The Mental Image Revealed by Gaze Tracking

Xi Wang
TU Berlin

Andreas Ley
TU Berlin

Sebastian Koch
TU Berlin

David Lindlbauer
TU Berlin / ETH Zurich

James Hays
Georgia Institute of Technology

Kenneth Holmqvist
Universität Regensburg

Marc Alexa
TU Berlin

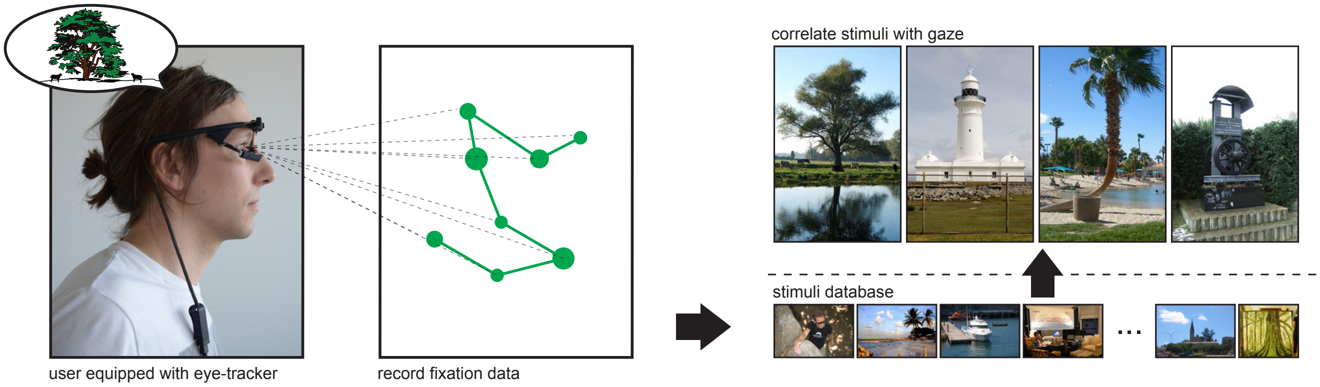


Figure 1: We use eye movements of a person recalling an image while looking at nothing (a white wall in this case) to retrieve a set of potentially matching images from a database.

ABSTRACT

Humans involuntarily move their eyes when retrieving an image from memory. This motion is often similar to actually observing the image. We suggest to exploit this behavior as a new modality in human computer interaction, using the motion of the eyes as a descriptor of the image. Interaction requires the user's eyes to be tracked, but no voluntary physical activity. We perform a controlled experiment and develop matching techniques using machine learning to investigate if images can be discriminated based on the gaze patterns recorded while users merely recall an image. Our results indicate that image retrieval is possible with an accuracy significantly above chance. We also show that these results generalize to images not used during training of the classifier and extends to uncontrolled settings in a realistic scenario.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; *Interaction techniques*; *Interaction paradigms*;

KEYWORDS

gaze pattern, mental imagery, eye tracking

ACM Reference Format:
Xi Wang, Andreas Ley, Sebastian Koch, David Lindlbauer, James Hays, Kenneth Holmqvist, and Marc Alexa. 2019. The Mental Image Revealed by Gaze Tracking. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300839

1 INTRODUCTION

Imagine that you are thinking about your vacation photos from last summer. After 5 seconds a photo appears in front of you, from that vacation, very similar to the moment you just recalled.

We believe such a seemingly magical task is possible through the use of eye tracking. It has been observed for over a century that humans move their eyes while thinking about images [21, 44, 49], more precisely, recalling images from memory. While this perhaps surprising effect has long been thought of as an odd superfluous expenditure of energy, it has been suggested in the 1960's that it is connected to the active construction and organization of the image [18, 46].

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05...$15.00
https://doi.org/10.1145/3290605.3300839

This hypothesis has since been rigorously and repeatedly proven [3, 23–25, 36, 37, 51, 56, 60]. It is now clear that eye movements during the recall of an image serve as a *spatial index*. This means gaze during recall is connected to the location of features in the image. This connection, in turn, suggests that gaze might contain enough information to retrieve a (known) image a user merely thinks about.

On the other hand, gaze data is coarse. It has to be expected that gaze data for different images may be very similar. To make matters even more difficult, it is known that eye movements during imagery are *not* reinstatements of those made during perception [24, 56]. Moreover, there is a large variability of eye movements during mental imagery across users. A common effect is that eye movements during imagery are of smaller magnitude compared to those during perception [3, 15, 22, 23]. We discuss more details in Section 2 and limitations imposed by these characteristics in Section 7.

In order to gauge the feasibility of image retrieval based on gaze data during mental imagery we set up a controlled lab experiment (see Section 3)[1]. Participants are viewing images while eye tracking is employed. Directly after an image is shown, the screen turns to gray and participants are asked to internally recall the image while keeping their eyes open. We record the eye motions during both image perception and recall using a video-based eye tracker. Based on the collected data we evaluate different retrieval scenarios. In all cases we essentially ask: how well can images be computationally discriminated from other images based on only gaze data.

The scenarios differ based on what gaze data is used as a query and what gaze data is used for matching. We consider different combinations which can be generally divided into two scenarios: in Scenario 1, we follow the idea of exploring available information contained in the data; in Scenario 2, we test a realistic setting in applications, which allows for the possibility of extension to new users.

We develop two types of retrieval algorithms for these scenarios. Restricting ourselves to using spatial histograms of the data, we consider an extended version of earth movers distance (EMD) and, at least for the scenarios that provide enough data, we also use deep neural nets. In general, we find that retrieval is feasible, albeit the data from looking at nothing is challenging, and the resulting performance varies significantly across scenarios and observers.

Based on the promising results in a lab setting we make a first step towards a real application (see Section 6): we sent several participants with a mobile eye tracker to a staged 'museum' exhibiting paintings. After their tour, we ask them to recall some of the images while looking at a blank whiteboard. We find that the median rank in a classification approach is small, showing that the idea is promising for practical use.

Despite the encouraging results, our proposed new modality still faces challenges that require further investigations. We discuss them in detail in Section 8 together with possible future applications.

2 BACKGROUND & RELATED WORK

Eye movements during imagery

Early studies [44, 49] reported a large amount of eye movement activity during visual imagery, which suggested a tight link between eye movements and mental images. Today, a large body of research has shown that spontaneous eye movements frequently occur when scenes are recalled from memory and that such gaze patterns closely overlap with the spatial layout of the recalled information [3, 23, 36].

It has been suggested that eye movements to blank spaces, or the processes driving them, function as "spatial indices" [1, 51] that may assist in arranging the relevant parts of a visualized scene [3, 18, 36, 46], and it has been argued that such looks at nothing can act as facilitatory cues during memory retrieval [14, 50].

But characteristics of eye movements during recalling pose four challenging issues on our idea of constructing an eye-movement based mechanism for image retrieval. Firstly, although imagery eye movements are functional, the eye movements during imagery are *not* reinstatements of those made during perception [24, 56]. Secondly, it is very possible that covert attentional movements occur during imagery that may account for the functional role of eye movements to nothing [55]. Thirdly, some people prefer to close their eyes during recall, an effect so far not fully understood [41, 62], which makes eye-movement recordings with video-based eye-trackers unpractical. Fourth, several studies have reported a large variability in the extent of the imagery eye movements of participants. Typically, some participants scale down the size of the area covered by imagery eye movements compared to the area covered by the original image [15, 23].

Gaze interaction

Despite the well-known 'Midas Touch' problem [20], gaze still offers an attractive option as input, since it is comparatively fast and accurate. Especially in hand-free interaction systems, gaze data provides an additional input modality. As an input method, gaze is used to indicate users' intention [19, 35], and to type words [40] and passwords [4]. Other types of applications are based on the characteristic of eye movements captured by dwell time [45] or smooth pursuit [13].

[1] Gaze data and source code can be found on the project page http://cybertron.cg.tu-berlin.de/xiwang/mental_imagery/retrieval.html

For a more comprehensive review of eye-based interaction, we refer readers to work by Majaranta and Bulling [39].

Most similar to our endeavor are attempts at learning something about the image or the observer by simply processing the gaze in a free viewing task. Fixations have been used to select images from a collection to indicate the desired attributes of a search target [54]; or gaze data is exploited to identify fine-grained differences between object classes [30]. Comparing to previous work, where users are mostly required to *consciously* move their eyes, the proposed interaction system relies more on *unconscious* eye movements.

Brain-computer interfaces

The idea of using brain activity for direct communication with the computer is intriguing. Early attempts based on electroencephalogram (EEG) go back to 1970s [61], and first consolidations of various research efforts [64] argue that *reading the mind* will be difficult and the focus of attention should be on individuals training to steer the EEG for providing certain commands. Today EEG is used in applications such as identity authentication [42] and studies about brain responses [16]. Despite the bandwidth of communication using EEG still seems to be very limited, the trend appears to be combining continuous EEG data with other modalities [38, 66].

Likewise, there have been attempts at reading a person's mental state using functional magnetic resonance imaging (fMRI) [17, 28]. The task of reconstructing mental images has generated much interest and shows promising results [11, 32, 47, 48, 58]. These results are similar to what we want to achieve: guess the image one is recalling from memory.

Many of EEG based applications are restricted to binary classification. Participants are instructed to imagine moving their left or right hand [2], which needs to be repeated many times to give a sufficient signal-to-noise ratio [43]. Yet it opens up a possibility of communication for patients with aggravating conditions [38, 66]. Participants in studies using fMRI are required to keep still so that acquired images can be aligned. Such restriction of movements during fMRI scanning poses another challenge (besides costs) in terms of practical usage. Our work is based on tracking the gaze patterns during mental imagery, which is more feasible for practical purposes.

Photo management and retrieval

Managing photos and image collections can be difficult for users (cf. [34]), oftentimes requiring context-based organization and sorting. Especially retrieval of individual images can pose a significant challenge and requires advanced interfaces and interactions (cf. e.g. [33, 59]). Sketches are probably the oldest way to communicate mental images. Unlike in image recall, gaze patterns are significantly distorted and incomplete compared to the real visual appearance of an object. Information in a sketch is much richer and more coherent than the data we have to deal with. Nevertheless sketches have been shown to work well for retrieving images [12, 53, 65]. In contrast to sketching, we explore how users can retrieve an image without explicit interaction, which can be beneficial in situations where sketching is not available (e.g. because the hands are occupied).

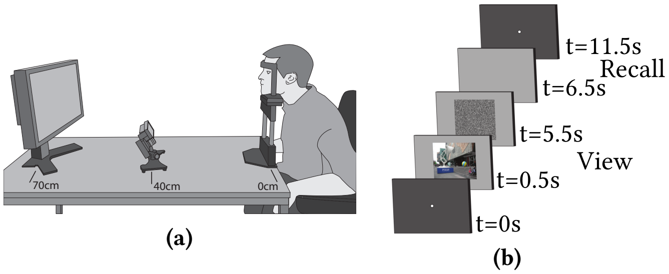


Figure 2: (a) Experimental apparatus. (b) Trial sequence.

3 EXPERIMENT

The two key components of our method are encoding an image while viewing it and then recalling it in front of a homogeneous background. The goal of this experiment is to collect a large dataset of pairs of eye movements during viewing and recall. The experimental design (timing, conditions) follows similar experiments in previous imagery studies [25, 36].

Setup

Apparatus. As visual stimuli, 100 images are randomly chosen from the MIT data set [27]. The selected dataset contains both indoor and outdoor scenes. Images were presented on a calibrated 24-inch display with a resolution of 1920 × 1200 pixels at a distance of $0.7m$. Gaze was tracked with an EyeLink 1000 in remote mode and standard 9-point calibration, using a chin and forehead rest to stabilize participants. The experiments were conducted in a quiet, dim and distraction-free room.

Participants. We recruited 30 participants (mean age = 26 years, SD = 4 years, 9 females). They had normal or corrected to normal visual acuity and no (known) color deficiencies. 10 participants wore contact lenses. No glasses were allowed due to concerns about eye tracking accuracy. Two observers failed to calibrate the eye tracker with the required accuracy, which left us with a dataset of 28 observers viewing and recalling 100 images. Importantly, all participants were naive with respect to the purpose of the experiment. Consent was given before the experiment and participants were compensated for their time.

Protocol. Each individual trial involved a $500ms$ fixation dot, followed by an image presented for $5000ms$, a $1000ms$ noise mask, and finally a $5000ms$ gray image for recall (Figure 2b). Participants were instructed to inspect the image when it is shown and then think about the image when the empty gray display is shown. All instructions for encoding and recall were given during an initial round of 10 trials for each observer. No further instructions regarding the task were given after this training.

The 100 images are shuffled in a random order for each observer and then divided in 5 blocks. At the onset and after each block of 20 trials, the eye tracker was calibrated. Calibration was accepted if accuracy in the following validation was below 0.5°. Each participant thus viewed five blocks of 20 trials and a total of 100 images and 100 subsequent recalls of the same images.

After finishing all five blocks, observers were asked to look at another 10 images, 5 of which were among the 100 they has seen, and determine whether they have seen them earlier in the experiment. Except for one observer who made one mistake in the memory test, all participants could correctly distinguish viewed images from new ones, and this suggests that the image contents were still in the observers' memory.

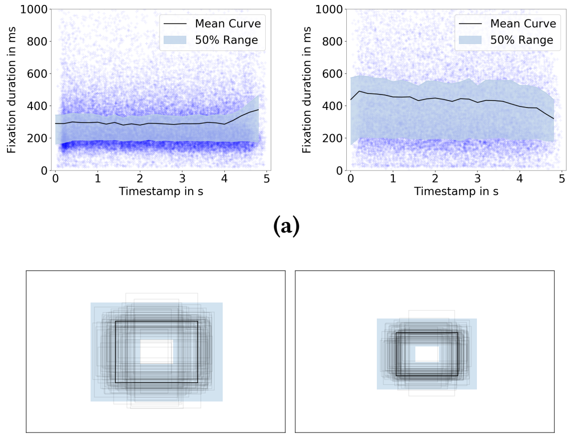
Analysis

Eye movements during visual imagery have characteristics different from eye movements during perception. As shown in Figure 3a, fixation durations on mental images ($452.2 \pm 308.0ms$) are longer ($t(6.7e4) = -57.29, p < .001$, Welch's t-test) than fixations on real images ($278.0 \pm 73.4ms$). Consequently, there are fewer fixations during recall (16 ± 2.8 in encoding vs. 11 ± 3.6 in recall, $t(5.6e3) = 62.77, p < .001$, Welch's t-test), possibly because memory retrieval requires additional processing time compared to perception [23, 25].

A comparison of the spatial distribution of fixations confirmed that observers tend to shrink the overall extent covered by their eye movements while thinking about an image, as shown in Fig. 3b. This is in line with previous studies [15, 22, 23]. Hence, most fixations during imagery/recall are inaccurate relative to the location of objects in the actual image, while in viewing/encoding, fixations are virtually always perfectly aligned with intended image features. This discrepancy implies that there is no obvious one-to-one correspondence between absolute fixation locations during encoding and recall.

4 RETRIEVAL

Retrieval is based on two steps: first the raw eye tracking data is turned into a more compact representation; then, we try to assign a query representation to one of the 100 image classes. For the second part, we consider different approaches, either measuring the distance between pairs of data representations using an appropriate distance metric, or using deep neural nets. Depending on what gaze data is used for query and matching, we have different retrieval scenarios. In particular, we consider the following combinations:



(a)

(b)

Figure 3: Eye movements characteristics. (a) Fixation durations (left encoding and right recall) are plotted as a function of starting time. The black curves indicate the mean durations and the center 50 percent intervals are depicted in blue. (b) Spatial distribution of fixations averaged over all observers plotted in screen coordinates (left encoding and right recall). Each bounding box of fixations in a single trial is visualized in light gray and the average is marked by the black box with a standard deviation shaded in blue.

(1) Within the same condition: the query gaze data and the gaze data used for matching are both taken from the observers while
  (a) looking at the images or
  (b) looking at nothing.
  In these situations we always consider matching across *different observers*. This means that when querying with data from a specific observer, his/her data is *not* included for matching.
(2) Across conditions: the query data is taken from an observer looking at nothing and it is matched against gaze data from looking at the images. Here we differentiate between matching the query data
  (a) against the gaze data from all other observers or
  (b) against the gaze data from the same observer.

Scenario 1 is meant to establish an idea about the available information contained in the data. Scenario 2, we believe, is a realistic setting in applications, where we want to use gaze data from imagining an image for retrieval while collecting viewing data from other users.

Data representation

We have experimented with a number of different representations and were unable to identify a representation that clearly outperforms others in terms of retrieval performance. For reasons of a clear exposition we use *spatial histograms*, as they consistently performed well, are simple to explain, and are a natural input for neural nets.

A spatial histogram is defined on a regular lattice with $k = m \times n$ cells. We denote the centers of the gird cells as $\{\mathbf{c}_i\}, 0 \leq i < k$. To each cell we associate a weight w_i. The weights are computed based on the *number* of *raw* gaze samples (returned from the eye tracker) that fall into each cell, i.e., that are closer to $\mathbf{c}_i$ than to any other center $\mathbf{c}_j, j \neq i$. Because of the grid structure, we may consider the set of weights $\mathbf{w}$ as intensities of a discrete image. This image is sometimes called *heat map* in the literature on eye tracking.

We are aware that it is common to first extract fixations from the raw gaze data before further processing. We have opted not to do this, as (1) fixations are not defined in a way that could be universally accepted, (2) the spatial histogram is dominated by the fixations anyways, and (3) saccade data may or may not be useful in classification. In an initial test using MultiMatch [10], a tool developed in the eye tracking community for comparing gaze sequences based on fixation data, we found that the retrieval tasks we are interested in performed only at chance level.

It has been observed that the sequence information could be useful [30] but we could not reproduce this observation. This makes sense, as it has been observed that the order of where people look during imagery is not identical to the order during looking at the image [22].

Distance measure

The nature of the data suggests that simple measures of distance are unable to capture the desired correspondences. Indeed, we found that simple Euclidean distances on the weight vectors of spatial histograms, i.e. $d(\mathbf{w}, \mathbf{w}') = \|\mathbf{w} - \mathbf{w}'\|$ are not promising. The reasons for this are that the fixations in one gaze sequence are commonly a subset of the fixations in the other sequence. Moreover, the data for mental images is spatially distorted and contains elements that are unrelated to the original stimulus.

Research on fixations and image saliency has investigated various ways to measure the agreement (or disagreement) of real fixations with computational predictions [26]. Bylinskii et al. [5] observe that the Earth Mover's Distance (EMD) [52] tolerates some positional inaccuracy. More generally, EMD is well known to be robust to deformation and outliers for comparing images.

In our context the weights $\mathbf{w}$ and $\mathbf{w}'$ are representative for the gaze data. The *flows* $\mathbf{F} = \{f_{ij}\}$ describe how much of the weight w_i is matched to w'_j. Based on the idea that $\mathbf{w}'$ is potentially a subset of $\mathbf{w}$ we scale the weight vectors so that

$$1 = \|\mathbf{w}\|_1 \geq \|\mathbf{w}'\|_1 = \sigma \tag{1}$$

The flow is constrained to relate all of the weights in $\mathbf{w}'$, but not more than w_i, to i:

$$\sum_i f_{ij} = w'_j, \quad \sum_j f_{ij} \leq w_i. \tag{2}$$

This implies $\sum_{ij} f_{ij} = \|\mathbf{w}'\|$. Among the flows satisfying these constraints (i.e., the flows that completely match $\mathbf{w}'$ to a part of $\mathbf{w}$) one wants to minimize the flow. Therefore we need to specify how far apart the elements i and j are by a distance *metric* $\{d_{ij}\}$. The resulting minimization is:

$$\arg\min_{\mathbf{F}} \sum_i \sum_j f_{ij} d_{ij}. \tag{3}$$

In many cases it is natural to use Euclidean or squared Euclidean distance between the cell centers $\mathbf{c}_i$ and $\mathbf{c}_j$ to define d_{ij}. On the other hand, we cannot be sure that the space in which the recall sequences 'live' is aligned with other such spaces, or the fixed reference frame of the images. In line of this, learning has been used to optimize the metric $\{d_{ij}\}$ for EMD [9, 63]. Our idea in this context is to restrict the potential mapping to be affine. This means we define the distances to be

$$d_{ij} = \|\mathbf{c}_i - \mathbf{T}\mathbf{c}_j\|_2^2 \tag{4}$$

where $\mathbf{T}$ is an arbitrary affine transformation. We optimize for the flows $\mathbf{F}$ and the transformation $\mathbf{T}$ in alternating fashion. With fixed transformation $\mathbf{T}$ this is the standard EMD problem and can be efficiently computed for the size of data we have to deal with. To optimize $\mathbf{T}$ we consider the matching pairs of viewing and recall sequences. Based on the given flows, computing an affine transformation for the *squared* distances is a linear problem. This procedure typically converges to a local minimum close to the starting point [7], which is desirable in our application, as we expect the transformation to be close to identity. In any case, the minimization indirectly defines the resulting distance as

$$d(\mathbf{w}, \mathbf{w}') = \frac{1}{\sigma} \sum_i \sum_j f_{ij} d_{ij} \tag{5}$$

Distance-based retrieval

Based on a distance measure d between spatial histograms represented by their weight vectors $\mathbf{w}, \mathbf{w}'$, we can perform retrieval, assuming $\mathbf{w}'$ represents the query data.

The simplest case is scenario 2b, where retrieval is restricted to a single observer. So we have 100 spatial histograms $\mathbf{w}_i$ representing the gaze sequences while looking at the images, and a single histogram $\mathbf{w}'$ representing a recall sequence as a query. Computing the 100 distances $d(\mathbf{w}_i, \mathbf{w}')$ allows ranking the images.

In the other three scenarios query data from a single observer is matched to the gaze data from other observers. In these scenarios we base our approach on *leave-one-out* cross validation, meaning we alway use all the data from the remaining 27 observers for matching. Let $\mathbf{w}_i^k$ be the spatial histogram for image i provided by observer k. For each image we compute the smallest distance of the query $\mathbf{w}'$ to each image across all observers:

$$d_i(\mathbf{w}') = \min_k d(\mathbf{w}_i^k, \mathbf{w}'). \tag{6}$$

Then the ranking is based on $d_i(\mathbf{w}')$. Note that the first rank in this case is the same as using the nearest neighbor in the space of spatial histograms with the distance measure defined above.

Convolutional Neural Networks

The data we have collected encompasses 2700 gaze data histograms–we felt this number justifies trying to learn a classifier using the currently popular *Convolutional Neural Networks* (CNN). We design the architectures to have few parameters to reduce overfitting on the data, which is still rather small comparing to the typical number of parameters in CNNs.

Network layout. The basic setup for the CNN is similar to the ones used for image classification and visualized in Figure 4a: Each convolution filters the previous layer with learned 3×3 filter kernels and produces a new image with as many channels as filter kernels are used. As is common, we combine each convolution with *Batch Normalization* (BN) and *ReLU* non-linearity. After two blocks of convolution, we perform *Max Pooling* to reduce spatial size. After two blocks of max pooling, the spatial size is down to 3×3 elements at which point we flatten and employ regular fully connected layers. The first of these two fully connected layers is again using BN and ReLU. The last one, however, directly feeds into a softmax layer to produce a probability distribution over the 100 image classes. To improve generalization, we employ dropout layers throughout the network with a dropout probability of 30% in the convolutional layers and 20% in the fully connected layer.

Application to scenario 1. The network can be directly used to perform retrieval within the same condition (scenarios 1a and 1b), as the query data is of the same type as the data used for generating the network. In this case we train using Cross Entropy Loss for 50 epochs using the Adam parameter update scheme and a batch size of 100. Classification accuracy is tested, as above, using leave-one-out cross validation.

Extension for scenario 2. For working with histograms coming from different processes, it seems better to learn independent encodings. Instead of mapping the histograms directly to their respective image index, and thus casting image retrieval as a classification task, we rather perform image retrieval by learning proper encodings and then comparing them based on distance. In other words, we train an embedding of the histograms from the two conditions into a low dimensional descriptor space such that matching pairs are close together, and non-matching pairs are far apart.

The network architecture (see Figure 4a) is similar to the classification architecture detailed above. The difference is in the lack of BN layers, as they could not be used in this training setup, the removal of the softmax output, and the reduction from 100 outputs to just 16. We simultaneously train two instances of this architecture, one for each condition. Both map histograms to 16 dimensional descriptors.

The *Triplet Loss* [57], employed during training, forces the networks to produce descriptors which, for matching pairs, are more similar (based on Euclidean distance) than for non-matching pairs. We used two triplet losses in tandem with four descriptor computations to balance the training (see Figure 4b).

A ranking for a query histogram can be computed by generating its representation in the low dimensional descriptor space and then using Euclidean distance. The procedure for selecting the best match is identical to the one explained above for the distances computed based on EMD.

5 RESULTS

Using the two retrieval methods we now present results for different scenarios. They are described based on the rank of the queried image among all images. Based on the ranks we form a curve that describes the percentage of the 100 different queries (one for each image) to be matched if we consider the first x ranks ($x \leq 100$). This is essentially a Receiver Operating Characteristic (ROC) curve. The leave-one-out cross validation generates one curve for each observer, so 28 curves in total. We consider the mean curve as the overall retrieval result and also provide the 50% interval in the visualization (see Figures 5, 6, and 7).

For a good retrieval algorithm, the ROC curve climbs up steeply at the beginning and quickly reaches the highest point. This would mean that for most queries the correct image match is among the top ranked retrieval results. The overall retrieval accuracy is then measured by the standard area under the curve (AUC). If retrieval was based on chance, the expected rank for retrieval would be 50 among the 100 images. The chance for being ranked first would be 1%, the chance to be among the first 5 would be 5%, and so on. This means the ROC curve for chance retrieval would be the bisector line with AUC = 0.5.

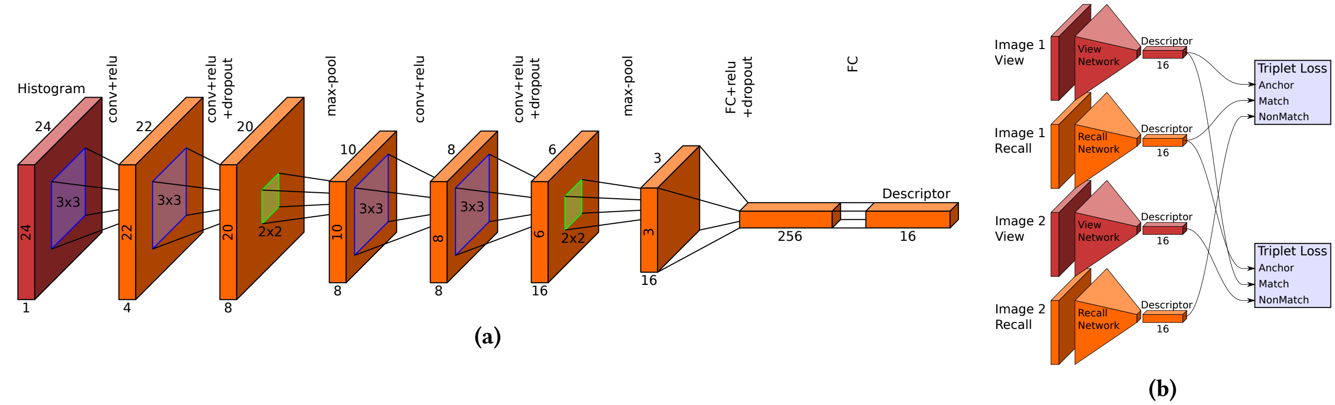


Figure 4: (a) CNN architectures employed for historgam based matching. Numbers around layers indicate width and height, as well as the number of channels (written below). (b) Descriptor learning setup with two triplet losses. For two different images, encoding and recall histograms were fed through their respective networks (truncated pyramids) to produce 16-dimensional descriptors. Two triplet losses were used to force matching descriptors to be closer to each other (based on Euclidean distance) than non-matching descriptors.

Both methods depend on the size of the histogram, and the distance measure provides the additional parameter σ. We also checked how enabling the affine transformation influences the results. In general, the results in terms of the AUC are not varying much with these parameters and we make the following recommendations:

- Choosing k too small slightly degrades the results and we suggest using histograms of size $k \geq 16 \times 16$.
- Allowing a partial match improves the result, i.e. choosing $\sigma = 1$ is suboptimal. We have settled for $\sigma = 0.5$.
- Allowing a global affine transformation shows improvements in retrieval rate for some of the scenarios. This means learning a global transformation matrix could potentially adjust the deformed recall sequences.

Scenario 1

The tentatively easiest case is retrieving an image based on the gaze data while looking at the image, matched against similarly obtained gaze data. Indeed, we find that the AUC for the distance measure is 96.3% (Fig. 5, U=9.73e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). In particular, the top ranked images is correct in 52.2% of the cases, and the correct match is among the top 3 in 72.4%. These results are largely independent of the choice of parameters and the use of the affine transformation.

The CNN performs only slightly better with AUC = 97.5% (Fig. 5, U=9.86e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test), and achieved top-1 and top-3 ranks of 61.3% and 79.1%.

This demonstrates that visual images can be easily discriminated based on eye movements from observers exposed to those images, at least for a database of 100 images.

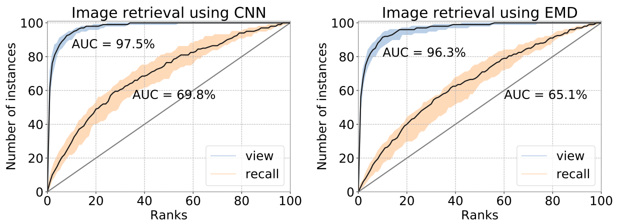


Figure 5: Retrieval performance using CNN (left) and EMD(right). The 50 percent center intervals of all ROC curves over all observers are depicted in light colors, viewing based retrieval in blue and recall based retrieval in orange.

If recall gaze data is matched against recall gaze data, the results are significantly worse. Using matching based on the distance measure we find AUC = 65.1% (U=7.08e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). The top-1 and top-3 ranks are 3.4% and 8.2% respectively. For the CNN we get AUC = 69.8% (Fig. 5, U=7.08e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). In 5.9% of the cases the system could correctly identify the image, and the top-3 rank dropped to 13.8% (chance would have been 1% and 3%). Notably, the variance of all ROC curves of retrieval based on recall gaze data is higher (SD = 7.13%, compared to SD = 2.42% for retrieval based on viewing gaze data).

Scenario 2

We perform cross-condition matching in scenario 2 by matching a recall query to viewing gaze data. When matching is performed against all other observers' gaze data, the performance drops to AUC = 68.4% as shown in Figure 6 on the left (U=6.937e3, n_1=n_2=100, P<10^{-5} two-tailed, Mann-Whitney rank test), with top-1 of 5.8% and top-3 of 12.8%.

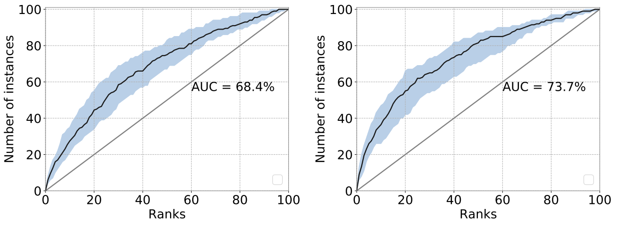


Figure 6: Retrieval performance in scenario 2. The left plot is matching result against viewing from all other observers and the right plot shows matching against data from the same observer. The 50 percent center intervals of all ROC curves over all observers are depicted in light colors.

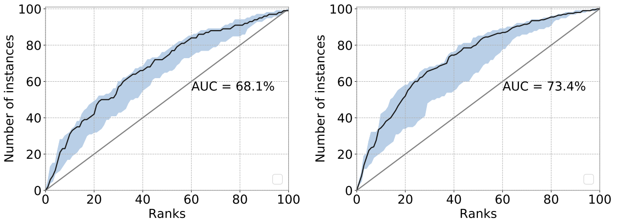


Figure 7: Retrieval performance in scenario 2 using CNN. The left plot is matching result against viewing from all other observers and the right plot shows matching against data from the same observer. The 50 percent center intervals of all ROC curves over all observers are depicted in light colors.

When matching is against the gaze data from the same observer, we achieved an AUC of 73.7% (U=7.47e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test) based on EMD distances (right plot in Figure 6). Top-1 and top-3 matches are 10.0% and 19.6% respectively.

Surprisingly, in this scenario the performance using CNN is not better than using EMD. Matching against data from all other observers gives AUC = 68.1% (U=6.91e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). Similar to the results based on EMD distance, the retrieval performance improves for matching against the data of the observer: AUC = 73.4% (U=7.44e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). We suspect this is due to the difficulty of the data in general and, in particular, due to the well known problem of CNN when dealing with global affine transformations. If there was a way to compensate for such global transformations in CNN, the results might be better. It would be interesting to continue with this idea in future work.

Discussion

Identifying an image based on gaze data during looking at the image works well, based on a variety of different approaches. It is clear, however, that this result is highly dependent on the set of images being used. If the 100 images evoke similar gaze patterns, they could not be discriminated based on gaze data. Our results indicate that gaze on the images we chose is different enough to allow for computational image retrieval. This is important, because if discrimination had been difficult based on gaze data from viewing, it would have been unlikely that gaze sequences during recall contained enough information for any task.

The results for retrieval based on gaze data from recall using data for recall to match indicate how severely distorted the recall data is. The performance significantly decreases and indicates that the task we believe is most important in applications, namely matching gaze data from recall against gaze data from viewing, may be very difficult.

The two scenarios we consider for matching recall data against viewing data show quite different results. Matching the recall data of an observer against their own viewing data works much better than matching against the viewing data from other observers. This indicates that viewing is idiosyncratic. The fact that observers agree more with themselves than with others is also consistent with the findings that fixations during recall are reenactments [1, 51].

To our knowledge, this is the first quantitative assessment of the amount of information in the recall gaze that can be used to identify images.

6 REAL-WORLD APPLICATION

We have established that the gaze pattern while only recalling an image could be used to identify the image using standard techniques from vision and learning. The data, however, was collected under artificial conditions. We are interested in performing a similar experiment, albeit this time under more realistic conditions.

A useful application could be retrieving one or more of the images seen from a museum visit. To explore if this is possible, we hung 20 posters in a seminar room (see Figure 8a), simulating a museum[2]. The images had slightly varying sizes at around $0.6m \times 1.0m$ on both portrait and landscape orientations. Their centroids were at a height of $1.6m$. For the recall phase, an (empty) whiteboard was used.

We recruited 5 participants (mean age = 32, 1 female) for the museum visit; only 1 had participated in the earlier experiment. Each of them was outfitted with a Pupil mobile eye tracker with reported accuracy of 0.6 degrees [31], equipped with two eye cameras at 120Hz and one front-facing camera at 30Hz. The eye tracker was calibrated on a screen prior to viewing the images. No chin rest was used during the calibration but participants were asked to keep their head still. As in the controlled experiment, we used a 9-point calibration and a display (1920 × 1200 pixels) placed at $0.7m$ distance.

[2] We initially planned to do this in cooperation with any of the large museums close by, but legal and administrative issues connected to the video camera in the eye tracker caused more complication then we felt the merely symbolic character was worth.

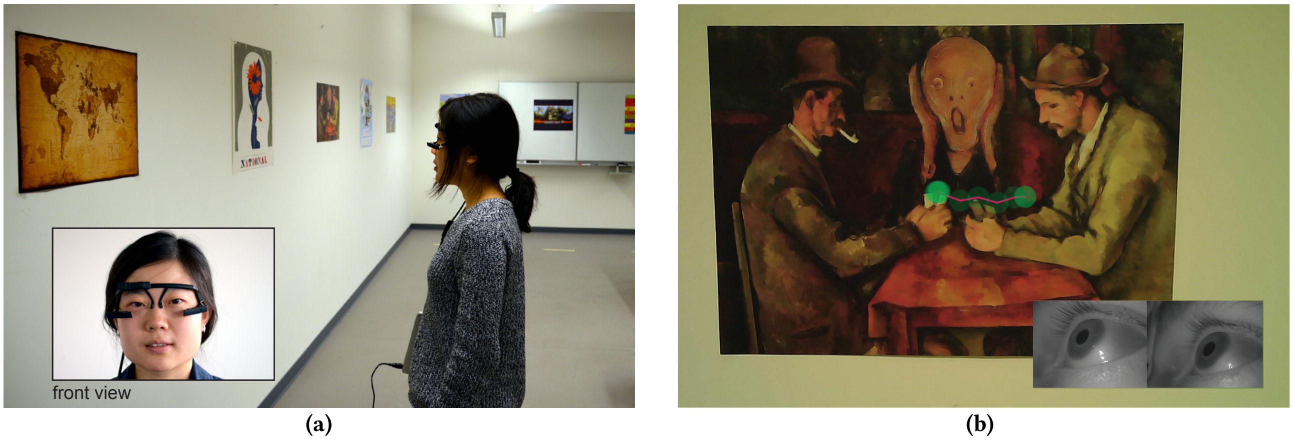


Figure 8: (a) The setup in the 'museum'-experiment. Participants view 20 images wearing an eye tracker. Later, they were asked to think about the images while looking at the (blank) whiteboard. Gaze patterns from this recall are used for identifying the image. (b) Example of video stream (scene and eye cameras). Green dots indicate fixations. Red lines represent the trajectory to the previous viewed position.

This also approximates the viewing distance between the visitor and the museum items. After calibration, positions of pupil center in eye cameras are mapped into the front-facing camera, yielding the resulting *gaze positions*. As long as the eye tracker stays fixed on the head, the calibrated mapping is valid. Participants were asked to inform us if they notice any displacement of the eye tracker.

After calibration the cameras were continuously recording. Participants were asked to view all the images, in no particular order or timeframe (but without skipping or ignoring any). Furthermore, participants were asked to obey markers on the floor indicating minimal distance to the images (similar to the rope in a museum). After viewing the images, participants were asked to recall images they liked, in any order, and as many as they wanted to recall. Each recall sequence started with instructions by the experimenter and ended with participants signaling completion.

We manually partitioned the resulting eye tracking data based on the video stream from the front facing camera. 20 viewing sequences are extracted for each participant and Figure 8b shows several frames of the data. The viewing duration varied greatly among participants, from few seconds to more than a minute per image. All participants recalled at least 5 images, with 10 being the highest number. Viewing and recall sequences are represented as variable-length gaze positions mapped in the front-facing camera. In total, each dataset from one participant contains gaze positions of 20 viewing sequences and five or more recall sequences.

For the analysis it turned out to be relevant what part of the viewing and recall sequences we would use. We considered the first 5, 10, 15, 20 seconds of the viewing sequence and the first 5, 8, 10, 12 seconds of the recall sequence. We represented the resulting raw data using k-means clustering with $k = 10$ and computed distances using EMD. The EMD based distance measure is able to compensate for global rigid transformation, which seems very important for the setting. The resulting ranks are above chance in all cases, yet different parameter settings worked best for different participants. The results did improve with optimization for a transformation. There was no significant advantage in including rotation or scaling, but translation was important.

The best median ranks for 5 recalled images were, in order from best to worst, 2, 2, 3, 4, 5 (without translation we found 2, 3, 3, 4, 7). So if duration can be adapted to each participant the results are surprisingly good. Applying any combination of duration to all 5 participants in the same way leads to worse overall results. Optimization for translation pays off in this situation as well, as the results get more stable against the choice of parameters: for a wide variety of combinations we find median ranks of 4 and 5. All the reported median ranks have to be seen in context of the median of a purely random retrieval being 10.

None of our participants had reported big movements during the experimental session and no secondary calibration was conducted. The camera rate of the eye tracker also seems to be irrelevant in our setting as long as meaningful eye movements are recorded. There is no special requirement for high-speed eye camera since saccades and other type of micro eye movements are not included in the analysis.

7 LIMITATION AND FURTHER WORK

By collecting a large eye movement dataset in the looking-at-nothing paradigm, we made the first attempt to computationally estimate the content of mental images from eye movements. Our results demonstrate that gaze patterns from observers looking at the photographic images contain a unique signature that can be used to accurately discriminate the corresponding photo. More excitingly, using gaze patterns of observers just thinking about an image, i.e. imagery/recall eye movements, achieved a reasonably good retrieval accuracy. While the proposed method introduces a novel interaction modality, a number of challenges and limitations have to be addressed to explore its full potential, discussed below.

Scalability

Retrieval depends on image content. Clearly, the more photos are added to the database, the more likely that several different photos will give rise to the same eye movement behavior, which inevitably affects performance. This is because image representations are essentially downsampled to gaze patterns, and similar representations lead to increased ambiguity. This would pose a limit to how many photos can be used with this retrieval method. In similar studies performed on fMRI data, Chadwick et al. [6] showed brain activity patterns can be used to distinguish among imagining only three film events and in [8] fMRI signals were used to reconstruct face images but only 30 images were used.

Moreover, it is very likely that if all participants would look around in the full extent of the monitor, when imagining the photo, the retrieval performance of the system would increase significantly. Maybe it would be possible to instruct participants to make more extensive eye movements than they would normally do, in order to help the computer find the right image? It is not unreasonable to expect a small effort on the part of the participant–which in many brain-computer interface applications is the norm.

Temporal stability

Our recall data were recorded immediately after image viewing. A longer delay might reduce the discrimination performance as the memory deteriorates. In future work, it would be interesting to explore the influence of memory decay and its effect on image retrieval from long-term memory.

Alternative sensing modalities

Our proposed technique relies on sensing the motion of the eyes over time. To gather this data, we used a video-based eye tracker, which relies on users' eye to be open, even when they look at a neutral surface during recall. In contrast, other sensing techniques, e.g. electrooculography (EOG), allow to sense eye movement when users' eyes are closed. This would also allow users to create a neutral background by simply closing their eyes, which might even increase the vividness of mental imagery [41, 62]. However, the lack of reference with closed eyes might introduce different types of distortions. We believe that exploring alternative sensing modalities such as EOG as replacement or additional data source will allow our concept to become more viable for everyday interactions. We aim to explore additional sensing modalities as well as their deployment in less controlled environments in the future.

Towards real-world applications

Our current work focuses on the evaluation of the proposed new interaction model along with the development of computational tools. We used image retrieval as an indicator for the success of understanding user's intention. From a practical point of view, it would be interesting to compare our method to existing techniques such as manual selection or speech-based interface.

How to accurately track eye movements using mobile eye trackers poses another challenge. In our museum visiting experiment, we did notice that the calibration accuracy gradually got worse. As the shifts of the eye tracker over time are rigid translations, our comparison methods should be able to compensate them. Such limitations reply on the further improvements of mobile eye tracking.

With the development in wearable devices, we believe tracking the motion of the eyes would be a natural by-product. Combination with other interaction modalities, such as the possibility offered by recent work in speech interface [29], offers a rich source of information. With additional sources of information, we believe that our interface would provide an improved interaction between users and software agents.

8 CONCLUSIONS

In this paper we present a new modality based on the involuntary activity of the ocular muscles during recall of a previously observed image. We have developed computational methods for matching eye movement data among different observers and different conditions. A controlled experiment together with an experiment in a museum-like setting have demonstrated the feasibility of the new modality.

Overall, this study provides evidence that eye-movement based image retrieval is computationally feasible. We have shown reasonable performance with naive participants, and we have good reason to believe that instructed participants who make a small effort to move their eyes more during imagery/recall can achieve very high accuracy levels for photo databases of 100 images or more.

REFERENCES

[1] Gerry TM Altmann. 2004. Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition* 93, 2 (2004), B79–B87.

[2] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Müller. 2008. Optimizing Spatial filters for

Robust EEG Single-Trial Analysis. *IEEE Signal Processing Magazine* 25, 1 (2008), 41–56. https://doi.org/10.1109/MSP.2008.4408441

[3] Stephan A Brandt and Lawrence W Stark. 1997. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience* 9, 1 (1997), 27–38. https://doi.org/10.1162/jocn.1997.9.1.27

[4] Andreas Bulling, Florian Alt, and Albrecht Schmidt. 2012. Increasing the Security of Gaze-based Cued-recall Graphical Passwords Using Saliency Masks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 3011–3020. https://doi.org/10.1145/2207676.2208712

[5] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2016. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605* (2016).

[6] Martin J. Chadwick, Demis Hassabis, Nikolaus Weiskopf, and Eleanor A. Maguire. 2010. Decoding Individual Episodic Memory Traces in the Human Hippocampus. *Current Biology* 20, 6 (2010), 544 – 547. https://doi.org/10.1016/j.cub.2010.01.053

[7] Scott Cohen and Leonidas Guibas. 1999. The Earth Mover's Distance under transformation sets. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2. 1076–1083. https://doi.org/10.1109/ICCV.1999.790393

[8] Alan S Cowen, Marvin M Chun, and Brice A Kuhl. 2014. Neural portraits of perception: reconstructing face images from evoked brain activity. *Neuroimage* 94 (2014), 12–22.

[9] Marco Cuturi and David Avis. 2014. Ground Metric Learning. *J. Mach. Learn. Res.* 15, 1 (Jan. 2014), 533–564. http://dl.acm.org/citation.cfm?id=2627435.2627452

[10] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. 2012. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior research methods* 44, 4 (2012), 1079–1100.

[11] Changde Du, Changying Du, and Huiguang He. 2017. Sharing deep generative representation for perceived image reconstruction from human brain activity. In *Neural Networks (IJCNN), 2017 International Joint Conference on.* IEEE, 1049–1056.

[12] Matthias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. 2011. Sketch-Based Image Retrieval: Benchmark and Bag-of-Features Descriptors. *IEEE Transactions on Visualization and Computer Graphics* 17, 11 (Nov 2011), 1624–1636. https://doi.org/10.1109/TVCG.2010.266

[13] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2015. Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 457–466.

[14] Fernanda Ferreira, Jens Apel, and John M Henderson. 2008. Taking a new look at looking at nothing. *Trends in cognitive sciences* 12, 11 (2008), 405–410.

[15] Joystone Gbadamosi and Wolfgang H Zangemeister. 2001. Visual imagery in hemianopic patients. *Journal of Cognitive Neuroscience* 13, 7 (2001), 855–866.

[16] Christiane Glatz, Stas S Krupenia, Heinrich H Bülthoff, and Lewis L Chuang. 2018. Use the Right Sound for the Right Job: Verbal Commands and Auditory Icons for a Task-Management System Favor Different Information Processes in the Brain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 472.

[17] Stephenie A Harrison and Frank Tong. 2009. Decoding reveals the contents of visual working memory in early visual areas. *Nature* 458, 7238 (2009), 632. https://doi.org/10.1038/nature07832

[18] Donald O Hebb. 1968. Concerning imagery. *Psychological review* 75, 6 (1968), 466.

[19] Rob Jacob and Sophie Stellmach. 2016. What you look at is what you get: gaze-based user interfaces. *interactions* 23, 5 (2016), 62–65.

[20] Robert JK Jacob. 1990. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 11–18.

[21] Edmund Jacobson. 1932. Electrophysiology of mental activities. *The American Journal of Psychology* 44, 4 (1932), 677–694.

[22] Roger Johansson, Jana Holsanova, Richard Dewhurst, and Kenneth Holmqvist. 2012. Eye movements during scene recollection have a functional role, but they are not reinstatements of those produced during encoding. *Journal of Experimental Psychology: Human Perception and Performance* 38, 5 (2012), 1289–1314.

[23] Roger Johansson, Jana Holsanova, and Kenneth Holmqvist. 2006. Pictures and Spoken Descriptions Elicit Similar Eye Movements During Mental Imagery, Both in Light and in Complete Darkness. *Cognitive Science* 30, 6 (2006), 1053–1079. https://doi.org/10.1207/s15516709cog0000_86

[24] Roger Johansson, Jana Holsanova, and Kenneth Holmqvist. 2011. The dispersion of eye movements during visual imagery is related to individual differences in spatial imagery ability. In *Proceedings of the Cognitive Science Society*, Vol. 33.

[25] Roger Johansson and Mikael Johansson. 2014. Look Here, Eye Movements Play a Functional Role in Memory Retrieval. *Psychological Science* 25, 1 (2014), 236–242. https://doi.org/10.1177/0956797613498260

[26] Tilke Judd, Frédo Durand, and Antonio Torralba. 2012. A Benchmark of Computational Models of Saliency to Predict Human Fixations. In *MIT Technical Report*.

[27] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to Predict Where Humans Look. In *IEEE International Conference on Computer Vision (ICCV)*. 2106–2113. https://doi.org/10.1109/ICCV.2009.5459462

[28] Yukiyasu Kamitani and Frank Tong. 2005. Decoding the visual and subjective contents of the human brain. *Nature neuroscience* 8, 5 (2005), 679.

[29] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977

[30] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. 2017. Gaze Embeddings for Zero-Shot Image Classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6412–6421. https://doi.org/10.1109/CVPR.2017.679

[31] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct*. ACM Press, New York, New York, USA, 1151–1160. https://doi.org/10.1145/2638728.2641695

[32] Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. 2008. Identifying natural images from human brain activity. *Nature* 452 (2008), 352–355. Issue 7185. https://doi.org/10.1038/nature06713

[33] Seon Joo Kim, Hongwei Ng, Stefan Winkler, Peng Song, and Chi-Wing Fu. 2012. Brush-and-drag: A Multi-touch Interface for Photo Triaging. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '12)*. ACM, New York, NY, USA, 59–68. https://doi.org/10.1145/2371574.2371584

[34] David Kirk, Abigail Sellen, Carsten Rother, and Ken Wood. 2006. Understanding Photowork. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 761–770. https://doi.org/10.1145/1124772.1124885

[35] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A Lee, and Mark Billinghurst. 2018. Pinpointing: Precise Head-and Eye-Based

Target Selection for Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 81.

[36] Bruno Laeng, Ilona M. Bloem, Stefania D'Ascenzo, and Luca Tommasi. 2014. Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition* 131, 2 (2014), 263 – 283. https://doi.org/10.1016/j.cognition.2014.01.003

[37] Bruno Laeng and Dinu-Stefan Teodorescu. 2002. Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science* 26, 2 (2002), 207–231. https://doi.org/10.1207/s15516709cog2602_3

[38] Brent J Lance, Scott E Kerick, Anthony J Ries, Kelvin S Oie, and Kaleb McDowell. 2012. Brain-Computer Interface Technologies in the Coming Decades. *Proc. IEEE* 100, Special Centennial Issue (May 2012), 1585–1599. https://doi.org/10.1109/JPROC.2012.2184830

[39] Päivi Majaranta and Andreas Bulling. 2014. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*. Springer, 39–65.

[40] Päivi Majaranta and Kari-Jouko Räihä. 2002. Twenty Years of Eye Typing: Systems and Design Issues. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications (ETRA '02)*. ACM, New York, NY, USA, 15–22. https://doi.org/10.1145/507072.507076

[41] Serena Mastroberardino and Annelies Vredeveldt. 2014. Eye-closure increases children's memory accuracy for visual material. *Frontiers in Psychology* 5 (2014), 241. https://doi.org/10.3389/fpsyg.2014.00241

[42] Nick Merrill and John Chuang. 2018. From Scanning Brains to Reading Minds: Talking to Engineers about Brain-Computer Interface. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 323.

[43] José del R. Millán, Rüdiger Rupp, Gernot Müller-Putz, Roderick Murray-Smith, Claudio Giugliemma, Michael Tangermann, Carmen Vidaurre, Febo Cincotti, Andrea Kubler, Robert Leeb, Christa Neuper, Klaus Müller, and Donatella Mattia. 2010. Combining Brain-Computer Interfaces and Assistive Technologies: State-of-the-Art and Challenges. *Frontiers in Neuroscience* 4 (2010), 161. https://doi.org/10.3389/fnins.2010.00161

[44] Charles S Moore. 1903. Control of the memory image. *The Psychological Review: Monograph Supplements* 4, 1 (1903), 277–306.

[45] Martez E Mott, Shane Williams, Jacob O Wobbrock, and Meredith Ringel Morris. 2017. Improving dwell-based gaze typing with dynamic, cascading dwell times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2558–2570.

[46] Ulric Neisser. 1967. *Cognitive psychology*. New York: Appleton-Century-Crofts.

[47] Dan Nemrodov, Matthias Niemeier, Ashutosh Patel, and Adrian Nestor. 2018. The Neural Dynamics of Facial Identity Processing: insights from EEG-Based Pattern Analysis and Image Reconstruction. *eNeuro* (2018). https://doi.org/10.1523/ENEURO.0358-17.2018 arXiv:http://www.eneuro.org/content/early/2018/01/29/ENEURO.0358-17.2018.full.pdf

[48] Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. 2011. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology* 21, 19 (2011), 1641 – 1646. https://doi.org/10.1016/j.cub.2011.08.031

[49] Cheves West Perky. 1910. An experimental study of imagination. *The American Journal of Psychology* 21, 3 (1910), 422–452.

[50] Daniel C Richardson, Gerry TM Altmann, Michael J Spivey, and Merrit A Hoover. 2009. Much ado about eye movements to nothing: a response to Ferreira et al.: Taking a new look at looking at nothing. *Trends in Cognitive Sciences* 13, 6 (2009), 235–236.

[51] Daniel C Richardson and Michael J Spivey. 2000. Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition* 76, 3 (2000), 269–295.

[52] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision* 40, 2 (2000), 99–121. https://doi.org/10.1023/A:1026543900054

[53] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies. *ACM Trans. Graph.* 35, 4, Article 119 (July 2016), 12 pages. https://doi.org/10.1145/2897824.2925954

[54] Hosnieh Sattar, Andreas Bulling, and Mario Fritz. 2017. Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2740–2748.

[55] Agnes Scholz, Anja Klichowicz, and Josef F Krems. 2017. Covert shifts of attention can account for the functional role of "eye movements to nothing". *Memory & Cognition* (2017), 1–14.

[56] Agnes Scholz, Katja Mehlhorn, and Josef F Krems. 2016. Listen up, eye movements play a role in verbal memory retrieval. *Psychological research* 80, 1 (2016), 149–158.

[57] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 815–823.

[58] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. 2017. Deep image reconstruction from human brain activity. *bioRxiv* (2017). https://doi.org/10.1101/240317

[59] Ben Shneiderman, Benjamin B. Bederson, and Steven M. Drucker. 2006. Find That Photo!: Interface Strategies to Annotate, Browse, and Share. *Commun. ACM* 49, 4 (April 2006), 69–71. https://doi.org/10.1145/1121949.1121985

[60] Michael J Spivey and Joy J Geng. 2001. Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological research* 65, 4 (2001), 235–241.

[61] Jacques J Vidal. 1973. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering* 2, 1 (1973), 157–180.

[62] Annelies Vredeveldt, Colin G. Tredoux, Kate Kempen, and Alicia Nortje. 2015. Eye Remember What Happened: Eye-Closure Improves Recall of Events but not Face Recognition. *Applied Cognitive Psychology* 29, 2 (2015), 169–180. https://doi.org/10.1002/acp.3092

[63] Fan Wang and Leonidas J. Guibas. 2012. Supervised Earth Mover's Distance Learning and Its Computer Vision Applications. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part I (ECCV'12)*. Springer-Verlag, Berlin, Heidelberg, 442–455. https://doi.org/10.1007/978-3-642-33718-5_32

[64] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. 2002. Brain-computer interfaces for communication and control. *Clinical Neurophysiology* 113, 6 (2002), 767 – 791. https://doi.org/10.1016/S1388-2457(02)00057-3

[65] Qian Yu, Feng Liu, Yi-Zhe SonG, Tao Xiang, Timothy Hospedales, and Chen Change Loy. 2016. Sketch Me That Shoe. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 799–807. https://doi.org/10.1109/CVPR.2016.93

[66] Thorsten O. Zander and Christian Kothe. 2011. Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *Journal of Neural Engineering* 8, 2 (2011), 025005. http://stacks.iop.org/1741-2552/8/i=2/a=025005