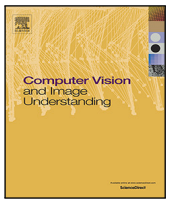
Computer Vision and Image Understanding 248 (2024) 104129

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



An egocentric video and eye-tracking dataset for visual search in convenience stores

Yinan Wang, Sansitha Panchadsaram, Rezvan Sherkati, James J. Clark *

Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada

ARTICLE INFO

MSC:
68T45
91E10

Keywords:
Attention
Saliency
Eye-tracking
Egocentric

ABSTRACT

We introduce an egocentric video and eye-tracking dataset, comprised of 108 first-person videos of 36 shoppers searching for three different products (orange juice, KitKat chocolate bars, and canned tuna) in a convenience store, along with the frame-centered eye fixation locations for each video frame. The dataset also includes demographic information about each participant in the form of an 11-question survey. The paper describes two applications using the dataset — an analysis of eye fixations during search in the store, and a training of a clustered saliency model for predicting saliency of viewers engaged in product search in the store. The fixation analysis shows that fixation duration statistics are very similar to those found in image and video viewing, suggesting that similar visual processing is employed during search in 3D environments and during viewing of imagery on computer screens. A clustering technique was applied to the questionnaire data, which resulted in two clusters being detected. Based on these clusters, personalized saliency prediction models were trained on the store fixation data, which provided improved performance in prediction saliency on the store video data compared to state-of-the art universal saliency prediction methods.

1. Introduction

The vast majority of research into human visual attention has focused on measuring and modeling human behavior during viewing of 2D imagery and video. But humans normally act in a complex 3D environment, and it is therefore important to study how humans allocate attention in these situations. In the computer vision field there has been an increasing emphasis on research involving egocentric video (e.g. Grauman et al. (2022)). Egocentric videos are obtained from video cameras mounted on the bodies of humans (usually head-mounted) while carrying out various activities such as cooking. These record the scenes as seen by observers and can be used for human-centric visual understanding.

There has also been much research into human attention in 3D Virtual environments (i.e. VR), including eye tracking studies. These have generally focused on studying eye movements and visual saliency in panoramic viewing with a fixed viewer position (Sitzmann et al., 2018). Other studies have concentrated on analyzing eye movements and saliency when viewing scenes with isolated 3D objects. Many of the latter studies model what is known as *mesh saliency*, which looks at eye fixations as mapped to mesh models of the 3D object structures (e.g. Ding and Chen (2022)).

In our work, we focus on the understanding of how people pay attention while moving about in unconstrained 3D environments such as grocery stores or convenience stores. Shopping is a complex task that involves navigation, visual search, decision making, human-object interaction and so on. Understanding shoppers' behavior in convenience stores is beneficial to the understanding of human visual attention mechanisms under different tasks. To this end, we created a dataset of egocentric videos with eye fixation information recording shoppers' visual inputs while purchasing different products in a convenience store. In our experiment, 36 participants were asked to fill out a demographic questionnaire after which they carried out a search for three target products, one at a time, in the convenience store while wearing a Tobii glasses-mounted binocular eye-tracking device. While the participants were searching for the products in the store, their eye positions were tracked and recorded, and the scene from their head viewpoint was recorded in an HD video.

This dataset is information rich, and should be useful for many and varied research programs. In this paper we describe two specific studies that make use of the dataset. First, we describe a study of the fixation characteristics, including the statistics of the fixation durations and post-fixation saccade amplitudes, as well as an analysis of the types of objects that are fixated during search. Analysis of the eye fixation metrics from our experiments suggest that cognitive processes underlying eye movement behavior is similar in exploration of 3D

* Corresponding author.
E-mail address: james.j.clark@mcgill.ca (J.J. Clark).

https://doi.org/10.1016/j.cviu.2024.104129
Received 31 May 2024; Received in revised form 15 August 2024; Accepted 18 August 2024
Available online 28 August 2024

environments as in 2D image viewing in the lab. Second, we use the fixation data to create saliency heatmaps and use these to train task- and population-specific saliency prediction models to improve saliency prediction in the store environment over what is provided by standard static image-based saliency prediction methods.

2. Related work

2.1. Egocentric video datasets

Egocentric videos are videos that are recorded from a first-person perspective, showing the scenes as recorded by a camera that is fixed on the observer's head. This video is different from what observers actually perceive since observers may also move their eyes to fixate onto various objects. While it is hard to get a camera that mimics the real time movement of eyes, we can record what the observers pay attention to by overlaying fixation information from an eye tracker on top of the scene camera videos. Various egocentric video datasets have been introduced over the past decades. For instance, the Georgia Tech egocentric activity datasets include: the GTEA dataset for seven types of daily activities (Fathi et al., 2011); the GTEA Gaze dataset for meal preparation activities, with no constraints on the participants (Fathi et al., 2012); the GTEA Gaze+ dataset for seven specific meal preparation activities (Fathi et al., 2012); and the Extended GTEA Gaze+ (EGTEA Gaze+) dataset (Li et al., 2018), which subsumes GTEA Gaze+ with 15 K hand masks and more than 15 K action instances from 200 action categories. The EPIC-KITCHENS dataset (Damen et al., 2018) records 32 participants' cooking activities in their kitchens and includes labels for action segments and object bounding boxes. The Ego4D dataset (Grauman et al., 2022) records daily activity videos for hundreds of scenarios, greatly expanding the diversity of publicly available egocentric footage. The Ego4D dataset also has abundant annotations supporting various complex tasks. The EGTEA Gaze+, EPIC-KITCHENS and Ego4D datasets have annotations of action segments alongside the egocentric videos. The egocentric dataset for browsing situations (Su and Grauman, 2016) includes scenarios for shopping in a market, window shopping in shopping mall, and touring in a museum. Participants determined on their own which activity to engage in, so that they would act naturally. Our dataset is complementary to this, in that it is captured in a controlled environment where the shoppers were shown the specific target products to buy before each visual search session. Other egocentric datasets include categorized activities such as dynamic outdoor sports (Kitani et al., 2011) and activities of daily living (Pirsiavash and Ramanan, 2012). These datasets record different activities from our dataset.

Egocentric video datasets can be used to train models for different tasks such as activity recognition (Kazakos et al., 2019; Ghadiyaram et al., 2019; Li et al., 2021; Wang et al., 2020; Zhou and Berg, 2015), human-object interaction (Darkhalil et al., 2022; Liu et al., 2020), activity anticipation (Kitani et al., 2012; Abu Farha et al., 2018; Furnari and Farinella, 2020; Girdhar and Grauman, 2021), video summarization (Lee et al., 2012; Del Molino et al., 2016; Lee and Grauman, 2015; Lu and Grauman, 2013) and so on. In contrast to most existing egocentric video datasets, our dataset focuses on controlled search tasks in a convenience store. Specifically, our dataset maintains the consistency of target products as well as the background environment, across each search trial with different participants. This control of variables enables identification of consistent and divergent behaviors among individuals during search tasks.

2.2. Eye movements in 3D VR and AR environments

Traditional saliency datasets capture eye fixation data of people looking at static images on computer screens. Notable examples include the MIT1003 dataset (Judd et al., 2009), the MIT300 dataset (Judd et al., 2012), and the CAT2000 dataset (Borji and Itti, 2015). These datasets all measure visual saliency for observers engaged in free-viewing (no specific visual task given). The SALIency in CONtext (SALICON) dataset (Jiang et al., 2015) is currently the largest dataset for saliency prediction, which contains 10000 training images, 5000 validation images and 5000 test images.

More recently, many saliency datasets have been collected in virtual reality (VR) or augmented reality (AR) environments. Salient360 (Rai et al., 2017) and AOI (Xu et al., 2021) are omnidirectional datasets for VR saliency prediction tasks. Duan et al. (2022) introduced the Saliency in AR Dataset (SARD), which contains visual saliency maps for background images, AR images, and superimposed images generated by overlaying AR images onto background images with different opacity levels.

Virtual Reality displays have enabled studies of visual behavior in simulated 3D environments of many kinds. These serve as an intermediate step in moving from studies of attention in 2D image viewing to those of attention in real-world 3D environments. Research on VR environments have provided data on many of the questions and issues that our work is concerned with, such as whether eye movement behavior is similar during 3D search or exploration in 3D environments as it is in 2D image viewing. In a key work, Song et al. (2023) asked the question "is 3D visual saliency an independent perceptual measure or is it a derivative of 2D image saliency". They approached this question by creating a new dataset, called 3DVisA, which consists of 540 view-dependent saliency maps for 180 3D object meshes. They develop a method for predicting 3D saliency for viewing single objects as well as for viewing of scenes constructed from multiple objects. They conclude that while prediction of 3D visual saliency for both a single objects and complex scenes can make use of 2D image saliency to some extent, it still requires from 3D specific information, such as depth maps.

Sitzmann et al. (2018) study how people explore virtual environments from a fixed viewpoint. They consider how models of 2D image viewing extend to the case of viewing in panoramic environments. They create a dataset of head orientations and gaze directions of 169 viewers of 22 panoramic scenes. Their study shows a high similarity between the saliency maps obtained for viewers viewing 2D images on a computer screen and the full VR immersive panoramic viewing. This suggests that eye movement control and processing mechanisms are similar in the 2D and 3D viewing situations. Their experiments also show that viewers of VR imagery exhibit a clear center bias, in this case in the form of a *horizon* bias, where fixation is likely to be near the center of the panoramic display.

Haskins et al. (2020) studied eye movements of viewers in a VR environment, viewing panoramic natural scenes. They looked at differences between passive viewing, in which the panoramic images were moved while the viewers head was fixed, and active viewing where observers were able to move their heads and bodies while the panorama image was fixed. They found that active viewers preferentially allocate their eye movements to semantically relevant scene features than in the passive case, and the eye movements were quicker and more exploratory.

Zhu et al. (2019) constructed a saliency dataset for omnidirectional videos with augmented bounding box contents. However, in real-world AR applications, the augmented contents are typically more complex than simple bounding boxes, as there are different superimposition levels of AR contents onto real-world scenes. Duan et al. (2022) showed that visual attention in AR is jointly and significantly influenced by multiple factors including background images, virtual contents (i.e., AR images), and the opacity of the virtual contents (i.e., superimposition levels). A higher opacity value generally leads to more attention to augmented contents, and a lower opacity value leads to more salient background scenes.

2.3. Visual attention during search

The human brain has limited computational resources (Marois and Ivanoff, 2005). Visual attention is a way for the brain to prioritize and focus computation when presented with multiple stimuli. In search tasks, observers are faced with targets and non-targets simultaneously, and need to direct their attention to the objects that might be task relevant.

Most computational models of visual attention during search tasks are based on the idea of a saliency map that highlights likely target areas (Treisman, 1988; Olshausen et al., 1993; Wolfe, 1994; Niebur and Koch, 1995; Itti et al., 1998; Itti and Koch, 2000). These models are mostly developed using data from experiments conducted on a 2D monitor, displaying a diverse range of image contents. These images include synthetic and natural images with various distortions. Modern visual search models are predominantly based on five factors that affect the search behavior in combination, namely bottom-up salience, top-down feature guidance, scene structure and meaning (Wolfe and Horowitz, 2017).

With the emergence of deep learning, there has been a shift towards using automatic data-driven approaches for visual saliency prediction. Deep neural networks have exhibited a remarkable performance in numerous computer vision tasks, including saliency prediction. These models take advantage of various deep learning architectures to automatically extract features from input images. For example, in 2014, Vig et al. introduced eDN, one of the first deep convolutional networks that could automatically extract features from 2D natural images and obtain a saliency map by combining the feature maps from different layers (Vig et al., 2014). Following eDN, Kümmerer et al. presented Deep Gaze I (Kümmerer et al., 2014), a CNN model based on AlexNet that used pre-trained weights on the ImageNet dataset (Deng et al., 2009) to boost its saliency prediction on the MIT1003 dataset (Judd et al., 2009). In 2016, Kümmerer et al. proposed DeepGaze II (Kümmerer et al., 2016), which used the features of VGG-19 network instead of AlexNet to predict the visual saliency on the MIT1003 dataset. However, relatively little work has been done on developing neural network saliency prediction methods specifically for visual search. Chen et al. (2021) introduced a model called DeepSearch, which fine-tunes a ResNet50 network pretrained on MS-COCO to predict fixation density maps (saliency maps). This is trained on fixation data from the COCO-Search18 dataset (Chen et al., 2021). Samiei and Clark (Samiei and Clark, 2022) introduced a dual-channel deep network that uses an input image of the search target, and predicts saliency of humans during search for that target. It provides similar performance to the DeepSearch approach.

In visual search it is well known that the ordering of fixation points (i.e. the scan-path) is more important than raw saliency values, as saliency changes during the search process, and different features become salient at different times. Because of this, more recent research on modeling of attention during search has focused on scanpath prediction rather than saliency. In addition, attention at the object level rather than at the pixel level has gathered increasing scrutiny by cognitive modelers (Cavanagh et al., 2023; Roth et al., 2023). This has led to the development of neural networks that predict scan-paths at the object level. A recent example of this is found in Fang et al. (2024), which presents the Object-level Attention Transformer (OAT), which predicts human scan-paths during search tasks. Finally, existing research on attention during search has mainly focused on modeling attention on 2D displays. There has been very little work modeling visual attention during search in 3D environments. This lack motivates the creation of a new dataset supporting studies of attention during search in 3D environments, as described in the next section.

Table 1
The participants' favorite colors.

Color	Number of participants
Blue	14
Green	8
Purple	4
Red	3
Cyan	2
Orange	2
Black	1
Yellow	1

3. Dataset creation : Experiment and data gathering

Motivated by the recent developments described in the previous sections, we set out to create a dataset that would support studies of visual attention during search in 3D environments. To support development of saliency and scan-path prediction techniques, we captured timestamped eye-fixation positions in the 3D space, as well as egocentric videos from the shoppers' perspective. To aid in studying object-centric attention, our dataset includes the identity of objects fixated. To provide a relatively constrained experimental setting, we limited our data gathering to monitoring of the visual attention of shoppers searching for a small set of products in a single convenience store.

The experimental data gathering process took place in the McGill University Retail Innovation Lab, which comprises a fully-operational convenience store managed by the *Couche-Tard* chain of stores. This store is located on the McGill University campus and supports research activity as well as normal retail operations.

Each participant in the experiment made three separate search runs in the store, each time looking for a predetermined single product. The order of the runs was the same for all participants and was ordered in increasing search difficulty. The three search tasks, in order, consisted of search for the following products: a bottle of orange juice, a KitKat chocolate bar, and a tin of canned tuna fish. Our pre-experiment expectation was that the search for the bottle of orange juice would be the easiest, as the refrigerator cases are clearly evident from the entrance of the store. The KitKat chocolate bar search would be next easiest, as this is a common item that people search for, but is made somewhat difficult due to the great variety of different chocolate bars nearby that could act as distractors. It is also harder to locate the shelves containing the chocolate bars than the orange juice refrigerator cases as there are many shelves with similar appearance. Finally, it is expected that the search for the canned tuna would be the most challenging, as it is a product not normally searched for in convenience stores, and occupies a small area of a single shelf in the store.

Each participant filled out a questionnaire (shown in Appendix) before their experimental runs. The questionnaire answers are of potential utility in relating shopper's attention patterns to demographic and preference information.

At the beginning of each of the three search runs, the participant put on the eye-tracking glasses (described below), and ran a brief calibration process. The participants were then shown an instance of the product to be searched for, and began the search process. They always started their run at the same spot, near the entrance to the store. The search process took varying amounts of time, ranging from 10 s to 320 s, depending on the difficulties the participant had in locating the search target.

There were 36 participants in the experiment, including 20 females and 16 males. Of these, 11 were in the age range of 20–24 years, 12 in the range 25–29, 8 in the range 30–39, and 5 with ages 40 and above. Participants' favorite colors are shown in Table 1 (one participant's response was excluded due to the selection of multiple colors):

The distribution of the participant occupations is given in Table 2. The demographic composition is quite diverse, encompassing a wide range of age groups and genders. Our coverage includes not only university students but also individuals from various other occupations.

Table 2
Occupation distribution among participants.

Occupation	Number of participants
Student	24
Teacher/Professor	4
Executive Director	1
Customer Service Representative	1
Administrator	1
Advancement Officer	1
Manager	1
Marketing	1
Program Administrator	1
Statistician	1

Fig. 1. The Tobii Pro Glasses 3 eye-tracker. *Image credit:* https://www.tobii.cn/products/eye-trackers/wearables/tobii-pro-glasses-3.

3.1. Eye-tracking device

During each experimental search run, the participants wore the *Tobii Pro Glasses 3* eye-tracker, shown in Fig. 1.

This lightweight glasses-mounted device enables the tracking of both eyes' gaze and fixation coordinates with respect to the video captured by a scene camera. This scene camera, located on the bridge of the glasses, captures videos at full-HD resolution with 106 degree field of view, recording what is in front of the participants. On each side of the glasses, there are eight infrared illuminators and two eye tracking cameras. The illuminators illuminate the eyes to assist the eye tracking sensors, and the eye tracking cameras record eye orientation and movements. A cable from the head unit connects to a separate recording unit that records and stores eye-tracking data and scene camera video on a removable SD card. The eye tracker also includes a microphone, a gyroscope (sampled at 100 Hz), an accelerometer (sampled at 100 Hz), and a magnetometer (sampled at 10 Hz). The eye position information is also sampled at 100 Hz, giving a time resolution of 10 ms.

3.2. Eye-tracking technique

When humans fixate on an object, they can only perceive fine details within the foveal vision area, a small central region of the retina densely packed with cone cells (Snowden et al., 2012). To construct a comprehensive visual representation, individuals must continuously perform saccadic eye movements, shifting their focus to different parts of the visual field. This dynamic process enables the brain to integrate information from various fixations, creating a detailed and cohesive view of the environment. The condition where the image falling on the fovea is steady is called a *fixation*. When transitioning from one fixation point to the next, the eyes perform rapid movements known as a *saccade*.

Visual information used for scene analysis and object identification is primarily acquired during fixations. The duration of a fixation typically varies from 100 to 600 ms, with occasional fixations of up to 3 s. The frequency of fixations is about 3 Hz. During saccadic movements between fixation locations, vision is largely suppressed. It typically takes between 150–175 ms for the brain to plan out a saccadic movement. Once the saccade is planned, the end point cannot be altered. The average duration of saccades is approximately 20–40 ms. These rapid eye movements can reach angular velocities of up to 900°/s.

The eye-tracking technique utilizes corneal reflection, dark pupil detection, and stereo geometry to track eye movements, thus allowing the detection of fixations and saccades. During the calibration process, the eye tracking system maps the participant's eyes onto a standard eye model. During the eye-tracking process, the eye tracking cameras capture the reflection of the illuminators on the cornea. The relative positions of the pupils and illuminator reflections are used to determine the orientation of the eyes. By tracking how eyes move and the duration of each movement, we can detect fixations and saccades.

The eye position measurement provided by the Tobii Pro Glasses 3 eye-tracker has an accuracy of 0.6 degrees of visual angle. The gaze recovery time and blink recovery time are both 1 video frame time (40 ms).

3.3. Raw eye-tracker data

The raw data captured by the Tobii eye-tracker for a single experimental run consists of the scene camera video file in mp4 format and several files readable by the Tobii Pro Lab software. Using the Tobii Pro Lab software, we export the eye tracker information as a single human readable .csv file. The exported file includes the information shown in Table 3. This information file is very rich and includes eye-related data such as video-frame related gaze position, pupil diameters, and 3D vergence point location information. It also includes non-eye-related sensor data such as accelerometer (linear acceleration), gyro (angular velocity), and magnetometer (compass heading) measurements from sensors mounted on the glasses.

At the same time the eye fixation information is being acquired by the eye-tracker, the scene camera captures video imagery at 24.95 fps (40.08 msec per frame) with a resolution of 1920 × 1080 pixels. An example of a video frame captured by the scene camera is shown in Fig. 2 (top).

The timestamp recorded by the eye-tracker allows us to align the scene video frames with the corresponding eye fixation coordinates. This synchronization enables us to overlay gaze data onto each video frame and visualize the participants' fixation points at any given recording timestamp. An example of such an overlay is shown in Fig. 2 (bottom).

3.4. Data post-processing

We perform two additional post-processing steps on the acquired scene video files. First, we trim the videos to remove the frames after the time when the search target has been found. This removes task-irrelevant frames. Secondly, we blur the appearance of any people that may be present in the video frames. This blurring is required by the McGill Research Ethics Board to maintain the anonymity of any and all people visible in the videos.

3.5. Public dataset description and access

The dataset will be made available on written request to the authors at the following URL: https://library.cim.mcgill.ca/data/datasets/Retail_Innovation_Lab_Egocentric_Video_Eyetracking_Dataset/.

The dataset consists of the following items:

Table 3
Raw eye-tracker data items present in the dataset .csv file. HUCS refers to the 3D head-centered coordinate system, determined using the vergence of the two eyes. MCS refers to the media coordinate system, which is the projection of the eye positions onto the 2D video frames.

Recording timestamp [μs]	The recording timestamp in microseconds
Computer timestamp [μs]	The computer timestamp in microseconds
Sensor	The sensor type. Available values: Eye Tracker/Gyroscope/Accelerometer
Participant name	The six-digit ID of each participant followed by the target object
Recording date UTC	The date when the recording was performed in UTC. Format: YYYY-MM-DD
Recording start time UTC	The start time of the recording in UTC. Format: HH.MM.SS.FFF
Recording duration [ms]	Total duration of the recording. Format: milliseconds
Recording Fixation filter name	The name of the fixation filter applied on the recording eye tracking data in the export
Event	Name of the event
Event value	The event value
Gaze point X, Y [MCS px]	Raw gaze coordinates for both eyes combined. Format: pixels (MCS)
Gaze point 3D X, Y, Z [HUCS mm]	The 3D vergence point of left and right gaze vectors. Format: millimeters (HUCS)
Gaze direction left X, Y, Z [HUCS norm]	Unit vector for left eye gaze direction. Format: Normalized coordinates (HUCS)
Gaze direction right X, Y, Z [HUCS norm]	Unit vector for right eye gaze direction. Format: Normalized coordinates (HUCS)
Pupil position left X, Y, Z [HUCS mm]	The 3D coordinates of the left eye pupil. Format: millimeters (HUCS)
Pupil position right X, Y, Z [HUCS mm]	The 3D coordinates of the right eye pupil. Format: millimeters (HUCS)
Pupil diameter left [mm]	Left eye pupil diameter. Format: millimeters
Pupil diameter right [mm]	Right eye pupil diameter. Format: millimeters
Pupil diameter filtered [mm]	The pupil diameter filtered. Format: millimeters
Validity left	Indicates if the left eye has been correctly identified. Available values: Valid/Invalid
Validity right	Indicates if the right eye has been correctly identified. Available values: Valid/Invalid
Recording media width, Height [px]	Dimension of the recording media. Format: pixels
Eye movement type	Format: Fixation, Saccade, Eyes Not Found, Unclassified
Gaze event duration [ms]	The duration of the current active eye movement. Format: milliseconds
Eye movement type index	Sequential number for each instance of an eye movement type
Fixation point X, Y [MCS px]	Coordinates of the fixation point. Format: pixels (MCS)
Ungrouped	The ungrouped data (empty)
Gyro X [°/s]	Angular velocity about the X axis. Format: degrees/second
Gyro Y [°/s]	Angular velocity about the Y axis. Format: degrees/second
Gyro Z [°/s]	Angular velocity about the Z axis. Format: degrees/second
Accelerometer X [m/s^2]	Linear acceleration along the X axis. Format: meters/second^2
Accelerometer Y [m/s^2]	Linear acceleration along the Y axis. Format: meters/second^2
Accelerometer Z [m/s^2]	Linear acceleration along the Z axis. Format: meters/second^2
Magnetometer X [μT]	Magnetic field along the X axis. Format: microteslas
Magnetometer Y [μT]	Magnetic field along the Y axis. Format: microteslas
Magnetometer Z [μT]	Magnetic field along the Z axis. Format: microteslas

Table 4
Summary statistics of video lengths.

Total length of all egocentric videos	6159.63 s
Average length of all egocentric videos	57.03 s
Minimum length of all egocentric videos	9.98 s
Maximum length of all egocentric videos	320.0 s
Total length of ‘kitkat chocolate bar’ search videos	2438.57 s
Average length of ‘kitkat chocolate bar’ search videos	67.74 s
Minimum length of ‘kitkat chocolate bar’ search videos	11.98 s
Maximum length of ‘kitkat chocolate bar’ search videos	283.97 s
Total length of ‘orange juice’ search videos	803.51 s
Average length of ‘orange juice’ search videos	22.32 s
Minimum length of ‘orange juice’ search videos	13.99 s
Maximum length of ‘orange juice’ search videos	51.02 s
Total length of ‘canned tuna’ search videos	2917.55 s
Average length of ‘canned tuna’ search videos	81.04 s
Minimum length of ‘canned tuna’ search videos	9.98 s
Maximum length of ‘canned tuna’ search videos	320.0 s

- 108 trimmed and sanitized egocentric scene videos in mp4 format, one per experimental run (36 participants with 3 search runs per participant).
- 108 eye-tracker data .csv files, as described in Table 3, one per experimental run.
- questionnaire .csv file with responses from all 36 participants.

As each experimental participant takes varying amounts of time to complete the search tasks, the video files are all of different lengths. The summary statistics of the video data lengths are shown in Table 4.

The total compressed size of the dataset is 4.6 Gbytes.

4. Dataset use Case 1: Analysis of fixations during visual search

In the next two sections of the paper we describe two studies that were done using the data from the dataset. In the first use case, we analyze the fixations collected from eye movements of the participants during their searches for orange juice, KitKat chocolate bar and canned tuna. Our primary goal in this study is to determine if there are measurable differences in eye fixation behavior between the three search tasks. In particular we wish to identify which types of objects people fixate on during search for a specific target, and whether the number of non-target object fixations depends on the relative difficulty of the search task.

A key aspect of visual search is the response of the searcher to search targets, anchor objects, or distractors. Anchor objects (Võ, 2021) are objects in the scene that provide information about other objects, in this case the target objects. This information may help guide the searcher to locate and identify a particular search target. For example, the refrigerator case is an anchor object for the orange juice bottle. Other candy bars may be anchor objects for the KitKat chocolate bar. Distractors are objects that are similar to the search target or objects that attract the searcher’s attention, but are non-informative to actually performing the search task. For example, someone searching for orange juice might be distracted by the coffee dispenser. Knowledge of distractor objects can potentially aid retail store operations by predicting future sales (of products attended to while searching for other products) or suggesting more effective layout of products in the store shelves. To this end, we analyze the scene video frames to identify the objects that are being fixated, and classify them as being search targets, anchor objects, or distractors, based on their fixation durations and object types.

Most studies of fixation durations concentrate on the role of fixations with durations less than 500 ms, with durations less than 250 ms indicating so-called *ambient* processing serving spatial localization, and

Fig. 2. (Top) An example video frame captured by the Tobii Pro Glasses 3 scene camera. (Bottom) The same video frame with the current fixation location indicated by a red circle.

those with durations between 250–500 ms indicating *focal* processing serving visual identification (Trevarthen, 1968; Unema et al., 2005; Eisenberg and Zacks, 2016). Such studies have been done almost exclusively on stationary viewing of static images or videos. In our study we wish to determine whether similar fixation duration patterns emerge during active task-specific activity in 3D environments.

4.1. Fixation duration

For each experimental run, the duration of each fixation was computed by the Tobii software. The normalized fixation duration histogram of each experimental run are shown overlaid in Fig. 3. The bin size is 60 ms. It is important to recall that each participant did the search tasks in the same order: orange juice run, KitKat chocolate bar run and canned tuna run, so that task effects are confounded with any fatigue and learning effects. However, the total time taken for all three runs was relatively short (about 5 min in total), suggesting that there was little fatigue or task learning. We observe that the chocolate bar and tuna can searches (runs 2 and 3) have more short fixations (less than 250 ms) than for the orange juice search (run 1). For the fixations that are greater than 250 ms, run 1 has a larger normalized frequency than run 2 and 3.

We see that there are similarities and differences between the three histograms. We see that each histogram appears to consist of a mixture of two distributions - a compact dense unimodal group of fixation durations with a peak around 100 ms, and a sparse group of fixation durations spread over a range from 1000 to 3000 ms. This long-tailed form of fixation duration distribution was also noted by Negi and Mitra (2020) who stressed the importance of considering the long tail of the distribution. This is seen in the histogram of Fig. 4, reproduced from Figure 1 of Negi and Mitra (2020), which shows the distribution of fixation durations for a person viewing an instructional video. The fixation

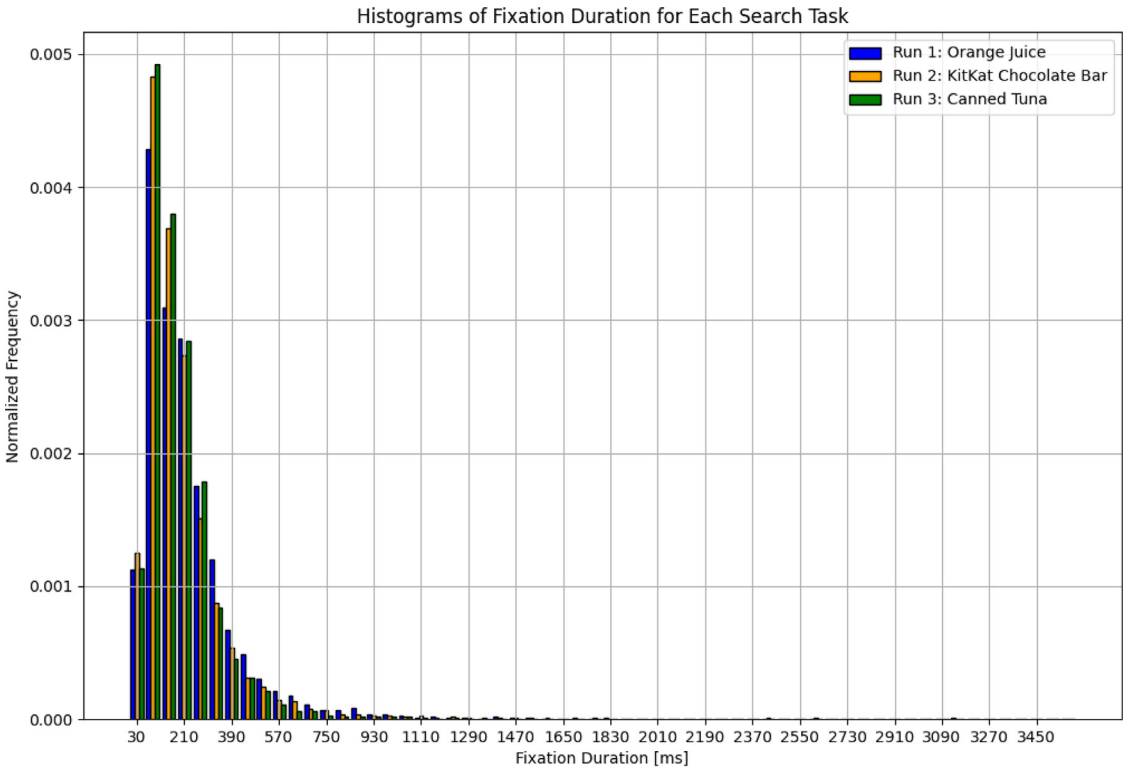


Fig. 3. Histograms of fixation duration for each search task.

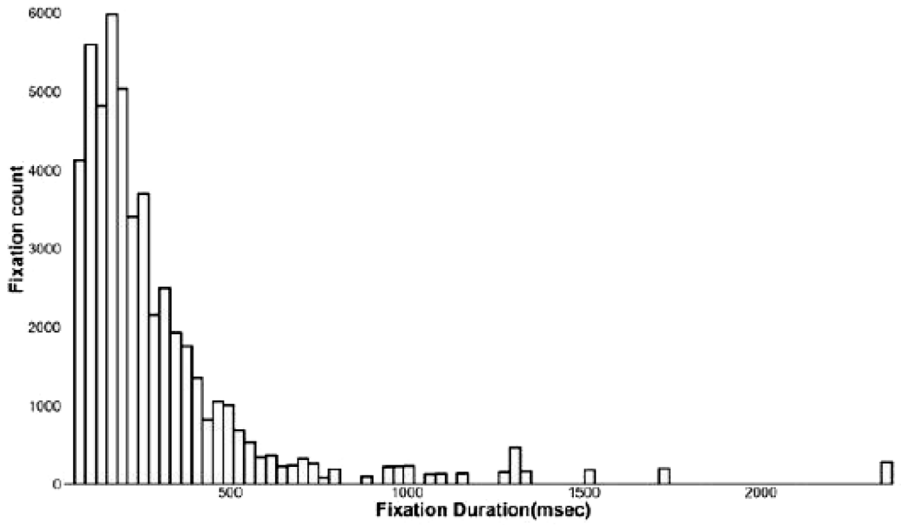


Fig. 4. Representative frequency distribution of raw fixation durations (n = 51238) from a single participant watching a long instructional video. *Source:* Reproduced with permission from Figure 1 of Negi and Mitra (2020).

duration distribution for 3D search is seen to be very similar to those seen in viewing of video imagery, indicating that visual processing in 3D dynamic environments is similar to that in viewing 2D images or videos.

It is evident that there are more fixations on average in the tuna can search runs than in the KitKat bar search runs, which again have more fixations on average than the Orange Juice search runs. This is due to the differences in the length of time taken for each task. From Table 4 we see that, on average, searching for the tuna cans takes longer than the search for the KitKat chocolate bars, which takes longer than the search for the orange juice. These average search times reflect the relative difficulty of the three search tasks.

4.2. Saccade amplitude

For each search task the amplitude of each saccade *after* a fixation was computed using the following equations (Miranda et al., 2018):

$$a = \sqrt{\left(g_x(t_1) - p_x(t_1)\right)^2 + \left(g_y(t_1) - p_y(t_1)\right)^2 + \left(g_z(t_1) - p_z(t_1)\right)^2} \quad (1)$$

$$b = \sqrt{\left(g_x(t_2) - p_x(t_2)\right)^2 + \left(g_y(t_2) - p_y(t_2)\right)^2 + \left(g_z(t_2) - p_z(t_2)\right)^2} \quad (2)$$

$$c = \sqrt{\left(g_x(t_2) - g_x(t_1)\right)^2 + \left(g_y(t_2) - g_y(t_1)\right)^2 + \left(g_z(t_2) - g_z(t_1)\right)^2} \quad (3)$$

$$\alpha = \arccos\left(\frac{a^2 + b^2 - c^2}{2ab}\right) \quad (4)$$

where $g_x(t)$, $g_y(t)$ and $g_z(t)$ are the gaze positions along the x, y and z directions at time t, and $p_x(t)$, $p_y(t)$ and $p_z(t)$ are the pupil positions along the x, y and z directions at time t. t_1 is the time when a saccade starts while t_2 is the time when the saccade ends (Miranda et al., 2018). α represents the saccade amplitude in degrees. Fig. 5 shows a histogram of saccade amplitude with a bin width of 1° for each search target. We observe that each histogram consists of an unimodal distribution with a peak at 2.5°. In target-driven searches, where participants look for a specific item, saccades tend to be directed towards the most probable target locations and participants engage in a more cautious search to not miss the target (Zelinsky, 2008). This often results in smaller, more frequent saccades. Indeed, as observed in Fig. 5, similar trends are seen in each histogram, where more than 30% of the data are within a range of 2 to 3°. The larger saccades may be caused by participants quickly moving their gaze towards salient features in the visual field, such as bright colors or high-contrast objects. These larger saccade amplitudes can also be interpreted as participants conducting a more confident search, eventually knowing where the search target is (Zelinsky, 2008).

We also studied the relationship between the fixation duration and the amplitude of the follow-on saccade (post-fixation). The paper of Negi and Mitra (2020) investigated this relationship in detail and summarized many different studies. They found that many studies observed that *ambient* fixations generally are followed by relatively large saccades, while focal fixations tend to be followed by short saccades. This follows the study of Velichkovsky et al. (2002) who investigated eye movements of people viewing dynamic 2D imagery in a driving simulator. They also found that fixation durations less than 250 msec were associated with follow-on saccades often larger than 4 degrees (i.e. parafoveal), while longer duration fixations tended to have shorter succeeding saccades, indicating attention guide exploration of the focal region. In Fig. 6 we show scatter plots of fixation duration and the amplitude of the post-fixation saccades for each search task in our study. We can see that the same trend observed in the 2D viewing studies occurs in our 3D viewing case. We see a wide spread in post-fixation saccade amplitudes for short fixation durations, including saccades larger than 5 degrees, and a narrow range of saccade amplitudes less than 5 degrees for the longer duration saccades. This provides additional evidence for the hypothesis that visual attention and eye movement control processes involved in 3D viewing are similar to that of 2D image viewing.

4.3. Labeling of fixated objects

For each fixation that occurs during each experiment search run, we identified the object type that the fixation lands on in the scene. To do this, we manually select a video frame from among those in the fixation period for which the object being fixated is clearly visible. If no fixated objects are visible (say due to motion blur), or there is no object at the fixation location, then we do not assign a label to that fixation. To speed up the manual labeling process, we ran Meta AI's Segment Anything Model (SAM) (Kirillov et al., 2023) on the video frames. The model was able to crop specific objects from each video frame. Among all the cropped objects, the image of the fixated object was preserved. This process was followed for each video frame of all 108 raw videos. Each fixated image object was then manually labeled. The labels include both general categories, such as soft drink, chips and candy as well as specific brands, such as Pepsi soft drink, Ruffles chips and Maynards candy. An example of a segmented video frame is shown in Fig. 7.

4.4. Analysis of fixated object type

Once the fixated objects have been identified we can filter them based on their fixation durations. We divide the fixations into three classes based on their durations - *ambient* (0–250 msec), *focal* (250–1000 msec), and *long* (1000+ msec). Following the literature, we operate under the assumption that the focal fixations are the most important ones for visual processing of the object identity. The ambient fixations serve shifting to and aligning with objects, but are not involved with detailed visual processing. We observe from the video data that the long duration fixations are almost always related to grasping of the target object. Based on these considerations, we concentrate our study on the focal fixations. In particular, we identify which object types are fixated frequently across all participants for the three search task types. The convenience store contains thousands of SKUs (stock-keeping units, or distinct products to be purchased), but only a small fraction of these are frequently fixated by searchers. In Tables 5–7 we see the average fixation duration and average number of fixations per run, averaged over the 36 participants, for the three search tasks. These tables only show the values for object types that have an average number of fixations per run greater than 2, and are sorted according to the average number of fixations per run.

We see that different types of objects had frequent focal fixations in the different search tasks. The orange juice search task was usually short and direct, and the objects fixated were all in the refrigerator cases and drink-related. For the KitKat chocolate bar search, most of the objects fixated were other types of chocolate bars. As there was some exploration of the store required in this task, some other objects were fixated, notably an advertising sign and other people in the store. But no unrelated products were fixated focally. The tuna can search task required significant exploration of the store, as the shelf containing the tuna can was small and hard to locate. In this case we see that the searchers often fixated on unrelated products, such as ice cream and cookies.

5. Dataset use Case 2: Clustered saliency prediction

In our second application of the dataset, we use it to train and evaluate a method for predicting the saliency maps of video frame images during the three different search tasks for groups of individuals.

It is well established that there are significant differences between the attention and eye fixation patterns between individuals. A recent study by De Haas et al. (2019) shows that much of the individual differences in saliency lies along semantic directions. That is, people tend to attend to different types of objects. However, most state-of-the-art saliency prediction methods have been based on pooled data from large groups of people. These so-called *universal* saliency methods then naturally predict the average, or population, attention allocation, and do rather poorly in predicting individual saliency. This observation has led to the development of *personalized salience* models and predictors such as those by Yu and Clark (2017), Li and Chen (2018), Xu et al. (2017a, 2018). Personalized saliency prediction takes into account the individual differences in attention, which can be influenced by personal preferences, prior knowledge, interests, and psychological or physiological traits. The development of personalized saliency models is greatly hampered by a lack of suitable training data. Collecting sufficient personalized data is challenging, as it requires tracking the attention of individual users over a wide range of images.

The main reason that the state-of-the-art saliency prediction methods are aimed at predicting the average behavior for large populations is that in this way a large amount of training data can be obtained. Some personalized methods address this problem by applying machine learning techniques that do not require much training data, such as few-shot learning or meta-learning techniques (Luo et al., 2022). In previous work (Sherkati and Clark, 2023), we proposed to ameliorate the data problem by predicting saliency for *groups* of people. The intuition here is that multiple people may have similar attention biases, and these may be discoverable through clustering along demographic and preference dimensions. Additionally, aggregating individuals into groups can aid in anonymization and enhance privacy protection.

There is a vast literature describing a wide variety of approaches to universal saliency prediction. In this paper we build upon, and compare

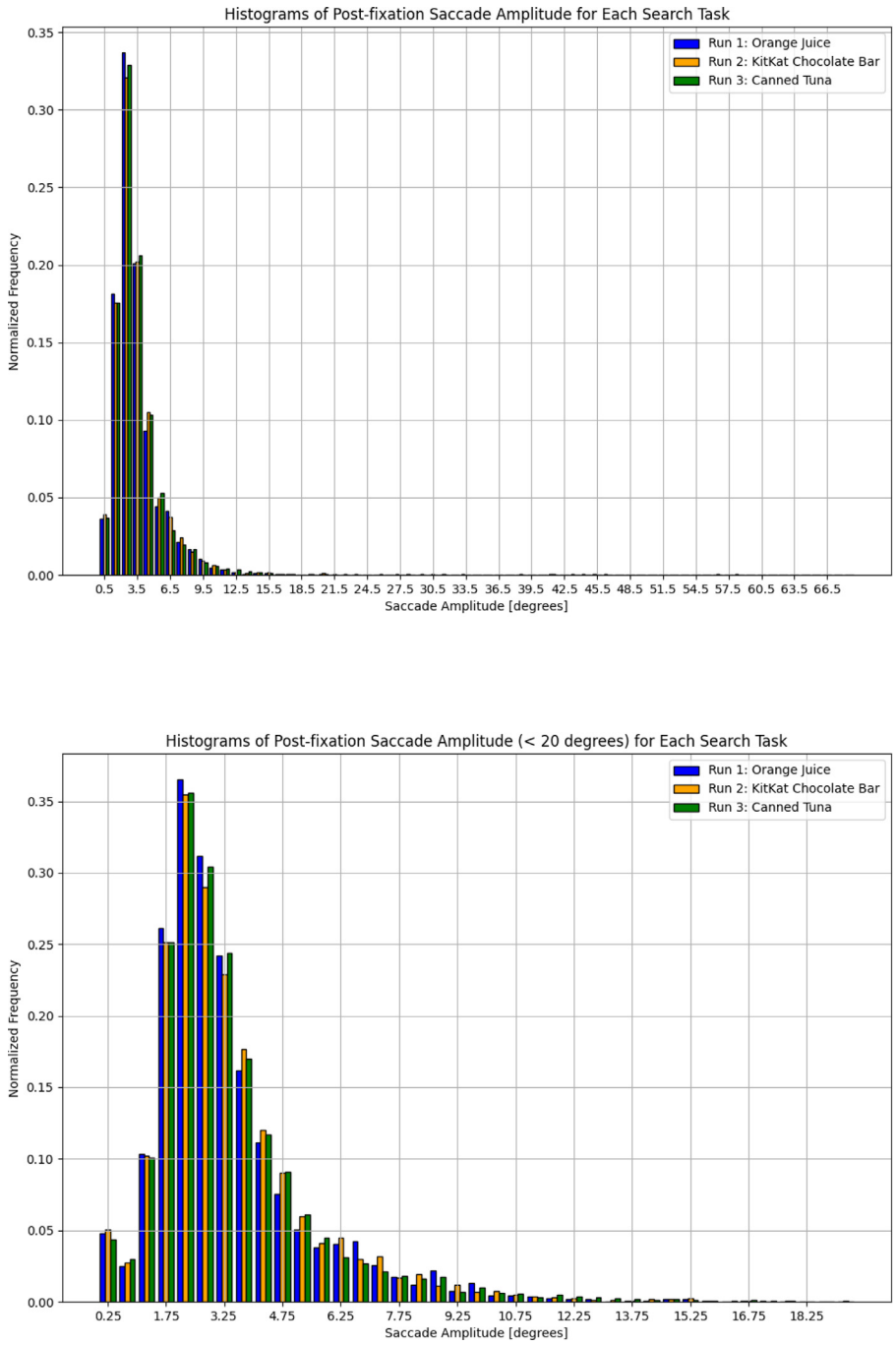


Fig. 5. (Top) Histograms of post-fixation saccade amplitude for each search task. (Bottom) Zoom-in on the saccade amplitudes less than 20 degrees.

to, two current state-of-the-art universal saliency prediction methods — DeepGaze IIE and ML-Net. DeepGaze I (Kümmerer et al., 2014) was a pioneering approach to saliency prediction, which applied transfer learning to the saliency domain. This approach has since evolved into DeepGaze II (Kümmerer et al., 2016), which was built on the VGG19 network. Following the lead of DeepGaze I, nearly all high-performing saliency models have adopted transfer learning, typically starting with networks pre-trained on ImageNet. DeepGaze IIE (Linardos et al., 2021) further improves upon DeepGaze II by replacing the VGG19 backbone with a ResNet50 network. DeepGaze IIE combines some of the state of the art ImageNet backbones, leveraging inter- and intra-model complementarity. In DeepGaze IIE a Gaussian prior is incorporated to account for center-bias, which is the natural tendency of human observers to focus on the center of an image. This network effectively converts the feature information into a probability distribution over the image, indicating the likelihood of gaze fixation at each pixel.

ML-Net (Cornia et al., 2016) is another powerful model for prediction of universal saliency maps. The architecture of ML-Net is based on a deep CNN backbone, such as VGG-16, to extract features from various levels of the image. ML-Net integrates features from different layers of the CNN to create a comprehensive representation of the image. The integrated representation is passed through additional convolutional layers to produce the final saliency map. Unlike many other saliency models such as DeepGaze IIE, ML-Net does not explicitly incorporate a center-bias prior.

The Clustered Saliency Prediction method (Sherkati and Clark, 2023) is a method for personalizing saliency prediction. It does this by first clustering observers into groups and then refining the predictions

Table 5
Object type at focal fixations, for the canned tuna search task. Only object types with more than 2 focal fixations per run are shown.

Object type	Average fixation duration (msec)	Average frequency of fixation per run
Ice Cream	402	8.33
Kraft Dinner	385	4
Cookies	310	4
Chewing Gum	332	3.833
Energy Drink	311	3.8
Noodles Pack	392	2.75
Canned Tuna (target)	392	2.56
Milk	276	2.5
Sunflower Seeds	285	2.5
Cheftop Oven	397	2.5
Cold Ready-made Meal	343	2.42
Person	402	2.28
Nuts	323	2.17

Table 6
Object type at focal fixations, for the KitKat chocolate bar search task. Only object types with more than 2 focal fixations per run are shown.

Object type	Average fixation duration (msec)	Average frequency of fixation per run
KitKat bar (target)	400	6.09
Advertising Signage	445	5.25
Floor	392	3
KitKat bar (non-target)	368	4
Maltesers chocolate	479	3.5
Nuts	312	3
Lindor chocolate bar	324	2.71
Other chocolate	358	2.35
Person	356	2.1

Table 7
Object type at focal fixations, for the orange juice search task. Only object types with more than 2 focal fixations per run are shown.

Object type	Average fixation duration (msec)	Average frequency of fixation per run
Energy drink	340	4.33
Orange juice (target)	400	3.54
Other bottled juice	413	2.87
Vitamin water	325	2.25

of a high-performing universal saliency model, such as DeepGaze IIE or ML-Net, to provide predictions for each group. The technique leverages a conditional generative adversarial network (cGAN) architecture for image-to-image translation, to map the universal predictions to the group predictions.

In this paper, we apply the Clustered Saliency Prediction method to the prediction of saliency of shoppers carrying out search for products in a convenience store. We cluster the participants into (two) groups using the approach outlined in Sherkati and Clark (2023). Then, we use saliency heatmaps derived from a portion of the eye-tracking dataset to independently train a neural network saliency predictor for each group, and for each search task. We validate our approach on the remaining part of the dataset, and compare the results to those obtained with a standard universal saliency model.

5.1. Salience heatmap generation

For each of the 108 videos in the dataset, between 5 to 10 keyframes were selected, choosing frames from among those that were not impacted by motion blur. The assisted mapping function in the Tobii Pro Lab software was used to automatically map the participants' gaze data from nearby video frames onto the keyframes and accumulate a heatmap of the overall gaze locations. The time interval over which the content of each keyframe appears is manually determined and the Tobii assisted mapping function is applied to that interval. The result is a set of 712 keyframe images and their corresponding fixation heatmaps. An example of one such keyframe and its associated heatmap is shown in Fig. 8. As the assisted mapping process combines a small number of discrete fixations of a single observer, the resulting heatmaps are quite compact and localized. To reflect the saliency maps that would be obtained with a large number of viewers, we follow standard practice (e.g. Gilani et al. (2015)) and obtain a fixation density estimate by Gaussian blurring with a circularly symmetric Gaussian kernel with standard deviation of 201‘pixels. This models the variability of fixation location that would be expected with multiple observers. Examples of blurred heatmaps can be seen in Fig. 11. These blurred heatmaps are used as the ground truth in training and testing our saliency prediction technique.

5.2. Grouping of participants

The dataset includes data from 36 participants. We group these participants based on their answers to the pre-experiment questionnaire. The grouping method used follows the clustering technique described by Sherkati and Clark (2023). To divide the participants into groups of individuals with similar traits, we create a complex weighted graph capturing the relationships between the 36 individuals in the dataset. Each individual is considered as a node in the graph. We connect every pair of individuals with an edge and assign a weight to this edge equal to the sum of the number of commonalities in their answers in the questionnaire. All questions have equal weight when summing the commonalities. After constructing the graph, we run the Louvain community detection method (Blondel et al., 2008) to determine the communities, or clusters, that have a high degree of connectivity or interaction among their members. The Louvain method is known for its scalability and ability to detect communities in large-scale graphs efficiently.

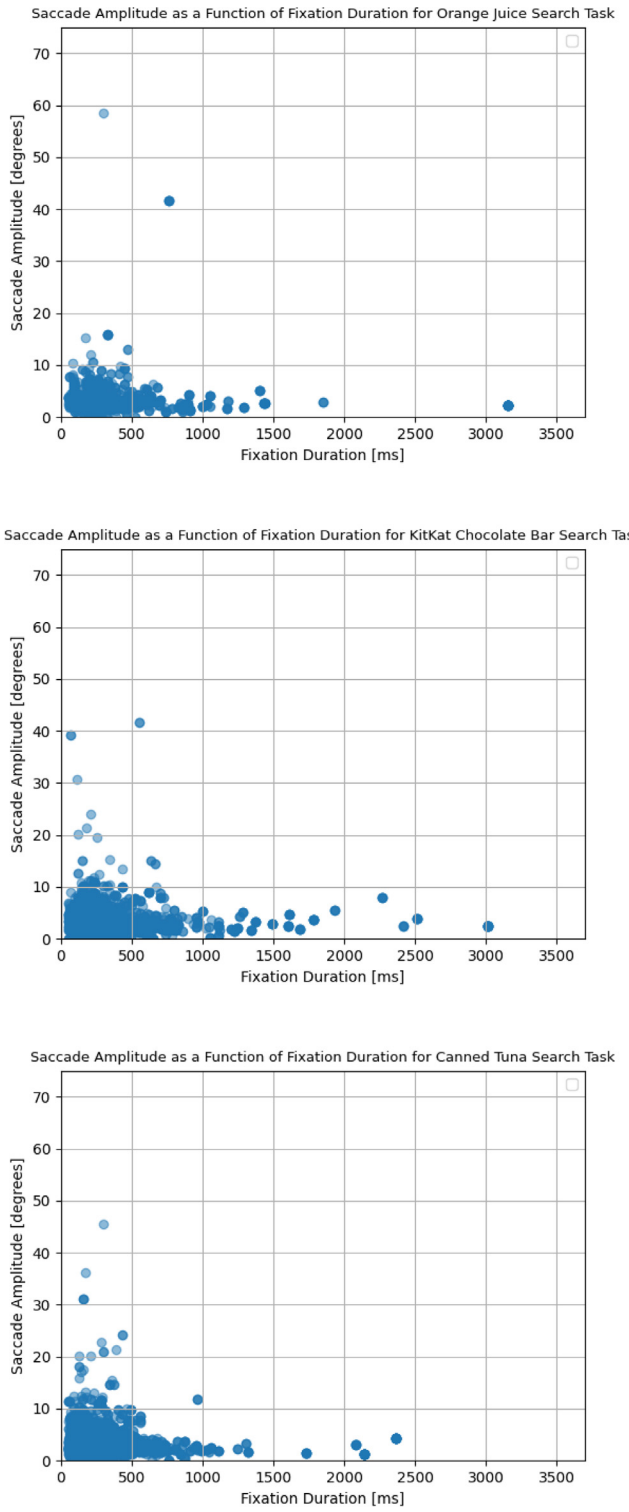


Fig. 6. Scatter-plots of post-fixation saccade amplitude and fixation duration. (Top) Orange juice search, (Middle) Kitkat chocolate bar search, (Bottom) Tuna can search task.

After running the community detection algorithm on the dataset, we obtain the network shown in Fig. 9. In the figure nodes of the same color belong to the same cluster. The ID of each node is the ID of the corresponding subject in the dataset. The red colored edge is the edge with highest weight, which connects the two individuals with the most similar chosen answers in the questionnaire. We see that the algorithm results in only two clusters.

Table 8
Comparison of average edge weights in each network, induced by the sets of nodes in the first column.

Network	Average of edge weights
Cluster 1	7.0784
Cluster 2	7.3595
All Subjects	6.5952

To compare the inter-observer similarity with the inter-group similarity, we focus on the average edges weights in the two clusters and average edge weights in the whole network. This is because the edge weight between two observers represents the similarity of their questionnaire answers. For each cluster we compute the average of the edge weights between each two individuals in that cluster. We also compute the average edge weight between each two individuals in the entire network. The results are shown in Table 8. We see that the average edge weight in each cluster is higher than the average edge weight in the whole network. This shows that individuals in each cluster have greater intra-cluster similarity than inter-cluster similarity.

For each question in the questionnaire, we compare the average similarity of each pair of individuals in each cluster vs. the whole network. The results can be seen in Table 9. We see that, for the age (Q1), education level (Q7), products preference (Q11) questions, the scores for each cluster are similar, and similar to that of all participants, indicating that the answers were similar in both clusters. The average similarity scores for the visual impairment question (Q3) are similar for both clusters, and higher than the average similarities over all the participants. This indicates that there is a significant separation between the clusters for this question. For the gender identity (Q2), eye surgery (Q4), color blindness (Q6), work (Q8), food identity (Q9) and color preference (Q10) questions one cluster had a significantly higher similarity than the other, indicating that that question was important in determining the membership of one of the clusters but not the other. In general, we conclude that the membership of cluster 1 is mainly driven by the answers to questions 3, 5, 8 and 10, while that of cluster 2 is mainly driven by the answers to questions 2, 3, 4, 6, and 9.

5.3. Prediction of saliency using the MDST approach

After clustering the individuals using the questionnaire, we use the Multi-domain Saliency Translation (MDST) technique from Sherkati and Clark (2023) to predict the saliency heatmaps at the cluster or group level. A general illustration of this framework is shown in Fig. 10. The MDST method is based on Conditional Generative Adversarial Networks (cGANs) and is an adaptation of the Pix2Pix image translation model of Isola et al. (2017). The MDST model incorporates a cluster-mapping network that takes the cluster label as input and produces a point in the class space. It consists of an embedding layer followed by four fully connected layers. The output of the cluster-mapping network for each cluster label is concatenated with the input image and its universal saliency map. This combined input serves as a latent code that is fed into a generator network that produces the personalized saliency map. The generator is based on the architecture of the U-Net generator used in the Pix2Pix model. The GAN discriminator uses the same architecture as the Pix2Pix model's discriminator. We concatenate the original image and its universal saliency map, along with the output generated by the generator, which is then passed to the discriminator.

In order to train the MDST network we first generate the ground truth saliency heatmaps from the video frames, as mentioned in Section 5.1. Then, for the individuals in each group obtained in Fig. 9, we create 3 sub-clusters with the sets of images obtained for each of the three search tasks (searching for Orange juice, Kitkat and Tuna cans). Considering that we have two groups of individuals in Fig. 9, we

Fig. 7. (Top) An example video frame segmented with the Segment Anything Model. The segmented images are used to speed up manual labeling of objects at fixation points.

Table 9
The average similarity of answers for each question in each cluster and the whole network.

	Cluster 1	Cluster 2	All subjects
Q1: How old are you?	0.2484	0.2353	0.2429
Q2: What is your gender identity?	0.4967	0.5752	0.4921
Q3: Do you have any form of visual impairment (hyperopia, myopia or astigmatism, etc..)	0.8889	1.0	0.4873
Q4: Have you had any form of eye surgery (corneal (e.g. LASIK, RK), cataract, intraocular implants)?	0.7908	1.0	0.8921
Q5: Do you have any eye movement or alignment abnormality (amblyopia, strabismus, nystagmus)?	0.8889	0.7908	0.8429
Q6: Do you have colorblindness?	0.8889	1.0	0.9444
Q7: Education level (last level obtained)	0.3922	0.3137	0.3286
Q8: Work/occupation	0.6928	0.3137	0.473
Q9: Do you feel you have a food identity? If so, which of the following do you identify with? Select all that apply.	0.4575	0.7843	0.6111
Q10: Which of the following colors do you prefer? (select one)	0.3333	0.183	0.2349
Q11: Which of the following products do you often buy at a convenience store? (select all that apply)	1.0	1.1634	1.046

have 6 training clusters in total (3 search specific sub-clusters for each of the 2 groups) $C_1, C_2, \ldots, C_6$. Then, in training the MDST network, for each cluster C_i such that $1 \leq i \leq 6$, we define the source images as the original video frames and the DeepGaze IIE (Linardos et al., 2021) universal saliency maps of the frames seen by the individuals in cluster C_i. The ground truth target images are the Gaussian blurred saliency heatmaps of the individuals for the corresponding frame and corresponding search task of cluster C_i. Note that in this method, as explained in Sherkati and Clark (2023), during training each set of video frame image and DeepGaze IIE saliency map of this frame in the source domain has a corresponding image in the target domain. Moreover, if we have multiple participants in the same cluster who observed the same frame image, for the target image we use the average of their corresponding saliency heatmaps.

Following the procedure outlined in Sherkati and Clark (2023), we train the MDST network for 200 epochs, with a batch size of 16. We use the Adam optimizer, with an initial learning rate of 0.0002 for 100 epochs, which then linearly decays to 0 over the remaining 100 epochs. The Adam optimizer uses momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and a weight decay of 0.00001 is applied to help prevent overfitting. We apply data augmentation techniques, specifically resizing the input images to 286×286 pixels, followed by a random crop to 256×256 pixels, and a random horizontal flip. This process is applied to the input images as well as the universal saliency maps and ground truth generated saliency heatmaps. The cluster label is transformed into a latent code of size 256×16 via an embedding layer, and this code is passed through four consecutive fully connected layers, each maintaining the input and output size of 256×16. The resulting output is then duplicated four times, concatenated, and resized to a shape of 256×256. We use random splits of all the images in each cluster with proportions of 80%, 10% and 10% for the train, validation and test sets. We trained the MDST network for 5 different independent random splits and averaged all the results.

Fig. 8. (Top) A key-frame from a video of a participant searching for a chocolate bar. (Bottom) The fixation heatmap generated by the Tobii Pro Lab assisted mapping function.

5.4. Experimental evaluation

For evaluation of the results, the Pearson's Correlation Coefficient (CC) and Similarity (SIM) metrics were used. The reason for selecting the CC and SIM metrics for evaluation is that they consider saliency heatmaps as the ground truth. As mentioned earlier, the saliency heatmaps for each key frame contains only a few fixation points. To account for the variability in fixation locations, we apply a Gaussian blur. This altered saliency heatmap is then used as the ground truth for our performance evaluations. Metrics such as Normalized Scan-path Saliency (NSS), which consider fixation points as the ground truth, cannot incorporate these modifications to the ground truth heatmap.

Pearson's Correlation Coefficient (CC): is a statistical measure used to assess the degree of correlation or dependency between two variables. It evaluates the linear relationship between variables, treating them as random variables. CC is symmetric and penalizes both false positives and false negatives equally, making it invariant to linear transformations. In the context of saliency maps, CC can quantify the linear relationship between predicted saliency maps and ground truth fixation maps.

Similarity (SIM): also known as histogram intersection, quantifies the similarity between two distributions considered as histograms. SIM is calculated by summing the minimum values at each pixel after normalizing the two saliency maps being compared. A SIM value of one signifies that the distributions are identical, whereas a SIM value of zero indicates no overlap between them. SIM is highly sensitive to missing data and penalizes predictions that do not encompass all aspects of the ground truth distribution.

In Table 10, adapted from Table 1 of Sherkati and Clark (2023), our MDST method shows 9.60% improvement in Correlation Coefficient (CC) metric and 7.06% improvement in Similarity (SIM) metric in prediction of saliency maps of subjects compared to direct application of the DeepGaze IIE model on the Personalized Saliency Maps (PSM)

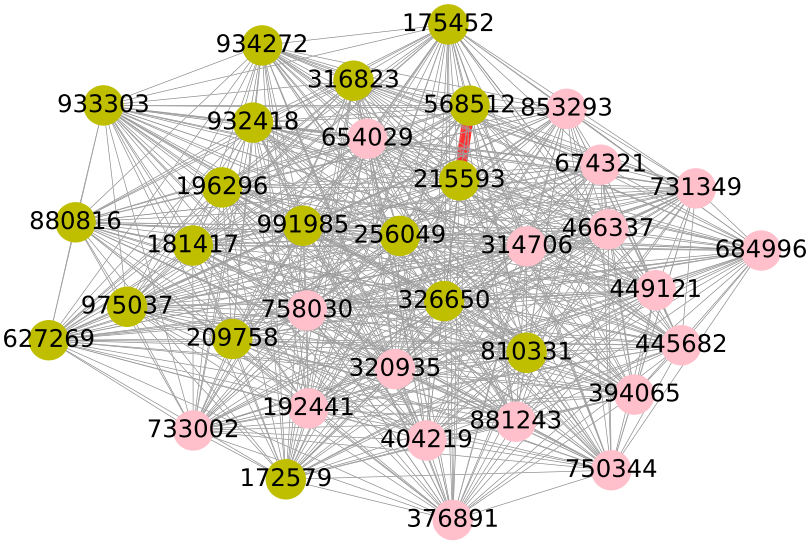


Fig. 9. Clustering network of the dataset participants. The red edge is the edge with the highest weight, indicating the two subjects that have the most answers in common. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

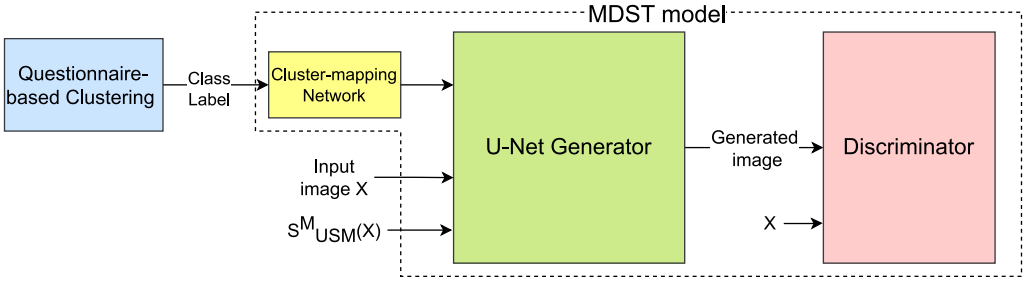


Fig. 10. A general framework of clustering and prediction of clustered saliency maps for our dataset.

dataset of Xu et al. (2017b). This model also shows significant improvement compared to the ML-Net (Cornia et al., 2016) universal saliency prediction model.

In evaluating the results the average metric values were measured in each cluster. Then the metric values for each cluster were averaged across all the clusters for the final performance. We also evaluated the test set results for training MDST on the setting with only 3 clusters rather than 6, where each cluster contains the saliency maps of all the individuals for one search task. We also obtained the average performance of DeepGaze IIE on these 3 clusters. The evaluations of the results are summarized in Table 11. We see in this table that the MDST model, on both the 6 cluster and 3 cluster cases, outperforms the DeepGaze IIE and ML-Net models. We also see that the MDST model on 6 clusters has higher performance than the MDST model on 3 task clusters. This suggests that our questionnaire based clustering method has a positive impact on the saliency prediction results. Furthermore, DeepGaze IIE's performance on all images within a single cluster is lower than its performance when images are divided into multiple clusters, providing additional evidence of the clustering method's effectiveness in grouping similar individuals. Note that in Table 11, the metrics for the DeepGaze IIE model in the 3 and 6 cluster cases are not the same. This is because all the 6 clusters of Section 5.3 have test sets of different sizes. While the Gaussian center-bias prior with $\sigma = 100$ performs better than ML-Net and DeepGaze IIE on this dataset, the MDST approach still outperforms the Gaussian center-bias prior. Previous research has demonstrated a significant center bias for saliency in 3D viewing (e.g. Sitzmann et al. (2018), Ding and Chen (2022)), which likely explains the high scores for the center-bias prior baseline. In Fig. 11 we see some examples of saliency predictions provided by the MDST and DeepGaze IIE approach.

Table 10
Mean performance of Clustered Saliency Prediction model of Sherkati and Clark (2023) for all subjects in Personalized Saliency Maps (PSM) dataset (Xu et al., 2017b) and comparison to the DeepGaze IIE and ML-Net universal saliency prediction methods. Adapted from Table 1 of Sherkati and Clark (2023).

Prediction method	CC	SIM
DeepGaze IIE based Clustered	**0.7418**	**0.6369**
DeepGaze IIE	0.6768	0.5949
ML-Net based Clustered	0.7115	0.6145
ML-Net	0.6504	0.5701

It should be noted that the performance metric values for all approaches tested on our dataset are relatively low as compared to their values on large free-viewing saliency datasets, where typical CC scores are above 0.7 (e.g. Table 10). We hypothesize that this is mainly due to the difference in saliency patterns for visual search tasks as compared to free-viewing tasks. DeepGaze IIE was trained on saliency maps acquired during free-viewing tasks. Similarly, the MDST approach takes in the DeepGaze IIE universal map as input, and will thus inherit its bias towards free-viewing saliency patterns. However, the MDST approach has the advantage of being trained on the task specific saliency data from the store experiments, and thus should do better than using DeepGaze IIE alone.

6. Conclusion

In this paper we present an egocentric video and eye-tracking dataset supporting studies of human attention during search in a complex 3D environment - a convenience store. The dataset includes 108

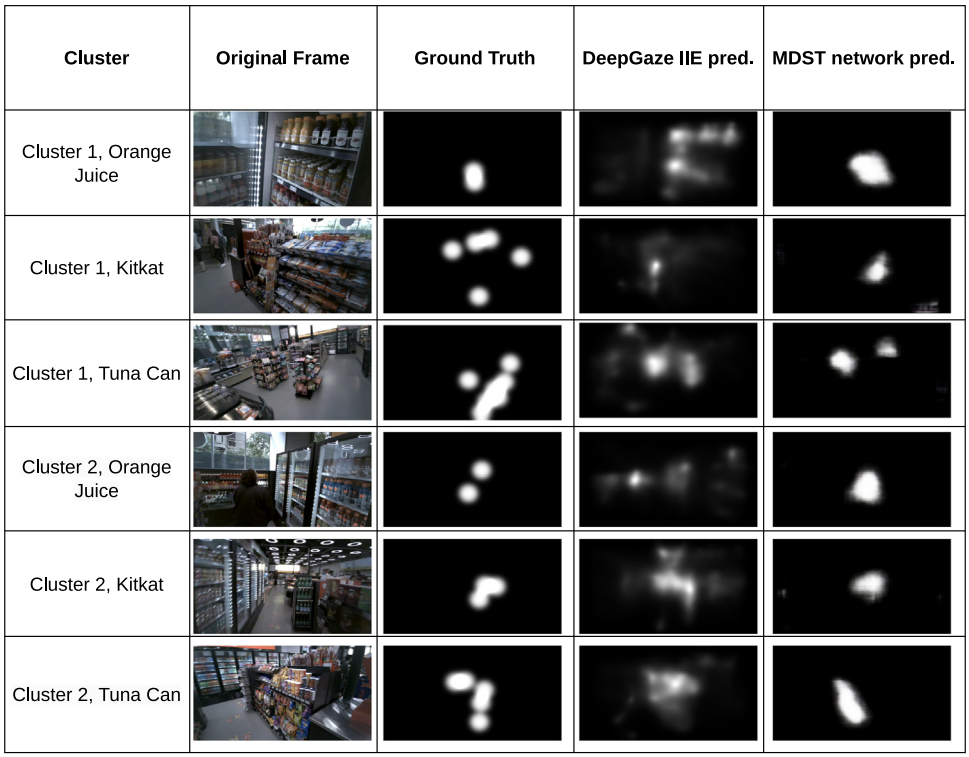


Fig. 11. The example outputs of the MDST and DeepGaze IIE models on the test set of one of the random splits, for all clusters.

Table 11
The comparison of average saliency prediction performance on the store dataset test set using the MDST (Sherkati and Clark, 2023), DeepGaze IIE (Linardos et al., 2021), ML-Net (Cornia et al., 2016) saliency prediction methods (averaged over 5 independent random splits). We also include comparisons to Gaussian center-bias prior maps with two different standard deviations. MDST model was trained both on 6 obtained clusters in Section 5.3 and also on only 3 clusters separated by tasks. For ML-Net, performance was averaged on the predictions for all the frames (of all subjects and all tasks) within the same cluster. CC: Pearson's Correlation Coefficient; SIM: Similarity Metric.

Method	CC	SIM
MDST on 6 clusters	0.3077	0.2782
MDST on 3 clusters (for only tasks)	0.2947	0.2715
DeepGaze IIE on 6 clusters	0.2839	0.2501
DeepGaze IIE on 3 clusters (for only tasks)	0.2854	0.2514
DeepGaze IIE (on all the frames in one cluster)	0.2795	0.2454
ML-Net (on all the frames in one cluster)	0.2128	0.1876
Gaussian center-bias prior, $\sigma = 100$	0.2969	0.2764
Gaussian center-bias prior, $\sigma = 80$	0.2572	0.2348

HD egocentric videos, with durations from 10 s to 320 s, taken from a glasses-mounted HD camera worn by a person engaged in search for three different products in a convenience store. The dataset also includes concurrent eye-tracking data aligned with the egocentric videos, as well as a demographic questionnaire. The dataset will be publicly available to other researchers on request.

We present two applications that make use of the dataset — an analysis of eye fixations during search in the store, and a training of a clustered saliency model for predicting saliency of shoppers engaged in product search in the store.

In the fixation analysis study we find that the distributions of low-level eye movement metrics – fixation duration and post-fixation saccade amplitude – during search in the 3D store environment are very similar to those observed in prior studies involving stationary viewers of 2D imagery, suggesting that similar visual processing is used in both situations. Examination of the so-called 'focal' fixations, i.e. those that have duration between 250–1000 msec, indicates that relatively few objects in the store are frequently attended to during the searches. For the easier searches, where objects are quickly found, the fixated objects are all related categorically to the search target. Conversely, for longer, more difficult, searches, unrelated objects are often fixated, suggesting increased levels of distraction. We observed in the case of the difficult search for a can of tuna, experiment participants often fixated on other food items such as ice cream and Kraft Dinner, perhaps suggesting that these are items frequently consumed by the participant.

We applied the MDST clustered saliency prediction approach of Sherkati and Clark (2023) to the problem of predicting fixation probability from the eye tracking data. The clustering was based on the answers provided in the demographic questionnaires. The clustering algorithm divided the 36 experiment participants into two groups, primarily based on the answers to questions on visual impairment, food identity and color preferences. Our hypothesis is that the two clusters would exhibit different attention biases that result in measurable differences in fixation patterns. We train the MDST network on saliency heat maps derived from the measured eye-track data separately for each cluster and each search task. The result is a saliency prediction for each combination of viewer cluster and search task type (6 different predictions in total). Our results show that the clustered saliency prediction models perform better than with no clustering, and these work better than using a universal saliency model (one trained on large numbers of people doing free viewing of images unrelated to the store) such as DeepGaze IIE (Linardos et al., 2021). The benefit is small, but does indicate that the store eye-tracking data can be used to provide fine-tuning of existing state-of-the-art pre-trained saliency models.

CRediT authorship contribution statement

Yinan Wang: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation. **Sansitha Panchadsaram:** Investigation, Formal analysis, Data curation. **Rezvan Sherkati:** Writing – review & editing, Validation, Methodology, Investigation. **James J. Clark:** Writing – review & editing, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset is available to the reviewers and will be made publicly available on request. The link to the data is given in the paper, and is password protected.

Acknowledgments

We acknowledge the Natural Sciences and Engineering Council of Canada (NSERC) and the Ministère de l'Économie, de l'Innovation et de l'Énergie (MEIE) for financial support. This research was enabled in part by computation support provided by Calcul Québec and the Digital Research Alliance of Canada. We also would like to thank Mme. Valerie Forget from Alimentation Couche-Tard and Mr. Jan Villaluz from the Bensadoun School of Retail Management at McGill University for their operational support.

Appendix. Questionnaire

Before the experimental runs, each participant is asked to fill out a questionnaire, with the following questions.

1. How old are you?

- (a) 16–19
- (b) 20–24
- (c) 25–29
- (d) 30–39
- (e) 40–60
- (f) 60+
- (g) Prefer not to say.

2. What is your gender identity?

- (a) Woman
- (b) Man
- (c) Non-binary
- (d) Prefer not to say

3. Do you have any form of visual impairment (hyperopia, myopia or astigmatism, etc..)

- (a) Yes
- (b) No
- (c) Prefer not to say

4. Have you had any form of eye surgery (corneal (e.g. LASIK, RK), cataract, intraocular implants)?

- (a) Yes
- (b) No
- (c) Prefer not to say

5. Do you have any eye movement or alignment abnormality (amblyopia, strabismus, nystagmus)?

- (a) Yes
- (b) No
- (c) Prefer not to say

6. Do you have colorblindness?

- (a) Yes
- (b) No
- (c) Prefer not to say

7. Education level (last level obtained)

- (a) Primary school
- (b) Secondary/CEGEP
- (c) Tertiary education
- (d) Bachelor's degree
- (e) Master's degree
- (f) Doctorate degree
- (g) Prefer not to say

8. Work/occupation

9. Do you feel you have a food identity? If so, which of the following do you identify with? Select all that apply.

- (a) No, I do not have a food identity.
- (b) Vegan
- (c) Vegetarian
- (d) Gluten-free
- (e) Organic
- (f) Keto
- (g) Other, please specify:
- (h) Prefer not to say

10. Which of the following colors do you prefer? (select one)

- (a) Red
- (b) Orange
- (c) Yellow
- (d) Green
- (e) Cyan
- (f) Blue
- (g) Purple
- (h) White
- (i) Black
- (j) Prefer not to say

11. Which of the following products do you often buy at a convenience store? (select all that apply)

- (a) Milk
- (b) Chocolate
- (c) Muffins
- (d) Coffee
- (e) Energy Drinks
- (f) Bottled Water
- (g) Soup
- (h) Juice
- (i) Noodles/ramen
- (j) Ice cream
- (k) Potato chips
- (l) Candy
- (m) Chewing gum
- (n) Canned goods
- (o) Salads
- (p) Sandwiches
- (q) Prefer not to say

References

Abu Farha, Y., Richard, A., Gall, J., 2018. When will you do what?-anticipating temporal occurrences of activities. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5343–5352.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008 (10), P10008.

Borji, A., Itti, L., 2015. Cat2000: A large scale fixation dataset for boosting saliency research. arXiv preprint arXiv:1505.03581.

Cavanagh, P., Caplovitz, G.P., Lytchenko, T.K., Maechler, M.R., Tse, P.U., Sheinberg, D.L., 2023. The architecture of object-based attention. Psychon. Bull. & Rev. 30 (5), 1643–1667.

Chen, Y., Yang, Z., Ahn, S., Samaras, D., Hoai, M., Zelinsky, G., 2021. COCO-Search18 fixation dataset for predicting goal-directed attention control. Sci. Rep. 11 (1), 1–11.
Cornia, M., Baraldi, L., Serra, G., Cucchiara, R., 2016. A deep multi-level network for saliency prediction. In: 2016 23rd International Conference on Pattern Recognition. ICPR, IEEE, pp. 3488–3493.
Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al., 2018. Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 720–736.
Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., Damen, D., 2022. Epic-kitchens visor benchmark: Video segmentations and object relations. Adv. Neural Inf. Process. Syst. 35, 13745–13758.
De Haas, B., Iakovidis, A.L., Schwarzkopf, D.S., Gegenfurtner, K.R., 2019. Individual differences in visual salience vary along semantic dimensions. Proc. Natl. Acad. Sci. 116 (24), 11687–11692.
Del Molino, A.G., Tan, C., Lim, J.-H., Tan, A.-H., 2016. Summarization of egocentric videos: A comprehensive survey. IEEE Trans. Hum.-Mach. Syst. 47 (1), 65–76.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.
Ding, X., Chen, Z., 2022. Towards mesh saliency in 6 degrees of freedom. Neurocomputing 502, 120–139.
Duan, H., Shen, W., Min, X., Tu, D., Li, J., Zhai, G., 2022. Saliency in augmented reality. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 6549–6558.
Eisenberg, M.L., Zacks, J.M., 2016. Ambient and focal visual processing of naturalistic activity. J. Vis. 16 (2), 5.
Fang, Y., Yu, J., Zhang, H., van der Lans, R., Shi, B., 2024. OAT: Object-level attention transformer for gaze scanpath prediction. arXiv preprint arXiv:2407.13335.
Fathi, A., Li, Y., Rehg, J.M., 2012. Learning to recognize daily actions using gaze. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12. Springer, pp. 314–327.
Fathi, A., Ren, X., Rehg, J.M., 2011. Learning to recognize objects in egocentric activities. In: CVPR 2011. IEEE, pp. 3281–3288.
Furnari, A., Farinella, G.M., 2020. Rolling-unrolling lstms for action anticipation from first-person video. IEEE Trans. Pattern Anal. Mach. Intell. 43 (11), 4021–4036.
Ghadiyaram, D., Tran, D., Mahajan, D., 2019. Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12046–12055.
Gilani, S.O., Subramanian, R., Yan, Y., Melcher, D., Sebe, N., Winkler, S., 2015. Pet: An eye-tracking dataset for animal-centric pascal object classes. In: 2015 IEEE International Conference on Multimedia and Expo. ICME, IEEE, pp. 1–6.
Girdhar, R., Grauman, K., 2021. Anticipative video transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13505–13515.
Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al., 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012.
Haskins, A.J., Mentch, J., Botch, T.L., Robertson, C.E., 2020. Active vision in immersive, 360 real-world environments. Sci. Rep. 10 (1), 14304.
Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.
Itti, L., Koch, C., 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. Vis. Res. 40 (10–12), 1489–1506.
Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20 (11), 1254–1259.
Jiang, M., Huang, S., Duan, J., Zhao, Q., 2015. Salicon: Saliency in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1072–1080.
Judd, T., Durand, F., Torralba, A., 2012. A benchmark of computational models of saliency to predict human fixations.
Judd, T., Ehinger, K., Durand, F., Torralba, A., 2009. Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision. IEEE, pp. 2106–2113.
Kazakos, E., Nagrani, A., Zisserman, A., Damen, D., 2019. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5492–5501.
Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026.
Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A., 2011. Fast unsupervised ego-action learning for first-person sports videos. In: CVPR 2011. IEEE, pp. 3241–3248.
Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M., 2012. Activity forecasting. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part IV 12. Springer, pp. 201–214.
Kümmerer, M., Theis, L., Bethge, M., 2014. Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045.
Kümmerer, M., Wallis, T.S., Bethge, M., 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563.
Lee, Y.J., Ghosh, J., Grauman, K., 2012. Discovering important people and objects for egocentric video summarization. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1346–1353.
Lee, Y.J., Grauman, K., 2015. Predicting important objects for egocentric video summarization. Int. J. Comput. Vis. 114 (1), 38–55.
Li, A., Chen, Z., 2018. Personalized visual saliency: Individuality affects image perception. IEEE Access 6, 16099–16109.
Li, Y., Liu, M., Rehg, J.M., 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 619–635.
Li, Y., Nagarajan, T., Xiong, B., Grauman, K., 2021. Ego-exo: Transferring visual representations from third-person to first-person videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6943–6953.
Linardos, A., Kümmerer, M., Press, O., Bethge, M., 2021. DeepGaze IIE: calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10–17, 2021. IEEE, pp. 12899–12908. http://dx.doi.org/10.1109/ICCV48922.2021.01268.
Liu, M., Tang, S., Li, Y., Rehg, J.M., 2020. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer, pp. 704–721.
Lu, Z., Grauman, K., 2013. Story-driven summarization for egocentric video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2714–2721.
Luo, X., Liu, Z., Wei, W., Ye, L., Zhang, T., Xu, L., Wang, J., 2022. Few-shot personalized saliency prediction using meta-learning. Image Vis. Comput. 124, 104491.
Marois, R., Ivanoff, J., 2005. Capacity limits of information processing in the brain. Trends Cogn. Sci. 9 (6), 296–305.
Miranda, A.M., Nunes-Pereira, E.J., Baskaran, K., Macedo, A.F., 2018. Eye movements, convergence distance and pupil-size when reading from smartphone, computer, print and tablet. Scand. J. Optom. Vis. Sci. 11 (1), 1–5.
Negi, S., Mitra, R., 2020. Fixation duration and the learning process: an eye tracking study with subtitled videos. J. Eye Mov. Res. 13 (6).
Niebur, E., Koch, C., 1995. Control of selective visual attention: Modeling the “where” pathway. Adv. Neural Inf. Process. Syst. 8.
Olshausen, B.A., Anderson, C.H., Van Essen, D.C., 1993. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci. 13 (11), 4700–4719.
Pirsiavash, H., Ramanan, D., 2012. Detecting activities of daily living in first-person camera views. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2847–2854.
Rai, Y., Gutiérrez, J., Le Callet, P., 2017. A dataset of head and eye movements for 360 degree images. In: Proceedings of the 8th ACM on Multimedia Systems Conference. pp. 205–210.
Roth, N., Rolfs, M., Hellwich, O., Obermayer, K., 2023. Objects guide human gaze behavior in dynamic real-world scenes. PLoS Comput. Biol. 19 (10), e1011512.
Samiei, M., Clark, J.J., 2022. Predicting visual attention and distraction during visual search using convolutional neural networks. arXiv preprint arXiv:2210.15093.
Sherkati, R., Clark, J.J., 2023. Clustered saliency prediction. In: 34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20–24, 2023. BMVA.
Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., Wetzstein, G., 2018. Saliency in VR: How do people explore virtual environments? IEEE Trans. Vis. Comput. Graphics 24 (4), 1633–1642.
Snowden, R.J., Snowden, R., Thompson, P., Troscianko, T., 2012. Basic Vision: an Introduction to Visual Perception. Oxford University Press.
Song, R., Zhang, W., Zhao, Y., Liu, Y., Rosin, P.L., 2023. 3D visual saliency: An independent perceptual measure or a derivative of 2d image saliency? IEEE Trans. Pattern Anal. Mach. Intell. 45 (11), 13083–13099.
Su, Y.-C., Grauman, K., 2016. Detecting engagement in egocentric video. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part V 14. Springer, pp. 454–471.
Treisman, A., 1988. Features and objects: The fourteenth bartlett memorial lecture. Q. J. Exp. Psychol. Sect. A 40 (2), 201–237.
Trevarthen, C.B., 1968. Two mechanisms of vision in primates. Psychol. Forschung 31 (4), 299–337.
Unema, P.J., Pannasch, S., Joos, M., Velichkovsky, B.M., 2005. Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. Vis. Cogn. 12 (3), 473–494.
Velichkovsky, B.M., Rothert, A., Kopf, M., Dornhöfer, S.M., Joos, M., 2002. Towards an express-diagnostics for level of processing and hazard perception. Transp. Res. Part F Traffic Psychol. Behav. 5 (2), 145–156.
Vig, E., Dorr, M., Cox, D., 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2798–2805.
Võ, M.L.-H., 2021. The meaning and structure of scenes. Vis. Res. 181, 10–20.
Wang, W., Tran, D., Feiszli, M., 2020. What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12695–12705.

Wolfe, J.M., 1994. Visual search in continuous, naturalistic stimuli. Vis. Res. 34 (9), 1187–1195.

Wolfe, J.M., Horowitz, T.S., 2017. Five factors that guide attention in visual search. Nat. Hum. Behav. 1 (3), 0058.

Xu, Y., Gao, S., Wu, J., Li, N., Yu, J., 2018. Personalized saliency and its prediction. IEEE Trans. Pattern Anal. Mach. Intell. 41 (12), 2975–2989.

Xu, Y., Li, N., Wu, J., Yu, J., Gao, S., 2017a. Beyond universal saliency: Personalized saliency prediction with multi-task CNN. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 3887–3893. http://dx.doi.org/10.24963/ijcai.2017/543.

Xu, Y., Li, N., Wu, J., Yu, J., Gao, S., 2017b. Beyond universal saliency: Personalized saliency prediction with multi-task CNN. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 3887–3893. http://dx.doi.org/10.24963/ijcai.2017/543.

Xu, M., Yang, L., Tao, X., Duan, Y., Wang, Z., 2021. Saliency prediction on omnidirectional image with generative adversarial imitation learning. IEEE Trans. Image Process. 30, 2087–2102.

Yu, B., Clark, J.J., 2017. Personalization of saliency estimation. arXiv preprint arXiv:1711.08000.

Zelinsky, G.J., 2008. A theory of eye movements during target acquisition.. Psychol. Rev. 115 (4), 787.

Zhou, Y., Berg, T.L., 2015. Temporal perception and prediction in ego-centric video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4498–4506.

Zhu, Y., Zhu, D., Yang, Y., Duan, H., Zhou, Q., Min, X., Zhou, J., Zhai, G., Yang, X., 2019. A saliency dataset of head and eye movements for augmented reality. arXiv preprint arXiv:1912.05971.