BV

BubbleView: an alternative to eye-tracking for crowdsourcing image importance

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

We present BubbleView, a methodology to replace eye-tracking with mouse clicks. 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. 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 can be used to successfully approximate eye fixations on different images, and that the regions where people click using BubbleView can also be used to rank image and design elements by importance. BubbleView is designed to measure which information people consciously choose to examine, and works best for defined tasks such as describing the content of an information visualization or measuring image importance. Compared to related methodologies based on a moving-window approach, BubbleView provides more reliable and less noisy data.

Additional Key Words and Phrases: human vision, visual attention, eye tracking, crowdsourcing, image saliency, image importance, visualizations, graphic designs, webpages, natural scenes, computer vision

* Equal contribution.

1. INTRODUCTION

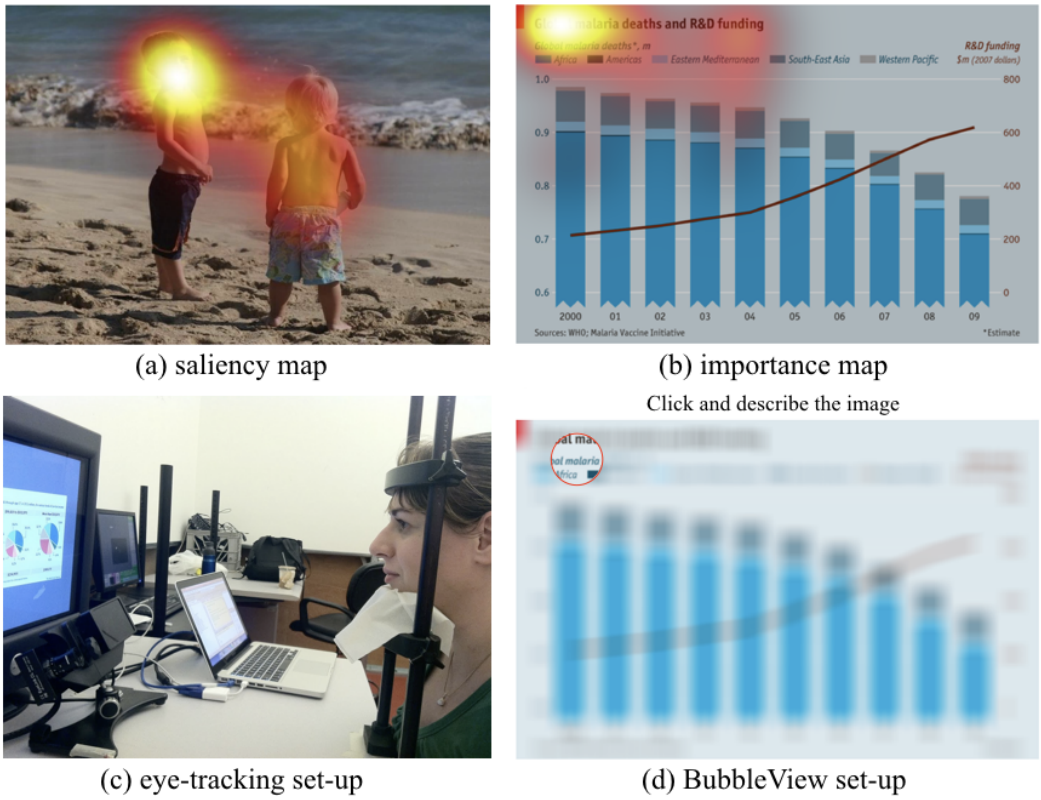
Eye-tracking is a technique to measure an individual's eye movements, visual attention, and focus. This experimental methodology has proven useful for studying the cognitive processes involved in visual information processing, including which visual elements people look at first and spend the most time 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 2010] or for designing gaze-based and attention-aware user interfaces [Majaranta and Bulling 2014; Lutteroth et al. 2015]. However, collecting accurate eye-tracking data often requires expensive eye-tracking equipment and time-intensive calibrations.

In this paper we present a methodology called BubbleView to collect clicks on static images as a proxy for eye fixations. Unlike the resource-heavy collection of eye movements, BubbleView can easily scale up data collection to more participants and images, and be launched remotely to enable online crowdsourcing.

Our interface presents blurred images and allows participants to click around to reveal small circular ("bubble") regions of the image at the original resolution, loosely approximating a blurred periphery and the confined area of focus of the human eye fovea.

Human eye movements are often collected as a proxy of visual attention, to study human perception, or to collect saliency datasets. The most common setting for eye movement experiments is free-viewing, whereby participants are not given a task but are instructed to freely look around the image. The pattern of eye fixations on a natural image can then be interpreted as a **saliency map** for the image (Figure 1). Where people look in a free-viewing setting and the resulting saliency maps tend to capture

bottom-up, pop-out image features, guided in part by the fast, unconscious movements of the eyes.



(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 is a resource-heavy way to collect human eye fixations (c), whereas the BubbleView interface can be launched online and feasibly scale up the collection of crowdsourced data (d).

BubbleView and related cursor-based methodologies naturally 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. As a result, the collected clicks more likely capture image regions that participants consciously choose to attend to. Because of this, we refer to the pattern of BubbleView clicks on an image as the **importance map** for the image. While importance can include saliency, we intend for importance to encapsulate image regions that are not only more attention grabbing initially, 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 for a directed task. 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 later. Like eye movements, BubbleView can provide important insights about human perception and cognition, but at a lower data collection cost than eye tracking.

In this paper, we extend our work from [Kim et al. 2015] to show that beyond just approximating task-based eye fixations on information visualizations, BubbleView also generalizes to approximating eye fixations on other image types under different task constraints. Specifically, we show that:

— BubbleView clicks can successfully approximate eye fixations on information visualizations, natural images, and websites, in a free-viewing condition and with a description task;
— Compared to related methodologies based on a moving-window approach, 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 data, as a proxy for eye fixations, or as a measure of image importance. For this purpose we offer an extensive analysis of BubbleView under different experimental settings and task constraints to examine the effect of different experimental parameters. We present the results of over 20 distinct experiments run with BubbleView on Amazon's Mechanical Turk. Our experiments are carried out on 5 different datasets, spanning information visualizations [Borkin et al. 2016], natural images [Xu et al. 2014; Jiang et al. 2015], static webpages [Shen and Zhao 2014], and graphic designs [O'Donovan et al. 2014]. We varied task type (free-viewing, describing) and task timing, 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 for more feasibly approximating eye movements 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

2.1. Eye Movements and Cognitive Tasks

There is a lot of work 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 [Just and Carpenter 1976; Hayhoe 2004; 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 [Duchowski 2002; Goldberg and Kotval 1999; Graf and Krueger 1989; Poole and Ball 2006; Jacob and Karn 2003], webpage 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. 2017; Pohl et al. 2009; Kim et al. 2012; Huang 2007]. 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. Alternative Techniques for Costly Eye-tracking

In order to work around the limitations of eye-tracking (i.e., expensive hardware equipment, tedious calibration for individual participants, difficulty of running a large-scale experiment), there has been a significant effort to find cheap, nonintrusive, and scalable alternatives to collect human attentional data. Most existing alternatives follow the moving-window approach in which a limited amount of information is visible through a variable size window contingent upon a point of interest [McConkie and Rayner 1975; Rayner 2014].

Inspired by the moving-window model, Jansen et al. developed a computer program called Restricted Focus Viewer (RFV) that takes an image, blurs it, and reveals only a restricted focus region of the image, allowing a user to move the region using a mouse [Jansen et al. 2003; Blackwell et al. 2000]. Commercial software for tracking user attention is also based on the same idea (http://www.attensee.com) [Schulte-Mecklenbeck et al. 2011]. The mouse-contingent methodology has been employed to investigate cognitive behaviors of users in diverse contexts such diagrammatic reasoning and program debugging, and to study the usability of web sites [Jansen et al. 2003; Bednarik and Tukiainen 2005; Tarasewich et al. 2005].

Recent studies have made further improvements. Jiang et al. [2015] implemented the real-time generation of multi-resolution blurred images to attempt to simulate the fall-off in acuity of peripheral vision. Similarly, Gomez et al. [2016] used image blending for a smooth transition between focus and blurred areas. On the other hand, Lagun and Agichtein [2011] directly preprocessed web search results to show one result and blur the rest of 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 involving hundreds to thousands of participants, proving the scalability of their methods.

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 [Rodden et al. 2008; Guo and Agichtein 2010; Huang et al. 2011; Huang et al. 2012]. 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.

Another line of work was devoted to using webcams and analyzing video frames to predict gaze locations [Lebreton et al. 2015; Xu et al. 2015; Papoutsaki et al. 2016]. This approach could continuously track the gaze locations even if a mouse cursor is idle, but requires participants to have quality webcams. More recently, Krafka et al. [2016] collected a massive amount of eye tracking data on mobile devices and developed a gaze prediction algorithm based on convolutional neural networks. The webcam-based eye-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]. 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. Lastly, one study investigated the validity of users' recall and prediction on eye movements as an alternative to eye tracking [Johansen and Hansen 2006], which is a purely qualitative approach and thus not appropriate for a large-scale computational model.

Our proposed approach, BubbleView, is similar to the moving-window approach as it reveals a focus area at normal resolution while the rest of the image is blurred. However, instead of continuously moving the focus area based on cursor position, our methodology uses discrete clicks to relocate and reveal the focus area. This enables a more explicit record of points of interest without the need for post-processing noisy mouse movement data. We build on our previous work [Kim et al. 2015] published before the two most recent and similar techniques [Jiang et al. 2015; Gomez et al. 2016]. In [Kim et al. 2015], we showed how BubbleView can be used to approximate eye fixations on information visualizations using the MASSVIS dataset [Borkin et al. 2016]. 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 visualization images). In this paper we provide more rigorous experimental results by comparing our methodology to eye movements on a diverse set of image stimuli and with different parameter settings. We aim to understand under which settings mouse clicks best approximate eye movements. Many of our insights are likely to hold for related methodologies as well.

2.3. Saliency Models and Eye-tracking Datasets

In addition to alternative techniques for eye tracking which require human participants, there has been a significant amount of effort building computational saliency models to predict human 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 human fixations [Bylinskii et al. 2016a; Bylinskii et al. 2014; Judd et al. 2012].

Models have typically been trained directly on human fixation data collected from eye tracking experiments [Judd et al. 2009; Kienzle et al. 2006]. Recently, a large dataset of mouse movements on natural images has been released for training computational models of saliency. This dataset, dubbed SALICON, was collected using the 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]. Collecting a dataset of this size using eye trackers in the lab would have been very costly at the least, if not infeasible.

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 web pages, 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]. Zhao and Shen developed a webpage saliency model based on the FiWI dataset [Shen and Zhao 2014], and recently improved the model by further taking into account high-level semantic features (e.g., positional bias and object detectors) [Shen et al. 2015]. O'Donovan developed a semi-automatic model of importance prediction for graphic designs by training on a crowdsourced dataset of importance annotations [O'Donovan et al. 2014]. The GDI dataset was collected by asking workers to annotate regions of importance on images using binary masks. An-

other recent study presented a computational model for predicting visual attention in user interfaces by using user interactions [Xu et al. 2016].

We draw on several existing datasets from this domain in saliency modeling 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 alternative to costly eye-tracking. It is inspired by the work of [Deng et al. 2013] in the computer vision community, in which the bubble model of [Gosselin and Schyns 2001] was used to discover image regions people pay most attention to when performing fine-grained object recognition. We first provide some background on human visual 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 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 the units used to measure 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, and size and resolution of the image on the screen[1]. The error of eye trackers is also measured in degrees of visual angle, and is commonly less than 1 degree.

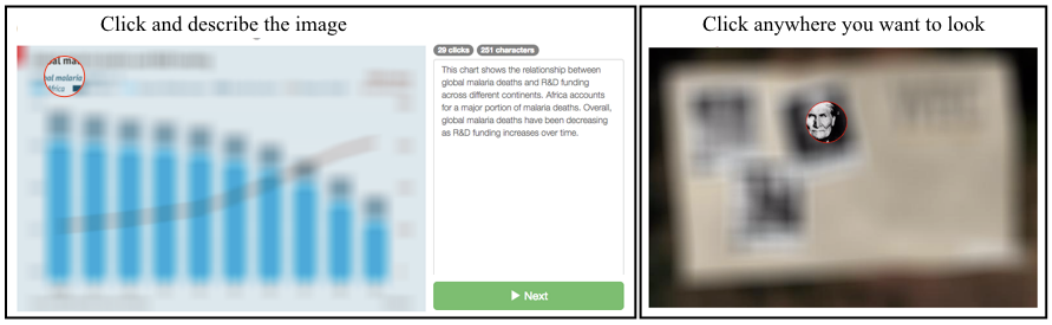
The pauses in eye movements are called **fixations**, and the transitions between successive fixations are called **saccades**. In this paper, we focus on fixations, since they give us the points of interest that the eye has stopped on to bring them into focus. The temporal sequence of fixations carries a lot of additional information but is beyond the scope of the present work.

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. 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.

In comparison to the moving-window approach which records continuous mouse movements, our approach records discrete mouse clicks since 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.

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



(a) description task (b) free-viewing task

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.

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

— **Image blur:** 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 constant across all images in the sequence. In our studies, we manually selected a blur value per image dataset to distort all the image text beyond recognition. We wanted the level of detail to be enough for reading only in regions of focus, i.e., in the clicked, bubble regions.
— **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 data.
— **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 (Sec. 6).
— **Time:** the viewing time per image. This should be task dependent. 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.

The experimenter may also choose the number of images displayed in sequence per experiment. 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.

3.3. Viewing collected data with BubbleView

We also developed a monitoring interface to inspect the results of the experiments (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 the experimenter to explore the temporal sequence and evolution of the bubble clicks and description text. 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.

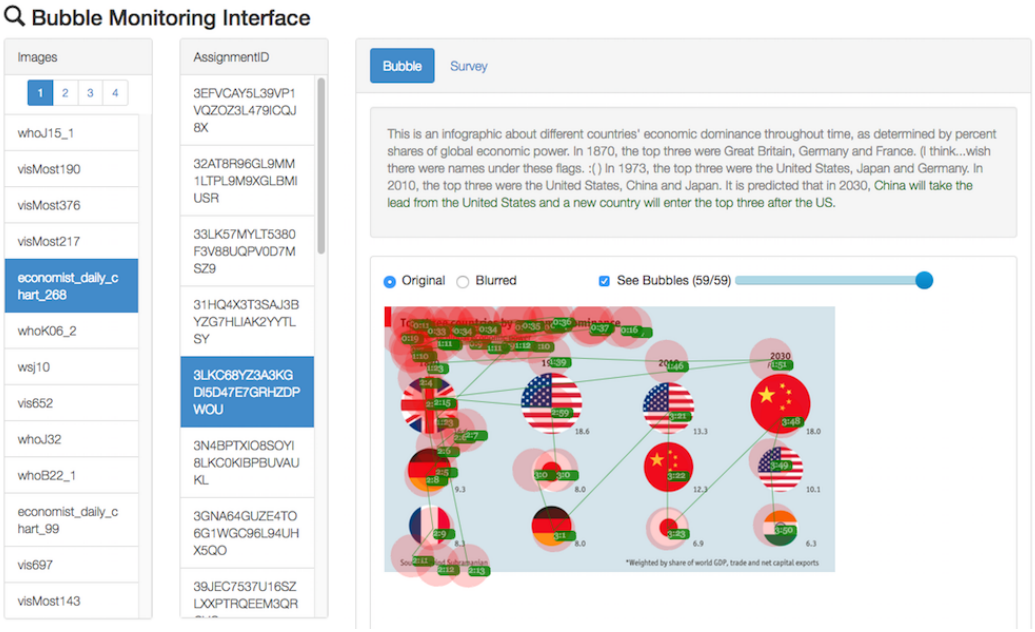


Fig. 3. Monitoring interface for manually inspecting the results of experiments. An experimenter can use a slider to explore the temporal sequence and evolution of bubble clicks and description text, for each image and participant.

4. EXPERIMENTS

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 still be well approximated in other 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.

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 a graphic design dataset [O'Donovan et al. 2014], as well as mouse movements on a natural image dataset [Jiang et al. 2015]. As summarized in Table I, we deployed over 23 separate experiments on Amazon's Mechanical Turk: 4 experiments for information visualizations

(4 different bubble radius sizes), 1 experiment for natural scenes, 6 experiments for static webpages (3 bubble radius sizes x 2 viewing times), 1 experiment for graphic designs, and 9 experiments for another dataset of natural scenes (3 image blur amounts x 3 bubble radius sizes). For some of these we also ran additional variations that we will describe in the experimental sections that follow.

Table I. Overview of BubbleView experiment settings including different image stimuli and parameters varied per experiment.

Exp.	Image Stimuli	Experimental Parameters			
	Dataset & Image Type	Task Type	Image blur (pixel sigma)	Bubble radius (pixel)	Time (sec)
1	MASSVIS [Borkin et al. 2016] Information visualizations	describe	40	16, 24, 32, 40	unlim.
2	OSIE [Xu et al. 2014] Natural scenes	free-view	30	30	10
3	FiWI [Shen and Zhao 2014] Static webpages	free-view, describe	50	30, 50, 70	10, 30 unlim.
4	GDI [O'Donovan et al. 2014] Graphic designs	free-view	30	50	10
5	SALICON [Jiang et al. 2015] Natural scenes	free-view	30, 50, 70	30, 50, 70	10

4.1. Tasks and Procedures Overview

We collected BubbleView data for 51 images selected out of each of the 5 datasets used, with additional images from the visualization 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 restricted 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-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, it took 8.5 min. 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 and to fill in a short voluntary demographic survey at the end. 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[2]. All participants were paid regardless of whether they completed the task successfully or not. See the Supplemental Material for a description of the data filtering procedure.

4.2. Analysis Overview

Across all the experiments comparing BubbleView clicks to eye fixations we use the same set of analyses, which we describe here. We compare 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 observer consistency 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 the fovea and the measurement error of the eye tracker (a common evaluation choice [Le Meur and Baccino 2013; Bylinskii et al. 2016a]). 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. When we intend to refer to the fixation or click maps more generally in this paper, we call them **importance maps**.

Measuring the similarity between clicks and fixations
We use two different similarity metrics to measure how well BubbleView clicks approximate eye fixations: Pearson's Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS). We first convert BubbleView clicks on an image into a click map. To obtain the **CC score** for the 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. Unlike CC, 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 predict each other's eye fixations very well. In these cases, BubbleView clicks will

[2]Our original estimates were that the description task would take 1.5 min per image, whereas in reality it took an average of 3.18 min per image.

not be able to accurately predict the fixations either. For a fair evaluation, we normalize the BubbleView prediction performance by the computed 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; Wilming et al. 2011; Le Meur and Baccino 2013]. We measure IOC using the NSS score.

We first compute the NSS score of the aggregate 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 (also computed using NSS). The resulting **normalized NSS** can be interpreted as: the percent of the eye fixations accounted for, or predicted by, the BubbleView click maps.

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, and 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. From the fitting, we also obtain 95% confidence bounds which we include in our result tables. 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.

4.3. Experiments comparing BubbleView clicks to eye fixations

Experiment 1: comparison to eye fixations on information visualizations

Building off our initial experiments in [Kim et al. 2015], 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, 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 in order to determine under which bubble setting we best approximate eye fixations, and how many participants are enough for a good quality approximation.

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 eye-tracking study [Borkin et al. 2016]. In this 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.

We selected a total of 202 target images from the 393 target images, spanning infographic, news media, and government publication categories (see sample images in Fig. 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

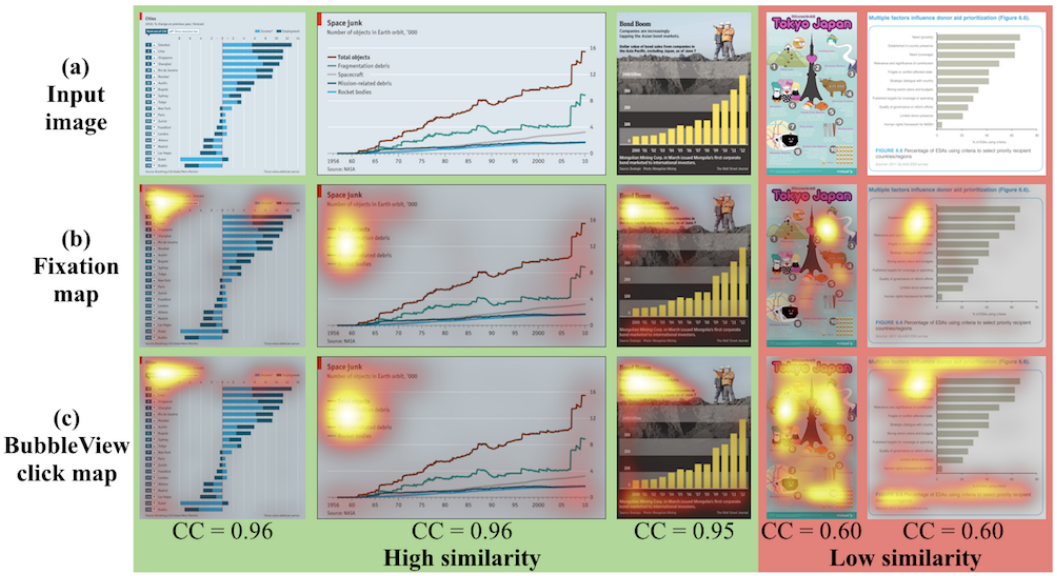


Fig. 4. Example images from the MASSVIS dataset. Dataset images (a), with corresponding ground-truth fixation maps (b) and BubbleView click maps (c). We show cases where BubbleView maps have high similarity, and cases with low similarity, to fixation maps.

images were blurred with a sigma of 40 pixels, which we found distorted the text in these images beyond legibility [Borkin et al. 2016; Kim et al. 2015].

Method

We ran a series of experiments progressively to 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 target 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, so as to account for individual differences in the amount of time for writing image descriptions.

For Exp.1.1., we requested enough HITs so that each image would be seen by an average of 40 participants. After receiving complaints about the task difficulty at a bubble radius of 16 pixels, we terminated data collection for this bubble size after gathering data from an average of 20 participants per image. From Exp.1.1. we found that 15 participants are sufficient for good data quality, and proceeded to collect an average of 15 participants worth of BubbleView click data for each image in Exp.1.2. and Exp.1.3.

Results on bubble size

We computed the similarity between the BubbleView click maps and ground truth fixation maps across all the information visualizations for all settings of bubble radius (Table II). To make scores comparable, we fixed the number of participants used to construct the BubbleView click maps at 10. This was the minimum number of participants available across all the information visualizations.

Table II. We evaluated BubbleView at approximating ground-truth eye fixations on the MASSVIS dataset by varying the bubble radius. We ran 3 different sets of experiments on different subsets of the MASSVIS dataset. We compared BubbleView click maps to ground truth fixation maps and averaged over all images tested, to obtain an aggregate cross-correlation (CC) score. The CC score has an upper bound of 1. We evaluated how well BubbleView click maps predict discrete fixation locations, averaged over all the images tested, to obtain a normalized scanpath saliency (NSS) score. The NSS upper bound depends on the ground-truth data, so we include the inter-observer consistency (IOC) score of the eye-tracking participants (measured in NSS). Normalizing the NSS score of the BubbleView maps by the IOC of the ground-truth 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, the minimum number available across all the images. We used all available participants per experiment to extrapolate performance to the limit of infinite participants. We report an upper bound in gray and provide a 95% confidence interval.

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	1.29 1.34 [1.30, 1.38]	91% 94%
	24	0.86	1.28 1.35 [1.32, 1.37]	90% 95%
	32	0.86	1.28 1.30 [1.30, 1.30]	90% 92%
Exp. 1.2: 51 visualizations description task (ground-truth IOC: 1.33)	24	0.82	1.20 1.28 [1.26, 1.29]	90% 96%
	32	0.84	1.20 1.27 [1.25, 1.28]	90% 95%
	40	0.83	1.19 1.25 [1.20, 1.31]	89% 94%
Exp. 1.3: 202 visualizations description task (ground-truth IOC: 1.35)	32	0.84	1.21 1.27 [1.26, 1.28]	90% 94%

The similarities between the BubbleView click maps and the fixation maps were close across all bubble radius sizes (CC = 0.82 - 0.86). The normalized NSS score was also very similar across all bubble radius sizes, with BubbleView clicks accounting for an average of 89-91% of eye fixations with 10 participants, and reaching 92-96% in the limit. In other words, bubble size did not seem to significantly affect the resulting BubbleView click maps (Figure 5). After receiving a number of participant complains about task difficulty at a bubble radius of 16 pixels, we discontinued the use of this bubble radius in future experiments. In summary, with bubble sizes in the range 24-40 pixels, BubbleView clicks serve as good approximations to eye fixations on visualizations with a description task.

Results on number of participants

In Exp.1.1, we collected about 40 participants of BubbleView clicks per image to investigate how BubbleView maps change with the number of participants (Figure 6). We found that after about 10-15 participants, the similarity of BubbleView click maps to ground truth fixation maps did not significantly improve. At 10 participants, the NSS score for a bubble radius of 24 pixels is already 1.28, and goes up to 1.30 with 39 participants; for a bubble radius of 32 pixels, the NSS score at 10 participants is 1.28 and goes up to 1.29 with 40 participants. Extrapolated performances by fitting power curves improve slightly upon these values, going up to an expected NSS score of 1.35 for a bubble radius of 24, and 1.30 for a bubble size of 32. For a bubble radius of 16, although we stopped the experiment early (at 15 participants per image), the NSS score did not improve from 10 to 15 participants. As a result of these analyses, we collected an average of 15 participants worth of BubbleView clicks per image for

(a) input image (b) bubble radius = 16 (c) bubble radius = 24 (d) bubble radius = 32

Fig. 5. We found few differences in the resulting click maps from different settings of the BubbleView experiment. 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 makes, and the longer the task takes to complete. Overall, the same regions of interest tended to be clicked on despite differences in bubble sizes.

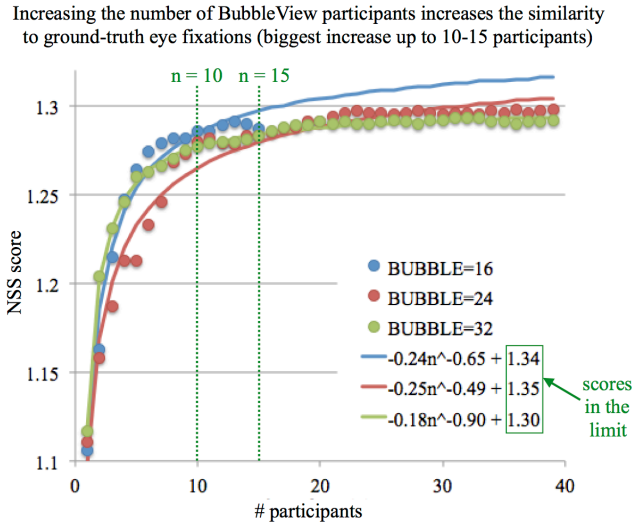


Fig. 6. The NSS score of BubbleView click maps computed with different numbers of participants, when used to predict discrete fixation locations. Each point represents the score obtained at a given number of participants, averaged over all 51 images used in Exp.1.1. We include data points from 3 different bubble radius sizes. Past about 10-15 participants, scores begin to saturate. 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.

all future BubbleView experiments.

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 labels (e.g., title, axis, legend, etc.) available in the MASSVIS dataset [Borkin et al. 2016]. For each of the 202 visualizations from Exp.1.3, we overlapped the element labels 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 [Jiang et al. 2015; Bylinskii et al. 2016b]. We averaged the element scores across all 202 visualizations to obtain an aggregate importance score for each type of element. We repeated this using the BubbleView click maps of the visualizations to get another set of importance scores for the same elements. As can be seen in Figure 7, the ranking of

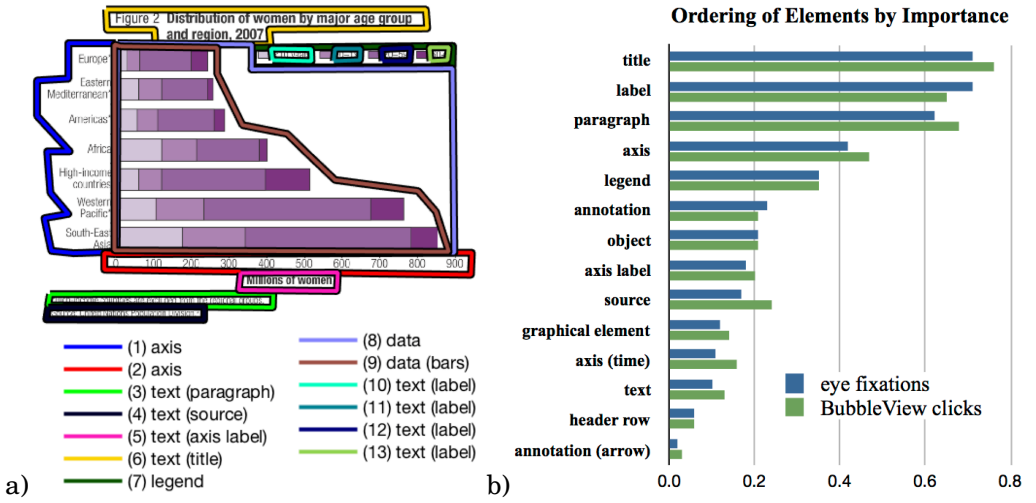


Fig. 7. (a) An example of a labeled visualization from the MASSVIS dataset. (b) By overlapping fixation maps and BubbleView click maps with such element annotations, we obtain an importance score for each element in each visualization. By averaging across 202 visualizations, we obtain an aggregate importance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

elements by importance scores according to BubbleView clicks is almost identical to the ranking according to eye fixations (Spearman correlation = 0.96).

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 of natural images is a more common setting for human perception studies and 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 to generate a good approximation to eye fixations: a bubble size around 30 pixels and 15 participants worth of clicks.

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. In this eye-tracking setup, images were presented at a resolution of 800x600 pixels and 1 degree of viewing angle corresponded to 24 pixels. For our study, we randomly sampled 51 OSIE images (see sample images in Fig. 8), downsized them to 640x480 pixels, and blurred them with a sigma of 30 pixels.

Method

We ran a BubbleView experiment asking participants to free-view a series of images and to click anywhere they want to look within a viewing time of 10 sec per image. We used a bubble radius of 30 pixels, corresponding to about 1.5 degrees of visual angle in the original eye-tracking study. Although the original viewing time for the eye-tracking experiment was 3 sec per image, we increased this time for the BubbleView experiment to account for the longer time and effort of clicking a mouse. We piloted different viewing times and determined 10 sec to be appropriate. We collected

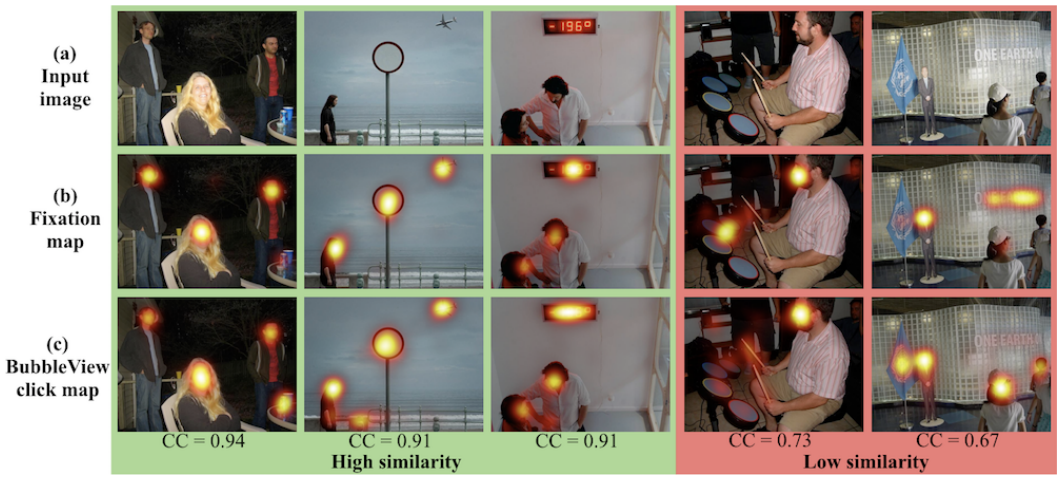


Fig. 8. Example images from the OSIE dataset. Dataset images (a), with corresponding ground-truth fixation maps (b) and BubbleView click maps (c). We show cases where BubbleView maps have high similarity, and cases with low similarity, to fixation maps.

an average of 15 participants worth of BubbleView click data for each image.

Results

The similarity between the BubbleView click maps and ground truth fixation maps with free-viewing on natural images was smaller (CC = 0.81, Table III) than for the visualizations with the description task for the same number of participants (Table II). Even though eye-tracking participants are quite consistent with each other on the OSIE dataset (IOC = 3.35), BubbleView participants are not as good predictors of eye-tracking participants in this case. In the limit, BubbleView clicks can only explain up to 82% of fixations, indicating that a systematic difference remains in this domain between BubbleView clicks and eye fixations. Although BubbleView clicks are not as good an approximation to free-viewing eye fixations on natural images as to task-based eye fixations on visualizations, 10 BubbleView participants can already account for 78% of eye fixations on natural images.

In Exp. 5 (Table V), we also show that a related methodology based on the moving-window approach [Jiang et al. 2015] is no better at approximating ground-truth eye fixations on this dataset. In fact, to achieve the same performance as BubbleView it actually requires more participants. This means that BubbleView can serve as a cheap and reasonable alternative to eye tracking. In cases where a lot of data needs to be gathered, and running a large number of eye-tracking experiments on many participants is infeasible, BubbleView can be used for studying human perception and collecting large-scale saliency datasets (as in [Jiang et al. 2015]).

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 [Shen and Zhao 2014; Shen et al. 2015; Buscher et al. 2009; Nielsen and Pernice 2010; Rodden et al. 2008; Chen et al. 2001]. For this reason, we wanted to test the generalizability of the BubbleView methodology to webpages, to see if it could also feasibly replace costly eye-tracking for such studies. Because the static webpage images we used were quite a bit denser in visual and information content than the information visualizations and natural im-

Table III. We evaluated BubbleView at approximating ground-truth eye fixations on the OSIE dataset. BubbleView maps were computed with 10 participants, for comparison to Exp.1. 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. In gray, we provide the extrapolated performance by taking the number of participants to the limit. We provide an upper bound and a 95% confidence interval. Normalized NSS is calculated by normalizing the NSS score by the inter-observer consistency (IOC) of the eye-tracking participants.

Exp. 2: natural scenes (ground-truth IOC: 3.35)	Time (sec)	Bubble Radius (pixel)	CC	NSS	Normalized NSS
Free-viewing	10	30	0.81	2.61 2.75 [2.70,2.80]	78% 82%

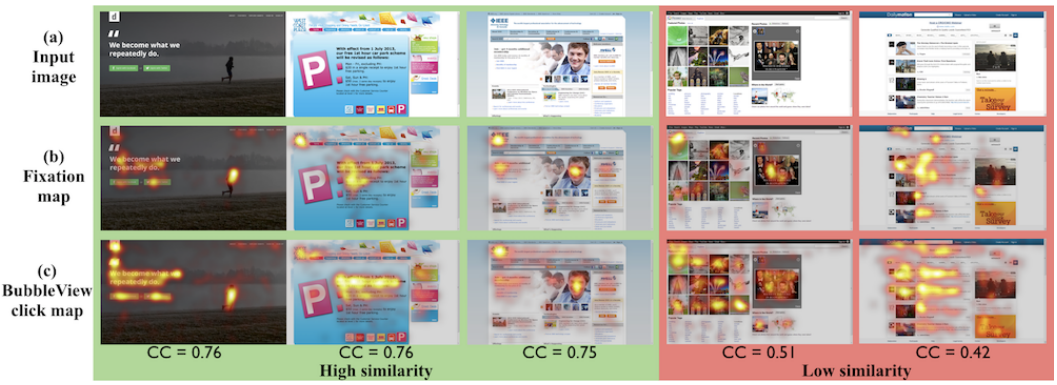


Fig. 9. Example images from the FiWI dataset. Dataset images (a), with corresponding ground-truth fixation maps (b) and BubbleView click maps (c). We show cases where BubbleView maps have high similarity, and cases with low similarity, to fixation maps.

ages 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 under a free-viewing task. Similar to Exp. 1, we also tried a description task with unlimited task time.

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. In this eye-tracking setup, 1 degree of the 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 (see sample images in Fig. 9). We downsized the images from 1360x768 pixels to 1000x565 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 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 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
Under all experimental settings, the aggregate similarity between BubbleView click maps and ground truth fixation maps on webpages was lowest of all image stimuli tested so far (Exp.1-3), as measured by CC scores (Table IV). However, the inter-observer consistency of eye-tracking participants was also low on webpages (IOC = 1.85). For instance, IOC was highest on the all text webpages (NSS = 1.97), followed by the pictorial (NSS = 1.77) and mixed (NSS = 1.80) webpages. BubbleView click maps were most predictive on the text webpages than the pictorial and mixed webpages.

After accounting for IOC, the normalized NSS scores show that BubbleView clicks can account for over 75% of eye fixations across all webpages. In the limit of infinite observers and with 30 seconds of viewing, BubbleView clicks can account for up to 96% of eye fixations, indicating that with enough participants, methodology differences can be significantly reduced. Because the fixation and click data is noisier on webpages (more differences between participants), more participants are required to obtain robust importance maps.

With a viewing time of 10 seconds, BubbleView click maps were most similar to ground truth fixation maps when a bubble radius of 50 pixels was used (Table IV). With a viewing time of 30 seconds, a bubble radius of 30 pixels was best, with scores decreasing with increasing bubble sizes. Therefore, with less time for viewing, a larger bubble radius is more effective; while with a longer viewing time, a smaller bubble radius can be afforded. Overall, BubbleView click maps generated with longer task durations of 30 seconds or longer (in the case of the description task) better approximated eye fixations than when only 10 second viewing times were used. From this we conclude that information-dense images like websites require longer viewing times.

For a small number 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 (see Supplemental Material). The difference between the tasks is larger for smaller number of participants, and decreases with each extra participant. In the limit, the 30-second free-viewing task generates BubbleView maps that are closer to the ground truth eye-tracking data, which was also collected in a free-viewing setting (Table IV). In other words, the click data tends to converge faster when a targeted task like description is used; but in the limit, systematic differences emerge when the task used for the BubbleView experiment does not match the task used for the original eye-tracking experiment. There is a trade-off between number of participants and data quality: with fewer participants, a description task can generate cleaner data; but when more participants are available, the free-viewing condition leads to BubbleView click maps that are more similar to free-viewing eye fixations.

4.4. Experiments comparing BubbleView to related methodologies

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 a dataset that comes with explicit importance annotations, where participants were asked 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.

Table IV. We evaluated BubbleView at approximating ground-truth eye fixations on the FiWI dataset. BubbleView maps were computed with 12 participants, to make all the scores below comparable. 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. In gray, we provide the extrapolated performance by taking the number of participants to the limit. We provide an upper bound and a 95% confidence interval. 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 1.67 [1.40,1.93]	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 1.77 [1.58,1.97]	76% 96%
Free-viewing	30	70	0.57	1.32	71%
Description	unlim.	30	0.63	1.46 1.55 [1.50,1.60]	79% 84%

Stimuli

The Graphic Design Importance (GDI) dataset contains a set of 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 particular 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 in a design with binary masks, and these masks were averaged over all the 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 the elements in a design.

We sampled 51 images from the GDI dataset at the original resolution of 600x400 pixels (see sample images in Fig. 10). We blurred the images with a sigma of 30 pixels, manually chosen to distort text beyond 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 under each task.

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 importance maps generated by our methodology. By construction, the

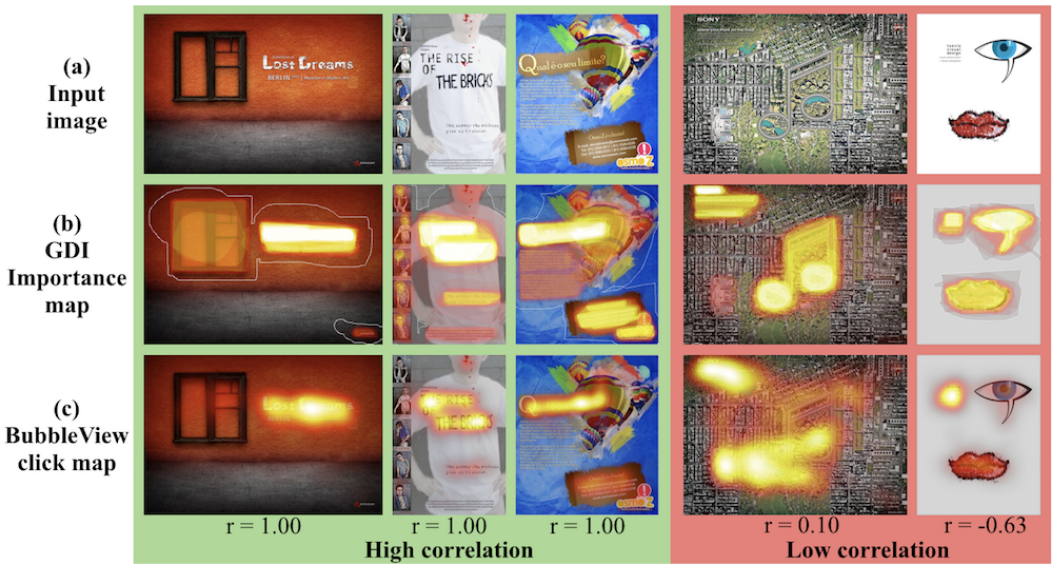


Fig. 10. Example images from the GDI dataset. Images from the dataset (a), along with the provided explicit importance annotations (b). We show cases where BubbleView maps have high correlation, and cases with low correlation, to the importance annotations, in terms of how design elements are ordered by importance (c).

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.

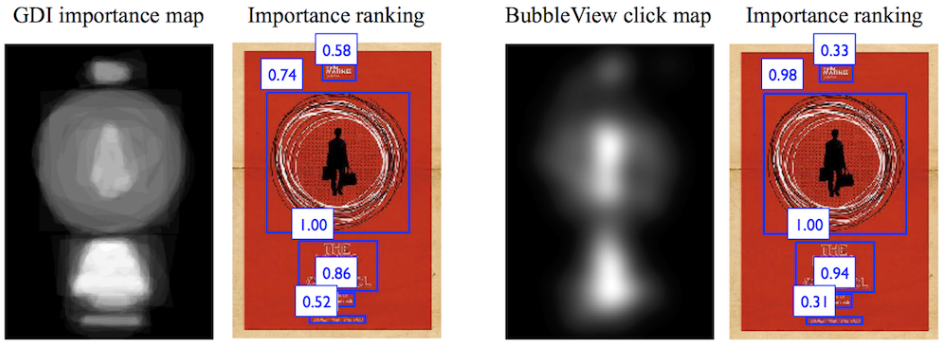


Fig. 11. Importance maps were overlapped with element bounding boxes (outlined with blue boxes) and the maximum map value per box was taken to be the importance score for that element. Note that importance maps were first normalized to have values between 0 and 1, so the importance scores for all the graphic design elements also lie 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.

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. Following the approach of [Jiang et al. 2015; Bylinskii et al. 2016b], we took the maximum value of each map within an element's bounding box as the importance score of that element (Figure 11). We correlated the importance scores assigned by both methodologies to the elements in each design.

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 Fig. 10 where certain visual elements were not clicked on in the BubbleView methodology (the note because it blended into the background in the blurred version; the eye because it was already too clear in the blurred version).

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[3]. 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 kernel recomputed for each cursor position. We investigated whether BubbleView produces similar importance maps on images as SALICON, when mouse clicks and mouse movements are averaged over participants using the respective methodologies. Because in the SALICON methodology blur is multi-resolution and adaptive, we experimented with different blur sigmas and bubble sizes in the BubbleView methodology to find a fixed setting of parameters that best approximates the SALICON viewing conditions. We also measured which methodology is better able to approximate eye fixations collected in a controlled lab setting, since both methodologies are presented as alternatives to eye tracking.

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 around. In the SALICON methodology, the image blur is adaptively regenerated for each location of the mouse cursor. We randomly sampled 51 images at the original image size of 640x480 pixels from the SALICON dataset (see sample images in Fig. 12).

Method
In the first set of experiments, **Exp. 5.1**, we used the free-viewing task with a 3 by 3 factorial design (blur sigma: 30, 50, and 70 pixels; bubble radius: 30, 50, and 70 pixels; as in Figure 13). Participants were presented with a series of images for 10 seconds each and asked to click anywhere they want to look.

[3]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, respectively.

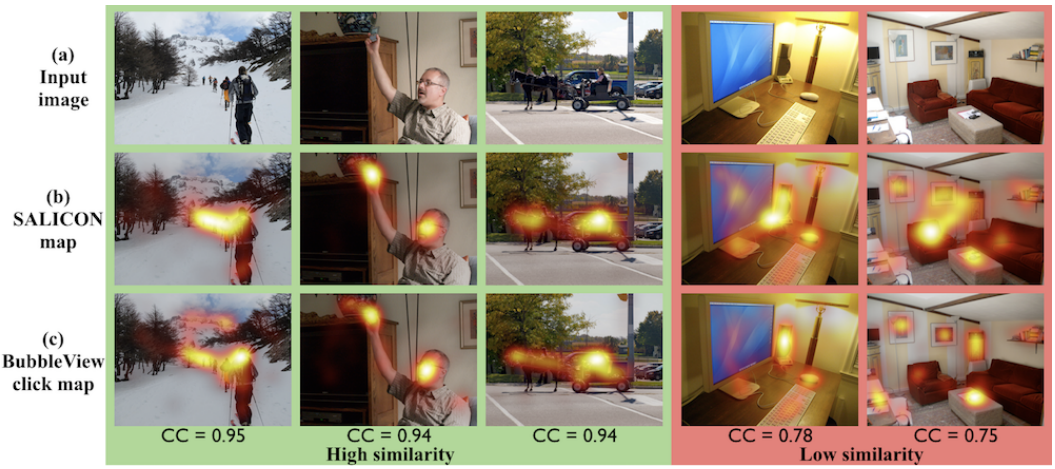


Fig. 12. Example images from the SALICON dataset. Example dataset images (a), and ground truth mouse movements collected by SALICON (b). We show cases where BubbleView maps have high similarity, and cases with low similarity, to SALICON maps (c).

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

Finally, we also re-ran the BubbleView methodology on the OSIE images from Exp. 2, similarly replacing mouse clicks with the moving-window approach. This was done to facilitate a direct comparison between the BubbleView and SALICON methodologies at approximating eye fixations gathered in a controlled lab setting.

Results on using BubbleView to approximate SALICON

Using all 9 combinations of blur and bubble radius from Exp. 5.1, 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 (see Supplemental Material). The normalized NSS score in the limit ranged from 92% to 100%, implying that BubbleView clicks can account for the mouse movements collected in the SALICON study quite well, and in some cases, the BubbleView data is more robust than the inter-observer consistency in the SALICON dataset. This is due to the comparatively lower consistency scores (IOC) between SALICON participants than BubbleView participants. 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, and thus lower consistency, into the data (Figure 14). This suggests that a different threshold might be more effective at converting continuous SALICON mouse movements into discrete points of interest. An advantage of the BubbleView methodology is that no such post-processing is necessary: BubbleView 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. With the moving-window BubbleView setting,

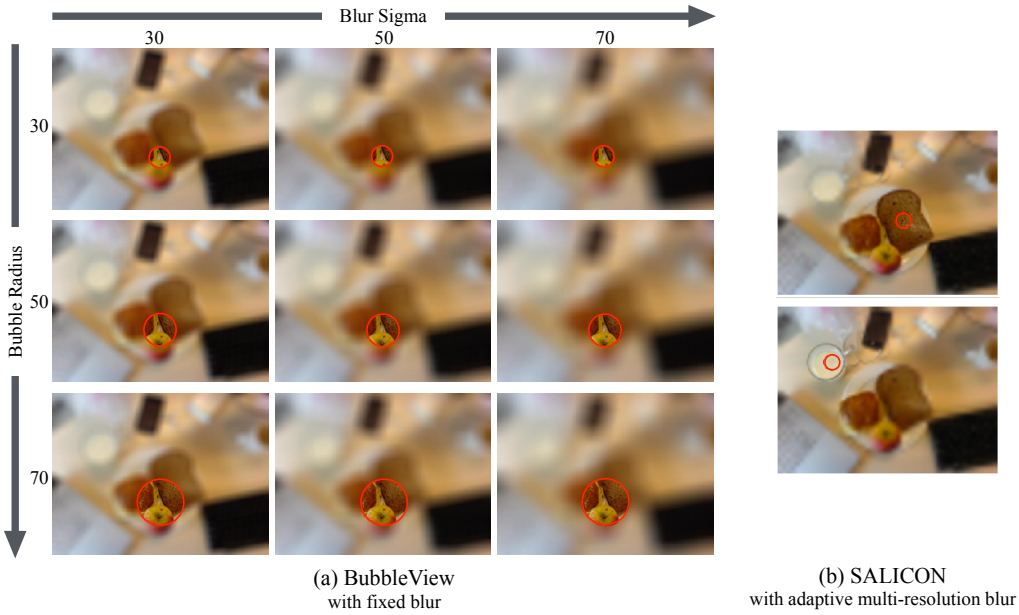


Fig. 13. (a) The 9 different parameter settings used in the BubbleView experiments, on images from the SALICON dataset. We wanted to find a fixed setting of bubble size and blur to mimic (b) the adaptive multi-resolution blur used in the SALICON methodology. The rightmost figure is from [Jiang et al. 2015].

the aggregate similarity scores measured using CC slightly improved (to 0.87-0.88), but the normalized NSS scores were not significantly different. In other words, BubbleView can approximate SALICON with or without mouse movements. The detailed results can be found in the Supplemental Material. 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].

Results on using both methodologies to approximate eye fixations

We compared BubbleView click data from Exp. 2 to SALICON mouse movement data on the same set of 51 OSIE images. The BubbleView click maps (with 12 participants and a bubble radius of 30 pixels) achieve CC = 0.81 at predicting ground-truth fixation maps, compared to SALICON mouse movement maps (with 12 participants) which achieve CC = 0.77. In fact, it takes over 45 SALICON participants to achieve the same similarity to fixation maps as 12 BubbleView participants. Replacing BubbleView clicks with mouse movements actually worsens performance for small numbers of participants, but in the limit of infinite participants, mouse movements approximate eye fixations better than mouse clicks by 2 percentage points. In other words, for smaller numbers of participants, continuous mouse movements introduce more noise than mouse clicks; however, in the limit, they offer slightly better approximations to natural eye movements.

Data was also available for 12 in-lab participants who used SALICON in a controlled lab setting [Jiang et al. 2015]. The resulting aggregate mouse movements, which generate the in-lab SALICON maps, achieve a CC = 0.81 when compared to fixation maps, the same score as our BubbleView maps (Table III). In the limit, lab-based SALICON data appears to account for more of the ground-truth fixations than either the Bubble-

(a) points of interest from SALICON mouse movements

(b) points of interest from BubbleView mouse clicks

Fig. 14. When participants can move the mouse anywhere on the image without having to click, the collected data will contain 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).

View or SALICON data from online participants (Figure 15), but at the expense of a controlled lab setting which we aim to avoid.

Table V. We compared BubbleView and SALICON at approximating ground-truth eye fixations on the OSIE dataset. For fair comparison, we only used 12 participants per image per study. However, we use all available participants per study to extrapolate performance to infinite participants and obtain the upper bounds listed in gray (along with 95% confidence intervals).

Exp. 2: natural scenes (ground-truth IOC: 3.35)	CC	NSS	Normalized NSS
BubbleView (clicks)	0.81	2.61 2.75 [2.70,2.80]	78% 82%
BubbleView (movements)	0.81	2.51 2.81 [2.58,3.05]	75% 84%
SALICON	0.77	2.39 2.84 [2.80,2.88]	71% 85%
In-lab SALICON	0.81	2.61 3.79 [2.88,4.69]	78% 100%

5. DISCUSSION

Similarity of BubbleView clicks to eye fixations: We extended our initial findings from [Kim et al. 2015] to show 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

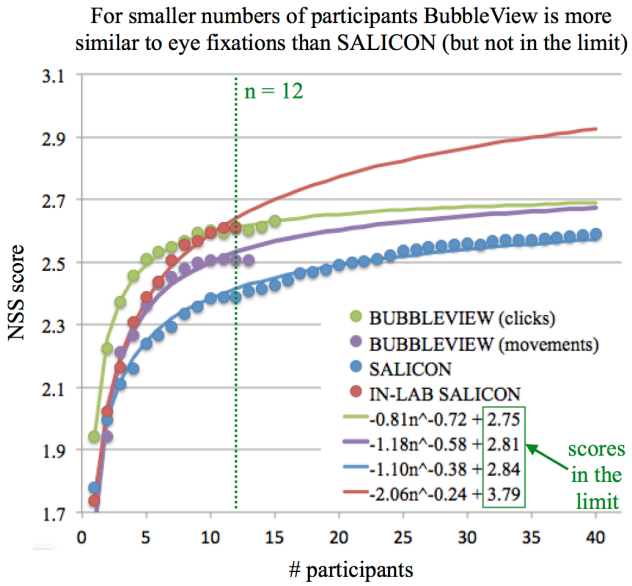


Fig. 15. The NSS score obtained by comparing mouse clicks and mouse movements to ground truth eye fixations on natural images in the OSIE dataset. Here we're comparing mouse clicks gathered using BubbleView on MTurk (green), mouse movements gathered using SALICON on MTurk (blue), and mouse movements gathered using SALICON in a controlled lab setting (red). Each point represents the score obtained at a given number of participants, averaged over all 51 images used. Since only 12 participants worth of data was collected for the in-lab SALICON experiment, the results in Table V are reported for 12 participants per experiment, for fair comparison. Here we use all available data for each experiment to extrapolate performance by fitting a power function of the form $an^b + c$ to each set of points.

best approximation to eye fixations on information visualizations with a description task, accounting for up to 91% of eye fixations with only 10 participants (Table II).

The gap between BubbleView and eye-tracking could be reduced further by considering a larger number of BubbleView participants. We found that taking the number of participants to the limit, BubbleView clicks could account for up to 96% of eye fixations on information visualizations and webpages (Tables II,IV), and up to 82% of eye fixations on natural images (Table III). With more complex visual content like static webpages, which tend to contain many more elements, a larger number of participants is required to account for the larger variation in observer behavior. These results show that there is a trade-off between the number of BubbleView participants and similarity to eye-tracking data. Compared to eye-tracking however, BubbleView does not require expensive eye-trackers, time consuming calibration procedures, and can be launched remotely to enable online crowdsourcing. Despite requiring more participants, BubbleView can still serve as a cheap alternative to eye-tracking.

Remaining differences between BubbleView clicks and eye fixations: Part of the remaining gap between BubbleView and eye-tracking 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 [Tatler 2007; Borji et al. 2013; Bylinskii et al. 2015], 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 fixa-

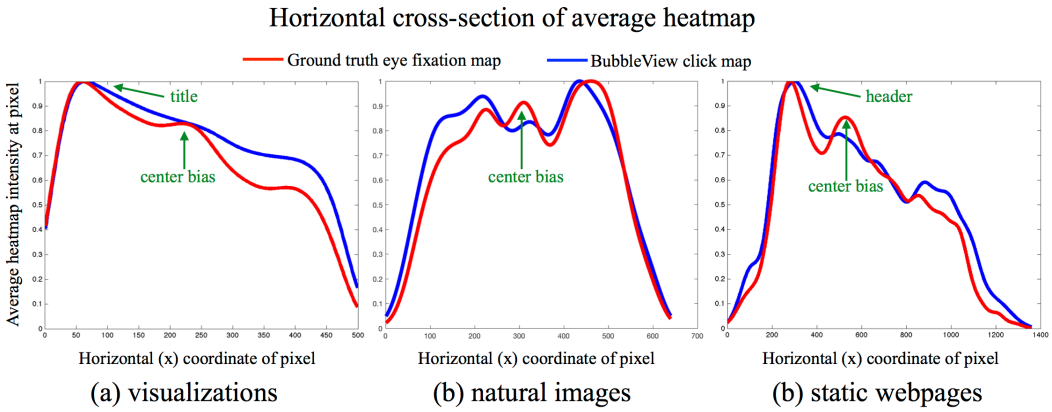


Fig. 16. Taking a horizontal cross-section of the average BubbleView click map and the average fixation map across 51 images on 3 datasets, we see the the fixation map has a consistent center bias. This replicates the analysis used by [Tatler 2007] to report on human fixation bias in natural images. 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.

tions, but not the BubbleView clicks (Figure 16). 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.

Effect of BubbleView parameters: The BubbleView click maps were quite robust across different parameter settings. We did not find large effects of image blur and 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 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.

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 500x500 to 1000x600 pixels). Here "best" refers to the ability of BubbleView clicks to most closely approximate fixations on images with the smallest number of participants. Much smaller bubble sizes were found to be too tedious by online participants. 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 makes sense because it 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 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 tested on 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 much harder to objectively define what should be described.

Finally, there is a trade-off between the number of participants and data quality: with fewer participants, the description task results in cleaner data (better approximation of BubbleView clicks to free-viewing eye fixations); but when more participants can be used, the BubbleView task should match the eye-tracking task (Exp. 3).

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 although in the limit, SALICON comes closer to ground truth eye fixations, for any number of participants less than about 45, BubbleView is a better approximation (Fig. 15). BubbleView provides cleaner data with fewer artifacts, such as the byproducts of continuous mouse movements (Fig. 14). 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, due to the considerably higher effort of clicking on an image area rather than passing a mouse over it, fewer image areas will be explored in the same amount of time. Over many participants, mouse movements rather than clicks will come closer to ground-truth eye fixations. If the focus of the study is to select the most important regions in an image, then BubbleView should suffice. Where gathering large amounts of participants per image is infeasible, BubbleView with clicks is also more effective. In Table VI we summarize the tradeoffs between the two methodologies. We note additionally that we were able to approximate the multi-resolution adaptive blur that SALICON relies on with a single, fixed blur (Exp. 2,5) to achieve similar performances at much lower computational cost.

Table VI. 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 then 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. Aside from the filtering of noise, no additional post-processing is required. The collected BubbleView data is less noisy and converges faster, but the SALICON data can converge closer to human eye fixations in the limit of many observers.

Property	BubbleView	SALICON
Speed of convergence to eye fixations	faster	slower
Number of participants required	less	more
Similarity to eye fixations in the limit	worse	better
Time per task (cost per image)	higher	lower
Post-processing	less	more
Computational cost	less	more

BubbleView for image importance: We showed that 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 human eye movements (Exp. 1.3). 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: which design elements are most important, how are design elements distributed across an image, how is importance distributed? This knowledge can in turn can be leveraged for design applications.

Data quality and filtering: Overall, we found that 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, enough to account for over 75% of eye fixations (Exp. 1-3).

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 that we recommend has the additional benefit of providing another filter 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 be performed manually by the experimenter, or if not feasible, 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. Because the average hourly rate for Amazon's Mechanical Turk is $6/hour, we also 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, by collecting clicks from 10-15 participants, which we have found leads to relatively consistent results. We provide a breakdown of costs in Table VII that experimenters can use as guidelines.

Table VII. Total computed costs per image, for obtaining the BubbleView clicks of 10-15 participants on the image (both ends of the range are computed). 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 fix the task time to be 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 current image. During piloting, we estimated time per image for clicking and describing to take about 1.5 minutes. In reality, it took on average 3.18 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	Assignments/HIT	Cost/Image
Free-viewing	10 sec	17	0.3	10-15	$0.18 - $0.26
Free-viewing	30 sec	17	0.9	10-15	$0.53 - $0.79
Description	180 sec	3	0.5	10-15	$3.34 - $5.00

Methodology limitations: Compared to a moving-window approach, clicking takes more time and effort, resulting in longer overall task timings and higher costs. The effort of clicking serves as a kind of energy barrier for participants, so certain image regions whose content might not be as relevant to the task might never be clicked on, even though they may have otherwise 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. How to encourage a more diverse sampling of image regions while maintaining all the other advantages of BubbleView is a question for future investigations.

6. CONCLUSION AND FUTURE WORK

In this paper we presented BubbleView, an alternative methodology to replace eye-tracking using mouse clicks. We evaluated BubbleView by conducting a series of experiments on different image stimuli and comparing clicks to eye fixations, importance maps, and mouse movements. We have shown that BubbleView can reasonably approximate fixations, be used to collect image importance driven by human perception, and has some advantages compared to the moving-window approach, including computational simplicity and higher consistency between study 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 timing, 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 for another participant to answer by using BubbleView to click on an image. In this setting, the first participant queries and supervises the responses of the second participant. As a result, 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. In particular, the 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. While numerous saliency models on natural images have been developed, and numerous natural-image saliency datasets exist, graphic designs and visualizations are much less explored. 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 to increase the saliency of a specific region (e.g., [O'Donovan et al. 2014; 2015]).

Measuring information content: Clicking on an image region takes more effort than mousing over, and in turn glancing at, the image region. There likely exists 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 seems to lower this energy barrier: participants become less selective of where they're clicking when they can expose more of the image's information content with each click. Changing the image blur can also change whether or not an 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-sized 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[4]. 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 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 an alternative to eye-tracking for the more efficient gathering of visual attention patterns 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 their data and methodology. The authors would also like to acknowledge Aaron Hertzmann, Bryan Russell and Jean-Daniel Fekete for helpful input.

This work has been made possible through support from Google, Xerox, the Department of Defense through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program, the NSF Graduate Research Fellowship Program, the Natural Sciences and Engineering Research Council of Canada Postgraduate Doctoral Scholarship (NSERC PGS-D), and the Kwanjeong Educational Foundation.

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

REFERENCES

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. DOI:http://dx.doi.org/10.1145/1056808.1056877

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. DOI:http://dx.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. DOI:http://dx.doi.org/10.1109/TPAMI.2012.89

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. DOI:http://dx.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. DOI:http://dx.doi.org/10.1109/TVCG.2015.2467732

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. DOI:http://dx.doi.org/10.1145/1518701.1518705

Zoya Bylinskii, Michelle A. Borkin, Nam Wook Kim, Hanspeter Pfister, and Aude Oliva. 2017. 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. DOI:http://dx.doi.org/10.1007/978-3-319-47024-5_14

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

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, 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. DOI:http://dx.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.

Abhishek Das, Harsh Agrawal, C 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. DOI:http://dx.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. DOI:http://dx.doi.org/10.1145/1658349.1658355

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.

Steven Gomez, Radu Jianu, Ryan Cabeen, Hua Guo, and David H. Laidlaw. 2016. Fauxvea: Crowdsourcing Gaze Location Estimates for Visualization Analysis Tasks. *IEEE Transactions on Visualization and Computer Graphics* PP, 99 (2016), 1–1. DOI:http://dx.doi.org/10.1109/TVCG.2016.2532331

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. DOI:http://dx.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 guiding attention guides thought. *Psychological Science* 14, 5 (2003), 462–466.

Rebecca Grier, Philip Kortum, and JT 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. DOI:http://dx.doi.org/10.1145/1753846.1754025

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

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. DOI:http://dx.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. DOI:http://dx.doi.org/10.1145/1978942.1979125

Weidong Huang. 2007. Using eye tracking to investigate graph layout effects. In *APVIS '07*. 97–100. DOI:http://dx.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. DOI:http://dx.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. DOI:http://dx.doi.org/10.1109/CVPR.2015.7298710

Sune Alstrup Johansen and John Paulin Hansen. 2006. Do We Need Eye Trackers to Tell Where People Look?. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. ACM, New York, NY, USA, 923–928. DOI:http://dx.doi.org/10.1145/1125451.1125630

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*. ACM, 43–49.

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, Frdo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*. 2106–2113. DOI:http://dx.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, Bernhard Schölkopf, and Matthias O. Franz. 2006. A Nonparametric Approach to Bottom-up Visual Saliency. In *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06)*. MIT Press, Cambridge, MA, USA, 689–696. http://dl.acm.org/citation.cfm?id=2976456.2976543

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. DOI:http://dx.doi.org/10.1145/2702613.2732934

Sung-Hee Kim, Zhihua Dong, Hanjun Xian, B. 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. DOI:http://dx.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. DOI:http://dx.doi.org/10.1145/1357054.1357127

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. DOI:http://dx.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. DOI:http://dx.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. DOI:http://dx.doi.org/10.3758/s13428-012-0226-9

Pierre Lebreton, Toni Mki, Evangelos Skodras, Isabelle Hupont, and Matthias Hirth. 2015. Bridging the gap between eye tracking and crowdsourcing. *Proc. SPIE* 9394 (2015), 93940W–93940W–14. DOI:http://dx.doi.org/10.1117/12.2076745

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. DOI:http://dx.doi.org/10.1145/2638728.2641688

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. 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. DOI:http://dx.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. DOI:http://dx.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. DOI:http://dx.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. DOI:http://dx.doi.org/10.3758/BF03203972

Jakob Nielsen and Kara Pernice. 2010. *Eyetracking web usability*. New Riders.

David Noton and Lawrence Stark. 1971. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision research* 11, 9 (1971), 929.

P. 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. DOI:http://dx.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*. ACM, 1221–1224.

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*. ACM, 147–154.

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.

Mathias Pohl, Markus Schmitt, and Stephan Diehl. 2009. Comparing the readability of graph layouts using eyetracking and task-oriented analysis. In *Computational Aesthetics in Graphics, Visualization and Imaging*. 49–56.

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. DOI:http://dx.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.

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. DOI:http://dx.doi.org/10.1145/1358628.1358797

Michael Schulte-Mecklenbeck, Ryan O. Murphy, and Florian Hutzler. 2011. Flashlight – Recording information acquisition online. *Computers in Human Behavior* 27, 5 (2011), 1771 – 1782. DOI:http://dx.doi.org/10.1016/j.chb.2011.03.004 2009 Fifth International Conference on Intelligent ComputingICIC 20092009 Fifth International Conference on Intelligent Computing.

C. Shen, X. Huang, and Q. 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. DOI:http://dx.doi.org/10.1109/TMM.2015.2483370

Chengyao Shen and Qi Zhao. 2014. *Webpage Saliency*. Springer International Publishing, Cham, 33–46. DOI:http://dx.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. *IJHCI* 19, 1 (2005), 35–54.

B. Tatler. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *JoV* 7, 14 (2007).

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. DOI:http://dx.doi.org/10.1145/985692.985733

Niklas Wilming, Torsten Betz, Tim C. Kietzmann, and Peter Knig. 2011. Measures and Limits of Models of Fixation Selection. *PLOS ONE* 6, 9 (09 2011), 1–19. DOI:http://dx.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. DOI:http://dx.doi.org/10.1167/14.1.28

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. DOI:http://dx.doi.org/10.1145/2858036.2858479