BubbleView: an interface for crowdsourcing image importance maps and tracking visual attention

NAM WOOK KIM*, Harvard SEAS
ZOYA BYLINSKII*, MIT CSAIL
MICHELLE A. BORKIN, Northeastern CCIS
KRZYSZTOF Z. GAJOS, Harvard SEAS
AUDE OLIVA, MIT CSAIL
FREDO DURAND, MIT CSAIL
HANSPETER PFISTER, Harvard SEAS

In this paper, we present BubbleView, an alternative methodology for eye tracking using discrete mouse clicks to measure which information people consciously choose to examine. BubbleView is a mouse-contingent, moving-window interface in which participants are presented with a series of blurred images and click to reveal "bubbles" - small, circular areas of the image at original resolution, similar to having a confined area of focus like the eye fovea. Across 10 experiments with 28 different parameter combinations, we evaluated BubbleView on a variety of image types: information visualizations, natural images, static webpages, and graphic designs, and compared the clicks to eye fixations collected with eye-trackers in controlled lab settings. We found that BubbleView clicks can both (i) successfully approximate eye fixations on different images, and (ii) be used to rank image and design elements by importance. BubbleView is designed to collect clicks on static images, and works best for defined tasks such as describing the content of an information visualization or measuring image importance. BubbleView data is cleaner and more consistent than related methodologies that use continuous mouse movements. Our analyses validate the use of mouse-contingent, moving-window methodologies as approximating eye fixations for different image and task types.

CCS Concepts: • **Human-centered computing** → *Interaction techniques*;

Additional Key Words and Phrases: human vision, visual attention, eye tracking, crowdsourcing, saliency, image importance, mouse-contingent interface, natural scenes, information visualizations, graphic designs, websites.

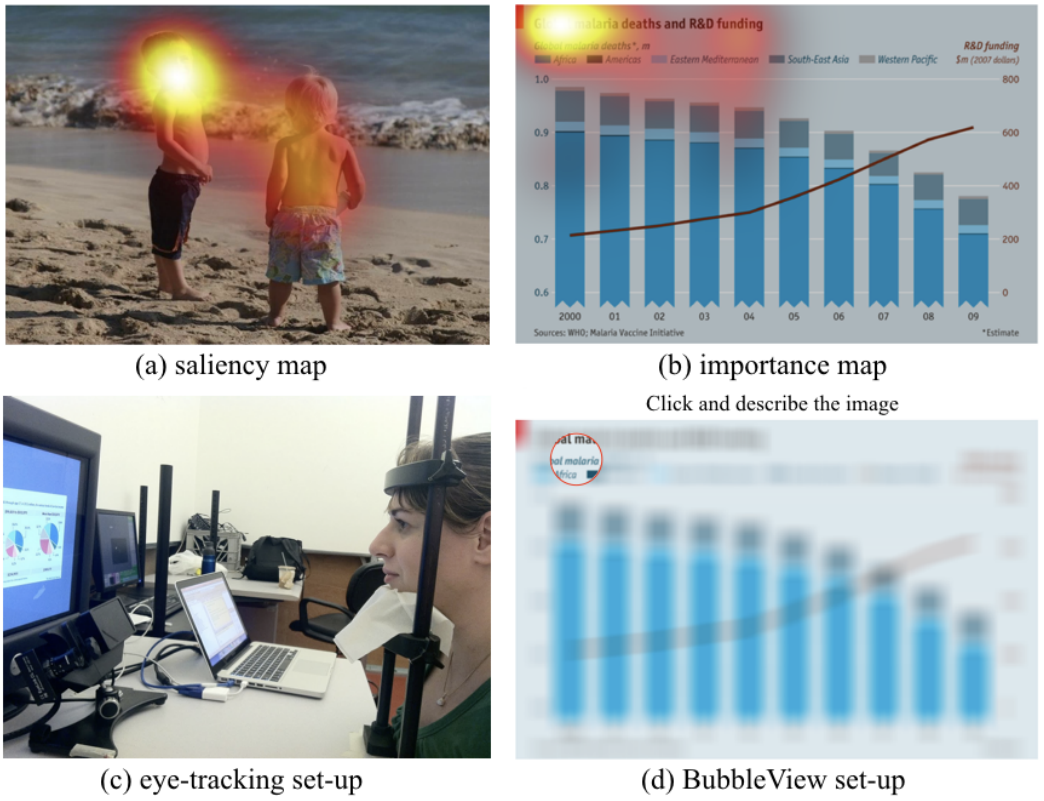
ACM Reference format:
Nam Wook Kim*, Zoya Bylinskii*, Michelle A. Borkin, Krzysztof Z. Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2017. BubbleView: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Trans. Comput.-Hum. Interact.* 24, 5, Article 36 (October 2017), 40 pages.
https://doi.org/10.1145/3131275

1 INTRODUCTION

Eye tracking is a technique to measure an individual's eye movements, visual attention, and focus. This experimental methodology has proven useful for studying the cognitive processes involved in

*Equal contribution.

1073-0516/2017/10-ART36 $15.00
https://doi.org/10.1145/3131275



(a) saliency map (b) importance map

(c) eye-tracking set-up (d) BubbleView set-up

Fig. 1. Just as the pattern of human eye fixations can be used as a heatmap of saliency for an image (a), the pattern of BubbleView clicks can be used as a heatmap of importance for an image (b). An eye tracking set-up (pictured: EyeLink1000) is a way to collect human eye fixations in the lab setting (c), whereas the BubbleView interface can be launched online and feasibly scale up the collection of crowdsourced data (d).

visual information processing, including which visual elements people look at first and spend the most time on [Jacob and Karn 2003; Majaranta and Bulling 2014]. Eye tracking is widely used for conducting usability studies for human-computer interfaces [Jacob and Karn 2003; Nielsen and Pernice 2009], for designing gaze-based and attention-aware user interfaces [Lutteroth et al. 2015; Majaranta and Bulling 2014] or for collecting gaze data to build saliency prediction models [Borji and Itti 2015; Judd et al. 2012].

Commercial eye-trackers mostly use specialized hardware such as advanced infrared sensors and high-quality cameras to accurately track eye positions and movements [Al-Rahayfeh and Faezipour 2013]. However, they often require high-cost equipment and invasive calibrations (e.g., EyeLink, ISCAN), which means it is difficult to scale to large scale studies beyond controlled lab environments. Recent appearance-based methods attempt to address this issue by enabling eye tracking on affordable cameras built into personal devices [Huang et al. 2015; Krafka et al. 2016; Xu et al. 2015]. However, these methods have not yet seen widespread adoption, as they still suffer in accuracy and robustness, and impose set-up constraints (camera quality, lighting conditions).

On the other hand, cursor-based attention tracking is based on the correlation between gazes and cursor locations [Guo and Agichtein 2010; Huang et al. 2012; Rodden et al. 2008] and reduce

the need to handle variations in real-world settings in camera-based methods; e.g., calibrations, ambient lighting, etc. The most popular cursor-based approach uses a moving window continuously following the position of the cursor to reveal a portion of the screen in normal resolution [Jansen et al. 2003].

Our **BubbleView** methodology is a cursor-based, moving-window approach to collect clicks on static images as a proxy for eye fixations. BubbleView presents blurred images and allows participants to click around to reveal small circular "bubble" regions of the image at the original resolution (Figure 1). This is intended to loosely approximate a blurred periphery and the confined area of focus of the human eye fovea.

Compared to natural viewing, BubbleView and related cursor-based methodologies slow down the exploration patterns of participants, because choosing where to move the mouse and click is a slower cognitive process than moving eyes around an image. Because of this, we refer to the pattern of BubbleView clicks on an image as the **importance map** for the image. We intend for importance to encapsulate image regions that are not only more attention grabbing initially (salient), but also regions that people spend more time on because they are more relevant, or interesting, to the task-at-hand.

BubbleView is especially well suited to capturing image regions of most importance when a directed task is provided (as compared to free viewing). Our initial target setting, first presented in Kim et al. [2015] was to show that BubbleView clicks can provide a good approximation for eye movements when participants are asked to describe the content of information visualizations (graphs, charts, tables). In Borkin et al. [2016], we further showed that knowing where people look can provide clues about what they store in memory and recall about an information visualization. Like eye tracking, BubbleView can provide important insights about human perception and cognition, but at a lower data collection cost than eye tracking. It can easily scale up data collection to many participants and images, and be launched remotely to enable online crowdsourcing.

In this paper, we validate that BubbleView generalizes to approximating eye fixations on different image types and under different task constraints. Specifically, we show that:

- ➔ BubbleView clicks can successfully approximate eye fixations on information visualizations, natural images, and websites, in both a free-viewing condition and with a description task;
- ➔ Compared to related methodologies based on a moving-window approach [Jiang et al. 2015], BubbleView clicks provide more reliable and less noisy data;
- ➔ The number of BubbleView clicks in different image regions can be used to measure the relative importance of those image regions.

We present the BubbleView methodology with the interested experimenter in mind who may consider it for crowdsourcing an experiment, or for an evaluation that would typically be conducted with an eye tracker in a conventional laboratory setting. While prior work contains some initial validation that a cursor-based interface can serve as a proxy for eye tracking [Bednarik and Tukiainen 2007; Jiang et al. 2015; Kim et al. 2015], we conducted an extensive quantitative analysis by running 10 experiments with 28 different parameter combinations, on Amazon's Mechanical Turk. Our experiments were carried out on 5 different datasets, spanning information visualizations [Borkin et al. 2016], natural images [Jiang et al. 2015; Xu et al. 2014], static webpages [Shen and Zhao 2014], and graphic designs [O'Donovan et al. 2014]. We varied task type (free-viewing, describing) and task duration, image blur kernel, and bubble radius. We compared BubbleView clicks not only to eye fixations [Borkin et al. 2016; Xu et al. 2014], but also to mouse movements [Jiang et al. 2015], and to explicit importance annotations [O'Donovan et al. 2014]. Our contributions include:

(1) The BubbleView interface which can be launched online for the cheap, feasible collection of crowdsourced data, provided at massvis.mit.edu/bubbleview;

(2) A thorough analysis of how different experimental parameters affect BubbleView click data, and guidelines about how to choose an appropriate setting of parameters for a given experiment;
(3) A discussion of how BubbleView can be used to approximate eye fixations collected in a controlled lab setting;
(4) A proposed list of applications of the BubbleView methodology, including for the measurement of image importance, image-based question-answering tasks, and training computational models of saliency/importance.

2 RELATED WORK

The original idea of "bubbles" comes from the work of Gosselin and Schyns [2001] who displayed masked images, punctured by randomly-located Gaussian windows (termed bubbles), and measured participant performances on categorization tasks to determine image regions important to the tasks. Deng et al. [2013] modified this methodology to allow participants to control the location of bubbles on blurred images, and reveal image regions in order to complete a fine-grained object recognition task. We further extended the bubbles technique by having participants click to expose image regions in order to describe information visualizations [Kim et al. 2015]. In this paper we evaluate the BubbleView methodology with both description and free-viewing tasks on information visualizations, natural images, and graphic designs, to validate that it can be used for discovering relevant image regions. In these settings, BubbleView is similar to other cursor-based, moving-window approaches which expose image regions depending on user-defined cursor positions [Jansen et al. 2003; Schulte-Mecklenbeck et al. 2011]. Because we vary the resolution of the image depending on where the cursor is located, BubbleView can also be classified as a mouse-contingent, multiresolutional display (similar to Jiang et al. [2015], which we compare to in this paper). In this section we review other gaze tracking techniques, including eye tracking, cursor-based and appearance-based approaches. We also discuss the relationship of gaze tracking and mouse tracking to saliency.

2.1 Eye movements and cognitive tasks

A significant amount of research has been conducted on the connection between eye movements and various cognitive tasks: the eyes can provide important clues about how visual perception proceeds as a human looks at images [Hayhoe 2004; Holmqvist et al. 2011; Just and Carpenter 1976; Kowler 1989; Noton and Stark 1971]. This area of research is so established and diverse that we refer the reader to some representative papers reporting on the utility of eye movements for studying human perception and cognition in the context of user interfaces [Bergstrom and Schall 2014; Bruneau et al. 2002; Duchowski 2002; Goldberg and Kotval 1999; Graf and Krueger 1989; Jacob and Karn 2003; Poole and Ball 2006; Rensink 2011], web search [Cutrell and Guan 2007; Goldberg et al. 2002], web browsing [Cowen et al. 2002; Josephson and Holmes 2002; Pan et al. 2004], problem solving [Grant and Spivey 2003], reading [Rayner 1998], advertisements [Rayner et al. 2001], and visualizations [Borkin et al. 2016; Bylinskii et al. 2017a; Huang 2007; Kim et al. 2012; Pohl et al. 2009]. These papers show that aside from providing information about how human perception proceeds, eye movements can also provide insights about the effectiveness of different visual content, or the usability of interfaces. Because of all the potential use cases, researchers have also sought ways to more efficiently collect eye movements without having to rely on standard eye tracking.

2.2 Cursor-based attention tracking

There has been a significant effort to find cheap, nonintrusive, and more scalable alternatives to collect human attentional data. Cursor-based techniques are a particularly suitable alternative for scaling to large web-based studies.

The **moving-window** approach is a popular cursor-based technique in which a limited amount of information is visible through a variable size window continuously following a cursor position [McConkie and Rayner 1975; Rayner 2014]. Inspired by the moving-window model, Jansen et al. [2003] developed a computer program called Restricted Focus Viewer (RFV) that takes an image, blurs it, and reveals only a restricted block of the image, allowing a user to move the region using a mouse [Bednarik and Tukiainen 2007; Blackwell et al. 2000; Jansen et al. 2003; Tarasewich et al. 2005]. Commercial software for tracking user attention has also built on the same idea (e.g., Attensee[1]). The mouse-contingent methodology has been employed to investigate cognitive behaviors of users in diverse contexts such as diagrammatic reasoning and program debugging, and to study the usability of web sites [Bednarik and Tukiainen 2005; Jansen et al. 2003; Tarasewich et al. 2005].

Recent studies have made further improvements. SALICON [Jiang et al. 2015] implemented moving-window, multi-resolution blur on images to attempt to simulate the fall-off in acuity of peripheral vision. On the other hand, Lagun and Agichtein [2011] directly preprocessed web search results to show one result and blur the other results based on a user's viewport; however, this method is not intended to approximate the human fovea as it shows an entire DOM element at a time. All these recent studies were conducted online with hundreds to thousands of participants, proving the scalability of their methods.

There is also a rich history of work in the space of gaze-contingent multiresolutional displays, where the moving-window approach is guided by gaze. We refer the reader to a review by Reingold et al. [2003]. These approaches complement, rather than replace, standard eye-tracking techniques, and have different motivations: bandwidth and processing savings. However, this line of work contains a related investigation of multiresolutional blur to approximate the peripheral visual system. Whether cursor-based or gaze-based, a moving-window approach slows down visual exploration patterns relative to natural viewing and can be used to discover the most important or relevant image regions.

Aside from the moving-window model for image exploration, other works also investigated the relationship between cursor movements and gaze positions, mostly focusing on web browsing [Chen et al. 2001] and search tasks [Guo and Agichtein 2010; Huang et al. 2012, 2011; Rodden et al. 2008]. Chen et al. [2001] found a high correlation between cursor and gaze locations. Rodden et al. [2008] found that cursor and gaze are better aligned along the vertical dimension, while Guo and Agichtein [2010] also found a similar result in their study of predicting eye-mouse coordination. Huang et al. [2012] found that people's cursors lag behind their gazes and there are individual differences in the distance between the cursor and gaze positions.

These cursor-based techniques have succeeded in providing an affordable and scalable alternative to eye tracking, but prior work has two key limitations. First, a moving window approach requires complicated post-processing of mouse movement data to extract mouse positions (e.g., SALICON). Second, evaluations of existing techniques have been mostly limited to simple aggregate comparisons with ground-truth eye tracking data on a specific set of images (i.e., natural images or webpages) with a fixed setting of parameters (e.g., blur kernel, bubble size).

[1] Attensee (http://www.attensee.com) is a commercial solution based on the idea of Flashlight [Schulte-Mecklenbeck et al. 2011], an open-source research tool: https://github.com/michaelschulte/flashlight.

With BubbleView, we overcome the first limitation by collecting discrete clicks instead of continuous mouse trajectories. This enables a more explicit record of points of interest without the need for post-processing noisy mouse movement data. To address the second limitation, we also systematically evaluate the effect of different parameters and task settings on the ability of a cursor-based methodology to approximate eye movements. We compare our methodology to eye fixations on a diverse set of image stimuli with different parameters to find the best settings under different task conditions. Our findings are likely to generalize to other related mouse-contingent displays.

2.3 Appearance-based gaze tracking

Another line of work has been devoted to non-intrusive, appearance-based gaze estimation, where images of the eyes are post-processed using computer vision techniques to determine gaze location. This type of gaze estimation often involves collecting a training dataset with a standard eye-tracker, training a computer vision model to map eye images to gaze coordinates, and using this model at test-time to directly infer gaze positions from a video stream of the eyes (e.g., captured via a webcam). At test time, these approaches do not require specialized eye tracking hardware (i.e., high quality special cameras, infrared sensors, and head mounting devices) and allow users to move their heads freely.

Early gaze tracking models were mostly based on relatively small training datasets collected through lab studies. For example, Baluja and Pomerleau [1994] collected 2000 images of the eyes for four postures by instructing a participant to visually track a moving cursor and built a neural network model to estimate gaze locations. Recent methods attempt to build gaze tracking models on large datasets to improve accuracy as well as to work in real-world settings. Funes Mora et al. [2014] constructed a database to enable comparison across different gaze tracking algorithms for variations including head poses, individual differences, and ambient and sensing conditions. Zhang et al. [2015] developed an appearance-based gaze estimation method using multimodal convolutional neural networks. Their model was trained on a hundred thousand images from 15 laptop users for several months using built-in cameras in laptops, accounting for realistic variability in illumination and appearance. Huang et al. [2015] similarly built a large gaze dataset and a gaze tracking algorithm for tablet users. While the two studies are still limited to datasets collected through labs, other works leverage online crowdsourcing to further extend the scale of gaze datasets. Xu et al. [2015] developed a webcam-based eye tracking game running in a browser on a remote computer. Their crowdsourced experiments could collect gaze data cheaper and faster than lab studies. Papoutsaki et al. [2016] also designed a similar webcam-based eye tracking system. Krafka et al. [2016] collected eye tracking on over 2.5M frames using a mobile application and online, and developed a gaze prediction algorithm based on convolutional neural networks, while achieving state-of-the-art results.

All of the above approaches have yet to reach the level of tracking accuracy and robustness possible with dedicated eye tracking hardware. These approaches also depend on either some initial calibration or have constraints on a participants' set-up: network connection, camera quality, and restricted range of face location relative to screen. As a result, we have not yet seen widespread adoption of appearance-based gaze tracking. Additionally, the camera-based gaze tracking approaches have the downside of requiring the capture of participants' face images throughout the study, which comes with privacy concerns [Liebling and Preibusch 2014].

2.4 Saliency models and eye tracking datasets

In addition to alternative techniques for eye tracking which require human participants, significant progress has been made building computational saliency models to predict eye fixations. Many

saliency models are motivated by psychological and neurobiological theories, and make use of both low-level image features (e.g., intensity, color, and orientation) and high-level semantic features (e.g., scenes, objects, and tasks) to approximate the human visual system [Borji and Itti 2013; Frintrop et al. 2010]. The performance of these models is usually evaluated against ground-truth eye fixations [Bylinskii et al. 2014, 2016a; Judd et al. 2012].

Models have typically been trained directly on fixation data collected from eye tracking experiments [Judd et al. 2009; Kienzle et al. 2007]. However, good models require large quantities of data, larger than what is practical to collect using conventional eye tracking techniques. To overcome this challenge, a large dataset of mouse movements on natural images was recently released for simulating the natural viewing behavior and subsequently training computational saliency models. This dataset, dubbed SALICON, was collected using a moving-window methodology [Jiang et al. 2015]. Since then, many neural network models of saliency trained on this data [Jiang et al. 2015; Kruthiventi et al. 2015; Pan et al. 2016] have achieved state-of-the-art performances on standard saliency benchmarks [Bylinskii et al. 2014]. Tavakoli et al. [2017] have recently shown that saliency models trained on mouse movements can generalize well to predicting eye fixations.

While most saliency models are focused on predicting eye fixations on natural scenes, there are relatively few studies that have looked at other image types including webpages, graphic designs, and information visualizations. These images are different from natural images in that they usually contain rich semantic data (e.g., texts, charts, and logos) or different viewing patterns such as top-left bias [Buscher et al. 2009] and banner blindness [Grier et al. 2007]. Shen and Zhao [2014] developed a webpage saliency model based on the FiWI dataset, and then improved the model with high-level semantic features (e.g., positional bias and object detectors) [Shen et al. 2015]. O'Donovan et al. [2014] developed a semi-automatic model of importance prediction for graphic designs by training on a crowdsourced dataset of importance annotations. The GDI dataset was collected by asking workers to annotate regions of importance on images using binary masks. Xu et al. [2016] presented a computational model for predicting visual attention in user interfaces with user interactions.

We draw on several existing datasets with accompanying attention data and look at how well BubbleView clicks can approximate fixations, mouse movements, and explicit importance annotations. We used the FiWI dataset [Shen and Zhao 2014] (static webpages), OSIE dataset [Xu et al. 2014] (natural scenes), and the MASSVIS dataset [Borkin et al. 2016] (information visualizations) to evaluate the degree to which BubbleView clicks can approximate eye fixations. We used the SALICON dataset [Jiang et al. 2015] to compare our methodology against the moving-window approach. We also used the GDI dataset [O'Donovan et al. 2014] (graphic designs) to see whether BubbleView clicks can be used to rank design elements by importance.

3 BUBBLEVIEW METHODOLOGY

BubbleView is an experimental methodology for collecting mouse clicks on images as an approximation to eye fixations. We first provide some background on human eye movements and perception, before discussing how BubbleView was designed to approximate eye tracking, and how it can be used for running perception experiments.

3.1 Background: human eye movements and perception

The human eye consists of light receptor cells that are differently distributed throughout the eye. The clearest and most detailed vision is in the central, **foveal area**, of the visual field, and blurrier vision is in the larger part of the visual field, which is called the **peripheral area**. The foveal area captures about 1-2 degrees of visual angle which constitutes less than 8% of the visual field, but

makes up 50% of the visual information sent to the brain [Tobii 2010]. When we move our eyes, we place the foveal region of the eye on different regions of the visual field, bringing them into focus.

Visual angles are units for measuring the projection of the visual field, as images, on our retina. For a given experimental viewing setup, visual angles can be computed by taking into account the distance to the screen, size and resolution of the image on the screen[2]. The error of professional-grade eye trackers (e.g., EyeLink) is also measured in degrees of visual angle, and is commonly less than 1 degree.

The pauses in eye movements are called **fixations**, and the transitions between successive fixations are called **saccades**. In this paper, we focus on fixations, since they give us the points of interest that the eye has stopped on to bring them into focus. The temporal sequence of fixations, fixation duration, saccade length, and other features of eye movements carry a lot of additional information about human perception [Bylinskii et al. 2017a; Holmqvist et al. 2011; Jacob and Karn 2003; Just and Carpenter 1976] but are beyond the scope of the present work. We concentrated on the location of fixations, which are most straightforward to analyze [Bruneau et al. 2002; Jacob and Karn 2003; Tobii 2010] and to model computationally [Bylinskii et al. 2016a].

3.2 Designing experiments with BubbleView

The BubbleView methodology is intended to approximate a blurred periphery, and users click on images to reveal small, circular regions ("bubbles") at the original resolution (Figure 2). This is similar to having a confined area of focus like the eye fovea. Different blur levels and bubble sizes can be used to approximate different eye tracking setups, with different visual angles (Figure 14).

In comparison to the moving-window approach which records continuous mouse movements, our approach records discrete mouse clicks where each click represents a conscious choice made by the user to reveal a portion of the image. As the clicks correspond to individual points of interest, we directly compare them to eye fixations.

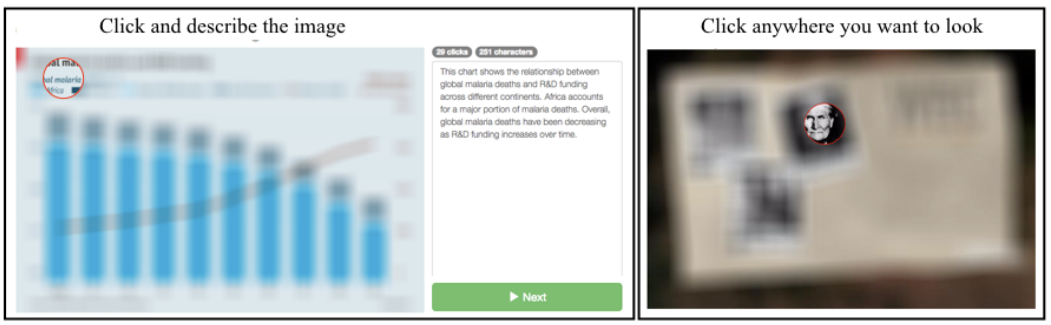


Fig. 2. Two different versions of the BubbleView interface for two task types, for gathering task-based (a) and task-free (b) clicks, as approximations to similar eye tracking experiments.

Tasks and image types for attention experiments

We evaluated our BubbleView interface on two tasks: free-viewing and description, and four image types: natural scenes, information visualizations, static webpages, and graphic designs. Here we discuss the motivations behind these design choices.

[2] https://github.com/cvzoya/saliency/tree/master/computeVisualAngle

During **free-viewing**, participants are not given a task but are instructed to freely look around the image. Free-viewing is commonly used in eye tracking experiments to study the human perception of natural scenes, because it can avoid large task-dependent effects. It is often assumed the eyes are drawn to conspicuous image elements, and attention proceeds in a bottom-up manner, guided by the image features rather than a high-level task[3]. This assumption has motivated the use of free-viewing for collecting ground truth data for saliency datasets [Koehler et al. 2014], where the pattern of eye fixations can be interpreted as the **saliency map** for the image (Figure 1). Most saliency datasets have been collected using free-viewing[4]. Computational models are in turn trained and tested on saliency datasets as a proxy for human attention. We are similarly motivated by the computational applications that can be built by training models on large attention datasets (e.g., [Bylinskii et al. 2017b]).

Compared to natural viewing, cursor-based moving-window methodologies naturally slow down visual exploration patterns. By providing a cognitively-demanding task, these exploration patterns can be slowed down further to bring more intentionality to each click. In the **description** task, participants are required to type a description of the image while using the BubbleView interface to explore the image. The descriptions naturally depend on the image regions clicked on. This task is well suited to images with an underlying message or concept that needs careful examination to decipher. We used the description task with visualization images from the MASSVIS [Borkin et al. 2016] dataset[5], and website images from the FiWI [Shen and Zhao 2014] dataset. We also tested the free-viewing task with the FiWI images, because the eye-tracking data from this dataset was collected with free-viewing, and we wanted to approximate the original experiment. For the same reason, we ran the free-viewing task on natural images from the OSIE [Xu et al. 2014] dataset. For the graphic designs in the GDI dataset [O'Donovan et al. 2014], which have importance annotations rather than eye fixations, we chose a free-viewing task. We chose this task because we found that the graphic designs could not be easily summarized by a description (i.e., some images required further context, not all were English, some had few visual elements, etc.).

Tasks deviating from description and free-viewing are beyond the scope of this paper, although they are common in user interface research [Bergstrom and Schall 2014; Cutrell and Guan 2007; Goldberg et al. 2002; Jacob and Karn 2003]. For instance, for testing websites or application interfaces, participants may be asked to perform tasks such as searching for a particular element or option, navigating to a particular region of the image or page, or answering questions. Related moving-window methodologies have previously been validated in the context of web navigation, program debugging, and question-answering [Bednarik and Tukiainen 2007; Jansen et al. 2003; Lagun and Agichtein 2011; Schulte-Mecklenbeck et al. 2011; Tarasewich et al. 2005]. These tasks can be quite specific to the interface being evaluated. We used two task types that can generalize (without modification) to a large collection of different image types. Our BubbleView tool is available to the research community so future work can investigate the generalizability of this tool for other tasks.

3.3 Implementation

We implemented a web-based BubbleView interface that takes a directory of images as input and displays a subset of the images in random sequence, blurring each one. Participants receive

[3]Alternative views posit that free-viewing is not task-free, but permits participants to choose their own internal agendas/-tasks [Parkhurst et al. 2002; Tatler et al. 2005, 2011]. Even under this interpretation, averaging data over many participants, each of which may have their own agenda, has the effect of averaging out the task and providing an approximately task-independent aggregate measurement.

[4]A list of eye tracking datasets and their attributes is available at: http://saliency.mit.edu/datasets.html

[5]In the MASSVIS eye-tracking set-up participants also provided image descriptions, but they did so at the end, not during, the viewing session. This is because memorability was part of the original study, whereas it is not here.

a set of task instructions and can click to reveal bubble regions (Figure 2). A demo is available at massvis.mit.edu/bubbleview.
The experimenter has a choice of parameters:

- ➔ **Task type:** the instructions given to participants. We used two different versions of the interface for a description task with an input text field (Figure 2a), and a free-viewing task with no additional inputs from participants (Figure 2b). Alternative tasks are possible.
- ➔ **Time:** the viewing time per image, which depends on the task. For the description task, we did not constrain the time. For the free-viewing task, we fixed time per image to be either 10 or 30 seconds, depending on the experiment.
- ➔ **Blur sigma:** the size of the Gaussian blur kernel (in pixels) to apply to each image to mimic peripheral vision. This is a fixed quantity over the whole image, and is constant across all images in the sequence. In our studies, we manually selected a blur value per image dataset to distort image text beyond recognition. We wanted the level of detail to be sufficient for reading only within regions of focus.
- ➔ **Bubble radius:** the size of the focus area (in pixels) that is deblurred during a click to mimic foveal vision. In our studies, we varied this size depending on other task constraints, but often stayed within 1-2 degrees of visual angle of the eye tracking setups used for the ground-truth eye movement datasets.
- ➔ **Mouse modality:** although we originally designed BubbleView for collecting mouse clicks, we extended it to allow bubble regions to be exposed during continuous mouse movements (as in Jiang et al. [2015]). We discuss the differences between the two modalities in Section 7.

The experimenter may also choose the number of images displayed in a sequence. In our description task, participants were able to continue to the next image after writing a minimum number of characters (150 in our experiments). In the free-viewing task, once the fixed time per image elapsed, the next image in the sequence was presented.

We also developed a monitoring interface to inspect experimental results (Figure 3). The purpose of the interface is to take a quick glance at the bubbles collected, before the main analysis. For each image, the experimenter can see the bubbles and (if applicable) text descriptions generated by each participant. Adjusting the slider allows exploration of the temporal sequence and evolution of bubble clicks and description text over time. The experimenter can also see how the blurred image looked to the participant to investigate why a region may have been clicked. This interface can be used to check if an experiment is running as intended in real time.

4 EXPERIMENTS AND ANALYSIS OVERVIEW

We first presented the BubbleView methodology as a way to approximate eye fixations on information visualizations under a description task [Kim et al. 2015]. The present paper is an extension that more systematically explores the BubbleView methodology and measures how varying parameters such as bubble size, image blur, and task timing affects the resulting clicks, and how the number of participants affects the quality of the resulting data. We wanted to test BubbleView for generalizability (i.e., on other image and task types) to see if eye fixations could be well approximated under different settings.

For our experiments, we used **Amazon's Mechanical Turk (MTurk)**, an online crowdsourcing platform that makes it easy for experimenters to collect data from as many participants as desired. **Human Intelligence Tasks (HITs)** are first posted by experimenters. Then participants (MTurk workers) complete the HITs, the results are saved, and payments are issued. Participants remain anonymous to the experimenters.

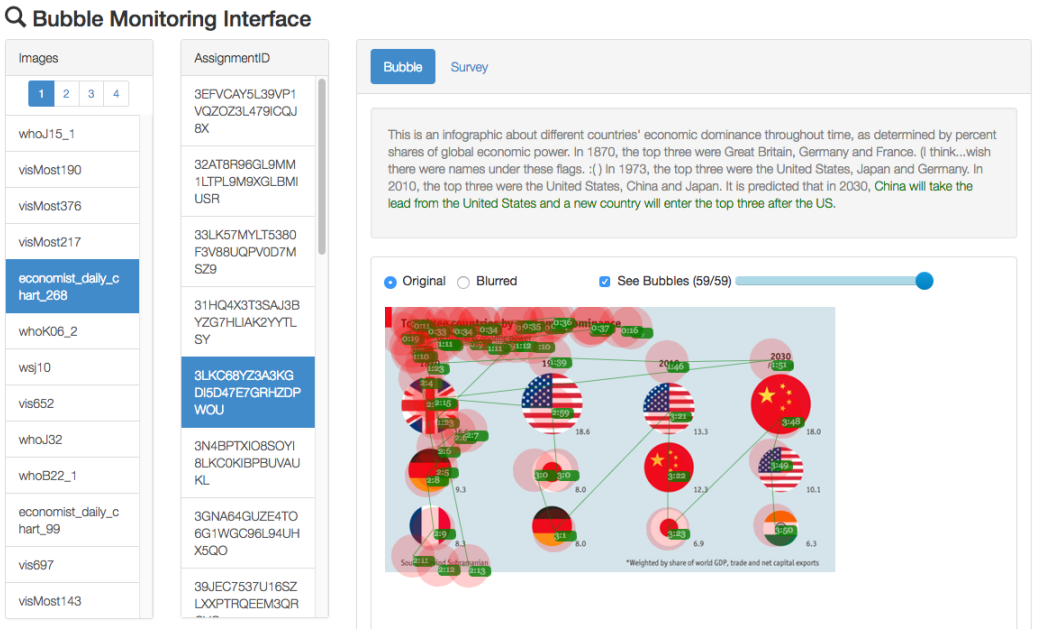


Fig. 3. Monitoring interface for manually inspecting the results of experiments. An experimenter can use a slider to explore the temporal sequence and evolution of bubble clicks and description text, for each image and participant.

We compared BubbleView clicks to eye fixations on information visualizations [Borkin et al. 2016], natural scenes [Xu et al. 2014], and webpages [Shen and Zhao 2014]. We also analyzed the relationship between BubbleView and related crowdsourcing methodologies: explicit importance annotations on graphic designs [O'Donovan et al. 2014], as well as mouse movements on natural images [Jiang et al. 2015]. We deployed 10 experiments with 28 different parameter combinations on MTurk (Table 1): 7 experiments with information visualizations (testing 4 different bubble radius sizes on 3 different image subsets), 2 experiments with natural scenes (comparing mouse clicks to mouse movements), 7 experiments with static webpages (3 bubble radius sizes x 2 viewing times with free-viewing, and a separate description task), 1 experiment with graphic designs, and 11 experiments with another dataset of natural scenes (3 image blur sigmas x 3 bubble radius sizes with mouse clicks, and 2 blur sigmas with mouse movements).

4.1 Tasks and Procedures Overview

We collected BubbleView data for 51 images selected out of each of 5 datasets (with additional images from the MASSVIS dataset). We used two different tasks with the following instructions: 1) description: "click and describe the image", 2) free-viewing: "click anywhere you want to look" (Figure 2). The description task required at least 150 characters to ensure that participants completed the task with enough thoroughness. For the free-viewing task, the image description was not required but the time for viewing each image was fixed to either 10 sec or 30 sec. The description task is most appropriate for image types containing sufficient textual content to describe. The description task was used for information visualizations, while the free-viewing task was used for natural images, to make the BubbleView task instructions as close as possible to the original eye

Dataset & Image Type	Exp.	Experiment Parameters				
		Task type	Blur sigma	Bubble radius	Time (sec)	Mouse modality
MASSVIS [Borkin et al. 2016] Information visualizations	1.1	describe	40	16, 24, 32	unlim.	click
	1.2	describe	40	24, 32, 40	unlim.	click
	1.3	describe	40	40	unlim.	click
OSIE [Xu et al. 2014] Natural scenes	2.1	free-view	30	30	10	click
	2.2	free-view	30	30	5	move
FIWI [Shen and Zhao 2014] Static webpages	3.1	free-view	50	30, 50, 70	10, 30	click
	3.2	describe	50	30	unlim.	click
GDI [O'Donovan et al. 2014] Graphic designs	4	free-view	30	50	10	click
SALICON [Jiang et al. 2015] Natural scenes	5.1	free-view	30, 50, 70	30, 50, 70	10	click
	5.2	free-view	30, 50	30	5	move

Table 1. Overview of BubbleView experiment settings including different image stimuli and parameters varied per experiment. We deployed a total of 10 experiments with 28 different parameter combinations. Bubble sigma and bubble radius are in pixels. Mouse modality corresponds to whether an image was revealed by discrete clicks, or by continuously moving the mouse cursor.

tracking experiments (Borkin et al. [2016] and Xu et al. [2014], respectively). We compared both task types on website images.

The free-viewing task had 17 images per HIT, for a total of 3 different HITs to cover all 51 images (no overlap of images among HITs). With a 10 sec viewing time, a single HIT was timed to take 2.8 minutes to complete; with 30 sec of viewing time, 8.5 minutes. For the description task, since time per image was estimated to be significantly longer, there were only 3 images per HIT, for a total of 17 different HITs to cover all 51 images. No explicit time constraints were placed on this task. On average, the description HITs took about 9 minutes to complete.

To accept one of our HITs, a participant had to have an approval rate of over 95% and live in the United States. After acceptance, the participant was asked to sign the informed consent before participating in the study. All participants were paid with approximately \$0.1/min rate which we translated to \$0.3 for the free-viewing task with 10 sec of viewing, \$0.9 for the free-viewing task with 30 sec of viewing, and \$0.5 for the description task[6]. All participants were paid regardless of whether they completed the task successfully or not. Some participant data was filtered out (see Supplemental Material) which is why the original number of participants recruited is not always equal to the final number of participants used for analysis.

[6]Our original estimates were that the description task would take 1.5 min per image, whereas in reality it took an average of 3.2 min per image. We later issued additional bonuses to compensate participants.

4.2 Analysis Overview

Across all the experiments comparing BubbleView clicks to eye fixations we used the same set of analyses, which we describe here. We compared how well the distribution of BubbleView clicks approximates the distribution of eye fixations, using two metrics commonly used for saliency evaluation: Pearson's Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS) [Bylinskii et al. 2016a]. While the two metrics provide complementary evidence for our conclusions, the NSS metric also allows us to account for differences in attentional consistency between participants (inter-observer congruency) across datasets.

Converting clicks and fixations into maps

Given a set of eye fixations on an image, we generate a **fixation map** by blurring the fixation locations with a Gaussian, with a sigma equal to one degree of the visual angle to approximate both the eye fovea and the measurement error of the eye tracker (a common evaluation choice [Bylinskii et al. 2016a; Le Meur and Baccino 2013]). This produces a continuous map which, when properly normalized, can be interpreted as a 2D distribution containing the probability of participants looking at each image region. Similarly, given a set of BubbleView mouse clicks on an image, we compute a **BubbleView click map** by blurring the click locations with a Gaussian with the same sigma as for the ground truth fixation maps. We used a sigma of 10 for the OSIE dataset, and a sigma of 25 for the MASSVIS and FiWI datasets. More generally, we refer to both fixation and click maps in this paper as **importance maps** for an image.

Measuring the similarity between clicks and fixations

We use two different metrics to measure the similarity scores between BubbleView clicks and eye fixations: Pearson's Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS). We compute similarity at the image level, by comparing the distributions of all clicks and fixations, across participants, per image. We then average all the per-image scores to obtain the similarity scores for a dataset.

To obtain the **CC score** for an image, we measure how well the click map predicts the fixation map, as a correlation between the two maps (see the Supplemental Material for details). The CC score is 0 when the two maps are not correlated, and 1 when they are identical. To obtain the **NSS score** for an image, we measure how well the click map predicts the discrete fixation locations (in this case, we do not compute a fixation map). We compute the average click map value at the fixated locations, after normalizing the click map. A map that is at chance at predicting fixation locations would receive an NSS score of 0, while a positive NSS indicates predictive power.

The advantage of the CC score is that it is bounded between 0 and 1, and can provide a simple, interpretable summary score that is ambivalent to the number of fixations that were used to generate the fixation map. The advantage of the NSS score is that it is computable for different numbers of eye tracking participants, and we use it for finer-grained analyses to examine how performance changes as we increase the number of participants. NSS is not bounded; to turn NSS into a bounded score, we can normalize it by inter-observer consistency, as described below.

Accounting for inter-observer consistency

If different eye tracking participants look at different regions of the image, they can not be used to predict each other's fixations. In these cases, BubbleView clicks will also not be as predictive of the fixations. For a fair evaluation, we normalize the BubbleView scores by the consistency of the eye tracking participants in a given dataset.

Consistency between eye tracking participants is measured in the following way: the fixations of all but one observer (i.e., N-1 observers) are aggregated into a fixation map which is used to predict the fixations of the remaining observer. This is repeated by leaving out one observer at a time, and then averaging the prediction performance to obtain the resulting inter-observer congruency (**IOC**) or inter-subject consistency [Borji et al. 2013; Le Meur and Baccino 2013; Wilming et al. 2011]. We measure IOC using the NSS metric.

We first compute the NSS score of the BubbleView click map at predicting all the eye fixations collected on an image, across all the observers. Then we normalize this score by the IOC of the eye tracking participants on that dataset. The resulting **normalized NSS** score can be interpreted as: the percent of the eye fixations accounted for, or predicted by, the BubbleView clicks.

Measuring performance in the limit
We consider performance when the number of study participants is taken to the limit, to get an upper bound on performance and determine if any systematic differences exist between methodologies that can not be reduced by gathering more data. To do this, we measure the ability of BubbleView click maps to predict ground-truth fixation locations, for different numbers of BubbleView participants. We obtain an NSS score for different numbers of participants n, by randomly selecting n participants for each of 10 splits, and averaging the results. Then we fit these scores to the power function $f(n) = a * n^b + c$, constraining b to be negative. Taking n to the limit, c is the NSS score at the limit. In cases where the total number of BubbleView participants for a particular experiment is not enough for a robust model fitting, we omit this analysis.

5 EXPERIMENTS COMPARING BUBBLEVIEW CLICKS TO EYE FIXATIONS

5.1 Experiment 1: comparison to eye fixations on information visualizations

We began by exploring how well BubbleView clicks on information visualizations gathered on MTurk approximate eye fixations collected in a controlled lab setting. In the initial experiments in Kim et al. [2015], we had gathered BubbleView data on 51 visualizations with a bubble radius size of 16 pixels. Here we extended these experiments to explore the effect of bubble radius size and number of participants on the quality of BubbleView data. We varied the bubble radius between 16 and 40 pixels, and collected up to 40 participants worth of clicks per image.

Motivating questions
➔ How does bubble radius size affect performance?
➔ How many BubbleView participants is enough?

Stimuli
The MASSVIS dataset contains over 5,000 information visualization images, of which 393 "target" images contain the eye movements of 33 participants free-viewing each image for 10 seconds as part of a memory test at the end of the study [Borkin et al. 2016]. In the eye tracking set-up, images were shown full-screen with a maximum dimension of 1000 pixels to a side, where 1 degree of viewing angle corresponded to 32.6 pixels. Participants made on average 39 fixations per image, or 3.9 fixations/sec.

We selected 202 from the total 393 target images, spanning infographic, news media, and government publication categories (Figure 4). We chose visualizations that had sufficiently large text and enough context to understand them without requiring specialized knowledge. We resized the images to half their original size with a maximum dimension of 500 pixels to a side. The images were blurred with a sigma of 40 pixels, which we found distorted the text in these images beyond

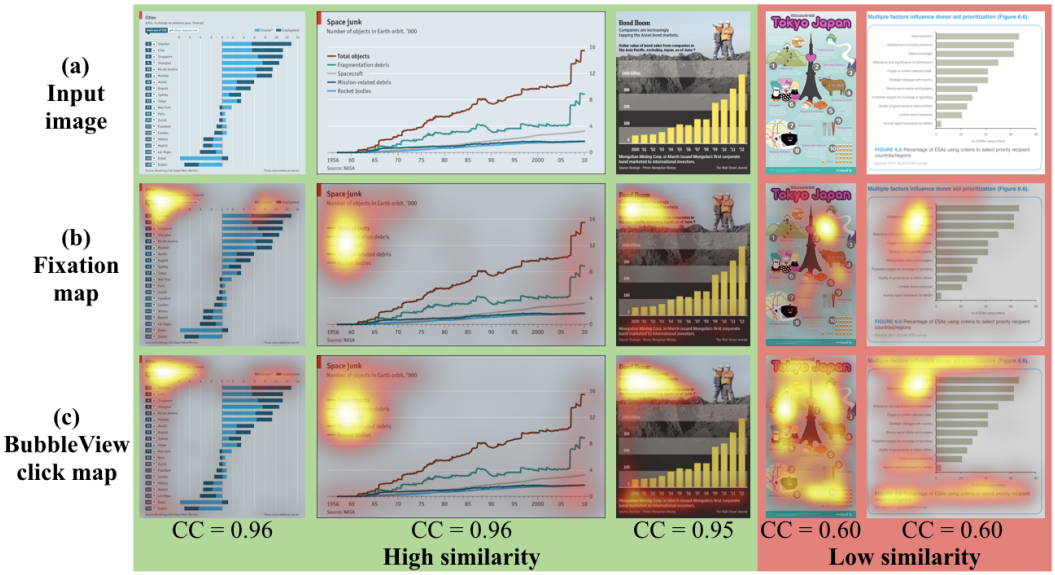


Fig. 4. Example images from the MASSVIS dataset. Dataset images (a), with corresponding ground-truth fixation maps (b) and BubbleView click maps (c). We show cases where BubbleView maps have high similarity, and cases with low similarity, to fixation maps.

legibility [Borkin et al. 2016; Kim et al. 2015].

Method

We ran a series of experiments to progressively find a bubble radius that best approximates eye fixations: **Exp. 1.1** with one set of 51 images and bubble radius sizes of 16, 24, and 32 pixels respectively, **Exp. 1.2** with another set of 51 images and bubble radius sizes of 24, 32, and 40 pixels, and **Exp. 1.3** with the remaining 100 images with a bubble radius of 32 pixels, which we determined from the first two experiments to produce good data quality. A bubble radius of 32 pixels corresponds to about 2 degrees of visual angle in the eye tracking studies on the original-sized images.

In a single HIT, participants were shown a random sequence of 3 images, and asked to describe each image with no time constraints on the task, allowing for individual differences in the time to write image descriptions.

For Exp. 1.1, we requested enough HITs so that each image would be seen by an average of 40 participants. From this experiment we found that 10–15 participants are sufficient for achieving high similarity scores to eye fixations, and proceeded to collect an average of 10–15 participants for each image in Exp. 1.2 and Exp. 1.3.

Results on bubble size

Participants explored each image for an average of 3 minutes, iterating between clicking around and typing text. As bubble size increased, the number of clicks and total task time monotonically decreased. Participants made an average of 103 clicks per image (0.5 clicks/sec) with a bubble radius of 16 pixels, 65 clicks (0.3–0.4 clicks/sec) with a bubble radius of 32, and 55 clicks (0.3 clicks/sec) with a bubble radius of 40 pixels. Depending on the bubble size, participants spent 15–30% of the task time clicking, and the rest of the time typing a description. After receiving a number of

participant complaints about task difficulty at a bubble radius of 16 pixels, we discontinued the use of this bubble radius in future experiments.

We computed the similarity between the BubbleView click maps and ground truth fixation maps across all images for all settings of bubble radius (Table 2). To make scores comparable, we set the number of participants $n = 10$ when computing the BubbleView click maps (the common denominator across all experiments). The similarities between the BubbleView click maps and the fixation maps were close across all bubble radius sizes (CC = 0.82–0.86). Because the different subsets of the MASSVIS dataset used in Exp. 1.1–1.3 had different inter-observer consistency (IOC) values[7], normalized NSS scores are more comparable across experiments than raw NSS scores. The normalized NSS score was very similar across all bubble radius sizes, with BubbleView clicks accounting for an average of 89–90% of eye fixations with 10 participants, and climbing up to 92% for larger numbers of participants ($n \geq 18$). Running a one-way ANOVA with bubble size as the factor, we did not find any significant effects of radius size on the similarity of clicks to fixations, under either of CC and NSS scores ($F < 1$ for all comparisons). Although the number of clicks changed, the overall pattern of bubble clicks remained the same (Figure 5).

Take-aways: no significant differences were found between bubble sizes in terms of similarity of BubbleView clicks to eye fixations. Bubble sizes in the range 24–40 pixels were found appropriate. Smaller bubble sizes increased the task time and effort.

Fig. 5. We found few differences in the resulting click maps from different settings of the bubble radius. Plotted here are the clicks of 3 participants (b-d) who explored the same image (a) with BubbleView, but with a different bubble size: 16, 24, and 32 pixel radius, respectively. The smaller the bubble, the more clicks a participant made, and the longer the task took to complete. Overall, the same regions of interest tended to be clicked on, despite differences in bubble sizes.

Results on number of participants

In Exp. 1.1 we collected an average of 40 participants of BubbleView clicks per image to investigate how BubbleView maps change with the number of participants (Figure 6). As described in Section 4.2, we fit power functions to the NSS scores for different numbers of participants to extrapolate performance. We found that after about 10–15 participants, the similarity of BubbleView click maps to ground truth fixation maps was already 97-98% of the performance achievable in the limit. The NSS score was extrapolated to increase to 1.31 in the limit (95% C.I. [1.312, 1.315]) with a bubble size of 16, 1.32 in the limit (95% C.I. [1.320, 1.324]) with a bubble size of 24, and 1.31 in the limit (95% C.I. [1.306, 1.310]) with a bubble size of 32. As a result of these analyses, we used an average of 10–15 participants for all future BubbleView experiments.

[7]This is an artifact of the images being different in the different subsets. In particular, Exp. 1.1 ended up containing more news media images and less government and infographic images than Exp. 1.2.

Exp. 1: visualizations	Bubble Radius (pixel)	CC	NSS	Normalized NSS
Exp. 1.1: 51 visualizations Description task (ground-truth IOC: 1.42)	16	0.86	1.27	89% ($n = 10$)
		0.87	1.30	92% ($n = 38$)
	24	0.86	1.27	89% ($n = 10$)
		0.87	1.30	92% ($n = 39$)
	32	0.86	1.27	89% ($n = 10$)
		0.87	1.29	91% ($n = 40$)
Exp. 1.2: 51 visualizations Description task (ground-truth IOC: 1.33)	24	0.82	1.20	90% ($n = 10$)
		0.84	1.22	92% ($n = 20$)
	32	0.84	1.20	90% ($n = 10$)
		0.85	1.22	92% ($n = 18$)
	40	0.83	1.19	89% ($n = 10$)
		0.84	1.19	89% ($n = 11$)
Exp. 1.3: 100 visualizations Description task (ground-truth IOC: 1.35)	32	0.84	1.21	90% ($n = 10$)
		0.84	1.21	90% ($n = 10$)

Table 2. We evaluated BubbleView clicks at approximating ground-truth eye fixations on the MASSVIS dataset by varying the bubble radius. We ran 3 sets of experiments on different subsets of the MASSVIS dataset. We measured the cross-correlation (CC) between BubbleView click maps and ground truth fixation maps, averaged over all images (CC has an upper bound of 1). The normalized scanpath saliency (NSS) score measured how well BubbleView click maps predict discrete fixation locations, averaged over all images. The NSS upper bound depends on the ground-truth data, so we included the inter-observer consistency (IOC) score of the eye tracking participants (measured using NSS). Normalizing the NSS score of the BubbleView maps by IOC allows us to report the percent of ground-truth fixations predicted by the BubbleView maps. To make the scores comparable across all the experiments, we fixed the number of participants to $n = 10$. In gray we report the results obtained by including all n participants that were collected for each experiment. The difference in CC and NSS scores with different bubble radius sizes was not significant ($F < 1$ for all comparisons).

Take-aways: 10–15 participants worth of BubbleView clicks already accounted for up to 97-98% of the performance achievable in the limit of the number of participants.

Results on ranking elements by importance

We also explored the relationship between BubbleView clicks and eye fixations at ranking visualization elements by importance. For this purpose we used the element segmentations (e.g., title, axis, legend, etc.) available in the MASSVIS dataset [Borkin et al. 2016]. For each of the 202 visualizations from Exp. 1, we overlapped the element segmentations with the fixation map of the visualization, and took the maximum value of the fixation map within the element's boundaries as its **importance score**, as in [Bylinskii et al. 2016b; Jiang et al. 2015]. We averaged the element scores across all 202 visualizations to obtain an aggregate importance score for each type of element (Figure 7). We repeated this computation using the BubbleView click maps of the visualizations to get another set of importance scores for the same elements. The ranking of elements by importance scores according to BubbleView clicks is highly correlated to the ranking according to eye fixations (Spearman correlation = 0.96).

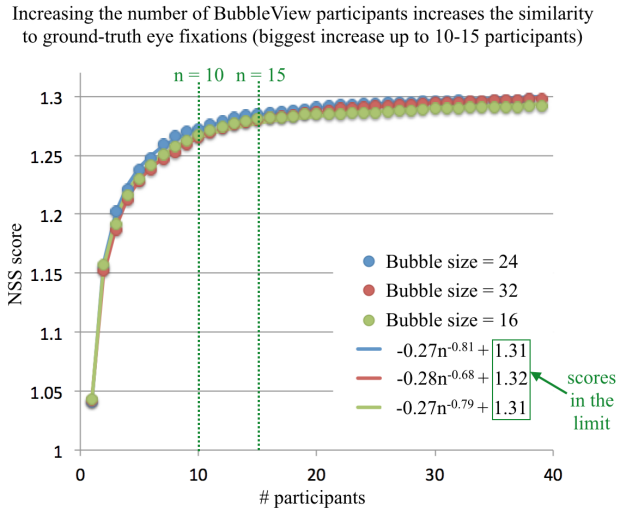


Fig. 6. The NSS score of BubbleView click maps computed with different numbers of participants, when used to predict discrete fixation locations on the MASSVIS dataset. Each point represents the score obtained at a given number of participants, averaged over 10 random splits of participants, and all 51 images used in Exp. 1.1. We include data points from 3 different bubble radius sizes. By fitting power functions of the form $an^b + c$ to each set of points, we find that these scores do not change significantly in the limit of participants ($n \to \infty$).

Take-aways: BubbleView can be used to rank visualization elements by importance, predicting how often people would fixate those elements during natural viewing.

5.2 Experiment 2: comparison to eye fixations on natural images

In Experiment 1 we found that BubbleView clicks offered a very good approximation to eye fixations on information visualizations with a description task. However, because free-viewing is a more common setting for human perception studies of natural images (specifically for saliency datasets), we wanted to determine if BubbleView clicks can also be used to approximate free-viewing fixations on natural images. We used similar BubbleView settings to the ones found in Exp. 1: a bubble size of 30 pixels and 15 participants worth of clicks.

Motivating questions
➔ Does BubbleView generalize to natural images with a free-viewing task?

Stimuli
The OSIE dataset contains 700 natural images with multiple dominant objects per image [Xu et al. 2014]. Eye movements on this dataset were collected by instructing 15 participants to free-view each image for 3 seconds. Participants made an average of 9.3 fixations per image (3.1 fixations/sec). In this eye tracking setup, images were presented at a resolution of 800 × 600 pixels and 1 degree of viewing angle corresponded to 24 pixels. For our study, we randomly sampled 51 OSIE images (Figure 8), downsized them to 640 × 480 pixels, and blurred them with a sigma of 30 pixels.

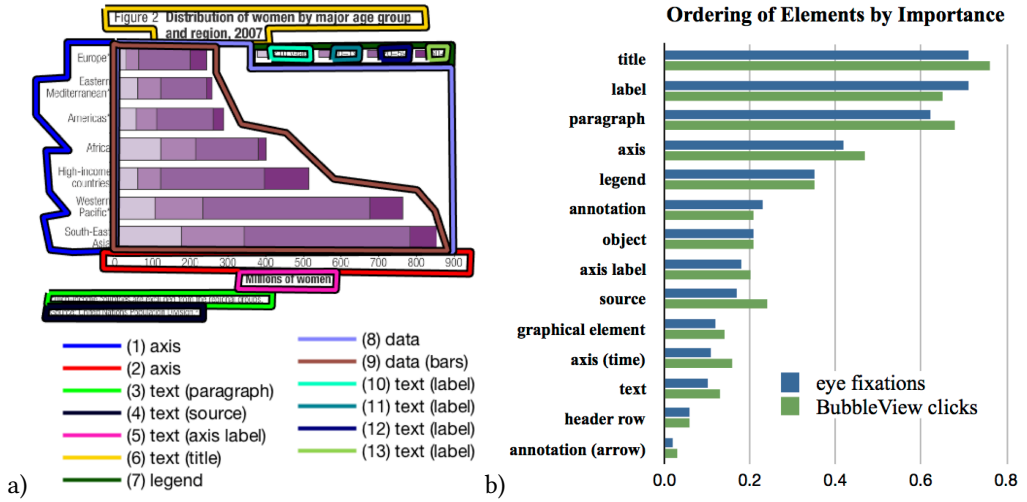


Fig. 7. (a) An example of a labeled visualization from the MASSVIS dataset. (b) By overlapping fixation maps and BubbleView click maps with such element annotations (and taking the maximum value of the map inside the element), we obtain an importance score for each element in each visualization. By averaging across 202 visualizations, we obtain an aggregate importance score per element type.

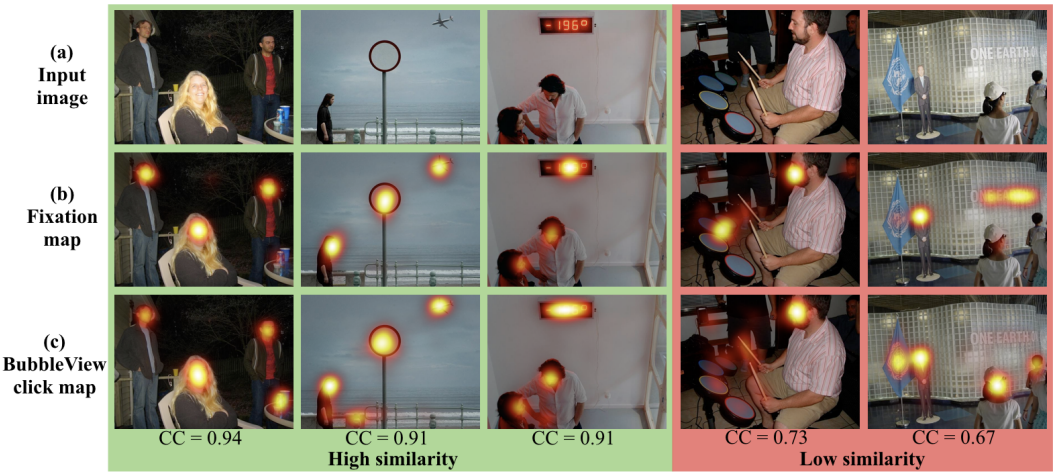


Fig. 8. Example images from the OSIE dataset. Dataset images (a), with corresponding ground-truth fixation maps (b) and BubbleView click maps (c). We show cases where BubbleView maps have high similarity, and cases with low similarity, to fixation maps.

Method

In **Exp. 2.1**, we asked participants to free-view a series of images and to click anywhere they want to look for 10 sec per image. We used a bubble radius of 30 pixels, equal to about 1.5 degrees of visual angle in the eye tracking study. Although the viewing time for the OSIE eye tracking study was 3 sec per image, we increased this time for the BubbleView experiment to account for the time of clicking a mouse. We piloted different viewing times and determined 10 sec to be appropriate

(clicking took about 3 times as long as natural viewing). We collected an average of 60 participants worth of BubbleView click data for each image.

Apart from ground truth eye fixations, mouse movements using the related SALICON methodology are also available for the OSIE dataset [Jiang et al. 2015]. To facilitate a direct comparison between BubbleView and SALICON, in **Exp. 2.2** we re-ran data collection with BubbleView, replacing mouse clicks with mouse movements, with a bubble radius of 30 pixels. As in SALICON, we used a task time of 5 seconds. The results of this experiment are discussed in Section 6.2, in the context of other comparisons to the SALICON methodology.

Results

During 10 seconds of viewing, participants made an average of 13.1 clicks, or about 1.3 clicks/sec - three times fewer clicks than fixations per second.

In Exp. 2.1, the similarity between BubbleView click maps and ground truth fixation maps with free-viewing on natural images was smaller (NSS = 2.61, CC = 0.81, Table 3) than in Exp. 1 with visualizations. Even though eye tracking participants are quite consistent with each other on the OSIE dataset (IOC = 3.35), BubbleView participants are not as predictive of eye tracking participants in this case. BubbleView clicks of 54 participants can predict 80% of eye fixations, while the projected performance in the limit only converges to 82% (95% C.I. [2.742, 2.754]). However, 10 BubbleView participants can already account for 78% of eye fixations.

Exp. 2.2 showed that a related methodology using a moving-window approach [Jiang et al. 2015] is no better at approximating ground-truth eye fixations on this dataset (Table 3). In fact, to achieve the same performance as BubbleView, SALICON actually requires more participants (Section 6.2). BubbleView can serve as an affordable and scalable alternative. When running a large number of eye tracking experiments is infeasible, BubbleView can be used for studying human perception and collecting large-scale saliency datasets (as in Jiang et al. [2015]; see Bylinskii et al. [2017b]).

Take-aways: Similarity between BubbleView clicks and eye fixations is lower on natural images with a free-viewing task than with visualizations with a description task. Despite this, 10 BubbleView participants can already account for 78% of eye fixations on natural images, so BubbleView can still serve as an affordable approximation to eye tracking.

5.3 Experiment 3: comparison to eye fixations on static webpages

Apart from natural images, webpages are another image type that frequently serve as the focus of eye tracking and usability studies [Buscher et al. 2009; Chen et al. 2001; Nielsen and Pernice 2009; Rodden et al. 2008; Shen et al. 2015; Shen and Zhao 2014]. For this reason, we wanted to test the generalizability of the BubbleView methodology to webpages. Because the static webpage images were denser in visual and information content than the information visualizations and natural images from the first two experiments, we evaluated a number of different BubbleView settings to try to find the best approximation to eye fixations. We varied bubble radius size and viewing time. As in the original FiWI eye-tracking experiment, we started with a free-viewing task. Similar to Exp. 1, we also tried a description task with unlimited task time.

Motivating questions

➔ Does BubbleView generalize to webpages?
➔ How do the task and viewing time affect performance?
➔ Does viewing time interact with bubble size?

Table 3. We evaluated BubbleView clicks at approximating ground-truth eye fixations on the OSIE dataset. We ran BubbleView data collection using mouse clicks (Exp. 2.1) and using mouse movements (Exp. 2.2). For comparison, we also include the performance of the SALICON methodology on the same dataset (Section 6.2). For fair comparison with in-lab SALICON, we only used $n = 12$ participants per image per study. The difference in scores at $n = 12$ participants was not significant [F(200)=1.81, n.s.]. In gray we report the results obtained by including all n participants that were collected for each experiment.

Exp. 5.3: natural scenes (ground-truth IOC: 3.35)	CC	NSS	Normalized NSS
BubbleView (clicks)	0.81 0.84	2.61 2.69	78% ($n = 12$) 80% ($n = 54$)
BubbleView (movements)	0.81 0.83	2.52 2.55	75% ($n = 12$) 76% ($n = 49$)
SALICON	0.81 0.84	2.52 2.61	75% ($n = 12$) 78% ($n = 92$)
In-lab SALICON	0.81 0.81	2.61 2.61	78% ($n = 12$) 78% ($n = 12$)

Stimuli

The FiWI dataset contains 149 screenshots of static webpages collected from various sources on the Internet and sorted into pictorial (dominated by pictures such as photo sharing websites), text (high density text such as encyclopedia websites), and mixed types [Shen and Zhao 2014]. Eye movements on this dataset were collected by instructing 11 participants to free-view each webpage for 5 seconds. Participants made an average of 17.9 fixations per image (3.6 fixations/sec). In this eye tracking setup, 1 degree of visual angle was approximately 50 pixels.

We sampled 17 images from each of the three categories (pictorial, text, mixed), resulting in a total of 51 images (Figure 9). We downsized the images from 1360 × 768 pixels to 1000 × 565 pixels to fit within a typical MTurk browser window, while preserving image aspect ratios. These webpages tended to have more varied font size compared to the images in Exp. 1–2. We manually selected a blur sigma of 50 pixels to distort the text on these images beyond legibility.

Method

We ran experiments with two task types where participants were asked to either free-view or describe each webpage. In **Exp. 3.1**, with the free-viewing task, we used a 2 x 3 factorial design (viewing time: 10 sec or 30 sec; bubble radius: 30, 50, or 70 pixels). In **Exp. 3.2**, with the description task, we used a bubble radius of 30 pixels and unlimited time. We collected an average of 15 participants worth of BubbleView click data for each image under each task.

Results on stimuli

In the free-viewing task (Exp. 3.1), participants made an average of 1.0–1.8 clicks/sec, while in the description task (Exp. 3.2), participants made an average of 0.5 clicks/sec, indicating that they spent more than half the time typing descriptions. Clicking took about 3 times longer than natural viewing. As in Exp. 1, the number of clicks per second monotonically decreased with increasing bubble size, even though viewing time was fixed (Exp. 3.1). Tripling the viewing time from 10 to 30 seconds did not quite triple the number of clicks, but increased them by 2.2–2.6 times.

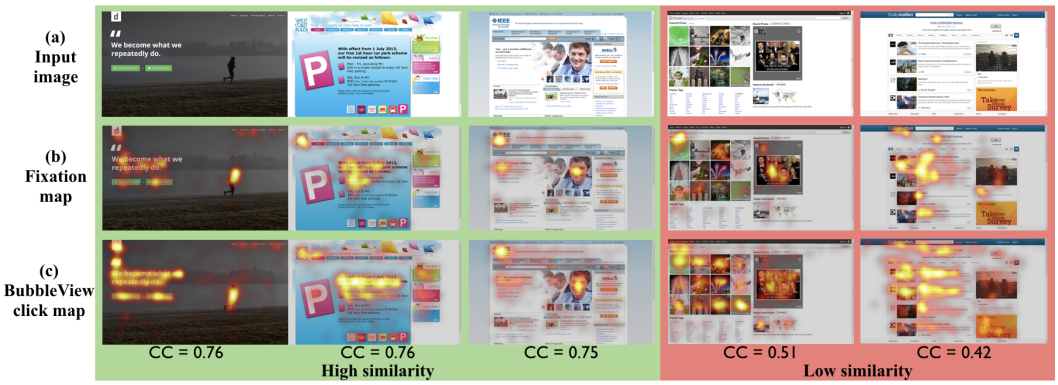


Fig. 9. Example images from the FiWI dataset. Dataset images (a), with corresponding ground-truth fixation maps (b) and BubbleView click maps (c). We show cases where BubbleView maps have high similarity, and cases with low similarity, to fixation maps.

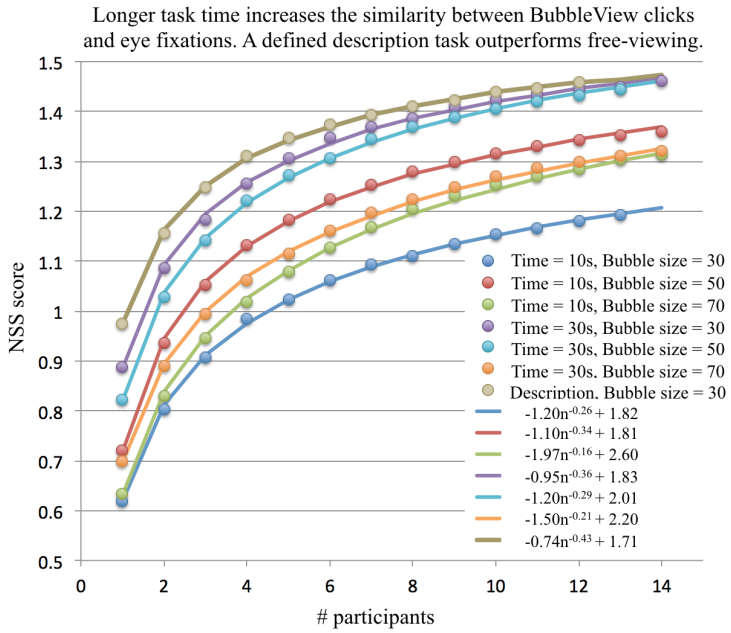


Fig. 10. The NSS score of BubbleView click maps computed with different numbers of participants, when used to predict discrete fixation locations on the FiWI dataset. Each point represents the score obtained at a given number of participants, averaged over 10 random splits of participants, and all 51 images.

The similarity between BubbleView click maps and ground truth fixation maps on webpages was lowest of all image types tested so far in Exp. 1–3 (Table 4). However, the inter-observer consistency of eye tracking participants is also lowest on webpages (IOC = 1.85). Recall that IOC between eye tracking participants serves as an upper bound for how well BubbleView clicks can predict eye fixations. After accounting for IOC, the normalized NSS scores show that BubbleView clicks can account for up to 78% of eye fixations on webpages, similar to the score on natural images (Exp. 2).

IOC was highest on the all-text webpages (NSS = 1.97), followed by the pictorial (NSS = 1.77) and mixed (NSS = 1.80) webpages. While the difference in NSS scores was not significant across webpage types for the similarity between BubbleView clicks and eye fixations, the NSS scores were consistently higher for the text webpages. Only for one case, with a bubble size of 30 pixels and 10 seconds of viewing, were the NSS scores for the pictorial webpages the highest (Supplemental Material). This provides evidence that clicks tend to be more consistent with fixations on text elements.

Take-aways: Both fixation and click data is more varied on webpages. Webpage images with lower IOC scores (more eye tracking variability) also had worse BubbleView similarity scores. Normalizing for IOC, BubbleView clicks can account for 78% of eye fixations on webpages (as for natural images).

Results on time, bubble size, and task

We ran a two-way ANOVA (time × bubble size) on Exp. 3.1. The main effect of time on the CC and NSS scores was significant [CC: $F(1,300)=19.25$, $p < .01$, NSS: $F(1,300)=9.65$, $p < .01$], respectively but the effect of bubble size was not [CC: $F(2,300)=1.92$, n.s., NSS: $F(2,300)=1.14$, n.s.]. The interaction effect between time and bubble size was significant [CC: $F(2,300)=6.95$, $p < .01$, NSS: $F(2,300)=3.35$, $p < .05$].

With a viewing time of 10 seconds, a bubble size of 30 pixels was too small, achieving significantly lower CC scores than bubble sizes of 50-70 pixels ($p < .05$). With a viewing time of 30 seconds, however, a bubble size of 70 pixels was too large, achieving significantly lower CC scores than bubble sizes 30-50 pixels ($p < .01$). No significant differences were found among the NSS scores (Table 4). There exists a trade-off: with a longer viewing time, a smaller bubble radius provides more consistent clicks among participants; when limited by a shorter time, a larger bubble size becomes necessary.

Given a bubble size of 30-50 pixels, the CC scores were significantly higher for a task duration of 30 seconds compared to 10 seconds ($p < .05$). The difference in NSS scores was only significant for the bubble size of 30 pixels. No significant differences were found with a bubble size of 70 pixels. Overall, BubbleView click maps generated with longer task durations of 30 seconds or longer (including with a description task) better approximated eye fixations than with a 10 second task duration. From this we conclude that information-dense images like websites require either longer viewing times or better defined tasks than free-viewing.

From Exp. 3.2, we found that for small numbers of participants ($n < 12$), the description task generated BubbleView click maps more similar to ground-truth eye fixations than the free-viewing task under all settings (Figure 10). The difference between the tasks is larger for smaller number of participants, and decreases with each extra participant. The click data tends to converge faster when a targeted task like description is used. However, this advantage disappears with more participants and a longer task time (30 sec, 30 pixel bubble radius). A description task takes longer and is more expensive to run, but might be a better choice when few participants are available.

Take-aways: the less viewing time available, the larger the bubble size should be in order to better approximate free-viewing fixations. For a study with fewer participants, a description task is better than a free-viewing task.

Table 4. We evaluated BubbleView clicks at approximating ground-truth eye fixations on the FiWI dataset. BubbleView maps were computed with 12 participants for all experiments below. The score of the BubbleView maps predicting the ground-truth fixation maps is reported in CC, and the score of the BubbleView maps predicting the discrete fixation locations is reported in NSS. Normalized NSS is calculated by normalizing the NSS score by the inter-observer consistency (IOC) of the eye tracking participants.

Exp. 3: webpages (ground-truth IOC: 1.85)	Time (sec)	Bubble Radius (pixel)	CC	NSS	Normalized NSS
Free-viewing	10	30	0.52	1.20	65%
Free-viewing	10	50	0.57	1.34	72%
Free-viewing	10	70	0.56	1.30	70%
Free-viewing	30	30	0.63	1.45	78%
Free-viewing	30	50	0.61	1.41	76%
Free-viewing	30	70	0.57	1.32	71%
Description	unlim.	30	0.63	1.46	79%

6 EXPERIMENTS COMPARING BUBBLEVIEW TO RELATED METHODOLOGIES

6.1 Experiment 4: comparison to importance annotations on graphic designs

We hypothesized that the regions on an image where participants click using the BubbleView methodology correspond to the most important regions of the image. To test this hypothesis, we used the GDI dataset [O'Donovan et al. 2014] which comes with explicit importance annotations, where participants were instructed to annotate the image regions they considered important in graphic designs. We used this dataset to evaluate whether the number of BubbleView clicks on image regions corresponds to explicit judgements of importance.

Motivating questions
- ➔ Does BubbleView generalize to graphic designs?
- ➔ Do BubbleView clicks correspond to regions of importance on graphic designs?

Stimuli

The Graphic Design Importance (GDI) dataset contains 1,075 single-page graphic designs (e.g., advertisements, flyers, and posters consisting of text and graphical elements), collected from Flickr [O'Donovan et al. 2014]. No eye movements were collected for this dataset. O'Donovan et al. [2014] highlighted two downsides of eye movements for this type of data: (1) fixations vary significantly over individual elements (like text blocks) even though those regions should have a uniform importance, and (2) eye fixations may occur in unimportant regions as a design is scanned and do not reflect conscious decisions of importance. Instead, 35 MTurk participants were asked to label important regions with binary masks, and these masks were averaged over all participants to produce a final importance map per design. O'Donovan et al. [2014] noted that although importance maps produced by individual users are noisy, the average map gives a plausible relative ranking over design elements.

We sampled 51 images from the GDI dataset at the original resolution of 600×400 pixels (Figure 11). We blurred the images with a sigma of 30 pixels, manually chosen to distort text beyond

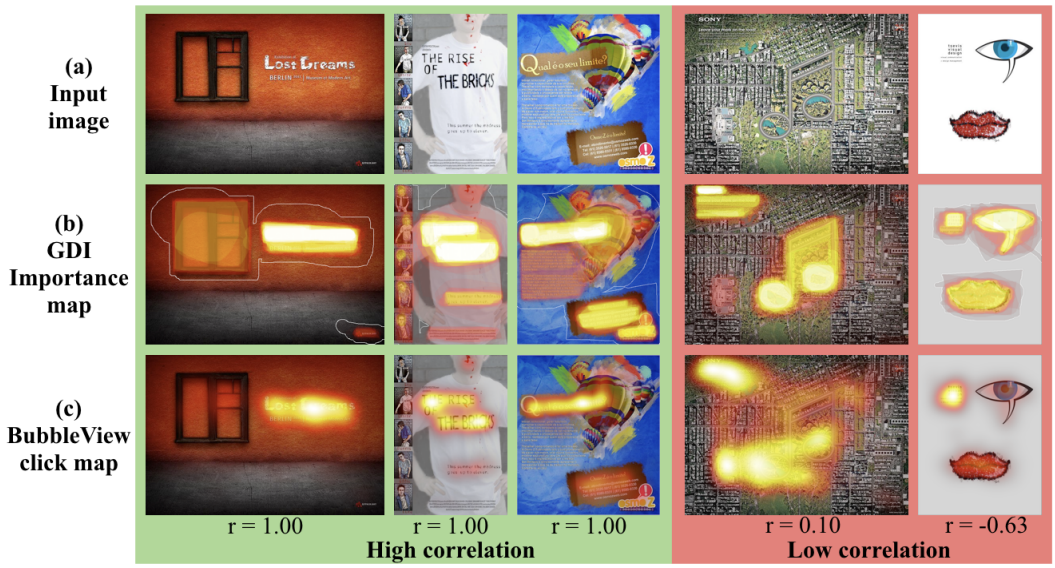


Fig. 11. Example images from the GDI dataset. Images from the dataset (a), along with the provided explicit importance annotations (b). We show cases where BubbleView maps have high correlation, and cases with low correlation, to the importance annotations, in terms of how design elements are ordered by importance (c).

recognition.

Method

We ran an experiment with a bubble radius of 50 pixels and viewing time of 10 seconds, in which participants were asked to free-view each graphic design. BubbleView strikes a balance between eye fixations and explicit importance judgements for these images: (1) like fixations, clicks are collected in a free-viewing setting and are not uniform over design elements, but (2) like explicit annotations, the decisions of where to click reflect conscious decisions of importance. We collected an average of 15 participants worth of BubbleView click data for each image.

Analysis

Unlike the quantitative evaluations in the previous sections, we did not directly compare the BubbleView click maps to the graphic design importance (GDI) maps. The spatial distributions of the explicit importance annotations in the GDI dataset are different from the click maps generated by our methodology. By construction, the importance annotations are uniform over design elements in the GDI dataset, while BubbleView clicks are not. For a fairer comparison, we computed the importance values each methodology assigns to different elements within each design (similar to the analysis at the end of Section 5.1).

We used bounding boxes to manually annotate all the elements in the 51 graphic designs chosen. For each design we normalized the GDI ground-truth importance map and the BubbleView click map. We took the maximum value of each map within an element's bounding box as the importance score of that element [Bylinskii et al. 2016b; Jiang et al. 2015]. We correlated the importance scores assigned by both methodologies to the elements in each design (Figure 12).

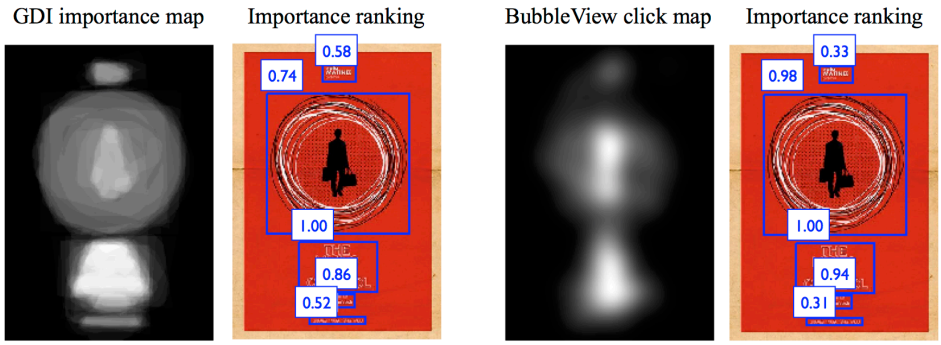


Fig. 12. Importance maps were overlapped with element bounding boxes (outlined in blue) and the maximum map value per box was taken to be the importance score for that element (scores are the numbers above each box). Maps were first normalized to have values between 0 and 1, so the importance scores for all the graphic design elements also fall within the same range, where 1 corresponds to the most important element. In the case of the GDI importance map, MTurk workers made explicit judgements about aspects of the graphic design they considered the most important. A region of a graphic design has an importance score of 1 if all MTurk workers labeled that element as important. In the BubbleView study, MTurk workers clicked a blurred graphic design to expose small regions of the design at full resolution. A region of a graphic design has an importance score of 1 if the density of MTurk clicks in that region was highest.

Results

Across all 51 graphic designs, we achieved an average Pearson correlation of 0.66 and an average Spearman (rank) correlation of 0.60 between the element importance scores as assigned by BubbleView versus the original GDI annotations. Over 70% of graphic designs had a correlation over 0.4. BubbleView importance maps can reasonably approximate explicit importance judgements for ranking elements of graphic designs, although there are some differences. For instance, the blurring of the image may interfere with visual features seen at different scales, as in the last two example images in Figure 11. Depending on the blur, certain visual elements might not be clicked on (e.g. in Figure 11, the *note* because it blended into the background when blurred; the *eye* because it was already visible in the blurred version).

Take-aways: BubbleView can be used to rank graphic design elements by importance. However, due to the varied feature sizes, blurring might significantly impact which design regions are clicked.

6.2 Experiment 5: comparison to mouse movements on natural images

The most similar methodology to BubbleView is SALICON [Jiang et al. 2015], which was introduced at roughly the same time[8]. SALICON is also intended to be used in a crowdsourcing setting to approximate eye fixations [Jiang et al. 2015]. The differences are that SALICON captures continuous mouse movements, instead of clicks, and images are blurred adaptively, with a multi-resolution blur, recomputed for each cursor position. We investigated whether BubbleView click maps are similar to SALICON mouse movement maps, when averaged over multiple participants. Because

[8]The SALICON and BubbleView methodologies were introduced a few months apart, but to different communities: Jiang et al. [2015] to computer vision and Kim et al. [2015] to human-computer interaction.

the SALICON blur is multi-resolution and adaptive, we experimented with different blur sigmas and bubble sizes in BubbleView, to find a fixed setting of parameters that best approximates the SALICON viewing conditions. We also compared SALICON and BubbleView at approximating eye fixations collected in a controlled lab setting, since both methodologies are presented as alternatives to eye tracking.

Motivating questions
➔ Under what settings does BubbleView most closely match SALICON?
➔ Which methodology better approximates eye fixations on natural images?

Stimuli
The SALICON dataset consists of mouse movements collected on 20K MS COCO (Microsoft Common Objects in Context) natural images [Lin et al. 2014]. In the original study, mouse movements were collected on Amazon's Mechanical Turk by presenting images to participants for 5 seconds each and allowing them to freely explore each image by moving the mouse cursor. We randomly sampled 51 images at the original image size of 640 × 480 pixels from the SALICON dataset (Figure 13).

Method
In **Exp. 5.1**, we used a 3 × 3 factorial design (blur sigma: 30, 50, and 70 pixels; bubble radius: 30, 50, and 70 pixels; see Figure 14). Using a free-viewing task, we had participants explore each image for 10 seconds each. We wanted to account for longer times to click, rather than move, the mouse.

To disentangle the influence of mouse clicks/movements versus fixed/adaptive blur on the methodology differences between SALICON and BubbleView, we ran **Exp. 5.2**, using BubbleView with a moving-window approach like SALICON, but maintaining a fixed blur kernel. In this setup participants used mouse movements to reveal image regions at normal resolution. We had two experiment conditions (bubble radius sizes of 30 and 50 pixels) with a fixed blur sigma of 30 pixels (found appropriate in Exp. 5.1) and viewing time of 5 seconds (as in SALICON). We collected an average of 15 participants worth of BubbleView click data for each image under each condition.

Results on using BubbleView to approximate SALICON
We ran a two-way ANOVA (blur × bubble size) on Exp. 5.1. The main effect of bubble size was not significant [CC: F(2,450)=2.28, NSS: F(2,450)=0.19, n.s.] (as found in Exp. 1 and 3.1). The main effect of blur on scores was significant [CC: F(2, 450)=19.97, $p < .01$, NSS: F(2,450)=6.86, $p < .05$]. BubbleView with a blur radius of 70 pixels achieved significantly lower CC scores than with other blur settings ($p < .01$ for all bubble sizes). We did not find an interaction effect between blur and bubble size [CC: F(4,450)=0.44, NSS: F(4,450)=0.04, n.s.]. We found highest similarity between BubbleView click maps and SALICON maps at bubble radius sizes of 30–50 pixels and blur sigma of 30–50 pixels (Table 5), for which the normalized NSS scores ranged from 77% to 82%.

What are the remaining differences? Using mouse movements, more points of interest are generated than using clicks. Many of the points sampled using mouse movements occur in the transition between regions in an image, and might be introducing noise into the data (Figure 15). This suggests that a different threshold might be more effective at converting continuous mouse movements into discrete points of interest. An advantage of the BubbleView clicks is that no such post-processing is necessary, since the clicks directly correspond to points of interest.

In Exp. 5.2, we modified BubbleView to collect continuous mouse movements and shortened the time per image to 5 sec, such that the only remaining difference with SALICON was the treatment of blur. We observed that the mean number of samples was 143.02 (SD=13.14) using the sampling rate

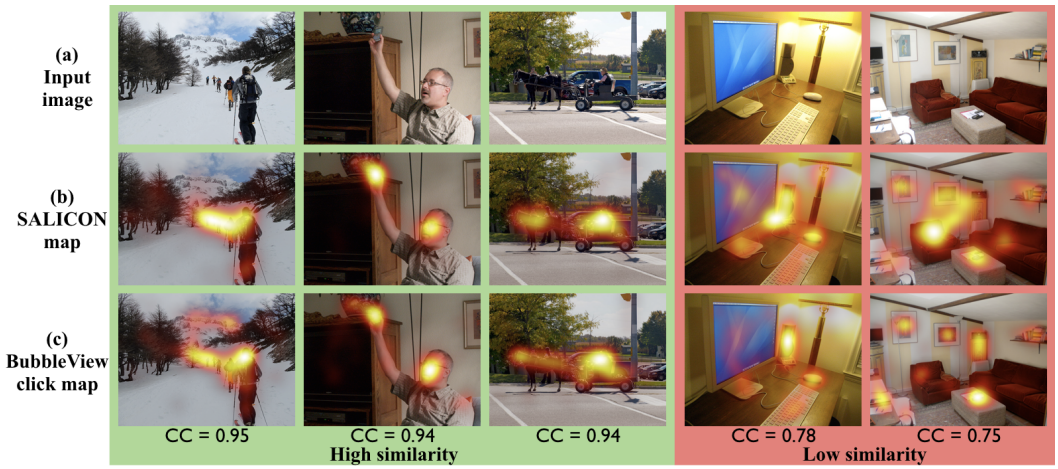


Fig. 13. Example images from the SALICON dataset. Example dataset images (a), and ground truth mouse movements collected by SALICON (b). We show cases where BubbleView maps have high similarity, and cases with low similarity, to SALICON maps (c).

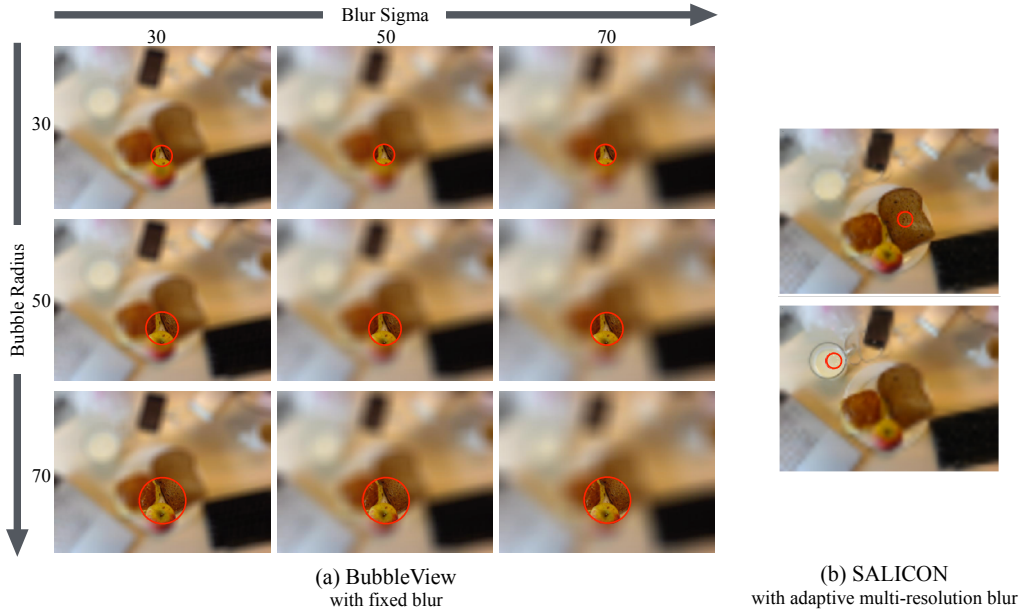


Fig. 14. We used 9 different parameter settings in our BubbleView experiments, on images from the SALICON dataset (a). We wanted to find a fixed setting of bubble size and blur to mimic the adaptive multi-resolution blur used in the SALICON methodology (b). The rightmost figure is from Jiang et al. [2015]; ©Martijn van Exel.

of 100 Hz, which translates to 14,302 raw samples, on average, per participant. This is significantly larger than the mean click count of 13.09 (SD=1.38) per participant in Exp 5.1.

Table 5. We evaluated BubbleView click maps (with $n = 12$ participants per image) at approximating SALICON mouse movements, measured using CC and NSS metrics. Normalized NSS is computed by taking into account the IOC of the SALICON participants (NSS = 1.50). Both bubble radius and blur sigma are measured in pixels. BubbleView with a blur radius of 70 pixels achieved significantly lower CC scores than with other blur settings ($p < .01$ for all bubble sizes). The other differences were not significant.

			Blur Sigma (pixel)		
			30	50	70
Bubble radius (pixel)	30	CC	0.84	0.84	0.78
		NSS	1.21	1.15	1.06
		Normalized NSS	81%	77%	71%
	50	CC	0.86	0.84	0.80
		NSS	1.23	1.15	1.04
		Normalized NSS	82%	77%	69%
	70	CC	0.84	0.84	0.79
		NSS	1.20	1.11	1.04
		Normalized NSS	80%	74%	69%

With the moving-window BubbleView setting the scores were: for bubble size 30: CC: 0.87, NSS: 1.21, normalized NSS: 81%; bubble size 50: CC: 0.88, NSS: 1.24, normalized NSS: 83%. Compared to the clicks, these scores were not statistically significantly different [$F(200) < 2.2$, n.s.]. In other words, BubbleView can approximate SALICON with or without mouse movements. Importantly, BubbleView can approximate SALICON without requiring a multi-resolution adaptive blur, simply with a single fixed blur setting. Our fixed blur setting is much less computationally expensive and does not require the pre-study system checks as in Jiang et al. [2015].

Take-aways: BubbleView with a bubble size of 30–50 pixels and a blur sigma of 30–50 pixels can approximate the continuous mouse movements and adaptive, multi-resolution blur of the SALICON methodology.

Results on using both methodologies to approximate eye fixations

In Exp. 2.2 we compared BubbleView clicks and mouse movements to SALICON mouse movements at approximating ground truth eye fixations on 51 OSIE images. The BubbleView click maps (with $n = 12$ participants, bubble radius of 30 pixels) achieve NSS = 2.61 (CC = 0.81) at predicting ground-truth fixation maps, compared to SALICON mouse movement maps which achieve NSS = 2.52 (CC = 0.81). It takes over 30 SALICON participants to achieve the same similarity to fixation maps as 12 BubbleView participants (Figure 16). Replacing BubbleView clicks with mouse movements actually decreases performance: NSS = 2.52 (CC = 0.81), but this drop in performance is not significant at the $p = .05$ level. For all feasible numbers of participants ($n < 60$ in Figure 16), BubbleView offers a better approximation to eye fixations than SALICON.

Data was also available for 12 in-lab participants who used the SALICON methodology to view images in a controlled lab setting [Jiang et al. 2015]. The in-lab SALICON maps, which capture these mouse movements, achieve NSS = 2.61 (CC = 0.81) when compared to fixation maps, the same score as our BubbleView maps (Table 3). From Figure 16 we can see that the performance of the in-lab SALICON is increasing at a greater rate than either BubbleView or online SALICON.

(a) points of interest from SALICON mouse movements

(b) points of interest from BubbleView mouse clicks

Fig. 15. When participants can move the mouse anywhere on the image without having to click, the collected data contains motion traces as byproducts (a). Instead of only capturing the points of interest in an image where an observer's attention stops, the moving-window approach also captures the transitions between these regions, which are less relevant and add noise to the data. Although these trajectories can be post-processed into discrete regions of interest, our approach is to directly collect participant mouse clicks on points of interest, with no further post-processing required (b).

However, more in-lab SALICON participants would be needed to see whether this trend continues. In any case, it requires a controlled lab setting, which we aim to avoid.

Take-aways: On a natural image dataset, BubbleView clicks better approximate eye fixations than SALICON mouse movements for all feasible numbers of participants ($n < 60$). BubbleView performed better with clicks than BubbleView with mouse movements.

7 DISCUSSION

Similarity of BubbleView clicks to eye fixations: We showed that across 3 different image types (information visualizations, natural images, and static webpages) and 2 types of tasks (free-viewing and description), BubbleView clicks provide a reasonable approximation to eye fixations collected in a controlled lab setting. Specifically, across all these image types BubbleView clicks accounted for over 75% of eye fixations when only 10–15 BubbleView participants were used (Tables II-IV). Of all settings, BubbleView clicks provided the best approximation to eye fixations on information visualizations with a description task, accounting for up to 90% of eye fixations with only 10 participants, and 92% with 20 participants (Table II). On both natural images and websites, BubbleView clicks could account for up to 78% of eye fixations with 10–12 participants (Tables III, IV). The fixations of eye tracking participants were much more consistent on the natural images than on the websites, so the viewing behavior on natural images should be easier to predict. Despite the remaining gap between BubbleView clicks and eye fixations for natural images and webpages, the fact that already 10–15 BubbleView participants achieves a reasonable approximation to fixations is promising for perception studies that might otherwise require specialized eye-tracking hardware.

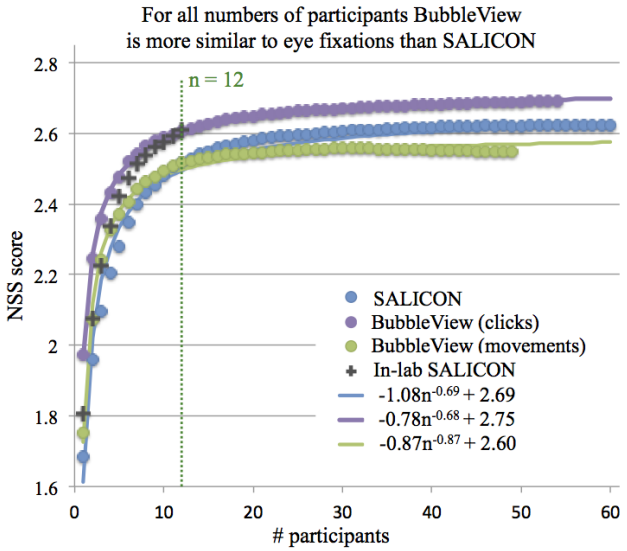


Fig. 16. The NSS score obtained by comparing mouse clicks and mouse movements to ground truth eye fixations on natural images in the OSIE dataset. We compare mouse clicks gathered using BubbleView on MTurk (purple), mouse movements gathered using BubbleView on MTurk (green), mouse movements gathered using SALICON on MTurk (blue), and mouse movements gathered using SALICON in a controlled lab setting (black crosses). Each point represents the score obtained at a given number of participants, averaged over 10 random splits of participants and all 51 images used.

Remaining differences between BubbleView clicks and eye fixations: Part of the remaining gap between BubbleView clicks and eye fixations is that BubbleView does not capture the unconscious movements of the eyes due to bottom-up, pop-out effects, or systematic biases. One such systematic bias commonly referred to in the eye tracking literature is center bias [Borji et al. 2013; Bylinskii et al. 2015; Tatler 2007], whereby a relatively high number of fixations occur near the center of the image. One explanation for such bias is that it is part of an optimal viewing strategy that is involved in planning successive fixations. By averaging fixation maps across dataset images, we can see a peak near the spatial center of the image emerge across the eye fixations, but not the BubbleView clicks (Figure 17). Because BubbleView naturally slows down the exploration task by making participants consciously decide where to click next, it captures higher-level viewing behaviors not as affected by systematic biases. We recommend using BubbleView with a well-defined task, like describing the content of the visual input, to measure which regions of that visual input are most important or relevant for the task.

A recent paper by Tavakoli et al. [2017] analyzes some of the semantic differences between eye fixations and mouse movements on the OSIE dataset, by taking into account annotated image regions. They find that there tends to be more disagreement between eye fixations and mouse movements in background regions of the image.

Effect of BubbleView parameters: The BubbleView click maps were quite robust across different parameter settings. We did not find significant effects of bubble radius on the resulting BubbleView clicks (Exp. 1,3,5). Across all our experiments (Exp. 1–5), we found that a blur kernel sigma in the range of 30–50 pixels was appropriate for all of our image types, where we manually selected a sigma value for each image dataset to ensure that text was unintelligible when blurred

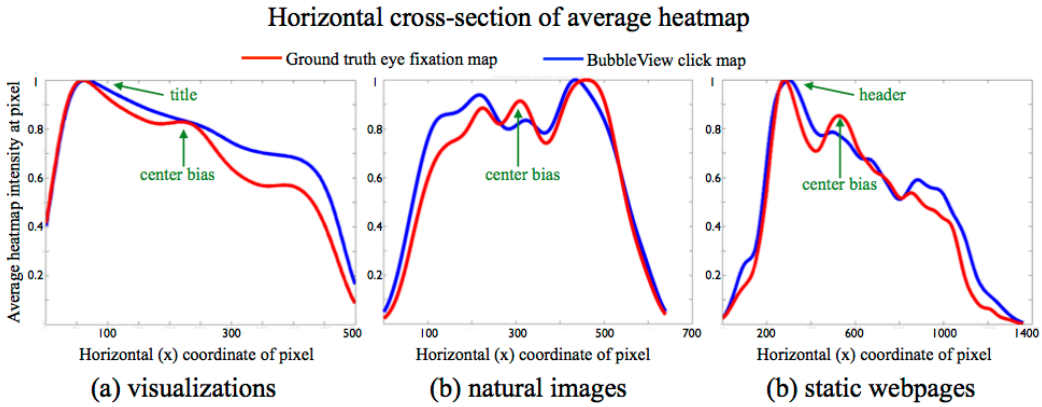


Fig. 17. Taking a horizontal cross-section of the average BubbleView click map and the average fixation map across 51 images on 3 datasets, we see the fixation map has a consistent center bias. This replicates the analysis used by Tatler [2007] to report on human fixation bias in natural images. This bias emerges as a peak near the center of an image, which corresponds to the midway point along the x-axis in each of these plots. The BubbleView click map does not have this bias, which accounts for some of the systematic differences observed between the click and fixation maps. At the same time, the bubble clicks tend to capture the same general characteristics as fixations, for instance of increased attention in the leftmost parts of visualizations and webpages, corresponding to the titles and headers.

and would require explicit clicking on to read. In other words, to mimic peripheral vision, the blur level was chosen to eliminate legible details beyond the focal region. However, a blur sigma with a 70-pixel radius was too high, and seemed to hinder exploration of the image by eliminating too much context, as similarity of BubbleView clicks to eye fixations significantly dropped for this blur level compared to a blur of 30–50 pixels (Exp. 5).

We found that a bubble radius in the range of 30 to 50 pixels seems to consistently work best for different image types and image sizes that comfortably fit within the browser window (ranging from 500×500 to 1000×600 pixels). Here "best" refers to the ability of BubbleView clicks to most closely approximate fixations on images with the smallest number of participants. Smaller bubble sizes lengthened the duration and effort for completing the task, for the same quantitative results. Our chosen bubble sizes typically corresponded to 1–2 degrees of visual angle as measured in the corresponding eye tracking experiments. A bubble size of 1–2 degrees of visual angle mimics the size of the foveal region during natural viewing.

However, bubble radius is also intricately related to task timing and image complexity (Exp. 3). The more content there is on an image to look at, the more time that is required; the smaller the bubble, the more clicks to explore all of the content. A larger bubble radius can compensate for less available time, because each click exposes more of the image. For best results, we recommend a smaller bubble radius but longer task time. In our studies, the longest time for free-viewing tasks was 30 seconds (Exp. 3). For description tasks, participants spent an average of 1.5–3 minutes per image, clicking and describing (Exp. 1,3).

The number of clicks participants made decreased with increasing bubble size, even though the time for the task stayed the same. We observed this trend across all of our experiments. On average, 1–1.5 clicks were made per second in the BubbleView setup, compared to an average of 2–3 fixations per second in eye tracking studies. The BubbleView setup (when implemented with

clicks) slows down visual processing so about half as many interest points are examined every second.

The best prediction performance overall occurs in the setting of a well-defined task, such as describing the visual content of an image. However, tasks must be well-matched to the images used. For instance, asking participants to describe an information visualization is well-defined because each of the visualizations we used had a main message that was being communicated (Exp. 1). On the other hand, we did not use the description task for the graphic designs (Exp. 4), because it was harder to objectively define what should be described.

Number of participants, task, and data quality: For our tasks, we found 10–15 participants sufficient, accounting for over 97% of the performance achievable with 40 participants (Exp. 1,2), where performance is measured by how many of the eye fixations can be approximated by clicks. The more participants, the better the data, as noisy clicks get averaged out. However, when there is a constraint on how many participants can be recruited/afforded, a more involved task (like asking the participant to provide a text description) can result in cleaner data (Exp. 3). Such a task adds an energy barrier to clicking: to minimize effort, participants are more likely to click on image regions informative for completing the task, rather than randomly.

Mouse clicks versus movements: We compared our methodology of collecting discrete mouse clicks to SALICON's moving-window approach [Jiang et al. 2015] in Exp. 5. We found that for any number of participants less than 60, BubbleView is a better approximation to ground truth eye fixations (Figure 16). This is similar to the task, data-quality trade-off discussed above. Clicks add an energy barrier to action: since clicking takes more effort than moving the mouse, participants are more selective about where they click. As a result, BubbleView provides cleaner data with fewer artifacts, such as the byproducts of continuous mouse movements (Figure 15). Furthermore, the moving-window methodology requires post-processing to differentiate mouse positions corresponding to points of interest from transitions. Collecting clicks directly eliminates such post-processing steps.

On the other hand, a byproduct of the higher effort of clicking on an image area rather than moving a mouse over it, is that fewer image areas will be explored by clicking. If the focus of the study is to select the most important regions in an image, then clicks should suffice. In Table 6 we summarize the tradeoffs between the two methodologies. We note additionally that we were able to approximate SALICON's multi-resolution adaptive blur with a single, fixed blur (Exp. 2,5) to achieve similar performances at much lower computational cost.

Table 6. Comparison of BubbleView and SALICON [Jiang et al. 2015]. SALICON consists of capturing continuous mouse movements on an image with adaptive multi-resolution blur. The blur is continuously recomputed for every mouse location at 100 Hz. Continuous mouse tracks are discretized into points of interest using experimenter-specified thresholds. In BubbleView, discrete mouse clicks are collected on an image with a fixed blur. This is easier to implement and has fewer computational limitations. No additional post-processing is required. The collected BubbleView data is less noisy and converges faster, although clicking takes more time.

Property	BubbleView	SALICON
Speed of convergence to eye fixations	faster	slower
Number of participants required	fewer	more
Time per task	higher	lower
Post-processing	less	more
Computational cost	less	more

BubbleView for image importance: The density of clicks in different image regions roughly corresponds to the importance of those regions. Specifically, across a collection of graphic designs, BubbleView clicks on different design elements correlated with explicit importance judgements made on the same designs (Exp. 4). BubbleView clicks ranked visualization elements similarly to eye fixations (Exp. 1). Thus, BubbleView can be used not only to derive conclusions about human perception (where people look), but also to make general conclusions about images and designs: how is importance distributed across an image? Which design elements are most important? This knowledge can in turn can be leveraged for design applications [Bylinskii et al. 2017b].

Data quality and filtering: BubbleView participants were quite consistent with each other in where they clicked, leading to a relatively fast convergence of the aggregate BubbleView click maps to ground truth eye fixation maps. For most of our experiments, we found about 10-15 participants provided enough click data to reasonably approximate eye fixations.

After collecting the BubbleView data, we performed a number of filtering steps, including throwing out participants who did not click a minimum number of times and additional clicking outliers. This filtering of participants and bubbles lead to a data reduction of only 2% on average, indicating that initial data quality was pretty high (Supplemental Material).

The description task has the additional benefit of providing another filtering layer: if a participant-provided description is evaluated as poor, we can assume that they did not do the task with sufficient thoroughness, or clicked in regions of the image that were irrelevant for the task. This filtering step can either be performed manually by the experimenter or implemented as a crowdsourcing task (e.g., by having Amazon Mechanical Turk workers rate descriptions by quality).

Cost: The price to obtain a BubbleView click map per image depends on the amount of time a participant spends on each image and the total number of participants recruited. The average hourly rate for Amazon's Mechanical Turk is \$6/hour, so we use \$0.1/min for our tasks. It is common to make MTurk tasks bite-sized (e.g., a few minutes to 10–15 min each) [Kittur et al. 2008]. Using these guidelines, we provide an approximate cost of obtaining a BubbleView click map per image using 10-15 participants. Table 7 contains a breakdown of costs that can be used as guidelines.

Table 7. Total computed costs per image for obtaining the BubbleView clicks of 10–15 participants (both ends of the range included). These costs depend on how long, on average, participants spend on each image, which in turn depends on the task used. In the free-viewing setting, we fixed the time to either 10 or 30 seconds per image. In the description task, time is unconstrained, and participants move on to the next image after submitting their description for the previous image. During piloting, we estimated time per image for clicking and describing to take about 1.5 minutes. In reality, it took on average 3.2 minutes per image. The description task is more expensive but provides higher-quality click data and an additional data source: the descriptions themselves. These descriptions also serve as quality-control: the clicks of participants who generated poor-quality descriptions can be discarded.

Task	Time/image	Images/HIT	Cost/HIT	Participants/HIT	Cost/Image
Free-viewing	10 sec	17	\$0.30	10–15	\$0.18–\$0.26
Free-viewing	30 sec	17	\$0.90	10–15	\$0.53–\$0.79
Description	180 sec	3	\$0.50	10–15	\$3.34–\$5.00

Methodology limitations: Compared to natural viewing or moving a mouse, clicking takes more time and effort, resulting in longer task timings and higher costs. Certain image regions which might not be as relevant to the task might never be clicked on, even though they may have received a quick glance in an eye tracking or moving-window setting. As a result, the image regions selected by clicks will tend to be more selective than the regions selected in these other settings.

As shown in this paper, the advantage of this selectivity is cleaner, more consistent results across participants. This can be used for determining the most important regions in an image (Exp. 4). But this comes at the potential disadvantage of certain image regions being missed, and other regions, like text, receiving disproportionate clicks (Exp. 1,3). How to encourage a more diverse sampling of image regions while maintaining all the other advantages of BubbleView is a question for future investigations.

8 CONCLUSION AND FUTURE WORK

In this paper we presented BubbleView, a mouse-contingent methodology to approximate eye fixations using mouse clicks. We validated BubbleView by conducting a series of experiments on different image stimuli and comparing clicks to eye fixations, importance maps, and mouse movements. We showed that BubbleView can reasonably approximate fixations, be used to collect image importance driven by human perception, and has a number of advantages compared to the moving-window approach, including better performance with fewer participants.

We analyzed BubbleView in the context of 4 image types (information visualizations, natural images, static webpages, and graphics designs), with 2 task types (free-viewing and description), with different task timings, image blur and bubble sizes, and different numbers of study participants. We provided the interested experimenter with some guidelines on how to use BubbleView for different tasks, how to select parameters, and which settings we found to work best under different conditions. Here we provide additional ideas of how BubbleView can be used and built on top of.

Integrating BubbleView into crowdsourcing pipelines: Unlike eye tracking experiments, BubbleView experiments can be feasibly ported online for the efficient and scalable collection of data using crowdsourcing. Large amounts of data call for data filtering and analysis methods that can scale as well. As shown in this paper, BubbleView clicks can be analyzed automatically. In cases where text input is also collected from participants, filtering and analysis may require additional manual effort. However, it is possible to consider crowdsourcing pipelines where the data collected from the BubbleView tasks is piped directly into filtering tasks.

Following the idea of question-answering tasks, BubbleView can be incorporated into multi-player crowdsourcing games (e.g., ESP Game [von Ahn and Dabbish 2004]). For instance, one participant can generate questions, while the other participant answers using BubbleView clicks. In this setting, the first participant queries and supervises the responses of the second participant. In such a way both data collection and data cleaning can be built into the game.

BubbleView data for training computational models: BubbleView can be used to generate large datasets for training computational models. BubbleView click maps on images can be used as importance maps for those images, and computational models can learn from this data to make predictions for new images [Bylinskii et al. 2017b]. While many saliency models have been developed for natural images, and many natural-image saliency datasets exist (Section 2.4), graphic designs and visualizations have not received as much attention. A saliency model based on these type of stimuli could open up many interesting applications such as extracting important information based on salient regions or providing design feedback [Bylinskii et al. 2017b; O'Donovan et al. 2014, 2015; Rosenholtz et al. 2011].

Measuring information content: Clicking on an image region takes more effort than mousing over, and in turn, glancing at it. There is likely a relationship between the information content of an image region and the likelihood with which it is clicked, moused over, and glanced at. Clicking imposes a kind of energy barrier on the image content that will be explored by participants. Given a targeted task such as describing an image, participants are motivated to click in as few regions as necessary to reduce the overall effort and total task time. As a result, they tend to click in the most informative regions. Increasing the bubble size lowers this energy barrier: participants become

less selective of where they're clicking when they can expose more of the image with each click. Changing the image blur also affects which image region will be clicked, based on its information content. More deeply studying the relationship between visual feature size, information content, image blur and bubble size is likely to provide some interesting insights. In the present study, by virtue of the images we selected for our experiments (e.g., to contain legible text) and the narrow range of image sizes we used, results were pretty stable across blur and bubble settings.

Extending BubbleView to other tasks: The interested experimenter may also choose to use BubbleView in settings and with parameters beyond the ones in this paper, which leaves many possibilities for future investigation. For instance, BubbleView can easily be extended to other visual attention tasks including visual search[9]. To implement a version of visual search using BubbleView, participants can be shown a blurred image and asked to find something in the image (e.g., an object in a natural scene, a specific piece of information in a graph, or an element in a graphic design). Task time can be either fixed, contingent on when the participant chooses to continue to the next image, or contingent on the participant's clicks (i.e., moving to the next image after the correct/expected location is clicked, or after a fixed number of clicks).

Another possible use for BubbleView is modifying the description task into a question-answering task. Participants can be asked to answer a specific question about the image by clicking around the blurred image to expose the content underneath. Each answer, correct, incorrect, or subjective, can be analyzed together with the sequence of clicks made (similar to Das et al. [2016]).

While we originally designed BubbleView as a more efficient alternative to collecting eye fixations on images, we have also shown in this paper that it can be used to measure the importance of different image regions. This idea can be pushed even further in the future, using BubbleView to narrow in on image regions most useful for answering specific questions, extracting particular insights, or completing specific visual tasks. We showed that BubbleView generalizes to different types of images, including natural scenes, visualizations, websites, and graphic designs. This can be expanded to new image types, for instance for studying medical images, geographical maps, user interfaces, slides and posters. For future explorations, we provide our tool and code for launching experiments at massvis.mit.edu/bubbleview.

ACKNOWLEDGMENTS

The authors would like to thank Peter O'Donovan for sharing his dataset and for helpful discussions, and Ming Jiang for answering questions about the SALICON data and methodology. The authors would like to acknowledge Aaron Hertzmann, Bryan Russell, Jean-Daniel Fekete, and Lester Loschky for helpful input. Feedback of anonymous reviewers has helped to significantly improve the quality and clarity of the writing.

This work has been made possible through support from Google, Xerox, the NSF Graduate Research Fellowship Program, the Natural Sciences and Engineering Research Council of Canada, and the Kwanjeong Educational Foundation. We also acknowledge the support of the Toyota Research Institute / MIT CSAIL Joint Research Center.

REFERENCES

Amer Al-Rahayfeh and Miad Faezipour. 2013. Eye Tracking and Head Movement Detection: A State-of-Art Survey. *IEEE Journal of Translational Engineering in Health and Medicine* 1 (2013). https://doi.org/10.1109/JTEHM.2013.2289879

Shumeet Baluja and Dean Pomerleau. 1994. Non-intrusive gaze tracking using artificial neural networks. (1994).

[9]Some examples of visual attention tasks with operational definitions and recommended evaluations are included in Bylinskii et al. [2015].

Roman Bednarik and Markku Tukiainen. 2005. Effects of Display Blurring on the Behavior of Novices and Experts During Program Debugging. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1204–1207. https://doi.org/10.1145/1056808.1056877

Roman Bednarik and Markku Tukiainen. 2007. Validating the restricted focus viewer: A study using eye-movement tracking. *Behavior research methods* 39, 2 (2007), 274–282.

Jennifer Romano Bergstrom and Andrew Schall. 2014. *Eye tracking in user experience design.* Elsevier.

Alan F. Blackwell, Anthony R. Jansen, and Kim Marriott. 2000. *Restricted Focus Viewer: A Tool for Tracking Visual Attention.* Springer Berlin Heidelberg, Berlin, Heidelberg, 162–177. https://doi.org/10.1007/3-540-44590-0_17

Ali Borji and Laurent Itti. 2013. State-of-the-Art in Visual Attention Modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1 (Jan. 2013), 185–207. https://doi.org/10.1109/TPAMI.2012.89

Ali Borji and Laurent Itti. 2015. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581* (2015).

Ali Borji, Dicky N. Sihite, and Laurent Itti. 2013. Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study. *IEEE Transactions on Image Processing* 22, 1 (Jan 2013), 55–69. https://doi.org/10.1109/TIP.2012.2210727

Michelle A. Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S. Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2016. Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan 2016), 519–528. https://doi.org/10.1109/TVCG.2015.2467732

Daniel Bruneau, M Angela Sasse, and JD McCarthy. 2002. The eyes never lie: The use of eye tracking data in HCI research. In *Proceedings of the CHI*, Vol. 2. 25.

Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. 2009. What Do You See when You'Re Surfing?: Using Eye Tracking to Predict Salient Regions of Web Pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 21–30. https://doi.org/10.1145/1518701.1518705

Zoya Bylinskii, Michelle A. Borkin, Nam Wook Kim, Hanspeter Pfister, and Aude Oliva. 2017a. Eye Fixation Metrics for Large Scale Evaluation and Comparison of Information Visualizations. In *Eye Tracking and Visualization: Foundations, Techniques, and Applications. ETVIS 2015*, Michael Burch, Lewis Chuang, Brian Fisher, Albrecht Schmidt, and Daniel Weiskopf (Eds.). Springer International Publishing, Cham, 235–255. https://doi.org/10.1007/978-3-319-47024-5_14

Zoya Bylinskii, Ellen M. DeGennaro, Rishi Rajalingham, Harald Ruda, Jinxia Zhang, and John K. Tsotsos. 2015. Towards the quantitative evaluation of visual attention models. *Vision Research* 116, Part B (2015), 258 – 268. https://doi.org/10.1016/j.visres.2015.04.007

Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. 2014. MIT Saliency Benchmark. (2014). http://saliency.mit.edu/

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2016a. What do different evaluation metrics tell us about saliency models? *CoRR* abs/1604.03605 (2016). http://arxiv.org/abs/1604.03605

Zoya Bylinskii, Nam Wook Kim, Peter O'Donovan, Sami Alsheikh, Spandan Madan, Hanspeter Pfister, Fredo Durand, Bryan Russell, and Aaron Hertzmann. 2017b. Learning Visual Importance for Graphic Designs and Data Visualizations. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software & Technology (UIST '17)*. ACM. https://doi.org/10.1145/3126594.3126653

Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. 2016b. Where should saliency models look next?. In *European Conference on Computer Vision*. Springer, 809–824.

Mon Chu Chen, John R. Anderson, and Myeong Ho Sohn. 2001. What Can a Mouse Cursor Tell Us More?: Correlation of Eye/Mouse Movements on Web Browsing. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI EA '01)*. ACM, New York, NY, USA, 281–282. https://doi.org/10.1145/634067.634234

Laura Cowen, Linden Js Ball, and Judy Delin. 2002. An eye movement analysis of web page usability. In *People and Computers XVI*. Springer, 317–335.

Edward Cutrell and Zhiwei Guan. 2007. What Are You Looking for?: An Eye-tracking Study of Information Usage in Web Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 407–416. https://doi.org/10.1145/1240624.1240690

Abhishek Das, Harsh Agrawal, Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? *arXiv preprint arXiv:1606.03556* (2016).

Jia Deng, Jonathan Krause, and Li Fei-Fei. 2013. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*. IEEE Computer Society, Washington, DC, USA, 580–587. https://doi.org/10.1109/CVPR.2013.81

Andrew T Duchowski. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers* 34, 4 (2002), 455–470.

Simone Frintrop, Erich Rome, and Henrik I. Christensen. 2010. Computational Visual Attention Systems and Their Cognitive Foundations: A Survey. *ACM Trans. Appl. Percept.* 7, 1, Article 6 (Jan. 2010), 39 pages. https://doi.org/10.1145/1658349.

1658355

Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 2014. EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*. ACM, New York, NY, USA, 255–258. https://doi.org/10.1145/2578153.2578190

Joseph H Goldberg and Xerxes P Kotval. 1999. Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics* 24, 6 (1999), 631–645.

Joseph H. Goldberg, Mark J. Stimson, Marion Lewenstein, Neil Scott, and Anna M. Wichansky. 2002. Eye Tracking in Web Search Tasks: Design Implications. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications (ETRA '02)*. ACM, New York, NY, USA, 51–58. https://doi.org/10.1145/507072.507082

Frédéric Gosselin and Philippe G. Schyns. 2001. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research* 41, 17 (2001), 2261 – 2271. https://doi.org/10.1016/S0042-6989(01)00097-9

W Graf and H Krueger. 1989. Ergonomic evaluation of user-interfaces by means of eye-movement data. In *Proceedings of the third international conference on human-computer interaction*. Elsevier Science Inc., 659–665.

Elizabeth R. Grant and Michael J. Spivey. 2003. Eye Movements and Problem Solving. *Psychological Science* 14, 5 (2003), 462–466. https://doi.org/10.1111/1467-9280.02454 arXiv:http://dx.doi.org/10.1111/1467-9280.02454

Rebecca Grier, Philip Kortum, and James Miller. 2007. How users view web pages: An exploration of cognitive and perceptual mechanisms. *Human computer interaction research in Web design and evaluation* (2007), 22–41.

Qi Guo and Eugene Agichtein. 2010. Towards Predicting Web Searcher Gaze Position from Mouse Movements. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 3601–3606. https://doi.org/10.1145/1753846.1754025

Mary Hayhoe. 2004. Advances in relating eye movements and cognition. *Infancy* 6, 2 (2004), 267–274.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Jeff Huang, Ryen White, and Georg Buscher. 2012. User See, User Point: Gaze and Cursor Alignment in Web Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1341–1350. https://doi.org/10.1145/2207676.2208591

Jeff Huang, Ryen W. White, and Susan Dumais. 2011. No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1225–1234. https://doi.org/10.1145/1978942.1979125

Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. 2015. TabletGaze: unconstrained appearance-based gaze estimation in mobile tablets. *arXiv preprint arXiv:1508.01244* (2015).

Weidong Huang. 2007. Using eye tracking to investigate graph layout effects. In *APVIS '07*. 97–100. https://doi.org/10.1109/APVIS.2007.329282

Robert JK Jacob and Keith S Karn. 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind* 2, 3 (2003), 4.

Anthony R. Jansen, Alan F. Blackwell, and Kim Marriott. 2003. A tool for tracking visual attention: The Restricted Focus Viewer. *Behavior Research Methods, Instruments, & Computers* 35, 1 (2003), 57–69. https://doi.org/10.3758/BF03195497

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. SALICON: Saliency in Context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1072–1080. https://doi.org/10.1109/CVPR.2015.7298710

Sheree Josephson and Michael E. Holmes. 2002. Visual Attention to Repeated Internet Images: Testing the Scanpath Theory on the World Wide Web. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications (ETRA '02)*. ACM, New York, NY, USA, 43–49. https://doi.org/10.1145/507072.507081

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

Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*. 2106–2113. https://doi.org/10.1109/ICCV.2009.5459462

Marcel Adam Just and Patricia A Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive psychology* 8, 4 (1976), 441–480.

Wolf Kienzle, Felix A. Wichmann, Matthias O. Franz, and Prof. Bernhard Schölkopf. 2007. A Nonparametric Approach to Bottom-Up Visual Saliency. In *Advances in Neural Information Processing Systems 19*, P. B. Schölkopf, J. C. Platt, and T. Hoffman (Eds.). MIT Press, 689–696. http://papers.nips.cc/paper/3122-a-nonparametric-approach-to-bottom-up-visual-saliency.pdf

Nam Wook Kim, Zoya Bylinskii, Michelle A. Borkin, Aude Oliva, Krzysztof Z. Gajos, and Hanspeter Pfister. 2015. A Crowdsourced Alternative to Eye-tracking for Visualization Understanding. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 1349–1354. https://doi.org/10.1145/2702613.2732934

Sung-Hee Kim, Zhihua Dong, Hanjun Xian, Benjavan Upatising, and Ji Soo Yi. 2012. Does an Eye Tracker Tell the Truth about Visualizations?: Findings while Investigating Visualizations for Decision Making. *IEEE TVCG* 18, 12 (2012), 2421–2430. https://doi.org/10.1109/TVCG.2012.215

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

Kathryn Koehler, Fei Guo, Sheng Zhang, and Miguel P. Eckstein. 2014. What do saliency models predict? *Journal of Vision* 14, 3 (2014), 14. https://doi.org/10.1167/14.3.14 arXiv:/data/journals/jov/932817/i1534-7362-14-3-14.pdf

Eileen Kowler. 1989. The role of visual and cognitive processes in the control of eye movement. *Reviews of oculomotor research* 4 (1989), 1–70.

Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2176–2184. https://doi.org/10.1109/CVPR.2016.239

Srinivas S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu. 2015. DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations. *CoRR* abs/1510.02927 (2015). http://arxiv.org/abs/1510.02927

Dmitry Lagun and Eugene Agichtein. 2011. ViewSer: Enabling Large-scale Remote User Studies of Web Search Examination and Interaction. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 365–374. https://doi.org/10.1145/2009916.2009967

Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods* 45, 1 (2013), 251–266. https://doi.org/10.3758/s13428-012-0226-9

Daniel J. Liebling and Sören Preibusch. 2014. Privacy Considerations for a Pervasive Eye Tracking World. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 1169–1177. https://doi.org/10.1145/2638728.2641688

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Christof Lutteroth, Moiz Penkar, and Gerald Weber. 2015. Gaze vs. Mouse: A Fast and Accurate Gaze-Only Click Alternative. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. ACM, New York, NY, USA, 385–394. https://doi.org/10.1145/2807442.2807461

Päivi Majaranta and Andreas Bulling. 2014. *Eye Tracking and Eye-Based Human–Computer Interaction*. Springer London, London, 39–65. https://doi.org/10.1007/978-1-4471-6392-3_3

George W. McConkie and Keith Rayner. 1975. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics* 17, 6 (1975), 578–586. https://doi.org/10.3758/BF03203972

Jakob Nielsen and Kara Pernice. 2009. *Eyetracking Web Usability* (1st ed.). New Riders Publishing, Thousand Oaks, CA, USA.

David Noton and Lawrence Stark. 1971. Scanpaths in Saccadic Eye Movements while Viewing and Recognizing Patterns. *Vision Research* 11, 9 (1971), 929 – IN8. https://doi.org/10.1016/0042-6989(71)90213-6

Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2014. Learning Layouts for Single-Page Graphic Designs. *IEEE Transactions on Visualization and Computer Graphics* 20, 8 (Aug 2014), 1200–1213. https://doi.org/10.1109/TVCG.2014.48

Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2015. DesignScape: Design with Interactive Layout Suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1221–1224. https://doi.org/10.1145/2702123.2702149

Bing Pan, Helene A. Hembrooke, Geri K. Gay, Laura A. Granka, Matthew K. Feusner, and Jill K. Newman. 2004. The Determinants of Web Page Viewing Behavior: An Eye-tracking Study. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications (ETRA '04)*. ACM, New York, NY, USA, 147–154. https://doi.org/10.1145/968363.968391

Junting Pan, Kevin McGuinness, Elisa Sayrol, Noel E. O'Connor, and Xavier Giró i Nieto. 2016. Shallow and Deep Convolutional Networks for Saliency Prediction. *CoRR* abs/1603.00845 (2016). http://arxiv.org/abs/1603.00845

Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI, 3839–3845.

Derrick Parkhurst, Klinton Law, and Ernst Niebur. 2002. Modeling the Role of Salience in the Allocation of Overt Visual Attention. *Vision Research* 42, 1 (2002), 107 – 123. https://doi.org/10.1016/S0042-6989(01)00250-4

Mathias Pohl, Markus Schmitt, and Stephan Diehl. 2009. Comparing the Readability of Graph Layouts Using Eyetracking and Task-oriented Analysis. In *Proceedings of the Fifth Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging (Computational Aesthetics'09)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 49–56. https://doi.org/10.2312/COMPAESTH/COMPAESTH09/049-056

Alex Poole and Linden J Ball. 2006. Eye tracking in HCI and usability research. *Encyclopedia of human computer interaction* 1 (2006), 211–219.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.

Keith Rayner. 2014. The gaze-contingent moving window in reading: Development and review. *Visual Cognition* 22, 3-4 (2014), 242–258. https://doi.org/10.1080/13506285.2013.879084

Keith Rayner, Caren M Rotello, Andrew J Stewart, Jessica Keir, and Susan A Duffy. 2001. Integrating text and pictorial information: eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied* 7, 3 (2001), 219.

Eyal M Reingold, Lester C Loschky, George W McConkie, and David M Stampe. 2003. Gaze-contingent multiresolutional displays: An integrative review. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 45, 2 (2003), 307–328.

Ronald A Rensink. 2011. *The management of visual attention in graphic displays.* Cambridge University Press, Cambridge, England.

Kerry Rodden, Xin Fu, Anne Aula, and Ian Spiro. 2008. Eye-mouse Coordination Patterns on Web Search Results Pages. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems (CHI EA '08).* ACM, New York, NY, USA, 2997–3002. https://doi.org/10.1145/1358628.1358797

Ruth Rosenholtz, Amal Dorai, and Rosalind Freeman. 2011. Do Predictions of Visual Perception Aid Design? *ACM Trans. Appl. Percept.* 8, 2, Article 12 (Feb. 2011), 12:1–12:20 pages. https://doi.org/10.1145/1870076.1870080

Michael Schulte-Mecklenbeck, Ryan O. Murphy, and Florian Hutzler. 2011. Flashlight - Recording Information Acquisition Online. *Comput. Hum. Behav.* 27, 5 (Sept. 2011), 1771–1782. https://doi.org/10.1016/j.chb.2011.03.004

Chengyao Shen, Xun Huang, and Qi Zhao. 2015. Predicting Eye Fixations on Webpage With an Ensemble of Early Features and High-Level Representations from Deep Network. *IEEE Transactions on Multimedia* 17, 11 (Nov 2015), 2084–2093. https://doi.org/10.1109/TMM.2015.2483370

Chengyao Shen and Qi Zhao. 2014. *Webpage Saliency.* Springer International Publishing, Cham, 33–46. https://doi.org/10.1007/978-3-319-10584-0_3

Peter Tarasewich, Marc Pomplun, Stephanie Fillion, and Daniel Broberg. 2005. The Enhanced Restricted Focus Viewer. *International Journal of HumanâĂŞComputer Interaction* 19, 1 (2005), 35–54. https://doi.org/10.1207/s15327590ijhc1901_4

Benjamin W. Tatler. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision* 7, 14 (2007), 4. https://doi.org/10.1167/7.14.4 arXiv:/data/journals/jov/932846/jov-7-14-4.pdf

Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist. 2005. Visual correlates of fixation selection: effects of scale and time. *Vision Research* 45, 5 (2005), 643 – 659. https://doi.org/10.1016/j.visres.2004.09.017

Benjamin W. Tatler, Mary M. Hayhoe, Michael F. Land, and Dana H. Ballard. 2011. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision* 11, 5 (2011), 5. https://doi.org/10.1167/11.5.5 arXiv:/data/journals/jov/933487/jov-11-5-5.pdf

Hamed R Tavakoli, Fawad Ahmed, Ali Borji, and Jorma Laaksonen. 2017. Saliency Revisited: Analysis of Mouse Movements versus Fixations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).

Tobii. 2010. *Tobii Eye Tracking: An introduction to eye tracking and Tobii Eye Trackers.* White paper. Tobii Technology AB.

Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04).* ACM, New York, NY, USA, 319–326. https://doi.org/10.1145/985692.985733

Niklas Wilming, Torsten Betz, Tim C. Kietzmann, and Peter KÃűnig. 2011. Measures and Limits of Models of Fixation Selection. *PLOS ONE* 6, 9 (09 2011), 1–19. https://doi.org/10.1371/journal.pone.0024038

Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. 2014. Predicting human gaze beyond pixels. *Journal of Vision* 14, 1 (2014), 28. https://doi.org/10.1167/14.1.28 arXiv:/data/Journals/JOV/933546/i1534-7362-14-1-28.pdf

Pingmei Xu, Krista A. Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. 2015. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. *CoRR* abs/1504.06755 (2015). http://arxiv.org/abs/1504.06755

Pingmei Xu, Yusuke Sugano, and Andreas Bulling. 2016. Spatio-Temporal Modeling and Prediction of Visual Attention in Graphical User Interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM, New York, NY, USA, 3299–3310. https://doi.org/10.1145/2858036.2858479

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4511–4520.