What Do You See When You're Surfing? Using Eye Tracking to Predict Salient Regions of Web Pages

Georg Buscher[1], **Edward Cutrell**[2], **Meredith Ringel Morris**[2]

[1] DFKI, Knowledge Management Dept.
Kaiserslautern, Germany
georg.buscher@dfki.de

[2] Microsoft Research
One Microsoft Way, Redmond, WA 98052
{cutrell, merrie}@microsoft.com

ABSTRACT

An understanding of how people allocate their visual attention when viewing Web pages is very important for Web authors, interface designers, advertisers and others. Such knowledge opens the door to a variety of innovations, ranging from improved Web page design to the creation of compact, yet recognizable, visual representations of long pages. We present an eye-tracking study in which 20 users viewed 361 Web pages while engaged in information foraging and page recognition tasks. From this data, we describe general location-based characteristics of visual attention for Web pages dependent on different tasks and demographics, and generate a model for predicting the visual attention that individual page elements may receive. Finally, we introduce the concept of fixation impact, a new method for mapping gaze data to visual scenes that is motivated by findings in vision research.

Author Keywords

Eye tracking, Web design.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

The World Wide Web has become an information platform of tremendous importance. Estimates from 2006 suggest that the average U.S.-based user viewed 120 Web pages per day [6]. The ability to model *what parts* of those Web pages receive the most visual attention could offer several benefits to both end-users and Web page authors.

From an end-user perspective, there is great value in being able to model both what users have already viewed in the past and what parts of a page they are likely to view in the future. With regards to revisitation, research shows that 50% [13, 26] to 80% [5] of all Web surfing behavior involves pages that users have visited before. Re-finding previously viewed websites can be quite challenging, and current browser history mechanisms are inadequate for this common task [26]. If one knew what regions of a Web page people use to recognize previously seen pages, one could create compact visual representations of Web pages that emphasize or contain only those areas most relevant for page recognition, thus assisting re-finding [27]. Similarly, a model of what parts of a page are most likely to be looked at for a given page could be used to construct compact "previews" of pages that could assist a user when triaging many as-yet-unexplored Web pages, e.g., during an investigational search task.

Figure 1: Heat map visualization of viewing behavior across 20 participants during a page recognition task.

Web page authors could also benefit from a model of visual attention to improve page layout and design, e.g., arranging page elements in such a way that users' attention is focused on the aspects that the author considers most important. And, of course, the value to advertisers of knowing how users' direct their attention is quite obvious!

We conducted an eye-tracking study in which 20 participants viewed 361 distinct Web pages while conducting both information foraging (information search, analysis, and synthesis) and page recognition tasks. First, we describe the general characteristics of gaze behavior across regions of different pages in the context of each task (see Figure 1 for an example page for the recognition task). We then introduce a methodology for associating eye gaze

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

data with elements from a Web page's Document Object Model (DOM), and use this method to construct a model of visual attention based on HTML- and rendering-based features of Web pages.

BACKGROUND & RELATED WORK

Viewing behavior of arbitrary images is dependent on the characteristics of the image itself, one's expectations about where to find information, and one's current task or information need. Web pages can be thought of as specific kinds of images. In this section, we first provide some background about viewing behavior on arbitrary images (e.g., photos), and then consider research specifically dealing with eye movements and Web pages.

Image Viewing Behavior

Eye movements are generally composed of fixations and saccades. A fixation is a time span of at least 80-100 ms, during which the eye steadily gazes at one point. A saccade is a rapid, ballistic eye movement from one fixation to the next. Visual information is generally only perceived during fixations, not during saccades (see [25] for more detail).

Much research has been conducted to shed light on where and when we fixate on images. There are three main factors that influence the placement of fixations: 1) salience of areas in the image; 2) memory and expectations about where to find information; and 3) task and information need at hand.

Salience of areas is typically computed based on low-level image characteristics, particularly contrast, color, intensity, edge density, and edge orientation (see [11]). The first fixation is typically placed on the most salient spot, and the following fixations are placed such that the information gain is maximized [11, 15, 28]. While salience may direct the first fixation, memory and expectations (e.g., about what is shown in the image and where to find specific objects) also play important roles in subsequent fixations [11, 22, 23].

Web Pages and the DOM

A Web page rendered and displayed by a browser can be thought of as a single complex image. Of course, this image is very different from scenic photos. Web pages often serve specific functions, i.e., to convey information to the user as in product descriptions, news, etc. Over the years, certain design patterns have been established and design guidelines have been created for Web page layout (see [2]). As a consequence, many Web pages contain certain elements at specific locations (e.g., logo, navigation bar, banners). Thus, users have general expectations about where to find certain pieces of information on a Web page [3].

Of course, Web pages are not just images displayed by the browser, but they are described in HTML. The browser reads the HTML code and transforms it internally into a formal representation, the document object model (DOM). All the different layout elements (e.g., links, images, text paragraphs, etc.) are different elements in the DOM.

Some research has focused on predicting abstract Web page characteristics based on the general page design. For example, Ivory & Hearst [16] developed measures based on the DOM that could be used to automatically predict the overall Web page quality. Fogg et al. [9] examined what features determine the credibility and trustworthiness of a Web page. Both of these studies are based on analyzing important features of a Web page based on descriptions of how they are perceived by the users. Dontcheva et al. [8] use direct interaction from users coupled with analysis of the DOM to bootstrap semi-automatic extraction of relevant content from Web pages to create browsing summaries. These studies suggest that there might be some elements on a Web page that most users notice (e.g., a logo) because they are indicators of some global page characteristics (such as quality and trustworthiness). Such characteristics may also be reflected in the eye movements on the page.

Web Page Viewing Behavior

Studies from search engine optimization [14] and Web usability consultancies [20] describe broad patterns of visual attention for a variety of Web pages. However, these studies generally do not provide a detailed quantitative analysis of how the gaze was distributed across pages, instead providing general descriptions such as the "golden triangle" (for Web search) or the "F-shaped pattern" for general Web content.

Pan et al. [24] studied the dependency of scan paths (i.e., repetitive sequences of fixations and saccades) on gender of the subjects, type of Web site (i.e., business, news, search, shopping), viewing order of Web pages, and task. They found significant differences in all variables except for task, which seemed to have no influence on viewing behavior. However, they used a weak task differentiation: remembering what was on a Web page vs. no specific task at all. In later work focused on Web search, they used stronger tasks (informational vs. navigational search) but again saw little difference in gaze behavior [19].

A study by Josephson & Holmes [17] suggested that people might follow habitually preferred scan paths over a Web page. They also suggested that other influential factors like specific features of the Web page or memory might play an important role. However, their study focused on only three Web pages, making the findings difficult to generalize. Furthermore, they only focused on scan paths, not on other measures like fixation duration or time to first fixation.

Goldberg et al. [10] studied eye movements on Web portals during search tasks. They found that header bars are typically not viewed before focusing the main part of the page. As a consequence, they suggest placing navigation bars on the left side of a page.

Beymer et al. [4] focused on a very specific feature on Web pages: images that are placed next to text content and how they influence eye movements during a reading task. They found significant influences, e.g., on fixation placement and duration. Those influences were dependent on how the

image contents related to the text contents (i.e., they showed either ads or text-related images).

A study by Cutrell & Guan [7] focused on viewing behavior on search result lists as created by commercial search engines, and investigated the effect of task and the information density of search results on gaze and search behavior. In particular, they focused on the composition of three elements for each result list entry (i.e., title, text content, URL). They found that the three elements influence each other and this has a differential effect on task performance. For example, the longer the text snippet was, the shorter was the time spent viewing the title or the URL; this improved performance on some tasks and hurt others.

For our study, we wanted to: 1) quantify how very different tasks or other variables may influence the pattern of gaze across a variety of pages; and 2) see if we could use the collection of gaze and Web page data to create a predictive model of which page elements are likely to be looked at when users explore a page.

METHODS

The fundamental premise of our study is that since gaze data can be seen as a proxy for attention, understanding how people look at Web pages may reveal something about the salience, recognizability, and importance of different areas that we can then use in a predictive manner. To generalize these findings to underlying elements or abstractions of all pages, we need two things:

- a mechanism that maps gaze data to elements of the HTML-induced document object model (DOM) as rendered by the browser; and
- a set of features that can be used to describe single DOM elements. The feature set is based on information in the HTML source and information about how the elements are rendered by the browser.

We can use our mechanism for mapping gaze data to DOM elements to build up a salience map of elements in different contexts. Machine learning techniques can then be used to try to learn and predict the attentional salience of a given element based on the set of features. In the end, we are aiming at a model that takes the DOM of an HTML page as input and produces predicted salience values of each DOM element as output.

DOM-Based Feature Extraction

For describing single DOM elements, we derive two main classes of features, HTML-related and rendering-related features. In general, HTML-related features can be computed very easily just by looking at the HTML source code. In contrast, rendering-related features can only be calculated after the page has been rendered by a browser.

For HTML-related features, we created 44 simple binary features for the most frequent element names (e.g., "A", "DIV", "H1", etc.). For example, for an A element (i.e., a link) the feature "A" has the value 1, all the other features are 0. Furthermore, we computed 3 more abstract features:

- DOMTreeLevel is the level of an element in the DOM (i.e., the BODY element is always the root element with level 0; all other elements are below having a higher level based on their hierarchical nesting depth).
- LogoImage is a binary feature only applicable to image elements ("IMG"). It tells whether the file name of the image contains the substring "logo".
- HomeLink is a binary feature only applicable to link elements ("A"). It tells whether the destination of the link is the top-level page of the entire Web site.

We also computed 12 rendering-related features:

- Size of the element, computed as *width · height*.
- AspectRatio is computed as *min(width, height) / max(width, height)*.
- 10 positional binary features, i.e., 3x3 equal-sized regions above the fold (TopLeft, TopCenter, …, BottomRight), and the entire area below the fold (BelowFold). The position of an element is decided by the position of its center point. The 10 positions are all computed with respect to the visible area of a page in the browser. All elements that are not visible without scrolling after opening a page are below the fold. In our experiment, the area above the fold amounted to 996 x 716 pixels at the top of a viewed page.

Experimental design and procedure

To help us understand Web page viewing behavior, we designed a user study to collect gaze data from participants engaging in Web tasks. To maximize the ecological validity of our tests, we had participants perform several different tasks. For our analyses, we collapsed these tasks into two broad categories: *information foraging* and *page recognition* tasks.

A common requirement in eye-tracking research is to decrease variance by insuring that many of the Web pages that people view are the same. To this end, we designed eight tasks with very specific information needs and provided participants with small sets of more or less relevant Web pages to work with. In order to simulate the common occurrence of page revisitation under different task needs, these tasks were constructed in pairs that were on the same topic; each pair used a common set of Web pages, for a total of four distinct sets of pages. We selected four different task topics: *cars*, *kite surfing*, *wind energy*, and *diabetes*. For each topic, we provided links to nine Web pages to be used to complete the tasks. The task descriptions are given in Table 1.

The nine pages for each topic were carefully selected and most contained at least some relevant information for each of the two tasks. Each set of 9 pages was constructed to include a variety of page types and layouts, such as pages from well-known domains, text-only pages, pages with lots of images, etc. Each set also contained pages of different types, e.g., news, product descriptions, home pages, and encyclopedia articles. We used a factorial design for the order of the task topics for each participant. When starting a

Topic	First task	Second task
Cars	Which of three given cars (Porsche, BMW, Audi) has the best performance?	Which of those cars is small enough to fit in a tiny garage?
Diabetes	What are risk factors for type 2 diabetes that cannot be controlled / changed?	In the U.S., what are estimated yearly costs for diabetes therapy?
Kite surfing	Find a kite surfing school where you don't have to have your own equipment.	What basic equipment is recommended for kite surfing?
Wind energy	What are drawbacks and problems of wind power generation?	Find detailed info about how much wind power is generated in the U.S.
"Free query"	Find information about places you might want to visit (e.g., a vacation)	Find some new equipment for your favorite hobby.

Table 1: Descriptions of the information foraging tasks.

new task, the task goal was displayed and a list of links to the nine preselected Web pages was presented in randomized order. Participants were free to choose which of the nine Web pages to open and in what sequence. Participants were given about five minutes to complete each task and were encouraged to use most of that time.

We also wanted to explore how participants freely navigate the Web in pursuit of their own interests, so we had two "free query" tasks (see Table 1). For these tasks, the participants were also given about five minutes each, but they were free to search and browse any Web pages.

Finally, to get a sense of how participants look at pages for recognition or revisitation, we included a task in which participants were asked to indicate their familiarity with all the pages they looked at in the experiment, as well as twelve other pages not seen in the experiment. These latter twelve included well-known pages like commercial Web portals (e.g., amazon.com), news portals (e.g., cnn.com), search engines (e.g., google.com), and entertainment portals (e.g., youtube.com), as well as more obscure sites such as a personal home page, web portals of public organizations, commercial pages of small companies, etc. For each page, participants gave two assessments on a 5-point scale (ranging from "never" to "frequently"): First, how often they had seen that specific page before, and second, how often they had seen any page from that Web site before (not including the page views during the study).

The experimental sequence was as follows:

1) Participants completed the first four tasks for each topic (*cars*, *kite surfing*, *wind energy*, and *diabetes*), with topic order balanced using a factorial design and link order randomized.
2) They then completed two "free query" tasks with very broad information needs. During this phase, the participants could freely surf the Web.
3) This was followed by the second four tasks for each topic.
4) Participants were then asked to enter a variety of demographic and Web experience information.
5) Finally, they were asked to state their familiarity with each previously seen page and the twelve other pages.

With this experimental design, participants may have viewed some of the pages in three different task contexts: The first time during phase 1, the second time during phase 3, and the third time during phase 5. We refer to the structurally similar tasks during phases 1, 2, and 3 (finding specific information somewhere on the page) as *information foraging* tasks. The main goal during phase 5 was recognition of pages and sites, so we refer to this as a *page recognition* task.

Apparatus

All Web pages were shown in Internet Explorer 7; the browser window was sized to 1024x741 pixels. The use of tabs or additional windows was prohibited. Eye tracking was performed using the Tobii x50 eye tracker (see http://www.tobii.se/) paired with a 17" LCD monitor (96 dpi) set at a resolution of 1024x768. The eye tracker sampled the position of users' eyes at the rate of 50Hz and had an accuracy of 0.5°. Gaze data was logged by Tobii Studio. Before starting the tasks, we performed a 9 point calibration of the eye tracker for each participant using Tobii Studio. After the last task, a manual 9 point calibration was applied to determine the average tracking error for each participant. This was used after the experimental run to manually correct for systematic tracking errors as much as possible. Using a browser plug-in, we took a screenshot of every viewed Web page on the fly and stored its DOM in a file. So, at the end of each experimental run, we had three associated data sets which we used for offline analysis: the gaze log file, the set of DOM files, and the set of screenshots.

Participants

Twenty participants (10 male) ranging in age from 18 to 69 years old ($\emptyset = 33.0$, $\sigma = 14.2$) with a diverse range of jobs, backgrounds and education levels were recruited for this study from a user-study pool. All participants were native English speakers. An experimental run took approximately 1 hour for each participant.

Gaze-Based Measures and DOM Elements

Processing gaze data on a Web page starts with detecting fixations and mapping them onto viewed pages. We used the software bundled with the eye tracker (Tobii Studio) for fixation detection and adjusting the gaze position for scrolling in the browser. A fixation was detected by Tobii Studio after steadily gazing in an area with a radius of 50 pixels for at least 100 ms. For each viewed Web page, Tobii Studio reported a stream of fixation coordinates and durations relative to the page coordinates. From this data, we computed several gaze-based measures in parallel for each DOM element. Our intention was to build prediction models for each of the four gaze-based measures and then to analyze whether there was a difference between these models. If the models are very similar, then this indicates

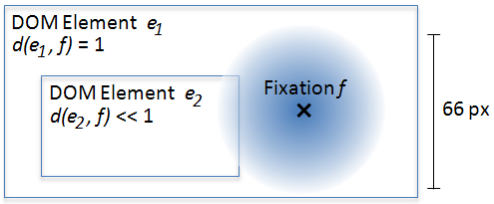

Figure 2: Fixation distance impact on nearby elements. The volume under the Gaussian distribution is represented by the color intensity around the fixation.

that all the measures have the same expressiveness concerning salience or importance of different features. We computed three measures which are described in detail below: *Median fixation impact*, *Viewing frequency*, and *Median time to first fixation*.

Fixation impact is probably the most unusual concept and bears some discussion. Human vision is characterized by a very narrow window of high acuity (the fovea) that covers about 2° of visual angle. When people fixate an area of the visual field, they also gather a smaller amount of information from the region around this point. Therefore, for each fixation we look at the surrounding area and determine the DOM elements that lie (partly) within this area. We chose the diameter of the area to be 2° of visual angle (i.e., matching the foveal area) which corresponded in our setting to 0.8 inches or 66 pixels on the display. Of course, this is a simplification since the focus of attention is not always in the center of the fixation; the locus of attention is also dependent on visual salience and context (e.g., while reading).

For a given fixation *f*, we first determine all DOM elements that intersect the circle around the fixation point. Then, we compute a Gaussian distribution with volume 1 and lay it over the circle around the fixation point. We calculate a distance impact value *d(e, f)* for each element *e* which is given by the volume of the Gaussian distribution above the element. So, if an element *e* completely covers the fixation circle, it gets a value of *d(e, f)*=1. If an element *e* covers the fixation circle only in parts, its value for *d(e, f)* is smaller (see Figure 2).

Having computed the distance impact value *d(e, f)*, we calculate the fixation impact value *i(e, f)* for a given element *e* by multiplying *d(e, f)* with the duration of the fixation *f* in milliseconds. So, an element that completely covers the fixation circle gets the full fixation duration as fixation impact value *i(e, t)*. This kind of computation is motivated by observations from vision research indicating fixation duration correlates with the amount of visual information processed. The longer a fixation, the more information around the fixation center is processed [21].

For each DOM element on a page, we keep track of all fixations and the fixation impact associated with it over all page views. A page view is the time between opening a Web page until closing it again by any participant, and a given participant may create several page views for the same page. Therefore, for each DOM element *e*, we calculate the following. Note that in all cases, an element *e* is "looked at" during a fixation *f* if it got some fixation impact $i(e, t) > 0$ from that fixation (i.e., if the element is close to the fixation point).

- *Median fixation impact*: *mi(e)*. We first computed the accumulated fixation impact on *e* for each page view and then stored these values in the set *I(e)*. So, each value in *I(e)* describes the accumulated fixation impact of *e* during one specific page view. *mi(e)* is the median of these values, i.e., the median across participants and page views.
- *Viewing frequency*: *p(e)*. The percentage of participants who looked at the element *e* on a page out of all participants that viewed that page at all.
- *Median time to first fixation*: *mt(e)*. Time-to-first-fixation is the time in milliseconds measured from opening the Web page until looking at the element *e*. This is the median of the time-to-first-fixation values across all participants and page views.

RESULTS

During the study, gaze data from 2,126 page views on 361 different Web pages was recorded. Each of the 9 pre-selected Web pages for the 4 task topics was viewed by 11.3 participants on average. This gave us a high overlap of gaze data across participants for 4x9=36 out of the 361 viewed Web pages.

For this analysis, we had two principal goals. First, we wanted to get an overview of the distribution of visual attention across Web pages. That is, we wanted to get a general sense of how users spatially allocate their visual attention at a high level for different tasks: which locations on a page generally attract most visual attention from users? Does this vary depending on tasks? Second, we wanted to see if we could create computational models based on the DOM of Web pages that can predict the visual salience of single elements on a page. Given an arbitrary Web page and HTML, can we predict what people will look at and how much?

General Characteristics of Web Page Viewing Behavior

Location-Based Overview

Figure 3 shows the median time to first fixation across all pages and page views for both information foraging and page recognition tasks. Here, each of 10 regions of the screen is represented in a corresponding rectangle: 9 equal-sized regions above the fold and everything else below the fold. We did not differentiate any further below the fold since it cannot be seen immediately after opening a page. Within each region is a circle proportional to the value in that region; smaller circles correspond to faster times to first fixation. The corresponding figures for median fixation impact (over the entire task duration) are shown in Figure 4. Here, larger circles correspond to greater fixation impact.

Figure 6 illustrates median fixation impact when we limit the data to the first second of viewing for each page. We

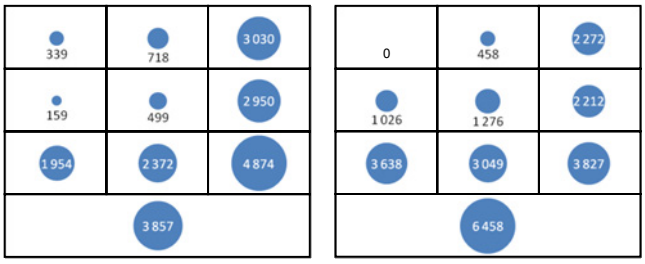

Figure 3: Median time to first fixation on the 10 page regions across all pages and page views (in milliseconds).

Figure 5: Viewing frequency of the 10 page regions across all pages during the first second of the page views.

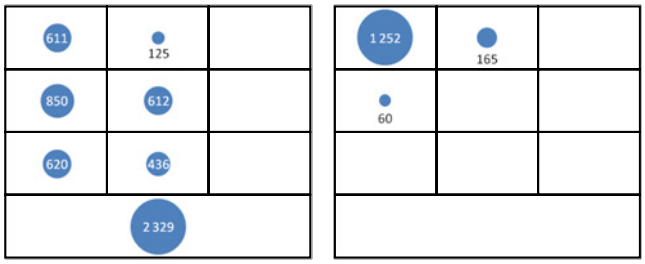

Figure 4: Median fixation impact on the 10 page regions across all pages and page views (in milliseconds) across the entire duration of tasks.

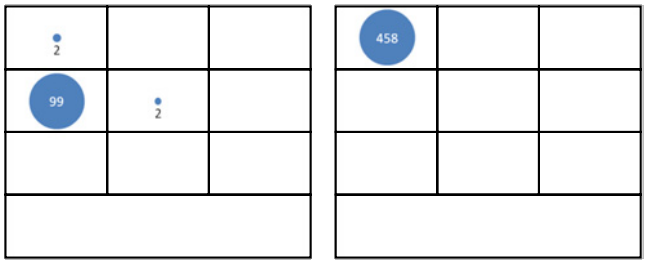

Figure 6: Median fixation impact on the 10 page regions across all pages and page views (in milliseconds) during the first second of page views.

included this analysis for the first second of each page view based on observations in [12]. They found that the first few fixations are controlled by visual features and *global semantic characteristics* of the visual scene. In our scenario of Web page viewing, such global semantic characteristics might result in expectations about where to find the most relevant information on a page before having seen the page. So, the first few fixations may reveal the locations where users expect to find relevant information on a page before they begin detailed exploration and reading. We did not exclude the very first fixation on a page despite the fact that it cannot have been influenced by the page's content, because it might still express the user's expectation about where to find important information.

These results illustrate several characteristics of viewing behavior. Some of these appear to apply to all viewing scenarios, while others seem to be task-dependent.

It is striking that the entire right side above the fold is neglected for both information foraging and page recognition tasks. Participants took about 3 seconds to fixate the three regions on the right side of a page for information foraging tasks, and about 2 seconds for recognition tasks. The median fixation impact on these regions is 0 for all tasks, indicating that even when they *do* look at these regions, participants don't spend much time there. In contrast, the four upper left regions generally attracted visual attention faster than all others. Independent of the task, the time to first fixation was lowest for these regions.

There are also interesting task-dependent differences in the relative importance of different regions of pages. This is particularly clear during the first second of each page view. For information foraging tasks, the center-left region attracts more and earlier attention than other regions. For recognition tasks, the top-left region attracts the most and earliest attention.

More generally (again mainly looking at the first second of page views and the time to first fixation), the three regions center-left, top-left, and center-center seem to be most important for information foraging tasks. For page recognition tasks, the top-left, top-center, and center-left regions appear to be most important. The very large difference in median fixation impact below the fold across the full duration of tasks (Figure 4), is likely due to the in-depth reading and exploration associated with our information foraging tasks.

Figure 5 shows the viewing frequency of the 10 regions during the first second of the page views (i.e., the percentage of people looking at the page who looked at a given region). Supporting the observations from above, the figures also indicate that the four top left regions are most important, especially for page recognition tasks. The bottom-left and bottom-center regions above the fold are viewed much more frequently during information foraging than page recognition tasks. Generally, the viewing behavior during the first second of page recognition tasks seems to be less diverse than during the same time span of information foraging tasks.

Differences Among Participants

To further explore the distribution of visual attention, we wondered if gender, age, Web site familiarity, or experience

in Web browsing might affect how different people viewed the 10 screen regions.

Generally, during page recognition tasks women tend to be more thorough, looking significantly longer at the page and at every region on it than men (mean women: 3432ms, mean men: 2866ms; two-tailed t-test: $t=2.1$, $p < 0.05$; mean age of both groups: 33.0). This is reminiscent of the findings of Lorigo et al. [19], who found large differences between the scanpaths of men and women for commercial Web search pages. During the first second of information foraging tasks, the differences for the top-center, top-right, and below-the-fold regions approached statistical significance (Bonferroni corrections of $\alpha=0.005$). Here, men tended to spend less time on the top-center region, but more time on the top-right and below-the-fold regions.

To explore the influence of age, we divided the participants in two groups: those under 30 years old (n=10; mean age: 22.3) and those older than 30 (n=10; mean age: 43.7). During the recognition task, participants older than 30 years looked significantly longer at the page and at every region than younger participants (mean older than 30: 3699ms, mean younger than 30: 2593ms, $t=4.1$, $p < 0.001$). During the first second of information foraging tasks, younger participants looked significantly longer at the center-center region, but significantly shorter at the center-left position.

Especially during the first second of page views of information foraging tasks there are significant differences with respect to Web site familiarity. When viewing pages from familiar Web sites participants looked significantly longer at the top-left, top-right, and bottom-left regions compared to pages from unfamiliar sites. At the same time, they looked for a shorter while at the center-center and bottom-center regions.

Most of our participants (17) were very experienced with Web browsing; only 3 stated that they didn't browse the Web multiple times a day. We found that these less-experienced participants generally looked longer at pages during page recognition tasks. However, this finding was not significant and we need to collect data from more people who only rarely use the Web to explore this effect.

Discussion

Our results suggest that for page recognition, users expect the most important features of Web pages to be in the top left-hand corner of the page. The very low times to first fixation and complementary large median fixation impacts and viewing frequencies for these regions support this finding (Figures 3-6, right). For our Web page recognition task, it is clear that the top left 4 regions are the most informative regions.

For more goal-directed, information-driven tasks, the first few fixations (i.e., during the first second of each page view) seem to be similar to the recognition task. Again, the 4 top left regions are looked at earliest (see Figure 3). However, the first few fixations remain longer in the center-left region (as opposed to the top-left region, as during the page recognition task) (see Figure 5). Here, fixations seem to be more directed by the user's intention to find task-specific information within a page.

We were somewhat surprised to see that the right third of Web pages attracted almost no visual attention at all during the first second of each page view. This suggests that our participants have low expectations of information content or general relevance on the right side of most Web pages. This seems plausible because many Web pages display advertisements on the right side. Most people seem to entirely ignore this region, only occasionally looking there. This finding seems consistent with results of previous studies that reported triangular- or F-shaped scan patterns on Web pages [14, 20].

In general, independent of the task they are engaged in, there seems to be a common orientation phase when people first view a Web page. In the first few moments, people quickly scan the top left of the page, presumably looking for clues about the content, provenance, type of information, etc. for that page. The elements that are normally displayed in the upper left third of Web pages (e.g., logos, headlines, titles or perhaps an important picture related to the content) seem to be important for recognizing and categorizing a page.

In the study we analyzed only the average viewing behavior with heat map-like techniques over a wide variety of page types, layouts, and designs. Presumably, certain layouts or designs have a specific influence on eye movements and there may be temporal behaviors (e.g., scan paths) that we did not detect here. However, in-depth analysis of these issues remains for future work.

Prediction of Salient Elements

In order to create a computational model for predicting salient DOM elements on a page we consider every DOM element as AOI (area of interest) and use sets of features to describe the AOIs and group them together. I.e., we have the set of HTML- and rendering-related features that can be used to describe a specific element (59 different features, see "Methods" section). In addition, we have the 3 different measures of gaze data described in the "Methods" section (fixation impact, viewing frequency, and time to first fixation). These gaze-based measures build the ground truth of salience for each DOM element (AOI) that we will attempt to predict.

A certain amount of preprocessing of the data was necessary prior to learning a prediction model. First, we removed all DOM elements larger than 200,000 pixels (e.g., 450x450) from the data set (6% of all elements). Very large elements, like the `BODY` element, that span the entire page are not interesting for us to predict and tend to produce outliers with respect to the gaze-based measures since they are associated with virtually any fixation on the page. Second, we normalized the values of all non-binary features and our gaze-based measures to (0, 1) intervals. All feature normalization was linear, except for `DOMTreeLevel`. This

was normalized logarithmically, since some of the Web pages had very deep DOM trees while others were very flat. Logarithmic normalization emphasized the differences between lower levels in the tree (i.e., closer to the root).

Overall, our dataset consisted of about 150,000 DOM elements coming from 361 different Web pages that were viewed by at least one participant during the experiment.

Feature Information Gains

As a first step we wanted to reduce the number of relevant features to a more tractable number, so we computed the information gain (as computed by the WEKA toolkit based on entropy, http://www.cs.waikato.ac.nz/ml/weka/) for each of the 59 features with respect to the 4 gaze-based measures. Therefore, we discretized the values of the measures to either 1 or 0 dependent on whether an element was assigned a value ≥0 or =0. Table 2 shows 10 features with the largest information gain based on median fixation impact, with task type and viewing time considered.

Seven of the top 10 features with the highest information gain are based on rendering-related information of the elements: their size, position on the page, and aspect ratio. The feature with the highest information gain of all HTML-based features is DOMTreeLevel.

For all 4 gaze measures, Size is among the features providing the highest information gains. This conforms to the intuition that bigger elements (like a DIV box spanning the entire navigation bar) are looked at more often that smaller elements (e.g., a specific link in the navigation bar).

Table 2 indicates that while the top features for information foraging and recognition tasks are very similar in many cases, there are some notable differences. This is particularly clear for AspectRatio, DOMTreeLevel, and some positional features. Note that the relative importance (represented by color intensity in Table 2) differs considerably for task type.

The relative differences of the feature information gains were very similar with respect to all 3 gaze-based measures. Further, all 3 measures could be predicted with comparable quality by two different prediction methods we used in the following sections. As a result, for the balance of the paper we concentrate on median fixation impact as our gaze-based measure of choice.

Linear Regression: Feature Weights

Having determined the top 10 features based on information gain, we wanted to see what influence each measure had on median fixation impact. The feature weights as computed by linear regression are presented in Table 3.

As suggested by information gain, Size is the most decisive factor and is positively related to fixation impact; larger elements accumulate more fixation impact. All positional features above the fold have positive weight and BelowFold has negative weight. In concordance with our findings regarding the region-based gaze distribution, the CenterCenter position is more important for information foraging tasks whereas the TopLeft and TopCenter positions are more important for recognition tasks. AspectRatio, DIV and A are generally not very useful in combination with the other features.

Interestingly, DOMTreeLevel has a negative weight for recognition tasks but a slight positive weight for the information foraging tasks. This indicates that elements that are deeper down the DOM tree are penalized more for recognition tasks. It can be assumed that elements on deeper levels in the tree are more (topically) specific than elements on higher levels. Thus, for recognition tasks, more general elements are important; for information foraging tasks more specific elements are important.

Performance of Prediction Methods

While there are many possible applications for predicting attentional salience of Web page elements, we decided to focus first on scenarios of page re-finding (e.g., Web history). Because information re-finding is particularly dependent on recognizing previously seen pages, we focus on the data from the recognition task. The information gains are always higher when we look at the entire time of each page view as opposed to just the first second (see Table 2), so we used median fixation impact from the entire time of page views for the recognition tasks for generating our

Task →	Info foraging		Recognition		Both	
Feature name	all time	1st sec	all time	1st sec	all time	1st sec
Size	0.075	0.032	0.099	0.062	0.06	0.03
BelowFold	0.042	0.037	0.098	0.057	0.047	0.025
AspectRatio	0.028	0.01	0.101	0.056	0.023	0.01
TopCenter	0.008	0.007	0.066	0.048	0.023	0.012
TopLeft	0.002	0.002	0.052	0.031	0.009	0.005
DOMTreeLevel	0.006	0.003	0.033	0.02	0.011	0.006
CenterCenter	0.015	0.014	0.003	0.001	0.008	0.004
DIV	0.002	0.001	0.007	0.004	0.004	0.002
A	0.005	0.002	0.003	0.002	0.004	0.002
CenterLeft	0.007	0.004	0.002	1E-04	0.003	8E-04

Table 2: Information gain of the top 10 features for median fixation impact. Color intensity represents relative importance compared to the other features.

Task →	Info foraging		Recognition	
Feature name	all time	1st sec	all time	1st sec
Size	0.318	0.231	0.291	0.262
BelowFold	-0.016	-0.013	-0.004	-0.002
AspectRatio	0.005	0.009	-0.009	-0.009
TopCenter	0.003	0.016	0.134	0.121
TopLeft	0.006	0.012	0.069	0.064
DOMTreeLevel	0.021	0.006	-0.038	-0.02
CenterCenter	0.055	0.084	0.033	0.032
DIV	-0.004	-0.003	0.006	-5.E-04
A	6.E-04	0.001	0.001	0.002
CenterLeft	0.023	0.021	0.022	0.017

Table 3: Feature weights of the top 10 features as determined by linear regression with respect to median fixation impact. Color and intensity correspond to sign and magnitude of the weights.

models of salience prediction. There are myriad machine learning techniques that we could potentially use for our predictions. Rather than immediately dive into sophisticated models, we decided to begin with fairly simple and easy-to-understand models of linear regression and decision trees to gauge the utility of the idea.

Linear regression for approximating median fixation impact based on the top 10 features from Table 2 yielded a correlation coefficient of 0.50 and a root mean squared error (RMSE) of 0.08 milliseconds. The quality measures were determined by 10-fold cross validation.

We further determined a decision tree (C4.5, pruned to 65 leaves) for predicting whether a DOM element has any fixation impact (yes or no). As determined by 10-fold cross validation, the decision tree had a Kappa coefficient of 0.59, a precision of 75%, and a recall of 53%.

Generally, the correlation and Kappa coefficients of the two prediction methods were promising, so we implemented a new prediction method based on a combination of these two methods: First, the decision tree from above is applied for deciding whether an element received any fixation impact at all. If this is the case, then the linear regression method is used to approximate the magnitude of the fixation impact on that element. This combined prediction method yielded a fairly good correlation coefficient of 0.69 and a relatively low RMSE of 0.08 milliseconds.

We also wondered how much the quality of prediction would drop if all rendering-based features were excluded. To predict salient elements based only on HTML-related features is especially interesting for search engines that cannot render every page while crawling. Based on the information gain seen in Table 2, it is not surprising that prediction was rather poor. Linear regression based only on the HTML-based features had a correlation coefficient of 0.28 and an RMSE of 0.10 ms. A decision tree performed with similarly poor quality.

To actually see the page elements and their predicted values, we visualized the actual and predicted median fixation impact for each Web page. In Figure 7, we show those visualizations for two representative Web pages. All DOM elements on the page are outlined by a black rectangle. The red color intensity of those rectangles represents the actual (left) and predicted (right) median fixation impact on each element, with deeper red color signifying more fixation impact.

Discussion

Generally, the prediction method seems to work well and finds the most important elements for recognizing a page. However, it is biased to prefer elements that are on the upper left-hand side of a page. As the region-based analysis has shown, this should be expected for the general case (e.g., the car review page in Figure 7).

For specific Web pages, however, the prediction method may miss important elements on the right side of the page, such as the CNN home page in Figure 7. This is especially apparent for pages from Wikipedia.org where illustrative images are frequently located on the far right side of the page. As noted above, however, our prediction models were (purposely) very simple. We believe that more sophisticated models could yield much higher accuracy for these cases.

Depending on the area of application, one could make use of the prediction in different ways. For example, to create recognizable small visual representations of a Web page (as motivated in the introduction), one could extract the image of a certain size that is predicted to be most salient together with the logo of the page and maybe a highest-ranked small text section. These three elements could be emphasized in a more intelligent thumbnail for a page, e.g., by enlarging them. With increasing size available for this "intelligent" visual representation (e.g., depending on the available space on the screen of desktop computers, laptops, mobile phones, etc.), one could emphasize more and more of the elements in the order of their predicted salience.

CONCLUSION

We have presented the methods and findings of a study aimed at understanding people's visual attention patterns when viewing Web pages. This work entailed several contributions, including:

- A method for mapping gaze data to Web page elements based on the concept of *fixation impact*.
- A model of the most salient regions on Web pages, taking into account task type (information foraging vs.

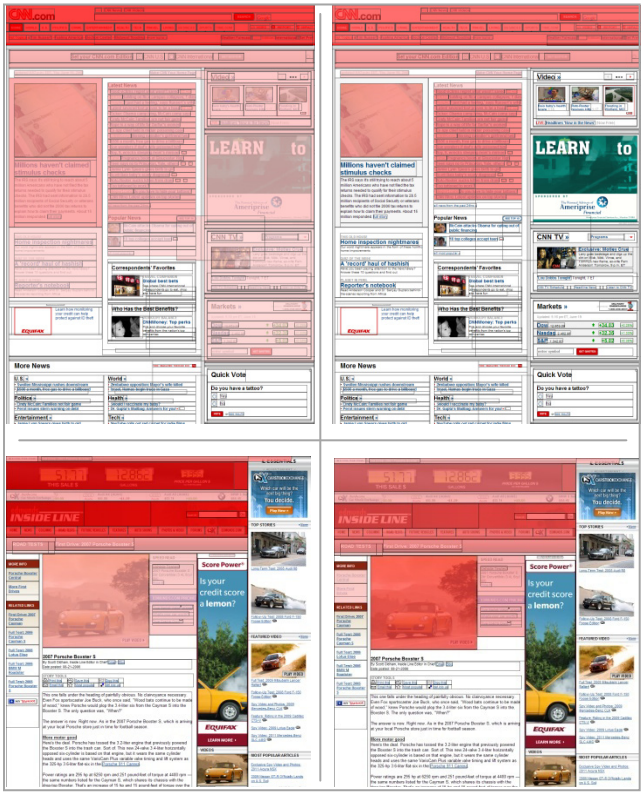

Figure 7: Measured median fixation impact and its prediction for CNN front page (top) and a car review page (bottom).

recognition) and demographics.

- A method for predicting salient DOM elements for a Web page.

This increased understanding of users' web-viewing behavior is valuable not only for improving Web page design, but also for creating new types of Web user interfaces. For example, compact representations of Web pages are desirable, but thumbnails of an entire page are unusable at small sizes [18]. Our model could be used to select the most salient regions of Web pages to create compact collages representing these pages. For information foraging tasks, our model could be used to create representations for page previews (i.e., thumbnails accompanying a search results list), while for recognition tasks it could be used to create representations for re-finding (i.e., thumbnails accompanying a bookmarks or history list).

In future work, we hope to explore more powerful methods of machine learning and classification to improve our prediction methods for DOM element salience. In addition, we would like to look at more complex models to explain why and when the eyes move to certain regions of a page. Those models should include a variety of visual features of Web pages like color, contrast, shape, etc, and they should also consider the temporal aspect of fixations. Finally, we would like to use the salient elements predicted by our model to automatically generate small visual representations of pages to help users recognize them for re-use.

ACKNOWLEDGMENTS

We would like to thank Mukund Narasimhan, Susan Dumais, Andreas Dengel, Steve Bush, and Raman Sarin for their valuable comments and great support. We are also grateful for the reviewers' thoughtful, detailed comments.

REFERENCES

1. Adar, E., Teevan, J., Dumais, S.T. Large scale analysis of web revisitation patterns. In *Proc. CHI 2008*, 2008, 1197-1206.
2. Beier, B., Vaughan, M.W. The bull's-eye: a framework for web application user interface design guidelines. In *Proc. CHI 2003*, 2003, 489-496.
3. Bernard, M.L. Criteria for optimal Web design (designing for usability). From *http://www.rocketface.com/archive/Criteria%20for%20optimal%20web%20design.html*, (Retrieved Sept. 8, 2008).
4. Beymer, D., Orton, P.Z., Russell, D.M. An Eye Tracking Study of How Pictures Influence Online Reading. In *Proc. INTERACT 2007*, 2007, 456-460.
5. Cockburn, A. and B. McKenzie. What do Web users do? An empirical analysis of Web use. *Int. J. of Human-Computer Studies*, 54(6), 2001, 903-922.
6. "Comcast Taps Hispanic Web Portal", http://www.mediapost.com/publications/index.cfm?fuseaction=Articles.showArticleHomePage&art_aid=40714, (Retrieved Sept. 15, 2008).
7. Cutrell, E., Guan, Z. What are you looking for?: an eye-tracking study of information usage in web search. In *Proc. CHI 2007*, 2007, 407-416.
8. Dontcheva, M., Drucker, S.M., Wade, D., Salesin, D. and Cohen, M.F. Summarizing personal Web browsing sessions.. In *Proc.of UIST*, ACM, 2006, 115-124.
9. Fogg, B.J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., Treinen, M. What makes Web sites credible?: a report on a large quantitative study. In *Proc. CHI 2001*, 2001, 61-68.
10. Goldberg, J.H., Stimson, M.J., Lewenstein, M., Scott, N., Wichansky, A.M. Eye tracking in web search tasks: design implications. In *Proc. of the 2002 symposium on Eye tracking research & applications*, 2002, 51-58.
11. Henderson, J.M. Human gaze control during real-world scene perception. In *TRENDS in Cognitive Sciences 7*, 11, 2003, 498-504.
12. Henderson, J.M., Hollingworth, A. High-level scene perception. In *Annu. Rev. Psychol.*, 50, 1999, 243-271.
13. Herder, E. Characterizations of user Web revisit behavior. In *Proc. Workshop on Adaptivity and User Modeling in Interactive Systems*, 2005.
14. Hotchkiss, G., Alston, S. and Edwards, G. Eye Tracking Study. *http://www.enquiro.com/eyetrackingreport.asp*, (Retrieved Sept. 15, 2008), 2006.
15. Itty, L. Models of Bottom-Up Attention and Saliency. In *Neurobiology of Attention*, Elsevier, 2005, 576-582.
16. Ivory, M.Y., Hearst, M.A. Statistical profiles of highly-rated web sites. *Proc. CHI 2002*, 2002, 367-374.
17. Josephson, S., Holmes, M.E. Visual attention to repeated internet images: testing the scanpath theory on the world wide web. In *Proc. of the 2002 symposium on Eye tracking research & applications*, 2002, 43-49.
18. Kaasten, S., Greenberg, S., Edwards, C. How people recognize previously seen Web pages from titles, URLs and thumbnails. In *Proc. HCI '02*, 2002, 247-265.
19. Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., Gay, G. The influence of task and gender on search and evaluation behavior using Google. In *Info. Processing and Management: an Int'l Journal. 42*, 4, 2006, 1123-1131.
20. Nielsen, J. F-Shaped pattern for reading Web content. http://www.useit.com/alertbox/reading_pattern.html, (Retrieved Sept. 15, 2008), 2006.
21. Ojanpää, H., Näsänen, R., Ilpo, K. Eye movements in the visual search of word lists. In *Vision Research 42*, Elsevier, 2002, 1499-1512.
22. Oliva, A., Torralba, A., Castelhano, M.S., Henderson, J.M. Top-down control of visual attention in object detection. In *Proc. ICIP 2003*, 2003, 253-256.
23. Over, E.A.B., Hooge, I.T.C., Vlaskamp, B.N.S., Erkelens, C.J. Coarse-to-fine eye movement strategy in visual search. In *Vision Research 47*, Elsevier, 2007, 2272-2280.
24. Pan, B., Hembrooke, H.A., Gay, G.K., Granka, L.A., Feusner, M.K., Newman, J.K. The determinants of web page viewing behavior: an eye-tracking study. In *Proc. of 2004 symposium on Eye tracking research & applications*, 2004, 147-154.
25. Rayner, K. Eye movements in reading and information processing: 20 years of research. In *Psychological Bulletin*, 124, 1998, 372-422.
26. Tauscher, L., S. Greenberg. How people revisit Web pages: Empirical findings and implications for the design of history systems. In *Int. J. of Human-Computer Studies*, 47(1), 1997, 97-137.
27. Teevan, J., Cutrell, E., Fisher, D., Drucker, S.M., Ramos, G., André, P., Hu, C. Visual Snippets: Summarizing Web Pages for Search and Revisitation. In *Proc. CHI 2009*. 2009.
28. Tseng, Y.-C., Howes, A. The Adaptation of Visual Search Strategy to Expected Information Gain. In *Proc. CHI 2008*, 2008, 1075-1084.