EG-SNIK: A Free Viewing Egocentric Gaze Dataset and Its Applications

Sai Phani Kumar Malladi, Jayanta Mukherjee, Mohamed-Chaker Larabi, Santanu Chaudhury

▶ **To cite this version:**

Sai Phani Kumar Malladi, Jayanta Mukherjee, Mohamed-Chaker Larabi, Santanu Chaudhury. EG-SNIK: A Free Viewing Egocentric Gaze Dataset and Its Applications. IEEE Access, 2022, 10, pp.129626-129641. 10.1109/ACCESS.2022.3228484 . hal-04025589

HAL Id: hal-04025589
https://hal.science/hal-04025589

Submitted on 2 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 November 2022, accepted 6 December 2022, date of publication 12 December 2022,
date of current version 16 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3228484

RESEARCH ARTICLE

EG-SNIK: A Free Viewing Egocentric Gaze Dataset and Its Applications

SAI PHANI KUMAR MALLADI[1], (Graduate Student Member, IEEE), JAYANTA MUKHERJEE[2], (Senior Member, IEEE), MOHAMED-CHAKER LARABI[3], (Senior Member, IEEE), AND SANTANU CHAUDHURY[4]

[1]Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
[2]Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur 721302, India
[3]Xlim Umr Cnrs 7252, University of Poitiers, 86000 Poitiers, France
[4]Department of Computer Science and Engineering, IIT Jodhpur, Jodhpur 342030, India

Corresponding author: Sai Phani Kumar Malladi (saiphani.malladi@gmail.com).

This work was supported by the Institute Challenge Grants, IIT Kharagpur, through the project titled as "Visual Attention Assisted Image and Video Compression" under Grant IIT/SRIC/CS/VIV_ICG_2017_SGCIR/2018-19/085.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institute Ethical Committee, IIT Kharagpur.

ABSTRACT Egocentric vision data captures the first person perspective of a visual stimulus and helps study the gaze behavior in more natural contexts. In this work, we propose a new dataset collected in a free viewing style with an end-to-end data processing pipeline. A group of 25 participants provided their gaze information wearing Tobii Pro Glasses 2 set up at a museum. The gaze stream is post-processed for handling missing or incoherent information. The corresponding video stream is clipped into 20 videos corresponding to 20 museum exhibits and compensated for user's unwanted head movements. Based on the velocity of directional shifts of the eye, the I-VT algorithm classifies the eye movements into either fixations or saccades. Representative scanpaths are built by generalizing multiple viewers' gazing styles for all exhibits. Therefore, it is a dataset with both the individual gazing styles of many viewers and the generic trend followed by all of them towards a museum exhibit. The application of our dataset is demonstrated for characterizing the inherent gaze dynamics using state trajectory estimator based on ancestor sampling (STEAS) model in solving gaze data classification and retrieval problems. This dataset can also be used for addressing problems like segmentation, summarization using both conventional machine and deep learning approaches.

INDEX TERMS Egocentric gaze dataset, museum exhibit, head movement compensation, representative scanpath, categorization.

I. INTRODUCTION

In the last decade, mobile (wearable) eye trackers have become a popular research tool after a few pioneering research works [1], [2]. With the increasing popularity of wearable tracker in recording our life experience, egocentric vision [2], captured from a first-person perspective [3], is seen an emerging field in computer vision. Although saccadic models during visual search may demonstrate an influence of visual saliency [4], the effect is overwhelmed in the presence of an assigned task. In real world applications, remote or desktop eye trackers have limited reachability in certain aspects than that of wearable ones. Furthermore, target motion is often restricted to two dimensions (stimuli displayed on a screen), and sometimes viewed monocularly. The wearable eye trackers allow movement of participants. This helps in few studies, unlike remote eye trackers, such as understanding human cognitive processes and the dynamics of gaze shifts [5], classroom teaching [6] and many more.

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Most of the eye gaze datasets [7] are acquired using remote eye trackers with a few tasks for participants. There are a very few datasets acquired using head mounted eye trackers, but have not post-processed the gaze and data streams [8]. The egocentric gaze data are prone to sudden head movements, missing gaze information due to eye tracker issues (sudden illumination changes, dilated pupil, etc), eye blinks, etc. Hence, we address these issues with a newly proposed egocentric gaze dataset collected from multiple viewers in a free viewing style with an end-to-end data processing pipeline. Such handy and ready-to-use datasets are necessary for addressing many vision related problems like semantic segmentation, classification, summarization, etc, using both conventional machine learning and deep neural networks (DNNs). In this work, we aim to address the following issues when building a new egocentric dataset:

1) Users' distraction by the surroundings: unwanted vision distractions may occur suddenly and capture irrelevant information futile for interpretation,
2) Huge and unnecessary head movements: since eye movements are identified by their velocities and duration, head movements add their part to that of eyes misleading fixations and saccades [9],
3) Clear identification of fixations and saccades: generally, Tobii's I-VT filters [10] are used for identifying fixations and saccades. They are sensitive to velocity threshold and need to be specifically tuned for an eye tracker, stimuli or environment [10], and
4) Building a generalized viewing trend from multiple users: the raw eye tracking data is then processed into scanpaths [11]. A pattern that reflects the attention synchrony of different human subjects plays an important role in understanding how humans explore their surroundings [11].

Subsequently, we employ a state space model (SSM) based statistical learning approach that characterizes the gaze pattern of human beings observing different category stimuli. This is part of our previous findings introduced as the state trajectory estimator based on ancestor sampling (STEAS) model and described in [12]. In this work, we perform certain parameter tuning for adapting STEAS model to EG-SNIK dataset. Since the computations with long and raw velocity vectors are computationally intensive, we uniformly sample gaze velocity from the available raw velocity vectors. This is done because the computations with the latter are intensive. STEAS model generates multiple estimates of state velocity vectors which are the inherent temporal gaze features associated with the category of visual stimuli. We also utilize such features for applications like classification of gaze data and video retrieval. It is to be noted that the main focus of this paper is to build an EC free viewing gaze dataset and show its diverse applications in a few studies. The adaptation of STEAS model to this dataset is done by optimizing a few SSM parameters and then used for classification and retrieval problems.

II. LITERATURE REVIEW

A. GAZE DATASETS BY REMOTE AND WEARABLE EYE TRACKERS

From literature, eye gaze datasets are grouped into two categories based on the type of the device being used:

1) USING REMOTE EYE TRACKER

Greene et al. [13] published eye-tracking data of 17 observers performing 4 different tasks (memorize, determine when the picture was taken, estimate if people know each other, estimate their wealth) on 20 different grayscale images for 60 sec viewing time. Greene et al. reported that replicating findings in [14] is harder than expected. Similarly, Koehler et al. [15] published a dataset from 158 participants on 800 natural pictures. The effect of task differences on the ability of three models of saliency and the performance of humans on novel datasets have been investigated. Behavior on the popular free viewing task was not best predicted by standard saliency models. Instead, they accurately predicted the explicit salient regions and eye movements made. Frame et al. published an eye tracking dataset [16] collected in a consistently well-lit environment using 2 display setup. The surveillance screeners are termed analysts here. Each display has full HD resolution and analysts were instructed to identify specific essential elements of information (EEI) from surveillance data. This helps in understanding problems that analysts face.

For medical applications, Castner et al. [7] published a dataset collected using a remote eye tracker. 57 dental students along with 30 experts were employed for dental radiograph interpretation. They were capable of distinguishing experts from novices with 93% accuracy while incorporating the image semantics. Using various categories of videos as stimuli, Coutrot and Guyader [17] published one with 60 video sequences belonging to four visual categories: 1) people or faces, 2) one moving object, 3) several moving objects, and 4) landscapes. They explained the probability distribution of eye positions across each video frame using Expectation Maximization. Through experimental and modeling results, they show that participants look more at people talking. It is to be noted that remote eye trackers do not have provision for unconstrained gaze data collection unlike wearable devices.

2) USING WEARABLE DEVICES

In a task specific data, Frame et al. [16] published a dataset consisting of 17 sequences, performed by 14 different participants. They were asked to sit and prepare some food. They also equip participants with markers on hands which help in action identification. The beginning and ending time of the actions are manually annotated. The authors proposed a method to understand if the participant is baking, boiling, toasting, etc. Cherto et al. [8] published EgoMon egocentric dataset consisting of 7 videos of 34 minutes each, on average. The videos were recorded in the city of Dublin (Ireland) in both indoor and outdoor environments. The authors have used these videos for evaluating the performance of a state-of-the-art visual saliency prediction model.

Lee et al. [18] published the Univ. of Texas at Austin Egocentric (UT Ego) dataset with 4 videos captured using head-mounted cameras by 4 participants. This is a first of its kind since there is no associated gaze data for these videos. The videos capture a variety of activities such as eating, shopping, attending a lecture, and driving. They proposed a method for generating a short summary of the day of camera wearer. The summary attends the most salient objects and people with which the human being interacts. Both UT Ego and our dataset are collected without imposing any viewing constraint on participants, whereas the former does not use eye-tracking device and hence does not provide eye gaze information unlike ours.

B. MODELING HUMAN GAZE DYNAMICS

Reported works in computational visual attention mostly estimate gaze shifts and estimate the image scanpath [19]. A model for detecting saliency using natural (*SUN*) statistics following a Bayesian framework has been introduced in [20]. The self information of the random variable is observed from the bottom-up saliency representing the pixel visual feature in an image. Handerson et al. [21] proposed that eye movements through such features help in categorizing while performing four different tasks, namely, scene search, scene memorization, reading, and pseudo-reading. The cognitive states are found to get linked by the dynamic human eye movements [22]. Building gaze-driven dynamic models has been an important research topic for many years. Therefore, dynamic eye movements could be exploited for extracting features that help understand human intentions.

A few studies use hidden markov models (*HMM*s) to analyze eye-tracking data. In a recent study by Haji-Abolhassani et al. [23], HMMs are used to predict the visual tasks performed by participants viewing a painting. Although not using eye-tracking data, the work proposed in [24] uses a similar algorithm to analyze computer mouse movements in adults. Coutrot et al. [25] also used HMMs to study patterns of eye movements involved in face recognition. A rather different HMM-based approach for classic multiple object tracking paradigm is proposed by Citorik [26]. Each stimulus object has been assigned a separate HMM where two states represent if the object is being tracked or not.

All the above studies share several features that contrast with our work. First, the stimuli presented were static images. Even though Coutrot et al. [25] used conversational video stimuli, the regions of interest, were essentially stationary relative to the display. In contrast, our stimuli are real world objects, and so the parameters of our model evolve over time while studying the gaze dynamics using a statistical method. Second, we consider free viewing style of data collection unlike restricting head movements.

III. DATA ACQUISITION

A. DEVICE COMPONENTS

We have used Tobii Pro Glasses 2 [27] for acquiring the gaze data of participants. Tobii Pro Glasses 2 provides information about the user's attention in real time while freely moving around in indoor or outdoor environments. Mainly, it consists of three parts: head unit (glasses), recording unit and controller software

1) HEAD UNIT

Following are main parts of the head unit as shown in Fig. 3:

- HD scene camera: captures a full HD video of what is in front of the participant with 90 degree field of view, and
- Eye tracking sensors: record eye gaze positions and gaze directions through IR sensors. They also record head and body orientations through accelerometer and gyroscopic information sensors. IR sensors focus IR radiation onto human eyes and the portion of light that is reflected from the retina is absorbed by sensors for computing parameters like gaze direction, pupil dilation, gaze position, etc.

2) RECORDING UNIT

The recording unit controls the head unit by recording and storing the eye tracking data. It also stores the scene camera video and sounds on a removable SD memory card.

3) CONTROLLER SOFTWARE

It manages participants, live viewing, and instant replay of recordings.

B. TECHNICAL SPECIFICATIONS OF THE DEVICE

Head unit has a scene camera with HD video resolution at 25 frames per sec. Therefore, ideally, we get 2 gaze positions per video frame. However, due to eye blinks, eye tracker faults, illumination issues, etc., we may get less than 2 gaze positions per video frame. Tobii pro glasses 2 has 4 eye tracking sensors combined with gyroscope and accelerometer sensors. An eye tracker with 50 Hz sampling rate captures gaze positions which can be identified as fixations and saccades only [28].

C. DETAILS OF THE EXPERIMENT

We have selected Nehru Museum of Science and Technology (NMST), IIT Kharagpur as our study location. The viewers can interact with the museum exhibits to understand them better. Hence, participants do simple tasks using their hands. We have considered two galleries: 1) Mathematics, and 2) Basic physics galleries containing 20 exhibits as a whole, for our work. A participant observes all the exhibits in a sequence without any break. We have 25 students aged between 18 and 30 from IIT Kharagpur as participants. All of them are certified having a normal or corrected to normal vision by an ophthalmologist at the local IIT Kharagpur hospital. We name it Egocentric Gaze dataset of Students collected at Nehru museum of IIT Kharagpur (EG-SNIK). EG-SNIK dataset will be made publicly available after acceptance.

To record eye tracking data, the head unit must be fitted onto the participant's head (similar to a standard pair of glasses). Then, we calibrate the system separately for each

participant by asking them to look at a calibration card held in-front for a few seconds. We then start the recording from the controller software. After the session, we stop the recording and remove head unit from the participant. All the interactions with eye tracker like adding participants, initiating calibration, starting/stopping recordings etc. are done through the controller software. It is to be noted that, for using Tobii Pro Glasses 2 with human subjects, we obtained the ethical clearance[1] for this study from the Institute Ethical Committee at IIT Kharagpur. We have also obtained the required informed consent from the human subjects before their gaze data collection.

D. RECORDED GAZE DATA

Following are the prime differences of EG-SNIK over many other datasets:

- A free viewing style of data collection, which means, no task is instructed on participants,
- No restrictions imposed on head or body movements,
- The museum is not sealed off for the experiment, and other people are present in the museum, and
- Unlike other datasets, there is no imposed time limit.

1) GAZE DATA

Table 1 contains various labels of the parameters provided by eye tracker in the gaze data (shown in Fig. 3):

2) VIDEO FRAMES

Fig. 1 shows 20 exhibits of the museum with highlighted gaze positions (as dark spots) of a participant. These gaze positions are acquired over the period of time the user observes the exhibit. All the gaze positions corresponding to a particular exhibit are shown on a single frame for quicker reference. However, going by the operating frequency of the eye tracker (50 Hz) and the frame rate of the scene camera (25 frames per second), the number of gaze points available per video frame would be 2. Gaze position values obtained as normalized gaze coordinates are scaled up to HD resolution. We consider the gaze positions occurring for the whole duration of a participant watching an exhibit. Every exhibit is shown as a video frame where gaze positions are accumulated around the object of interest and its corresponding description. Also, some outliers occur due to distractions from surroundings in the form of sounds, people, illumination changes, etc,.

IV. DATA PRE-PROCESSING

A. FILTERING THE GAZE DATA STREAM

1) SYNCHRONIZATION OF THE VIDEO AND GAZE DATA STREAMS

When we start the data recording, the eye tracker clock starts ticking and capture the pupil. But, the scene camera switches on a bit later. Hence, practically, in any eye tracker, there exists some delay in between the eye tracker clock and the scene camera clock [10]. This results in asynchronous gaze and video data streams with some redundant gaze information before the scene camera is turned on. It is important to filter that unnecessary information to make the gaze and video data synchronous with each other.

From Fig. 2, the offset between *ts* and *vts* is used to synchronize the video and gaze data streams. The portion highlighted in red shows the occurrence of *vts* as 0 from where the video recording actually starts and video stream is available. We subtract this particular value of *ts* from all the *ts* values for synchronization. From this time stamp, a video frame changes after every 40 μs since the frame rate of video stream is 25 fps. Whereas, a 20 μs step size is for identifying a new gaze position. Within 40 μs, all the occurring gaze positions (ideally two) correspond to the same frame.

2) FILTERING THE MISSING GAZE DATA STREAM

One may notice that there are a few status indicators (*s*) with non-zero values mentioning that there is some problem with gaze acquisition at that particular *ts*. The non-zero *s* value also happens when one (or both) of the eyes are not tracked. Hence, we do not consider the corresponding *ts* data for further processing. However, we do interpolate these missing data points one by one [10]. This is done in three steps: 1) a scaling factor (sf) is computed as follows:

$$sf = \frac{t_{tSR} - t_{tFSG}}{t_{tLSG} - t_{tFSG}} \tag{1}$$

where t_{tSR}, t_{tFSG}, t_{tLSG}, and t_{tFSG} represent time stamp of first sample to be replaced, time stamp of first sample after gap, time stamp of last sample before gap, time stamp of first sample after gap, respectively. 2) *sf* is multiplied with the position data of the first valid sample of the gap, and 3) the result is added to the position data of the last valid sample before the gap. In this process, we cannot interpolate samples with large separation in time. Based on [29], we consider 75 ms as the maximum separation as it is shorter than a normal blink (150 $\pm$ 107 ms). It might prolong or shorten the duration of fixations and saccades due to the linear interpolation of samples. And, the measured velocity between two samples within the gap will be identical. If velocity is below threshold, all samples within the gap will be classified as fixation samples or otherwise.

B. FILTERING THE VIDEO STREAM

In this step, we initially separate out video clips depending on the museum object and then compensate the head movements using the sensory data from gyroscope and accelerometer.

1) EXHIBIT-BASED VIDEO CLIPPING

Since we have 20 exhibits in the museum, every participant provides gaze and video data streams for those 20 objects at a stretch. However, our STEAS model learns the gaze pattern for different video clips containing different exhibits. Hence, initially, we separate the whole video stream into 20 different clips for all the participants using the egocentric sub-shot

[1]Ethical clearance reference no: IIT/SRIC/DR/2019, dated November 6, 2019.

TABLE 1. Labels with their descriptions of the parameters in gaze data.

Label	Description
ts	Time stamp (in μ seconds) of the data stream (eye tracker data)
gidx	Gaze-index counter for all messages that belong to a single eye tracking event or gaze event
s	Status indicator for messages in the stream. Any non-zero value indicates some kind of problem with the data
pc	With origin at the scene camera, pupil center is specified in 3D coordinates in mm for each eye (left and right)
pd	Pupil diameter in mm for each eye (left and right)
gd	Gaze direction is a unit vector with origin at the pupil center
gp	Gaze position is the position on the scene camera image where the gaze will be projected. Top left corner is $(0, 0)$, bottom right corner is $(1, 1)$
gp3	3D gaze position in mm, relative to the scene camera where the gaze is focused
vts	Video timestamps are used as reference to synchronize the gaze data with the video
pv	Pipeline version that changes every time the pipeline is restarted for some reason
ac	Acceleration along X, Y, and Z axes of the glasses (or head) in m/sec^2
gy	Gyroscopic or rotation data of the glasses along (or head) along X, Y, and Z axes in $degree/sec$

representation technique detailed in [30]. It tailors a sub-shot segmentation approach to egocentric data and detects generic categories of ego-activity. Specifically, it predicts whether the camera wearer is watching an exhibit without undergoing significant body or head motion, observing various parts of an exhibit by slightly moving physically, or in transition while shifting attention from one exhibit to the other. We manually label the parts of 5 videos as belonging to 20 exhibits. We also label the rest of the video frames of those 5 videos as other two categories from the three above mentioned categories.

Each frame is represented by features extracted from the optical flow and blurriness for characterizing 20 museum exhibits, especially dense optical flow [31]. Then, the flow angles and their magnitudes are computed and organized into 8 bins. Then, a histogram of flow angles weighed by their magnitude is formed, concatenated with a histogram of magnitudes. A frame is divided into a 3×3 grid and each cell is given a score based on its blurriness for computing blur features [32]. We train one-vs.-rest SVM classifier for identifying 20 exhibits. For a given input video, class likelihoods for each frame are estimated using the classifier. Each frame is connected to its neighbors with a temporal window of 11 frames by smoothing labesl using a markov random field (MRF). Depending on the similarity of color histograms, consecutive frames receiving different labels are penalized. The resulting smoothed labels define sub-shots: a sequence of consecutive labels with same label belong to the same sub-shot. This results in identified shots for all the 20 exhibits of an egocentric video. Our EG-SNIK dataset is challenging for the following reasons:

1) Two successive exhibits can be divided into two categories: 1) nonadjacent exhibits: exhibit-irrelevant frames existed between two events, and 2) adjacent exhibits: two exhibits appear seamlessly.
2) Exhibits numbered (5) and (6) (see Fig. 1) contain a knob to rotate which need similar head and hand movements. Also, exhibits with numbers (15), (19), and (20) contain a button to be pressed for their operation.
3) Exhibit numbered (8) just needs human attention without any considerable movement.

With all the above mentioned details and complexities, the approach of sub-shot representation technique is seen to provide desired results for video clipping. Red bounding boxes (in Fig. 4) surrounding a set of frames notify that those frames correspond to a clip from the same event. In Figs. 4 (a), (b), and (c), video clips of exhibits numbered as (20), (15), and (19) in Fig. 1, respectively, are shown. Though they share a common action, they are successfully identified as different events because of two reasons: 1) they are not observed one after the other, and 2) participant's gaze pattern is different because of different objects in them. Fig. 4 (d) shows video frames with two exhibits ((2) and (3)) occurring seamlessly and middle frame shows the presence of the next exhibit without any break. Those two exhibits are successfully identified as different events where frames occurring in between them belong to exhibit (3). Fig. 4 (e) shows video frames with two exhibits ((5) and (6)) occurring with exhibit-irrelevant frames between them. The middle frame shows a switch board and bio-data of a scientist which are not in the list of 20 objects. Those two exhibits are identified as two different events successfully.

Fig. 4 (f) shows two different exhibits occurring with a similar action being identified as a single clip. The algorithm in [30] handles the first and third challenges mentioned

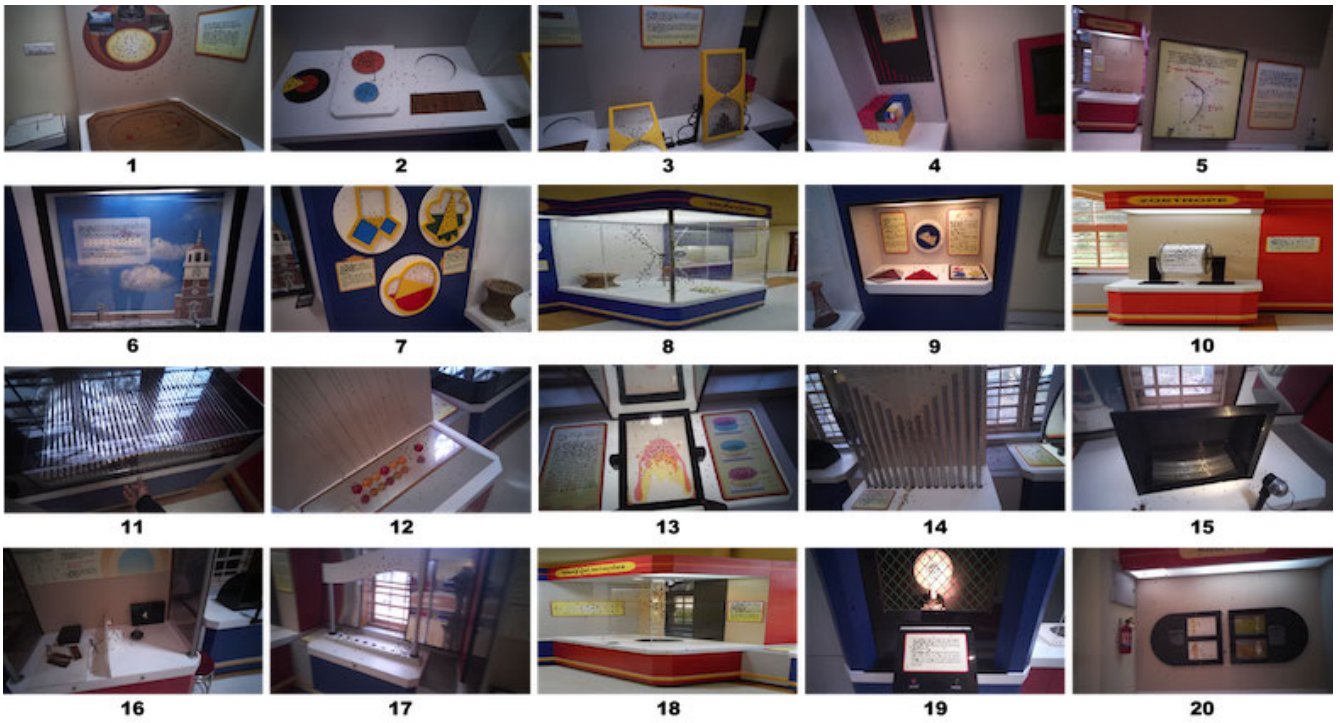


FIGURE 1. Exhibits viewed by participants with highlighted gaze points given in black dots: 1) elliptical carrom board, 2) area of a circle, 3) probability curve, 4) convergent series, 5) differentiation, 6) height and distance, 7) pythagoras theorem, 8) acrobatic stick, 9) traingle and polygons, 10) zeotrope, 11) wave motion, 12) newton's crade, 13) color subtraction, 14) musical pipes, 15) mechanical hologram, 16) mirror, prism, and lens, 17) music in the air, 18) nodes and anti-nodes, 19) stroboscope, and 20) reverse optics.


```
{"ts":1127350602,"s":0,"gidx":1690,"gp3":[572.43,328.47,1988.14]}
{"ts":1127410576,"s":0,"pts":2831719,"pv":1}
{"ts":1127410576,"s":0,"vts":0}
{"ts":1127370585,"s":1,"gidx":1691,"pc":[0.00,0.00,0.00],"eye":"left"}
```


FIGURE 2. Gaze data stream showing *vts* as 0 and its corresponding *ts* highlighted in red box.

above. However, when two different exhibits with similar motion occur seamlessly, it considers them as the one. Fig. 4 (f) shows frames of objects numbered (5) and (6) in Fig. 1, respectively. We find that detecting those two as two different events is a difficult task for the algorithm reported in [33]. This happens primarily for three reasons: 1) while operating the knobs for two exhibits, objects in them appear to move in a similar fashion, 2) action done by the participant's hand, is also similar for both the exhibits, and 3) since both of them are placed next to each other in the museum, continuity in participant's gaze pattern and action look similar for both the exhibits. This arises only when the participant watches exhibits (5) and (6) one after the other. Out of 25 participants, 16 have followed that order and hence, the clips corresponding to 20 exhibits are successfully extracted. However, the issue is with the other 9 clips which identify only 19 clips corresponding to 20 exhibits rather than 20 clips. Video frames corresponding to exhibits (5) and (6) are detected as a single clip which is then manually separated into two clips (Fig. 4 (f)). Since we obtain different video clips for 20 museum exhibits, we also partition the corresponding gaze data stream into 20 substreams for further processing.

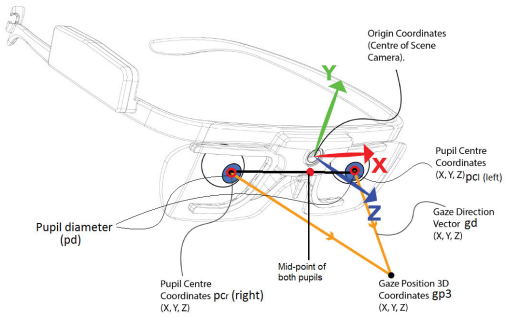


FIGURE 3. Reference coordinate system of Tobii pro glasses 2 [27].

2) COMPENSATION FOR HEAD MOVEMENTS

For head-mounted eye tracker (Tobii pro glasses 2), the eye movements are not just because of eyes but also due to head movements (shifts and rotations) [9]. The eye movement velocity or acceleration combined with that of the head movement results in a much higher value. This may mislead the correct identification of fixations and saccades. Hence, we compensate for the unwanted head movements for each of the video clips separated for 20 exhibits in the previous step.

a: USING GYROSCOPE SENSOR DATA [34]

Tobii pro glasses 2 gaze data stream also contains gyroscopic information along X, Y, and Z axes. The gyroscope records the head rotational velocities along the three axes.

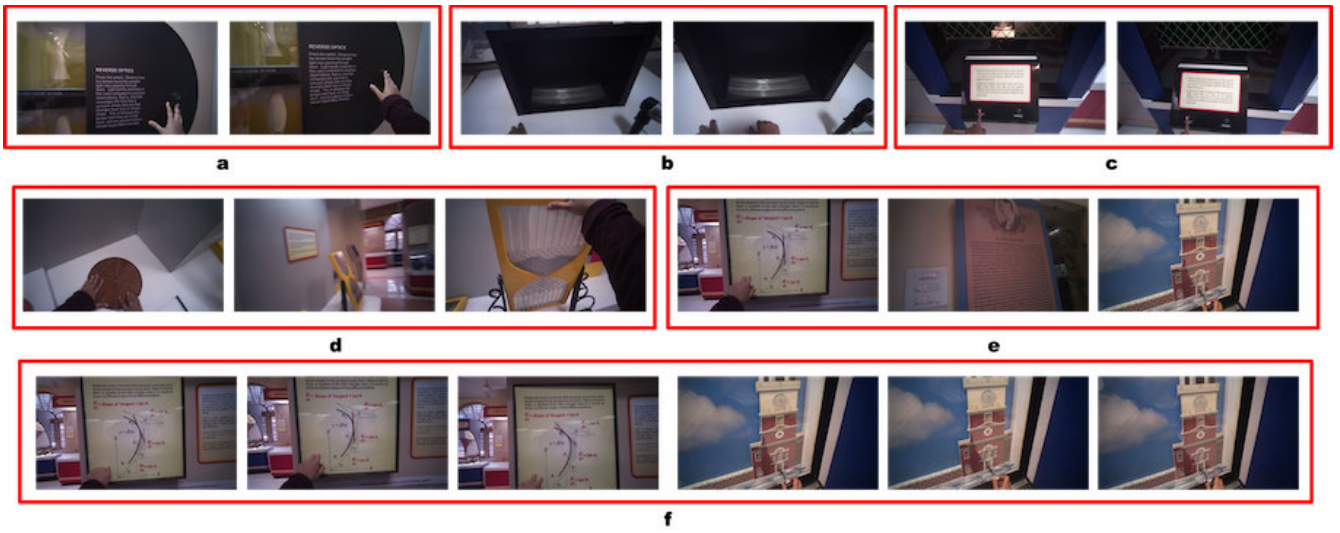


FIGURE 4. Results of exhbit based clipping using [33] for a few exhibits. Video frames of clips for exhibits a) (20), b) (15), c) (19) requiring similar action, d) two exhibits appear seamlessly, e) exhibit irrelevant frames between two events, f) exhibits (5) and (6) are in a single video clip.

The rotational velocities at time instant t found in the vectors $GyroX(t)$, $GyroY(t)$, and $GyroZ(t)$ are integrated over time to result in the corresponding Euler angles

$$\phi(t) = \int_{t-\delta t}^{t} GyroX(t)\,dt \tag{2}$$

$$\beta(t) = \int_{t-\delta t}^{t} GyroY(t)\,dt \tag{3}$$

$$\gamma(t) = \int_{t-\delta t}^{t} GyroZ(t)\,dt \tag{4}$$

where ϕ, β, and γ corresponds to pitch, yaw and roll rotations, respectively and $t - \delta t$ is the previous time instant. $GyroX(t)$, $GyroY(t)$, and $GyroZ(t)$ represent the gyroscopic sensor data (orientation) along X, Y, and Z axes, respectively. With the computed ϕ, β, and γ, a rotation matrix, R is calculated as:

$$R_G = R_G(\phi)R_G(\beta)R_G(\gamma) \tag{5}$$

where

$$R_G(\phi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos(\phi) & -sin(\phi) \\ 0 & sin(\phi) & cos(\phi) \end{pmatrix},$$

$$R_G(\beta) = \begin{pmatrix} cos(\beta) & 0 & sin(\beta) \\ 0 & 1 & 0 \\ -sin(\beta) & 0 & cos(\beta) \end{pmatrix}, \text{ and}$$

$$R_G(\gamma) = \begin{pmatrix} cos(\gamma) & -sin(\gamma) & 0 \\ sin(\gamma) & cos(\gamma) & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

b: USING ACCELEROMETER SENSOR DATA [35]

Tobii pro glasses 2 data stream also contains accelerometer information along X, Y, and Z axes. It records the head acceleration along the three axes. It is be noted that the accelerometers are affected by the linear acceleration and local gravitational field. They measure the difference between any linear acceleration in the accelerometer's reference frame and the earth's gravitational field vector [35]. The accelerometer measures the rotated gravitational field vector which can be used to determine the Euler angles, pitch and roll orientation, as follows:

$$\phi(t) = tan^{-1}\frac{AccY(t)}{AccZ(t)} \tag{6}$$

$$\gamma(t) = tan^{-1}\frac{-AccX(t)}{\sqrt{(AccY(t))^2 + (AccZ(t))^2}} \tag{7}$$

where $Acc(X(t))$, $Acc(Y(t))$, and $Acc(Z(t))$ represent the accelerometer sensor data along X, Y, and Z axes, respectively. The rotation matrix evaluated using the accelerometer is calculated as

$$R_A = R_A(\phi)R_A(\gamma) \tag{8}$$

where $R_A(\phi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos(\phi) & -sin(\phi) \\ 0 & sin(\phi) & cos(\phi) \end{pmatrix}$, and

$R_A(\gamma) = \begin{pmatrix} cos(\gamma) & -sin(\gamma) & 0 \\ sin(\gamma) & cos(\gamma) & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Furthermore, accelerometer signals do not contain information about the rotation around the vertical axis (yaw angle) and hence do not provide orientation information [36]. Hence, we need to fuse both sensor readings for compensating the head movements in egocentric videos.

c: FUSING GYROSCOPE AND ACCELEROMETER SENSOR DATA [37]

Accelerometer and gyroscope orientations are achieved by using rotational matrices. Since accelerometer and gyroscope sensors offer advantages by providing the motion and angular velocities respectively, a single estimate through the fusion of both sensors is estimated as follows:

$$R = R_A R_G \tag{9}$$

The 3D gaze information (gaze and eye positions) $a(t)$ at time instant t is multiplied with R to compensate for head movements. The resultant head compensated gaze position $a_{hc}(t)$ is given as

$$a_{hc}(t) = Rx(t)^T \tag{10}$$

We use bi-cubic interpolation to compute the output pixel value as a weighted average of pixels in the nearest 4-by-4 neighborhood in the input. These compensated gaze positions are subsequently used for computing the velocity of eye movements. Fig. 5 shows the results of head movement compensation using gyroscope information of Tobii pro glasses 2 data stream. Figs. 5 (a) and (c) are the input video clips with their corresponding head movement compensated videos being shown in (b) and (d) for two different exhibits. From Figs. 5 (a) and (b), we see that initial frames are compensated for clockwise head movements as all the selected human beings are right-handed [38]. They observe or perform actions from left to right which makes their heads turn clockwise while observing the museum exhibits also. Therefore, initial video frames undergo anti-clockwise rotation around the frame center (left zoomed portions in (b) and (d)). This gradually reduces in the middle frames since the participant observes the exhibit with a comparatively still head. This is because the initial and end frames focus on shifting their attention towards next exhibits. In the same way, we see that end frames are compensated for anti-clockwise head movements. This happens when the participant moves head towards an object placed on his right side. Therefore, end video frames undergo clockwise rotation around the frame center (right zoomed portions in (b) and (d)) for compensation. Hence, we successfully get all the 20 filtered and head movement compensated video clips corresponding to 20 different exhibits for all the participants. At the end, we obtain 25×20 (500) video clips and their corresponding gaze data streams.

C. IDENTIFICATION OF FIXATIONS AND SACCADES

After the data filtering, we use the I-VT filter [10] to calculate the visual angle and velocity using the 3D gaze and eye positions. The angular velocity calculation is done using the law of cosines [39]. For a sample at time instant t, the angle is calculated between it and the ones before and after at time t, t_1 and t_2 ($t_1 < t < t_2$). The velocity is calculated by dividing the angle between the sample at time t_1 and t_2 with the time between the two samples. To do this, we compute the vectors a, b, and c as

$$\vec{a}(t) = gp3(t_1) - pc(t_1) \tag{11}$$
$$\vec{b}(t) = gp3(t_2) - pc(t_2) \tag{12}$$
$$\vec{c}(t) = gp3(t_2) - gp3(t_1) \tag{13}$$

The vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$ are visualized in Fig. 6. As mentioned earlier, pupil centers for both eyes are available for every time instant which gives two gaze vectors for a single gaze sample. Hence, to obtain a single gaze vector considering both the eyes, we use a common mid-point for both the pupil centers $pc_l(t_1)$ and $pc_r(t_1)$ as $pc(t_1)$ which is given as

$$pc_x(t_1) = \frac{pc_{l,x}(t_1) + pc_{r,x}(t_1)}{2}, \tag{14}$$
$$pc_y(t_1) = \frac{pc_{l,y}(t_1) + pc_{r,y}(t_1)}{2}, \tag{15}$$
$$pc_z(t_1) = \frac{pc_{l,z}(t_1) + pc_{r,z}(t_1)}{2}, \tag{16}$$

where $pc_{(l/r,x/y/z)}(t_1)$ represents the 3D pupil center (left or right) for x, y, or z axis. Then, we use the law of cosines to compute angle α using

$$c^2 = a^2 + b^2 - 2 * a * b * cos(\alpha) \tag{17}$$

where $\vec{a}$, $\vec{b}$, and $\vec{c}$ are directional vectors as shown in Fig. 6. When the angle is known, we calculate the angular velocity $\vec{v}(t) = \Big(v_x(t), v_y(t), v_z(t)\Big)$ using the sampling time as follows:

$$\vec{v}(t) = \frac{|\vec{\alpha}(t)|}{|t_2 - t_1|}. \tag{18}$$

where $\vec{\alpha}(t) = \Big(\alpha_x(t), \alpha_y(t), \alpha_z(t)\Big)$. After velocity calculation, each eye movement is categorized as an event (saccade or fixation) by a velocity threshold value and shown as

$$Event(t) = A(\vec{v}(t)), \tag{19}$$

where A is the I-VT classification function [10], and *Event* contains 1 or 2 for fixation or saccade, respectively. The exception is for samples in the trailing or beginning of a recording where it has not been possible to calculate the velocity. Empirically, a velocity threshold of 90 degree per second [34] is found to achieve the best results for identifying fixations and saccades using Tobii pro glasses 2.

Then, we merge fixations located close to each other in space and time [10]. If a fixation does not belong to a set of consecutive fixation points with the minimum length of 60 ms, we discard that fixation. 60 ms is used because a fixation of that short duration is not meaningful for studying the user behavior due to the processing time of visual stimuli in the visual pathway and brain [10].

Figs. 7 and 9 show the scanpaths of only five participants for visualization. Different colored scanpaths correspond to different participants. Each circle corresponds to a fixation and its radius is proportional to the duration and the number of sample gaze points in it [11]. Larger the fixation duration, the higher the participant's attention towards that region. It can be seen that the parts of an exhibit touched by participants like knobs of exhibit (5 and 6), adjustable parts (2 and 13), and moving parts of an exhibit (11 and 19), obviously grab human attention for longer duration. Also, we observe that there is a similar pattern in viewing style of participants. For example, for exhibit 19 (stroboscope), participant reads the description first to operate it. Then, the light and fan of the stroboscope turns on to grab user's attention which forms a trend in viewing that exhibit. Hence, we categorize all the 20 exhibits into four different categories and a few scanpaths are shown in Figs. 7 and 9:

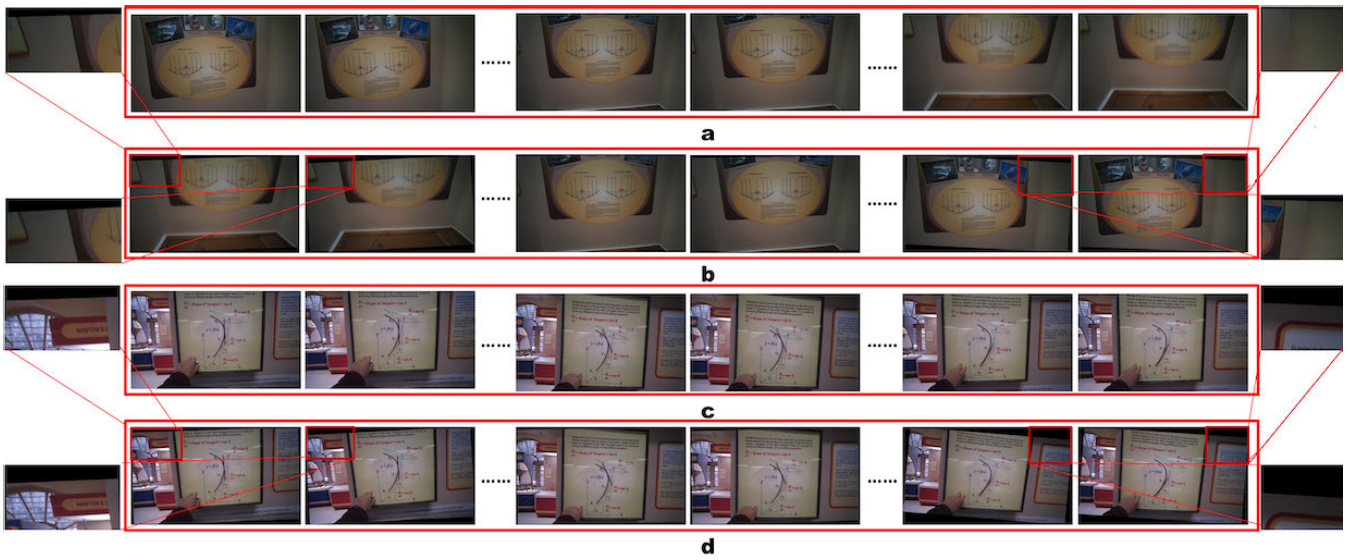


FIGURE 5. Results of head movement compensation using gyroscope information. (a), (c) input, and (b), (d) corresponding output video frames. The zoomed portions show the type of rotations happened after compensation.

1) Order of operation or observation may slightly vary from one to another: exhibits 2, 3, 4, 7, 9, and 13.
2) Observing or operating using a knob and viewing pattern remains almost similar: exhibits 5, and 6.
3) Observing and operating using buttons and viewing pattern remains almost similar: exhibits 11, 15, 16, 18, 19, and 20.
4) Observing and operating without any knob or button and viewing pattern remains almost similar: exhibits 1, 8, 10, 12, 14, and 17.

Out of these four categories, second, third and fourth categories are found to have similar scanpaths for all the participants. Whereas, the first category contains multiple objects to operate which makes participants follow different sequences. Thereby, the corresponding order of fixations vary slightly from a participant to the other.

D. REPRESENTATIVE SCANPATH IDENTIFICATION

We identified the gaze samples as either fixation or saccades and evaluated their scanpaths. At this stage, we identify a pattern in scanpaths considering all the participants gaze data for all museum exhibits separately. We call this *generalized viewing pattern* (*GVP*) of an exhibit as the representative scanpath. We follow a procedure similar to the one detailed in [11]. Though there are a few other works for this operation, we consider this since others have not discussed representative scanpaths using gaze duration analysis.

This procedure consists of three parts: 1) preprocessing of eye gaze data, 2) scanpath aggregation, and 3) gaze duration analysis. In order to identify the viewing pattern, we aim to exploit the fixation order, duration and position. The first step is divided into three substeps: 1) outlier removal, 2) area of interest (*AOI*) extraction, and 3) AOI center identification. The second step addresses the shape of scanpath

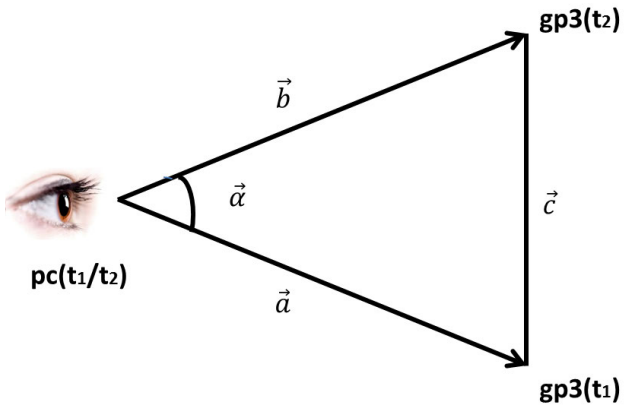


FIGURE 6. Angle calculation for $\alpha(t)$ between sample points, $gp3(t_1)$ and $gp3(t_2)$.

by aggregating multiple scanpaths into one. In the last step, we analyze such pattern in terms of gaze duration for obtaining the representative scanpath.

1) GAZE DATA PRE-PROCESSING

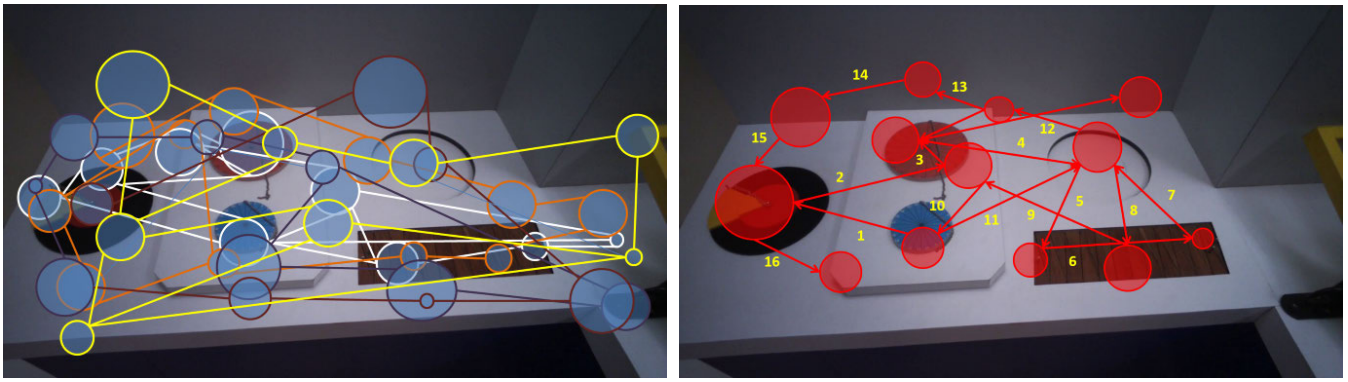
Following are the steps of gaze data pre-processing.

a: OUTLIER REMOVAL

The sequence of fixations, which reflect the actual viewing process, varies with different individuals even if the fixation distributions are similar. Hence, the position and order of the fixations cause inconsistency in scanpath. Dynamic time warping (DTW) [40] with boxplot is used to eliminate the influence of scanpaths on both the temporal order and spatial distribution.

b: AOI EXTRACTION

The fixations points are clustered by an algorithm in [41] by considering two point properties: local density ρ, and



(a) Exhibit 2 (b) Representative scanpath of (a)

FIGURE 7. **(a) Scanpaths of a few participants. Each color corresponds to a participant with a fixation radius proportional to its duration and number of gaze points in it, (b) The corresponding representative scanpath considering all the 25 participants' scanpaths. Red circles represent fixations with arrows as saccades and numbers indicate the order of their occurrence.**

distance from points with higher density δ. In order to determine the number of clusters, $\gamma = \rho \times \delta$ is computed and then sorted. Fixations with γ higher than the set threshold stick out with corresponding cluster number. A weighted geometric mean is computed for using it as threshold as follows:

$$threshold = \sqrt[\sum_{k=1}^{n}\alpha_k]{\prod_{k=1}^{n}\gamma_k^{\alpha_k}} \tag{20}$$

$$\alpha_k = 2^{logn - logk + 1} - 1 \tag{21}$$

where $\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_n$ have been sorted in decreasing order and n represents the number of clusters. Compared to the geometric mean, the weighted one emphasis more on the larger γ leading to fewer and lesser overlapped clusters.

c: CENTER IDENTIFICATION

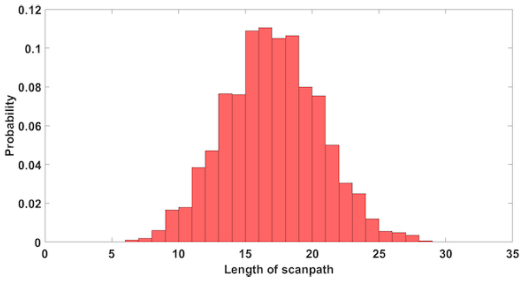
A random walk based method [42] is employed to identify AOI centers rather than simply averaging coordinates with large γ. The former method aims to generate a coefficient f for each fixation in the AOI, and computes the weighted center as the final AOI one. The following equation is used for updating the coefficient f:

$$f_{t+1}(i) = \frac{1}{\eta}\left(\sum_{j=1}^{n}(1-(1-\alpha))f_t(j)r(j,i)\right) + (1-\alpha)f_t(i)u(i) \tag{22}$$

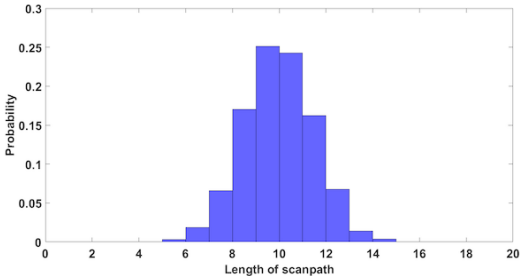
where $u(i)$ is the initial coefficient of fixation i defined by ρ, η is the normalizing parameter, $r(j,i)$ is the transition probability from fixation i to j.

$$r(j,i) = \frac{e^{-\sigma \times D(j,i)}}{\sum_{k=1}^{n} e^{-\sigma \times D(j,i)}} \tag{23}$$

where $D(j,i)$ is the Euclidean distance from fixation j to i, σ subtly influences the distribution of the center. AOI centers are also regarded as indicators of AOI distribution. Later, in the aggregation stage, candidate components of the representative scanpath are the AOIs with identified centers.



(a) Distribution of the lengths of original scanpaths



(b) Distribution of the lengths of representative scanpaths

FIGURE 8. **Distibutions of scanpath lengths.**

2) SCANPATH AGGREGATION

At this stage, we compute the barycenter [43] of all the available scanpaths, to assemble multiple scanpaths into a single one. That means, one that is closest to the individual scanpaths based on DTW distance is the representative scanpath. This is represented as follows:

$$r = \underset{s'}{argmin} \sum_{s \in sps} Dist(s', s) \tag{24}$$

where r is the representative scanpath, s' is the one to become the representative scanpath, s is an individual scanpath in the given set sps, and $Dist$ is a function calculating the distance or dissimilarity between two scanpaths. Basically, scanpaths are sequences of components with coordinates. Consider two scanpaths, $P = P_m = \langle p_1, p_2, p_3, .., p_m \rangle$, and $Q = Q_n = \langle q_1, q_2, q_3, .., q_n \rangle$, DTW is recursively

computed as:

$$DTW(P_i, Q_j) = \delta(p_i, q_j) + min \begin{cases} DTW & (P_{i-1}, Q_{j-1}) \\ DTW & (P_{i-1}, Q_j) \\ DTW & (P_i, Q_{j-1}) \end{cases} \quad (25)$$

where P_i, Q_j are subsequences of P and Q, p_i and q_j are components of scanpaths P and Q, respectively and $\delta()$ is the Euclidean distance function. The dissimilarity or distance between scanpath P and Q is:

$$Dist(P, Q) = DTW(P_m, Q_n) \quad (26)$$

In this work, Candidate-constrained DTW Barycenter Averaging (CDBA) algorithm is used for scanpath aggregation.

a: CDBA ALGORITHM

Initially, CDBA algorithm builds a set of candidate AOIs for each AOI. For a certain AOI, this set contains all the valid subsequent AOIs. In other words, candidate set AOIs of AOI_i is supposed to follow AOI_i in one individual scanpath at minimum. We extend scanpaths of 1 fixation to scanpaths of n fixations and enumerate all the scanpaths. A scanpath is extended by choosing an AOI from the candidate set of the last AOI on the scanpath and adding it to the end. If a certain AOI occurs maximum in the individual scanpaths, we remove that AOI and do not consider in later enumerated scanpaths. Then it defines an initial average scanpath as the reference scanpath and then updates the reference scanpath iteratively. Each iteration of CDBA contains two steps: (1) DTW between the reference scanpath and every individual scanpath is computed, and (2) reference scanpath components are updated. We repeat these steps until the reference scanpath stays without any changes.

3) GAZE DURATION ANALYSIS

The aggregated scanpath tells us both the areas drawing our attention and also their priority of attraction. Each scanpath of fixations is transformed into an AOI sequence of clusters to notify the attention duration of an AOI. Then, we statistically analyze the gaze duration of each AOI for all the individual scanpaths. In the aggregated scanpath, gaze duration of each AOI is obtained by averaging that of the same AOI in the individual scanpaths. Note that while analyzing the AOI duration, one and the same AOI appearing more than once in a sequence is regarded as different AOIs. Then, they are distinguished by their appearing order in the sequence.

Fig. 9 shows representative scanpaths for a few exhibits in second and fourth columns. The red circles, connecting arrows, and their numbers indicate fixations, directions of eye movements, and the order of fixations, respectively. Figs. 9 (b) and (d) contain multiple objects in an exhibit and hence the scanpaths are dispersed while viewing. Also, the participant may start with any object in the exhibit and thereby the scanpath contains fixations with multiple visits for a few fixations back and forth. The exhibits in Figs. 9 (f), (h), (i) and (l) contain objects to operate and participants go for them without much effort and thought. For these exhibits, going by the order of fixations, participants are found to go through the description first, then operate the object. Also, we find that a few outlier fixations are eliminated which occur due to participant's vision distraction because of ambience disturbances. In order to verify the effectiveness of this approach of scanpath aggregation, we compare and analyse the distribution of the fixation based scanpaths with that of the representative scanpath in the subsequent section.

a: ANALYSIS OF SCANPATH LENGTH

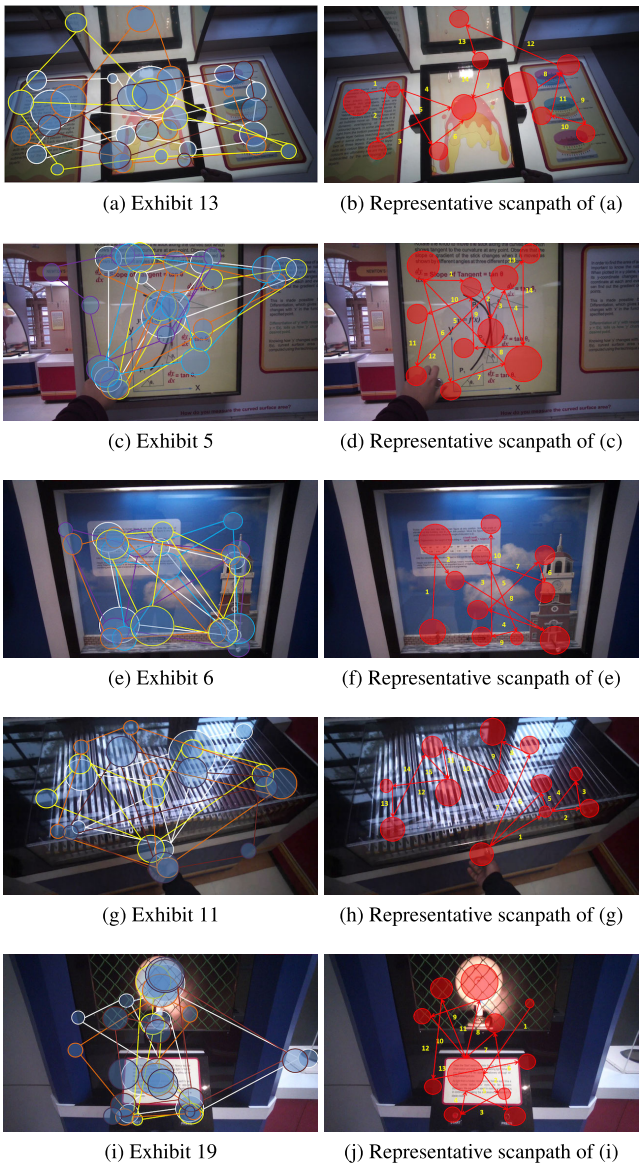
It is observed that the frequency of attention shift is reflected by the scanpath length. Fig. 8a shows that the normal distribution is followed by the length distributions of individual scanpaths. Thus, the representative scanpaths are supposed to follow the bell-shaped property, which we observe happening from Fig. 8b. Though the absolute values of those two scanpath lengths may be different, their bell-shaped property of length distribution is preserved for our data. Therefore, the representative scanpath pattern reflects the group trend and is visually consistent with individual scanpaths.

V. STATE TRAJECTORY ESTIMATOR BASED ON ANCESTOR SAMPLING (STEAS)

In this section, we demonstrate the application of EG-SNIK dataset on a statistical learning model for characterizing the inherent gaze dynamics of human beings. We have reported this model in [12] as STEAS. First, from EG-SNIK dataset, we plot the velocity distribution of a random viewer observing an exhibit in Fig. 10. It is observed that the fixations (in red) and saccades (in green) are identified with a threshold of 90 degree per sec (as explained earlier). We use a Gaussian Mixture Model (*GMM*) on this distribution and use the corresponding model parameters for approximating the gaze transitions empirically. A scanpath is shown in Fig. 11 corresponding to the data shown in Fig. 10.

A. VELOCITY DISTRIBUTION ANALYSIS USING N-WAY ANOVA

Fig.10 shows the velocity distribution for the gaze data of a random viewer watching a museum exhibit. We would like to observe whether the velocity distributions of saccades and fixations follow a typical pattern for any viewer's gaze information watching any exhibit. Hence, we randomly choose gaze data of 20 viewers and analyse their gaze velocity variations using N-way analysis of variance (*ANOVA*). It also ensures the velocity variations of fixations and saccades are the same for all the 20 viewers. A p-value greater than 0.05 indicates no significant difference between variations in three state velocities between velocity vectors of 4 groups. Table 2 shows the variations in their parameters. Table 3 shows the ANOVA results for the data in Table 2. We see that p-values are 0.75 and 0.79 for fixations and saccades, respectively. Hence, we understand that the velocity distributions of fixations and saccades remain intact irrespective of



(a) Exhibit 13 (b) Representative scanpath of (a)

(c) Exhibit 5 (d) Representative scanpath of (c)

(e) Exhibit 6 (f) Representative scanpath of (e)

(g) Exhibit 11 (h) Representative scanpath of (g)

(i) Exhibit 19 (j) Representative scanpath of (i)

FIGURE 9. A few more examples of the representative scanpaths. First column: scanpaths of a few participants, second column: corresponding representative scanpaths. Notation are similar to that of Fig. 7.

the museum exhibit (or stimulus) as shown in Fig. 10. This set of observations corroborates our earlier findings in [12].

B. SAMPLING FOR TRAINING, TEST AND GROUND TRUTH DATA

For an exhibit, we have 25 video clips of 25 participants and its corresponding velocity vector. Consider that the raw gaze velocity vector of a v^{th} participant for a c^{th} video clip is represented by a vector, $\vec{vel}_{v,c}$, where $v = 1..V$ and $c = 1..C$. Each element of $\vec{vel}_{v,c}$ is the overall gaze velocity at a particular time instant which is calculated as

$$vel_{v,c,i} = \sqrt{vel^2_{v,c,i,x} + vel^2_{v,c,i,y} + vel^2_{v,c,i,z}}, i = 1..L, \quad (27)$$

where L varies depending on the time spent by a participant in observing an exhibit.

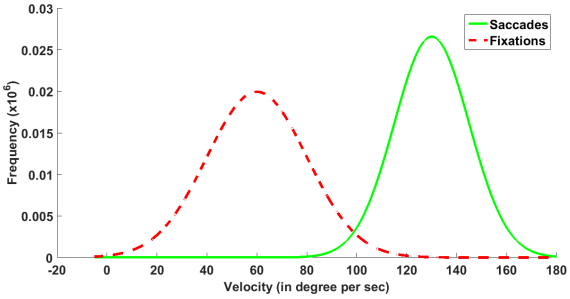


FIGURE 10. Velocity distribution of fixations and saccades of a random viewer watching an exhibit.

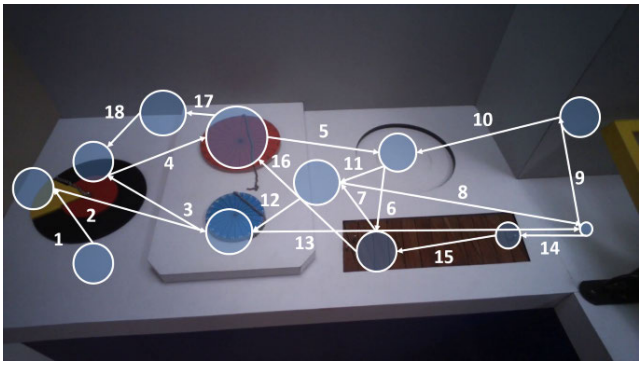


FIGURE 11. Scanpath corresponding to an exhibit.

TABLE 2. Means (μ) and standard deviations (σ) for 4 groups of gaze velocity vectors.

	Group 1		Group 2		Group 3		Group 4	
	μ	σ	μ	σ	μ	σ	μ	σ
Fixations	36.55	17.4	39.21	16.43	40.46	18.12	35.91	17.12
Saccades	123.82	27.4	124.12	18.43	123.31	28.12	129.40	20.12

TABLE 3. ANOVA results for the data provided in Table 2.

Fixations						Saccades					
Source of variations	SS	dof	Var	F	p	Source of Variations	SS	dof	Var	F	p
Between groups	349.55	3	116.51	0.39	0.75	Between groups	606.93	3	202.31	0.35	0.79
Within groups	28659.20	96	298.53			Within groups	54863.36	96	571.49		
Total	29008.76	99				Total	55470.30	99			

1) TRAINING SET

For training, we randomly select 20 clips for each exhibit out of 25 clips. As there are 20 exhibits, there are 400 clips chosen for training the model. A training vector is represented as $\vec{vel}_{m,n}$, where $m = 1..20$ clips and $n = 1..20$ viewers. For a vector $\vec{vel}_{m,n}$, we uniformly sample it K times ($K = 10$) with each sampling length from the set S (here, $S = \{1000, 2000, 3000, 4000, 50000, 6000\}$). We choose the sampling lengths of S empirically.

2) TEST SET

From each participant, the rest of the 5 clips and their gaze data are considered for the purpose of testing. Since there are 25 participants, there are 100 clips chosen for testing.

We evaluate our statistical model performance for gaze data classification and video retrieval on these test vectors.

3) SAMPLED GROUND TRUTH FOR TRAINING

We sample the velocity and label vectors of the clips corresponding to those sampled for building the training set. Those are 20 clips represented as $\vec{V}_{GT_{m,s}}$ and $\vec{E}_{GT_{m,s}}$, respectively for $m = 1..20$.

4) SAMPLED GROUND TRUTH FOR TESTING

After training, we obtain an evaluated parameter for sample length. We use this while sampling the ground truth for testing. We represent the sampled velocity and label vectors for 20 clips associated with 20 exhibits as $\vec{V}_{GT_g}$ and $\vec{E}_{GT_g}$ $g = 1..20$, respectively.

C. STEAS MODEL

STEAS model is a statistical learning model which characterizes the human temporal gaze pattern for different kinds of stimuli using SSMs. We build it by studying the various parameters of the human gaze data. We employ the same set of state and observation equations as detailed in [12] with a few alterations to their parameters. We design state and observation equations for *STEAS* model and fit a *GMM* to the training vector $\vec{vel}_{m,n}$. This gives us parameters corresponding to GMM, mean (μ_e) and standard deviation (σ_e), where $e \in [fixation, saccade]$. The state equation is given as

$$\vec{\theta}_t = \vec{\theta}_{t-1} + \vec{\eta}_t, \tag{28}$$

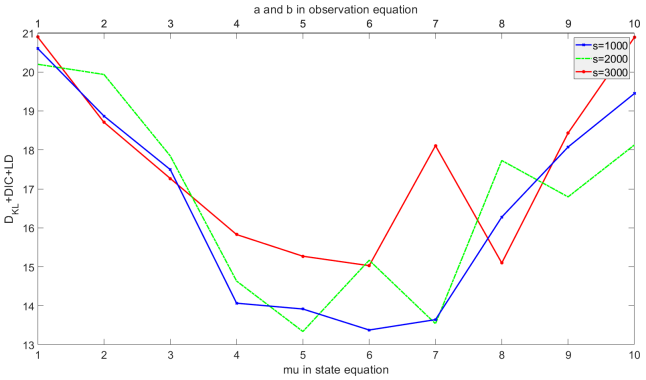
where $\vec{\theta}_t$ and $\vec{\eta}_t$ are velocity vectors of the same length at time instant t, while the latter is a randomly sampled vector from a normal distribution with parameters μ_t and σ_t. We empirically set $\sigma_t = 1$ [44] and found to be experimentally effective, and μ_t is the varying parameter. The observation equation is given as

$$\vec{v}_t = INC(\vec{\theta}_t, CI, \mu_e, \sigma_e) + \vec{v}_t, \tag{29}$$

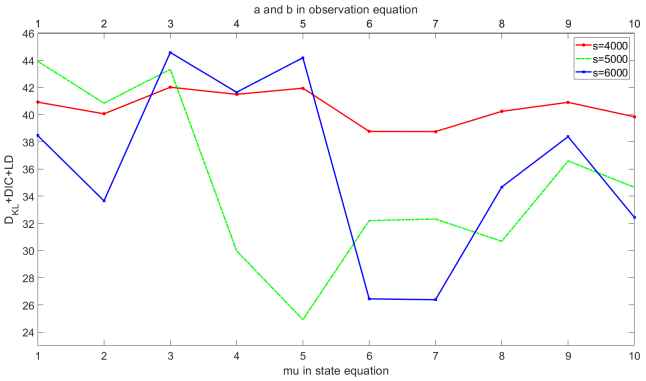
where $INC(\vec{\theta}_t, CI, \mu_e, \sigma_e)$ is the inverse normal cumulative distribution function. μ_t, a_t, and b_t are varied within $[1, 10]$ with an incremental step of 1 considering the noise variability in eye movements [45]. From Fig. 10, we observe that there are chances of a state being identified as another one since their velocity distributions are found to overlap. Hence, every element of $\vec{\theta}_t$ belongs to two states with certain probabilities based on (μ_t, σ_t). Then, a state with the maximum probability of belonging level is selected, and a velocity is sampled within 90% confidence interval. This ensures the sampled velocity value lies in certain state with a considerably good probability. Then, $\vec{w}_T$ is the weight vector evaluated at T^{th} time instant as

$$w_{T,i} = e^{-\frac{(v_{T,i} - V_{GT_{m,s,i}})^2}{2}}, T = 1000, i = 1..s \tag{30}$$

Then, we normalize $\vec{w}_T$ to unit norm. $\vec{v}_t$ has two parameters in it, a_t and b_t. Number of iterations and burn-in (BI) period for (28) and (29) are $N = 500$ and $BI = 100$, respectively.



(a) Sample lengths of velocity vectors 1000, 2000 and 3000



(b) Sample lengths of velocity vectors 4000, 5000 and 6000

FIGURE 12. $D_{KL} + DIC + LD$ **for varying sample lengths.**

BI period refers to the number of iterations STEAS model requires to reach the a state of equilibrium. Usually, the model output is discarded until it finished *BI* period [46]. This process is repeated for empirically varying the parameters μ_t, a_t and b_t over their defined ranges.

STEAS model parameters are optimized to find the best fit of them. We use Deviance Information Criteria (DIC) [47], KL divergence (D_{KL}) [48] and Levenshtein Distance (LD) [49] to find out the best suitable parameters. Considering the linear combination of these metrics, the objective function [50] constrained by the varying parameters is given as:

$$\begin{aligned} &\underset{\mu,a,b,s}{minimize} && DIC_s + D_{KL_s} + LD_s \\ &subjectto && 1 \leq \mu, a, b \leq 10, \quad s \in S, \end{aligned} \tag{31}$$

The set of state and observation equations fitted with the optimized parameters are used for gaze data classification and retrieval on the test set. During the classification, we consider a $K = 31$ in K-NN. Since we have 25 participants, we empirically consider K as 31 to make sure that K is greater than the number of participants. We also perform the video retrieval on the test set using *Mean Reciprocal Rank (MRR)* over D_{KL} and LD metrics separately. For a particular test velocity vector, the inverse of the position of its D_{KL} evaluated with its corresponding ground truth in a sorted array of 60 values is considered as the rank.

VI. RESULTS AND DISCUSSION

A. OPTIMIZED STEAS MODEL PARAMETERS

Fig. 12 shows the variations in the objective function over the empirically varying parameters. Figs. 12a and 12b are plotted for different ranges of sampling lengths in the discretized parametric space. They are shown in two different figures to have visual palatability. We see that STEAS model performs comparatively much better for sample lenghts in the range of 1000 till 3000 than for the other lengths. More particularly, we observe that the global minima of the objective function happens to be 14.87 for sample length $s = 3000$ with parameters μ, a, b being at 6. This denotes that STEAS model captures the inherent gaze dynamics better at certain model parameters. This says that STEAS model characterizes the representative nature of the raw gaze data with parameters being fine tuned [51]. It is to be noted that we perform uniform sampling on velocity vectors and do not impose any bias constraints for fixations and saccades. Hence, we say that the STEAS resultant vector is a representative or features of the raw velocity vector for identifying the visual category of the corresponding stimuli. This feature can model the raw data's dependencies to get incorporated into the sampled data with tuned model parameters [51].

B. GAZE DATA CLASSIFICATION AND VIDEO RETRIEVAL

1) CLASSIFICATION

Table 4 shows the performance comparison of STEAS model with other related models on EG-SNIK dataset. STEAS model is found to perform better than the other models. This happens since the state and observation equations of our statistical model are designed after studying the dynamics of our data and its noise variability. Thereby, the tuned parameters are optimized for this dataset resulting in its optimal performance during classification. However, it is not the case with the other models in Table 4. SubsMatch 2.0 [52] considers smaller subsequences from the gaze data of a viewer and understands how frequently they occur based on AOIs. These subsequences are used for training an SVM model providing certain ranks to the features based on their discriminative power. However, since our data has diverse exhibits with just two or three interesting objects in them, participant may not have very frequent visits to them, thus, SubsMatch 2.0 results in inferior performance. MinHash [53] also works in a similar style as that of SubsMatch 2.0. But, MinHash approximates the Jaccard Index considering only a few subsequences, and therefore cannot achieve the classification performance of SubsMatch. HMM based model performs comparatively closer to our approach since ours is also a statistical model. However, HMM based model involves a large group of parameters whose roles and weights are hard to interpret [25]. It is also one of the major drawbacks of the HMM based model approach. Thus, it results in slightly inferior performance on our EG-SNIK dataset. Hence, a promising value of classification accuracy highlights that the *STEAS* model extracts inherent features from the gaze data irrespective of the viewer. This helps us to identify the unknown gaze data with its corresponding visual category.

TABLE 4. Classification and retrieval performance comparison with other models on the test set of EG-SNIK dataset.

Models	Classification						Retrieval	
	Percentage accuracy		Precision		Recall		MRR	
	D_{KL}	LD	D_{KL}	LD	D_{KL}	LD	D_{KL}	LD
SubsMatch 2.0 (52)	69.34	71.5	0.65	0.672	0.623	0.645	0.66	0.668
HMM based model (25)	73.45	75.62	0.691	0.713	0.664	0.686	0.68	0.693
MinHash (53)	72.26	74.43	0.68	0.701	0.652	0.674	0.668	0.673
STEAS model (Ours)	**75.95**	**78.12**	**0.716**	**0.738**	**0.698**	**0.711**	**0.714**	**0.73**

2) VIDEO RETRIEVAL

Using D_{KL} and LD metrics, we have also evaluated the performance of STEAS for video retrieval of test velocity vectors. Table 4 shows the performance of various other models. As shown, the performance of our model is better than the others because of the reasons mentioned in Section VI-B1. However, for our STEAS model, MRR_{LD} is less than the classification accuracy using LD. The same is the case with MRR_{KL} also. This happens since the classification does not associate the given test vector to its corresponding video ground truth. Instead, it classifies the vector to one of the many classes. Hence, the *STEAS* model based classification and video retrieval can be used for viewer independent video indexing for providing viewers a way to access and navigate contents easily. We find that the *STEAS* model has performed considerably well on our dataset collected in a free viewing style of watching real world museum exhibits. This makes the model a more generalized one, resulting in better-optimized parameters and performance.

VII. CONCLUSION

In this work, we propose a novel egocentric vision dataset acquired at a technological museum in IIT Kharagpur (named as EG-SNIK dataset). We also propose an end-to-end pipeline of processing the data on its video and gaze data streams. A group of 25 students aged between 18 and 35 years with normal vision wear Tobii Pro Glasses 2 eye tracker and observe musuem exhibits. We then build a representative scanpath for each of the museum exhibits aggregating 25 viewers' filtered gaze data. We employ STEAS model to extract inherent gaze features from raw gaze data and optimize it while learning on the training set of EG-SNIK. The optimized model is evaluated on the EG-SNIK test set for gaze data classification and retrieval. We witness a superior performance of STEAS model over other techniques with 77%, 0.727, 0.705 of accuracy, precision and recall, respectively for classification and a MRR of 0.722 for retrieval. The proposed dataset is readily useful for many computer vision problems like object detection, semantic segmentation, etc. Our work has these advantages over others: 1) data collected using the state-of-the-art Tobii pro glasses 2, 2) addressed all possible issues and made the dataset ready-to-use, and 3) shown two vision related use-cases, i.e., classification and retrieval.

REFERENCES

[1] J. B. Pelz and R. Canosa, "Oculomotor behavior and perceptual strategies in complex tasks," *Vis. Res.*, vol. 41, nos. 25–26, pp. 3587–3596, Nov. 2001.

[2] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 744–760, May 2015.

[3] M. Cai, K. M. Kitani, and Y. Sato, "Understanding hand-object manipulation with grasp types and object attributes," in *Robotics, Science and Systems*, vol. 3. Ann Arbor, MI, USA: MIT Press, 2016.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[5] R. Kothari, Z. Yang, C. Kanan, R. Bailey, J. B. Pelz, and G. J. Diaz, "Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities," *Sci. Rep.*, vol. 10, no. 1, pp. 1-18, Feb. 2020.

[6] N. A. McIntyre and T. Foulsham, "Scanpath analysis of expertise and culture in teacher gaze in real-world classrooms," *Instructional Sci.*, vol. 46, no. 3, pp. 435–455, Jun. 2018.

[7] N. Castner, T. C. Kuebler, K. Scheiter, J. Richter, T. Eder, F. H. Uttig, C. Keutel, and E. Kasneci, "Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Berlin, Germany, 2020, pp. 1–10.

[8] M. Cherto, "EgoMon gaze and video dataset for visual saliency prediction," M.S. thesis, Dept. Signal Theory Commun., Universitat Politecnica de Catalunya, Barcelonatech, Spain, 2016.

[9] D. C. Niehorster, T. H. W. Cornelissen, K. Holmqvist, I. T. C. Hooge, and R. S. Hessels, "What to expect from your remote eye-tracker when participants are unrestrained," *Behav. Res. Methods*, vol. 50, no. 1, pp. 213–227, Feb. 2018.

[10] A. Olsen, "The Tobii I-VT fixation filter," *Tobii Technol.*, vol. 21, pp. 4–19, Mar. 2012.

[11] A. Li and Z. Chen, "Representative scanpath identification for group viewing pattern analysis," *J. Eye Movement Res.*, vol. 11, no. 6, pp. 125–136, Nov. 2018.

[12] S. P. K. Malladi, J. Mukhopadhyay, M.-C. Larabi, and S. Chaudhury, "Eye movement state trajectory estimator based on ancestor sampling," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.

[13] M. R. Greene, T. Liu, and J. M. Wolfe, "Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns," *Vis. Res.*, vol. 62, pp. 1–8, Jun. 2012.

[14] B. W. Tatler, N. J. Wade, H. Kwan, J. M. Findlay, and B. M. Velichkovsky, "Yarbus eye movements, and vision," *I-Perception*, vol. 1, no. 1, pp. 7–27, 2010.

[15] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein, "What do saliency models predict?" *J. Vis.*, vol. 14, no. 3, p. 14, Mar. 2014.

[16] M. E. Frame, R. Warren, and A. M. Maresca, "Scanpath comparisons for complex visual search in a naturalistic environment," *Behav. Res. Methods*, vol. 51, no. 3, pp. 1454–1470, Jun. 2019.

[17] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *J. Vis.*, vol. 14, no. 8, p. -5, Jul. 2014.

[18] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1346–1353.

[19] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, Nov. 2006.

[20] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.

[21] J. M. Henderson, S. V. Shinkareva, J. Wang, S. G. Luke, and J. Olejarczyk, "Predicting cognitive state from eye movements," *PLoS ONE*, vol. 8, no. 5, May 2013, Art. no. e64937.

[22] A. L. Yarbus, *Eye Movements and Vision*. Berlin, Germany: Springer, 2013.

[23] A. Haji-Abolhassani and J. J. Clark, "An inverse Yarbus process: Predicting observers' task from eye movement patterns," *Vis. Res.*, vol. 103, pp. 127–142, Oct. 2014.

[24] K. Kumar, S. Harding, and R. M. Shiffrin, *Inferring Attention Through Cursor Trajectories*, C. K. J. Rau and T. Rogers, Eds. Cambridge, MA, USA: MIT Press, 2018.

[25] A. Coutrot, J. H. Hsiao, and A. B. Chan, "Scanpath modeling and classification with hidden Markov models," *Behav. Res. Methods*, vol. 50, no. 1, pp. 362–379, Feb. 2018.

[26] J. Citorik, "Predicting targets in multiple object tracking task," M.S. thesis, Dept. Softw. Comput. Sci. Educ., Univerzita Karlova, Matematicko-Fyzik Aln 1 Fakulta, Karlova, Croatia, 2016.

[27] Tobii. (2020). *Tobii Pro Glasses 2*. Accessed: Feb. 5, 2021. [Online]. Available: https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

[28] TobiiHeadUnit. (2020). *Tobii Pro Glasses 2*. Accessed: Feb. 5, 2021. [Online]. Available: https://www.tobiipro.com/siteassets/tobii-pro/product-descriptions/tobii-pro-glasses-2-product-description.pdf/?v=1.95

[29] O. V. Komogortsev and A. Karpov, "Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades," *Behav. Res. Methods*, vol. 45, no. 1, pp. 203–215, Mar. 2013.

[30] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2714–2721.

[31] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[32] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: Perception and estimation with a new no-reference perceptual blur metric," in *Proc. SPIE*, vol. 6492, Feb. 2007, Art. no. 64920I.

[33] S. Huang, W. Wang, S. He, and R. W. Lau, "Egocentric temporal action proposals," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 764–777, Feb. 2018.

[34] A. Hossain and E. M. Eus, "Eye movement event detection for wearable eye trackers," M.S. thesis, Dept. Appl. Math., Link oping Univ., Sweden, 2016.

[35] M. Pedley, "Tilt sensing using a three-axis accelerometer," *Freescale Semicond. Appl. Note*, vol. 1, pp. 2012–2013, Mar. 2013.

[36] J. Barraza-Madrigal, R. M. Noz-Guerrero, L. Leija-Salas, and R. Ranta, "Instantaneous position and orientation of the body segments as an arbitrary object in 3D space by merging gyroscope and accelerometer information," *Revista Mexicana De Ingenier Ia Biom Edica*, vol. 35, pp. 241–252, Dec. 2014.

[37] H. J. Luinge and P. H. Veltink, "Inclination measurement of human movement using a 3-D accelerometer with autocalibration," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 12, no. 1, pp. 112–121, Mar. 2004.

[38] S. Knecht, M. Deppe, B. Dräger, L. Bobe, H. Lohmann, E.-B. Ringelstein, and H. Henningsen, "Language lateralization in healthy right-handers," *Brain*, vol. 123, no. 1, pp. 74–81, Jan. 2000.

[39] J. Mielikainen, "A novel full-search vector quantization algorithm based on the law of cosines," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 175–176, Jun. 2002.

[40] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 1, pp. 43–49, Feb. 1978.

[41] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

[42] X. Chen and Z. Chen, "Exploring visual attention using random walks based eye tracking protocols," *J. Vis. Commun. Image Represent.*, vol. 45, pp. 147–155, May 2017.

[43] J. Rabin, G. E. Peyr, J. Delon, and M. Bernot, "Wasserstein barycenter and its application to texture mixing," in *Proc. Int. Conf. Scale Space Variational Methods Comput. Vis.* Cham, Switzerland: Springer, 2011, pp. 435–446.

[44] R. J. Van Beers, "The sources of variability in saccadic eye movements," *J. Neurosci.*, vol. 27, no. 33, pp. 8757–8770, Aug. 2007.

[45] J. S. Mitchell and D. L. Rabosky, "Bayesian model selection with BAMM: Effects of the model prior on the inferred number of diversification shifts," *Methods Ecology Evol.*, vol. 8, no. 1, pp. 37–46, Jan. 2017.

[46] F. Lindsten, M. I. Jordan, and T. B. Schon, "Particle Gibbs with ancestor sampling," *J. Mach. Learn. Res.*, vol. 15, pp. 2145–2184, 2014.

[47] J. D. Hadfield, "MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package," *J. Stat. Softw.*, vol. 33, no. 2, pp. 1–22, 2010.

[48] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, Mar. 1951.

[49] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys.-Dokl.*, vol. 10, no. 8, pp. 707–710, 1966.

[50] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[51] A. Ünlu and M. Schrepp, "Toward a principled sampling theory for quasi-orders," *Frontiers Psychol.*, vol. 7, p. 1656, Nov. 2016.

[52] T. C. Kubler, C. Rothe, U. Schiefer, W. Rosenstiel, and E. Kasneci, "SubsMatch 2.0: Scanpath comparison and classification based on subsequence frequencies," *Behav. Res. Methods*, vol. 49, no. 3, pp. 1048–1064, Jun. 2017.
[53] D. Geisler, N. Castner, G. Kasneci, and E. Kasneci, "A MinHash approach for fast scanpath classification," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Jun. 2020, pp. 1–9.

SAI PHANI KUMAR MALLADI (Graduate Student Member, IEEE) received the M.S. (by Research) and Ph.D. degrees from the Advanced Technology Development Center, Indian Institute of Technology (IIT) Kharagpur, in 2017 and 2022, respectively. He joined IIT Kharagpur, as a Junior Research Fellow under MHRD sponsored project titled as "Predicting Cancer Treatment Outcomes of Lung and Colo-Rectal Cancer by Modeling and Analysis of Anatomic and Metabolic Images." During his Ph.D., he was working on static and dynamic human visual saliency. He worked under the guidance of Prof. Jayanta Mukherjee, Prof. Santanu Chaudhury, and Prof. Mohamed-Chaker Larabi. He was a Visiting Researcher at the University of Poitiers, France, in 2020. During that time, he received the Raman-Charpak Fellowship–2019 from the Indo–French Centre for the Promotion of Advanced Research (CEFIPRA). He worked as a Teaching Assistant for courses, such as machine learning, foundations of algorithm design, advanced digital image processing, and computer vision. He is a student member of IEEE Signal Processing Society and IEEE Young Professionals.

JAYANTA MUKHERJEE (Senior Member, IEEE) received the B.Tech., M.Tech., and Ph.D. degrees in electronics and electrical communication engineering from the Indian Institute of Technology (IIT) Kharagpur, in 1985, 1987, and 1990, respectively. He joined the Faculty of the Department of Electronics and Electrical Communication Engineering, IIT Kharagpur, in 1990, and later moved to the Department of Computer Science and Engineering, where he is currently a Professor. He also holds the office of Dean, Outreach and Alumni Affairs of the Institute. He served as the Head of the Computer and Informatics Center, IIT Kharagpur, from September 2004 to July 2007. He also served as the Head of the Department of Computer Science and Engineering, School of Information and Technology, from April 2010 to March 2013. He was a Humboldt Research Fellow at the Technical University of Munich, Germany, for one year in 2002. He also has held short term visiting positions at the University of California, Santa Barbara, University of Southern California, and the National University of Singapore. His research interests include image processing, pattern recognition, computer graphics, multimedia systems, and medical informatics. He has published more than 300 research papers in journals and conference proceedings in these areas. He is a fellow of the Indian National Academy of Engineering (INAE). He received the Young Scientist Award from the Indian National Science Academy, in 1992.

MOHAMED-CHAKER LARABI (Senior Member, IEEE) received the Ph.D. degree from Université de Poitiers, in 2002. He is currently an Associate Professor with Université de Poitiers. He is also the Deputy Scientific Director of the GdR-ISIS (French Research Group on Signal and Image Processing). He has participated in several national and international projects. He supervised more than 20 Ph.D. students and published more than 200 papers. His research interests include quality of experience and bio-inspired processing/coding/optimization of images and videos, such as 2-D, 3-D, HDR, and 360/VR/AR/MR. He has been elected to serve as a member of the IEEE SPS IVMSP, MMSP, and EURASIP TAC-VIP Technical Committees. He is also a member of the CIE, IS&T, and the MPEG and JPEG Committees. He served as the Chair for the JPEG Advanced Image Coding and the Test and Quality Subgroup. He acted as the French Head of Delegation for several years. He was the Program Chair of the EUVIP 2011 and 2018, the Plenary Chair of the EUVIP 2013, the Chair of the EI Image Quality and System Performance (2014–2016 and 2021–2023), the short courses Co-Chair of the Electronic Imaging Symposium (2016–2018), the Technical Co-Chair of the EUVIP 2018, the Special Sessions Co-Chair of ICIP 2016, the Publicity Chair of ICIP 2017 and 2021, the Workshop Co-Chair of ICME 2022, and the Exhibits and Demo Show Chair of ICIP 2022. He is a part of the Steering Committees of ICME, AVSS, and EUVIP. He played several roles in different conferences. He serves as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the Springer *Signal, Image and Video Processing*, the SPIE/IS&T *Journal of Electronic Imaging*, IEEE ACCESS, and Elsevier *Journal of Visual Communication and Image Representation* and *Signal Processing: Image Communication*.

SANTANU CHAUDHURY received the B.Tech. degree in electronics and electrical communication engineering and the Ph.D. degree in computer science and engineering from IIT Kharagpur, Kharagpur, India, in 1984 and 1989, respectively. He is a Professor with the Department of Electrical Engineering, IIT Delhi, New Delhi, India. He is currently serving as the Director with IIT Jodhpur, Jodhpur, India. Before joining IIT Jodhpur, he completed his tenure as the Director with the Central Electronics Engineering Research Institute, Pilani, India. He has authored or coauthored more than 300 research publications in peer-reviewed journals and conference proceedings, 15 patents, and four authored/edited books to his credit. His research interests include image and video processing, computer vision, machine learning, and embedded systems. He is a fellow of the Indian National Academy of Engineering, the National Academy of Sciences, and the International Association for Pattern Recognition. He was a recipient of the Distinguished Alumnus Award from IIT Kharagpur, the Indian National Science Academy Medal for Young Scientists in 1993, and the Advanced Computing and Communications Society-Centre for Development and Advanced Computing (ACCS-CDAC) Award for his research contributions in 2012.