presented as a circle at the center

3.3 Task procedure

At the preparation stage, we asked the participants to watch a 3-min introductory video clip about the virtual museum's activities. Thereafter, they proceeded to a demo scene and familiarized themselves with the VR headset and the joystick for 3 min. Each of them put on and fixed the VR headset, and looked at nine white dots sequentially displayed on the VR screen, fixing the relative positions of the participant's head and the VR headset to ensure the mapping accuracy.

At the navigation stage, a participant could interact with virtual objects in several ways. For example, she/he might walk through each corner in the virtual museum using the joystick while controlling the speed. She/He was able to use gazes and head movements to support the joystick-based navigation, as if in a real museum. It was also possible to use the gaze cursor to interact with the UIs, as shown in Fig. 4. When the participant gazed at a UI, a circle appeared with a progress bar, triggering an input operation after viewing the interaction area for 2 s.

We allowed participants to choose their own navigation paths, freely interacting with the exhibits, and encouraged them to create unique choices for diversity in the collected data. During the task, we recorded the eye-tracking movement at a frequency of 30 Hz. In addition, we interviewed participants about their experience and recorded their feedback.

3.4 Data collection

As shown in Table 1, we collected two types of data in the task: (1) the VR gaze data, via both the VR device and the eye-tracking gadget, referring to the 3D eye movement sequences with 11 features; (2) the pupil data, containing the 2D gaze information with 33 features.

The VR gaze data recorded participants' spatiotemporal activities in the 3D VR space. As

Table 1 Types of collected data

Type	Number of features	Frequency	Device
VR gaze	11	$>$ 30 Hz	VR headset, eye-tracking gadget
Pupil	33	30 Hz	Eye-tracking gadget

shown in Table 2, the timestamp feature represents the temporal dimension with a precision of 0.001 s. Three-dimensional PoG position features indicate the place that the participant is observing in the virtual museum. The camera's position features show the participant's position, and its orientation features describe a participant's head orientation.

The pupil data were recorded in a normalized coordinate system that is irrelevant to the virtual environment. As shown in Table 3, we employed 30 feature channels concerning gaze and the pupil from the eye-tracking gadget (Pupil Labs, 2020). The pupil detector's measurement confidence was also used as a feature.

3.5 Raw subset

Due to the different timestamps, we merged the data items with similar timestamps and combined both types of data into the raw subset, including the eye-tracking data of 63 participants. Because of the sampling frequency used in data collection, 30 data items were produced per second. Each data item had 44 features with no labels, and each corresponded to a unique timestamp.

3.6 Practical subset

The aim of context-aware interaction requires learning from user behaviors and predicting visual attention in the next moment, which enables the system to adapt to users accordingly and synchronously. For example, a UI is displayed in real time near the next objects in which the user may be interested. To achieve context-aware interaction, we need not only eye movements but also subsequent visual attention. Therefore, the raw subset was further processed into the practical subset, which included the above two pieces of information.

Table 2 Gaze data details

Feature	Range	Precision	Unit
Index	$[1,\infty)$	1	–
Timestamp	$(-\infty,\infty)$	0.001	s
3D PoG position, x_s	$[-23.4,-9]$	0.1	m
3D PoG position, y_s	$[0,3.7]$	0.1	m
3D PoG position, z_s	$[-9.7,0.4]$	0.1	m
Camera's position, x_O	$[-23.4,-9]$	0.1	m
Camera's position, y_O	$[0,3.7]$	0.1	m
Camera's position, z_O	$[-9.7,0.4]$	0.1	m
Camera's orientation, x_f	$[-1,1]$	0.1	m
Camera's orientation, y_f	$[-1,1]$	0.1	m
Camera's orientation, z_f	$[-1,1]$	0.1	m

Table 3 Pupil data details

Feature	Number of channels
Timestamp	1
Index	1
Confidence	1
Gaze parameters	2
Pupil parameters	28

For the previous eye movements, we sampled the raw subset data using a time window of 10 s. Each adjacent time window was 1 frame apart from the next window at the frequency of 30 Hz. We regarded each time window as an input instance and constructed the instance matrix $\boldsymbol{s}\in\mathbb{R}^{n\times f}$ as

$$\begin{bmatrix} \cdots & \boldsymbol{I}_1 & \cdots \\ \cdots & \boldsymbol{I}_2 & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \boldsymbol{I}_n & \cdots \end{bmatrix}, \tag{3}$$

where $\boldsymbol{I}_i$ $(i=1,2,\ldots,n)$ denotes eye movements, n represents the number of eye movements in 10 s (i.e., $n=300$), and f refers to the number of features (i.e., $f=40$). An instance matrix indicates the previous eye movements in a fixed duration.

As for subsequent visual attention, we divided the VR space into 12 areas: upper interface, central interface, lower interface, south open space, north pillar, east pillar, southwest exhibit, northwest exhibit, southeast exhibit, piano area, central floor, and central ceiling. The following analysis was conducted to map user fixations in these areas to obtain the position of visual attention.

There are two reasons behind the division of the VR space: (1) Because the EDVAM is the first 3D eye-tracking dataset in a virtual museum, no previous study contributed to the visual attention prediction in a 3D VR space. Hence, we resorted to the most related work on 360° videos (Fan et al., 2017) and proposed that a fixation prediction network could predict the future viewing probability of each video tile. We extended the concept of tile and divided the VR space into various 3D tiles. (2) Each area demonstrates a candidate region and stands for the spatial scope of the adaptation in a context-aware environment.

The next question is about determining the area that locates the fixation, especially when the fixation is at the junction of two areas. We devised a solution

to this challenge. Because the fixation included PoGs that depend on its start time and duration, it was possible to calculate the number of PoGs in each area and determine the one to which the fixation most likely belongs as

$$\max_{n=12}(N_1, N_2, \ldots, N_n), \tag{4}$$

where N_i $(i = 1, 2, \ldots, n)$ denotes the number of PoGs in the i^{th} area. The area with the maximum number of PoGs was regarded as the one containing the fixation.

We observed that the number of PoGs was 0 in an area that was not estimated as one to which the fixation belonged in most cases. This observation supports the reliability of the fixation-based approach to visual attention in the next moment.

We built the practical subset by matching each time window to a visual attention area according to the timestamp, and subdivided it into a training set and a test set, ensuring that the samples were drawn from different participants.

4 Predictive deep learning model

To predict the visual attention and validate the collected dataset, we devised a three-layer long short-term memory (LSTM) network deep learning model, because LSTM has shown satisfactory performance in classification, processing, and prediction of temporal data (Gers et al., 2000; Eck and Schmidhuber, 2002; Chen et al., 2015).

4.1 Feature extraction

Each input instance from the practical subset was a time window with 40 features. To reduce the workload and increase the accuracy, we adopted the most relevant 10 features via pre-experimental analysis: 3D PoG positions (3 feature channels), camera's position (3 feature channels), camera's orientation (3 feature channels) from the VR gaze data, and confidence (1 feature channel) from the pupil data.

The influence of previous eye movements on subsequent visual attention was enabled by appending the last 10-s time window with increasing weights. We assigned the highest weight to the latest sub-window based on the assumption that the more recent the experience is, the more influential it may be on the current trial for selecting the target virtual object again in the UI task, as the mechanism of human memory works (Li et al., 2018). Table 4 demonstrates that one frame was sampled per six frames during the first 5 s, and that the sampling was terminated during the latest 0.33 s.

4.2 Model design

Previous research in deep learning proposed that adding more layers to a neural network (NN) would improve its capacity to yield complex behaviors (LeCun et al., 2015). Accordingly, we added two hidden layers to an LSTM network to model high-dimensional eye movements. At each step, the network took as input an eye movement item $\boldsymbol{I}_n$ with 10 features, as shown in Fig. 5. We used 3×78 LSTM cells in the recurrent layer. A hidden state of the current cell propagated to the next cells and its output propagated to deeper cells, because past eye movements might be captured in the user's current behavior input and the hidden state of the previous steps. We set the hidden dimension to 20. The model predicted the area to be viewed, given the linear transformation $\boldsymbol{L}$ of the output from the last LSTM cell.

4.3 Experiments

We experimented with our model on the practical subset of the EDVAM, and performed training

Table 4 Temporal dimension sampling

Timestamp	Number of frames	
	Before sampling	After sampling
$[-10.00, -5.00]$	150	25
$[-5.00, -2.00]$	90	18
$[-2.00, -0.33]$	50	25
$[-0.33, 0.00]$	10	10
Present	–	–

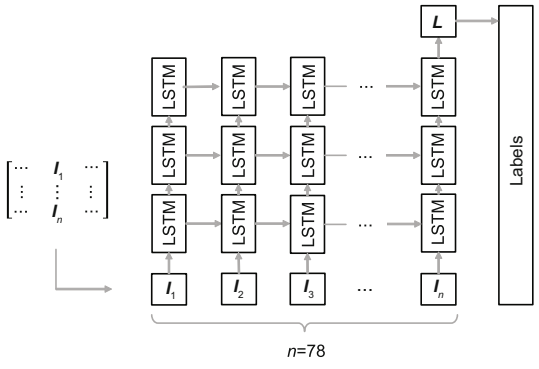

Fig. 5 Architecture of the three-layer LSTM network

and test tasks on a workstation with a 3.7-GHz CPU, 64-GB RAM, and an Nvidia Geforce GTX 1080 graphics card with 8-GB RAM. The model was trained using back-propagation on the training set and optimized via mini-batch gradient descent with a batch size of 128 for 40 epochs.

We computed a cross-entropy loss to measure the performance of the trained model on prediction as

$$\text{loss} = -\frac{1}{N}\sum_{i=1}^{N}\log_2\frac{\mathrm{e}^{f_{y_i}}}{\sum_j \mathrm{e}^{f_{y_j}}}, \tag{5}$$

where f_{y_j} denotes the j^{th} element of the label score vector, f_{y_i} refers to the scores of the correct labels, and N represents the batch size. The training was accomplished by minimizing the loss.

We validated the trained model on the test set. Each test instance comprised 10-s eye movements and the corresponding ground truth on visual attention. Fig. 6 illustrates the cross-entropy loss over 40 epochs in a declining trend. The loss converged rapidly in both training and validation. In particular, it underwent fluctuations from the 6^{th} to the 15^{th} epoch after the first five epochs. At this stage, the optimizer hovered around local minimums, looking for the global one. After 15 epochs, both training loss and validation loss converged to the global optimum. The prediction accuracy reached 78.94%, and predicting each sample required less than 0.02 ms, satisfying the real-time requirement.

In the example shown in Fig. 7, our model took as input the 3D PoG position sequences with confidences, camera position sequences, and camera orientation sequences, and then predicted that the next visual attention would be the southeast exhibit,

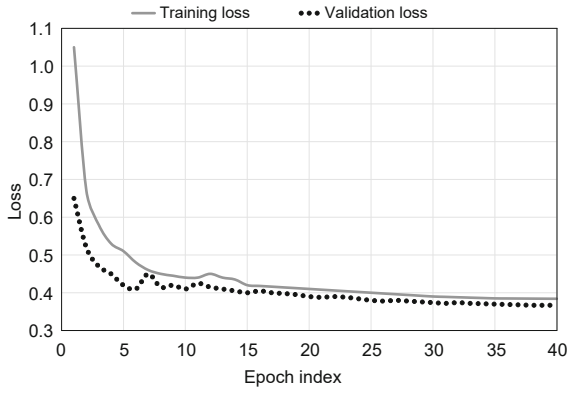

Fig. 6 Per-epoch loss when training and validating the LSTM network

based on the user's past behaviors around the piano area.

4.4 Analysis of gender effects

Although a previous study indicated that men and women devoted approximately the same amount of attention to a virtual environment (Felnhofer et al., 2012), we further explored whether the gender of a user would affect the prediction accuracy of our model. In particular, we trained it with data from either male users or female users in addition to the instance discussed in Section 4.3. Table 5 shows the performance of our model in the three cases. We observed no significant difference in the prediction accuracy regarding the user's gender in any instance. This can be interpreted as both female and male users having a similar visual attention pattern while navigating in a less gender-sensitive virtual environment like a museum.

5 Applications

This study, including the eye-tracking dataset and the predictive model, is expected to support a context-aware virtual museum environment with the eye-tracking dataset and the predictive model. Such an interaction system adapts itself based on the results of learning user behaviors. In particular, the adaptation mentioned in this study refers to the real-time and intelligent UI display near the objects that may interest users. Therefore, a potential research direction is to define the rules of the adaptive UIs and improve the model's prediction accuracy for context-aware interaction.

Saliency-aware rendering can also benefit from this study. An improved sense of immersion in VR demands higher HMD screen resolutions, whereas low-delay images can hardly support it. Previous studies implemented foveated rendering techniques for image synthesis with progressively fewer details outside the eye fixation region. With our predictive

Table 5 Prediction accuracy of the model regarding the user's gender

Model case	Accuracy (%)	
	Female	Male
As in Section 4.3	79.80	79.87
Trained on data from female users	77.99	77.15
Trained on data from male users	77.88	77.02

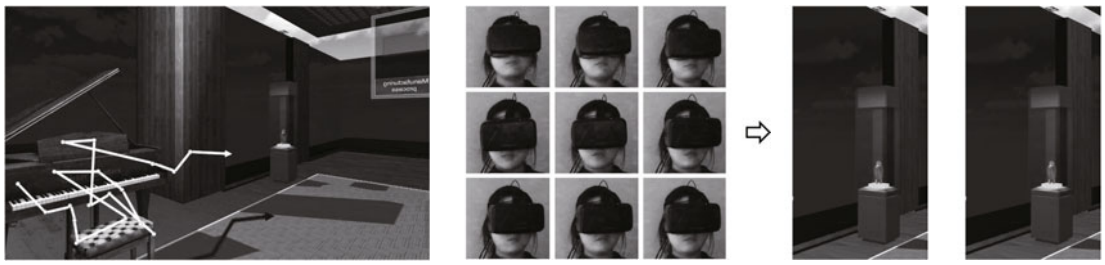

Fig. 7 An example of predicting the next visual area: (a) 3D PoG position sequences (white trajectories) and the camera's position sequences (black trajectory) visualized inside the virtual museum; (b) the camera's orientation sequences represented by the user's head motions; (c) the next visual area predicted by the model; (d) the corresponding ground truth

model for visual attention, it becomes possible to render the potential salient area and blur others in advance, reducing the workload of HMDs notably.

Researchers may exploit our dataset to explore the mechanism of 3D visual attention (e.g., identification and classification of eye movements in a virtual environment, the effects of eye movement features on prediction). Based on the learned knowledge about eye movements, we have confidence in making progress in context modeling, personalized interaction, and virtual museum design.

6 Conclusions and discussion

Previous visual attention studies and datasets have concerned us due to their 2D case limitations, inadequate freedom for users, and lack of consideration of temporal aspects, which are significantly different from the real world. We disagree that these studies are completely capable of enabling context-aware modeling in a 3D virtual environment.

In this paper, we introduced the EDVAM, the first 3D eye-tracking dataset in a virtual museum, to fill the gap, and proposed a predictive model for visual attention based on previous eye movements. Our model, based on the LSTM network, supports fundamental context-aware interactions in a 3D virtual museum. Overall, this study contributes to enabling a virtual museum's adaptiveness for a context-aware user experience. It helps users interact with virtual objects and adaptive UIs through a personalized virtual museum tour.

A significant limitation of this study lies in the devices used in the task. According to the participants' feedback, the Oculus Rift DK2 HMD has room for improvement in terms of precision and resolution. For example, some users complained about the coarse detail in the virtual museum caused by the HMD's low resolution. Although the participants interacted with virtual objects using a joystick and 3D PoGs in the task, the sense of presence still required other interactions (e.g., haptics and hand tracking). The use of VR HMDs with improved hardware and the introduction of multiple interaction methods can improve the 3D virtual environment and the quality of the collected data.

Currently, our model predicts only a limited number of visual areas. Alternatively, each visual area can be divided into more fine-grained subareas for training and improvement of our model's capability.

Despite the analysis showing no significant gender effect in the trained model instances, it is still worth investigating the generality of our approach (i.e., collecting eye-tracking data in different virtual museums) and the capability of our model to capture the potential individual differences in visual attention (i.e., conducting extended analysis with other user information including age, education, and cultural background).

We expect this study to serve as a reference for visual attention modeling and context-aware interaction in 3D virtual environments other than museums.

Contributors

Yunzhan ZHOU, Tian FENG, and Xiangdong LI designed the research. Yunzhan ZHOU and Shihui SHUAI processed the data. Yunzhan ZHOU and Tian FENG drafted the paper. Xiangdong LI, Lingyun SUN, and Henry Been-Lirn DUH helped organize the paper. Yunzhan ZHOU and

Tian FENG revised and finalized the paper.

Compliance with ethics guidelines

Yunzhan ZHOU, Tian FENG, Shihui SHUAI, Xiangdong LI, Lingyun SUN, and Henry Been-Lirn DUH declare that they have no conflict of interest.

Data availability

The dataset that supports the findings of this study is publicly available at https://github.com/YunzhanZHOU/EDVAM.

References

Alers H, Redi JA, Heynderickx I, 2012. Examining the effect of task on viewing behavior in videos using saliency maps. Proc SPIE 8291, Human Vision and Electronic Imaging XVII, p.82910X. https://doi.org/10.1117/12.907373

Azmandian M, Hancock M, Benko H, et al., 2016. Haptic retargeting: dynamic repurposing of passive haptics for enhanced virtual reality experiences. Proc CHI Conf on Human Factors in Computing Systems, p.1968-1979. https://doi.org/10.1145/2858036.2858226

Barbieri L, Bruno F, Muzzupappa M, 2018. User-centered design of a virtual reality exhibit for archaeological museums. *Int J Interact Des Manuf*, 12(2):561-571. https://doi.org/10.1007/s12008-017-0414-z

Beer S, 2015. Digital heritage museums and virtual museums. Proc Virtual Reality Int Conf, p.1-4. https://doi.org/10.1145/2806173.2806183

Bruce NDB, Tsotsos JK, 2006. Saliency based on information maximization. Proc 18th Int Conf on Neural Information Processing Systems, p.155-162.

Carmi R, Itti L, 2006. Visual causes versus correlates of attentional selection in dynamic scenes. *Vis Res*, 46(26):4333-4345. https://doi.org/10.1016/j.visres.2006.08.019

Carrozzino M, Bergamasco M, 2010. Beyond virtual museums: experiencing immersive virtual reality in real museums. *J Cult Herit*, 11(4):452-458. https://doi.org/10.1016/j.culher.2010.04.001

Cerf M, Harel J, Einhäeuser W, et al., 2008. Predicting human gaze using low-level saliency combined with face detection. Proc 20th Int Conf on Neural Information Processing Systems, p.241-248.

Chen K, Zhou Y, Dai FY, 2015. A LSTM-based method for stock returns prediction: a case study of China stock market. Proc IEEE Int Conf on Big Data, p.2823-2824. https://doi.org/10.1109/BigData.2015.7364089

Ciolfi L, Damala A, Hornecker E, et al., 2015. Cultural heritage communities: technologies and challenges. Proc 7th Int Conf on Communities and Technologies, p.149-152. https://doi.org/10.1145/2768545.2768560

Connor CE, Egeth HE, Yantis S, 2004. Visual attention: bottom-up versus top-down. *Curr Biol*, 14(19):R850-R852. https://doi.org/10.1016/j.cub.2004.09.041

David EJ, Gutiĺęrrez J, Coutrot A, et al., 2018. A dataset of head and eye movements for 360° videos. Proc 9th ACM Multimedia Systems Conf, p.432-437. https://doi.org/10.1145/3204949.3208139

Davis MM, Gabbard JL, Bowman DA, et al., 2016. Depth-based 3D gesture multi-level radial menu for virtual object manipulation. Proc IEEE Virtual Reality, p.169-170. https://doi.org/10.1109/VR.2016.7504707

de Jesus Oliveira VA, Nedel L, Maciel A, 2016. Speaking haptics: proactive haptic articulation for intercommunication in virtual environments. Proc IEEE Virtual Reality, p.251-252. https://doi.org/10.1109/VR.2016.7504748

Eck D, Schmidhuber J, 2002. Finding temporal structure in music: blues improvisation with LSTM recurrent networks. Proc 12th IEEE Workshop on Neural Networks for Signal Processing, p.747-756. https://doi.org/10.1109/NNSP.2002.1030094

Ehinger KA, Hidalgo-Sotelo B, Torralba A, et al., 2009. Modelling search for people in 900 scenes: a combined source model of eye guidance. *Vis Cogn*, 17(6-7):945-978. https://doi.org/10.1080/13506280902834720

Engelke U, Barkowsky M, Callet PL, et al., 2010. Modelling saliency awareness for objective video quality assessment. Proc 2nd Int Workshop on Quality of Multimedia Experience, p.212-217. https://doi.org/10.1109/QOMEX.2010.5516159

Fan CL, Lee J, Lo WC, et al., 2017. Fixation prediction for 360° video streaming in head-mounted virtual reality. Proc 27th Workshop on Network and Operating Systems Support for Digital Audio and Video, p.67-72. https://doi.org/10.1145/3083165.3083180

Fang YM, Zhang C, Li J, et al., 2016. Visual attention modeling for stereoscopic video. Proc IEEE Int Conf on Multimedia Expo Workshops, p.1-6. https://doi.org/10.1109/ICMEW.2016.7574768

Felnhofer A, Kothgassner OD, Beutl L, et al., 2012. Is virtual reality made for men only? Exploring gender differences. Proc Int Society for Presence Research Annual Conf, p.103-112.

Fu HZ, Xu D, Lin S, 2017. Object-based Multiple Foreground Segmentation in RGBD Video. *IEEE Trans Image Process*, 26(3):1418-1427. https://doi.org/10.1109/TIP.2017.2651369

Gers FA, Schmidhuber J, Cummins F, 2000. Learning to forget: continual prediction with LSTM. *Neur Comput*, 12(10):2451-2471. https://doi.org/10.1162/089976600300015015

Hadizadeh H, Enriquez MJ, Bajic IV, 2012. Eye-tracking database for a set of standard video sequences. *IEEE Trans Image Process*, 21(2):898-903. https://doi.org/10.1109/TIP.2011.2165292

Hirota K, Tagawa K, 2016. Interaction with virtual object using deformable hand. Proc IEEE Virtual Reality, p.49-56. https://doi.org/10.1109/VR.2016.7504687

Hou HT, Wu SY, Lin PC, et al., 2014. A blended mobile learning environment for museum learning. *Edu Technol Soc*, 17(2):207-218.

Hou XD, Zhang LQ, 2007. Saliency detection: a spectral residual approach. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1-8. https://doi.org/10.1109/CVPR.2007.383267

Itti L, 2000. Models of Bottom-Up and Top-Down Visual Attention. PhD Thesis, California Institute of Technology, Pasadena, USA.

Itti L, 2004. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans Image Process*, 13(10):1304-1318. https://doi.org/10.1109/TIP.2004.834657

Itti L, Koch C, Niebur E, 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Patt Anal Mach Intell*, 20(11):1254-1259. https://doi.org/10.1109/34.730558

Jian MW, Dong JY, Ma J, 2011. Image retrieval using wavelet-based salient regions. *Imag Sci J*, 59(4):219-231. https://doi.org/10.1179/136821910X12867873897355

Judd T, Ehinger K, Durand F, et al., 2009. Learning to predict where humans look. Proc IEEE 12th Int Conf on Computer Vision, p.2106-2113. https://doi.org/10.1109/ICCV.2009.5459462

Kadir T, Brady M, 2001. Saliency, scale and image description. *Int J Comput Vis*, 45(2):83-105. https://doi.org/10.1023/A:1012460413855

Kootstra G, de Boer B, Schomaker LRB, 2011. Predicting eye fixations on complex visual stimuli using local symmetry. *Cogn Comput*, 3(1):223-240. https://doi.org/10.1007/s12559-010-9089-5

Koskenranta O, Colley A, Häkkilä J, 2013. Portable CAVE using a mobile projector. Proc ACM Conf on Pervasive and Ubiquitous Computing Adjunct Publication, p.39-42. https://doi.org/10.1145/2494091.2494102

Kruthiventi SSS, Ayush K, Babu RV, 2017. DeepFix: a fully convolutional neural network for predicting human eye fixations. *IEEE Trans Image Process*, 26(9):4446-4456. https://doi.org/10.1109/TIP.2017.2710620

Lang CY, Nguyen TV, Katti H, et al., 2012. Depth matters: influence of depth cues on visual saliency. Proc 12th European Conf on Computer Vision, p.101-115. https://doi.org/10.1007/978-3-642-33709-3_8

LaViola JJJr, 2015. Context aware 3D gesture recognition for games and virtual reality. Proc ACM SIGGRAPH 2015 Courses, Article 10. https://doi.org/10.1145/2776880.2792711

LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. *Nature*, 521(7553):436-444. https://doi.org/10.1038/nature14539

Li Y, Bengio S, Bailly G, 2018. Predicting human performance in vertical menu selection using deep learning. Proc CHI Conf on Human Factors in Computing Systems, p.1-7. https://doi.org/10.1145/3173574.3173603

Liu HT, Heynderickx I, 2009. Studying the added value of visual attention in objective image quality metrics based on eye movement data. Proc 16th IEEE Int Conf on Image Processing, p.3097-3100. https://doi.org/10.1109/ICIP.2009.5414466

Lo WC, Fan CL, Lee J, et al., 2017. 360° video viewing dataset in head-mounted virtual reality. Proc 8th ACM on Multimedia System Conf, p.211-216. https://doi.org/10.1145/3083187.3083219

Lopes P, You SJ, Cheng LP, et al., 2017. Providing haptics to walls & heavy objects in virtual reality by means of electrical muscle stimulation. Proc CHI Conf on Human Factors in Computing Systems, p.1471-1482. https://doi.org/10.1145/3025453.3025600

Mathe S, Sminchisescu C, 2012. Dynamic eye movement datasets and learnt saliency models for visual action recognition. Proc 12th European Conf on Computer Vision, p.842-856. https://doi.org/10.1007/978-3-642-33709-3_60

Nielsen M, Toft C, Nilsson NC, et al., 2016. Evaluating two alternative walking in place interfaces for virtual reality gaming. Proc IEEE Virtual Reality, p.299-300. https://doi.org/10.1109/VR.2016.7504772

Pupil Labs, 2020. Pupil Labs Developer Documentation. https://docs.pupil-labs.com/developer/core/overview/ [Accessed on Sept. 27, 2020].

Rai Y, Gutiérrez J, Le Callet P, 2017. A dataset of head and eye movements for 360 degree images. Proc 8th ACM on Multimedia Systems Conf, p.205-210. https://doi.org/10.1145/3083187.3083218

Ramanathan S, Katti H, Sebe N, et al., 2010. An eye fixation database for saliency detection in images. Proc 11th European Conf on Computer Vision, p.30-43. https://doi.org/10.1007/978-3-642-15561-1_3

Riche N, Mancas M, Culibrk D, et al., 2013. Dynamic saliency models and human attention: a comparative study on videos. Proc 11th Asian Conf on Computer Vision, p.586-598. https://doi.org/10.1007/978-3-642-37431-9_45

Roth SD, 1982. Ray casting for modeling solids. *Comput Graph Image Process*, 18(2):109-144. https://doi.org/10.1016/0146-664X(82)90169-1

Shokoufandeh A, Marsic I, Dickinson SJ, 1999. View-based object recognition using saliency maps. *Image Vis Comput*, 17(5-6):445-460. https://doi.org/10.1016/S0262-8856(98)00124-3

Sitzmann V, Serrano A, Pavel A, et al., 2018. Saliency in VR: how do people explore virtual environments? *IEEE Trans Vis Comput Graph*, 24(4):1633-1642. https://doi.org/10.1109/TVCG.2018.2793599

Suma EA, Azmandian M, Grechkin T, et al., 2015. Making small spaces feel large: infinite walking in virtual reality. Proc ACM SIGGRAPH 2015 Emerging Technologies, p.16. https://doi.org/10.1145/2782782.2792496

Sun LY, Zhou YZ, Hansen P, et al., 2018. Cross-objects user interfaces for video interaction in virtual reality museum context. *Multimed Tools Appl*, 77(21):29013-29041. https://doi.org/10.1007/s11042-018-6091-5

Unity Technologies, 2019. Unity Documentation. https://docs.unity3d.com/ScriptReference/ [Accessed on Aug. 20, 2019].

Winkler S, Subramanian R, 2013. Overview of eye tracking datasets. Proc 5th Int Workshop on Quality of Multimedia Experience, p.212-217. https://doi.org/10.1109/QoMEX.2013.6603239

Xu PM, Ehinger KA, Zhang YD, et al., 2015. TurkerGaze: crowdsourcing saliency with webcam based eye tracking. https://arxiv.org/abs/1504.06755

Zhao Q, Koch C, 2012. Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost. *J Vis*, 12(6):22. https://doi.org/10.1167/12.6.22

Zhou YZ, Feng T, Shuai SH, et al., 2019. An eye-tracking dataset for visual attention modelling in a virtual museum context. Proc 17th Int Conf on Virtual-Reality Continuum and its Applications in Industry, Article 39. https://doi.org/10.1145/3359997.3365738

Zhu JY, Wu JJ, Xu Y, et al., 2015. Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE Trans Patt Anal Mach Intell*, 37(4):862-875. https://doi.org/10.1109/TPAMI.2014.2353617