How many words is a picture worth? Attention allocation on thumbnails versus title text regions

Chaitra Yangandul
University of Florida
Gainesville, Florida
chaitray@ufl.edu

Sachin Paryani
University of Florida
Gainesville, Florida
sachinparyani@ufl.edu

Madison Le
Google
San Bruno, California
madisonle@google.com

Eakta Jain
University of Florida
Gainesville, Florida
ejain@ufl.edu

ABSTRACT

Cognitive scientists and psychologists have long noted the "picture superiority effect", that is, pictorial content is more likely to be remembered and more likely to lead to an increased understanding of the material. We investigated the relative importance of pictorial regions versus textual regions on a website where pictures and text co-occur in a very structured manner: video content sharing websites. In our study, we tracked participants' eye movements as they performed a casual browsing task, that is, selecting a video to watch. The fixations were coded as falling on one of two areas of interest: thumbnail image or title text region. We found that participants allocated almost twice as much attention to thumbnails as to title text regions. They also tended to look at the thumbnail images before the title text, as predicted by the picture superiority effect. These results have implications for both user experience designers as well as video content creators.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **User studies**; **Interaction paradigms**; **Web-based interaction**; **Empirical studies in HCI**;

KEYWORDS

User experience, Eye tracking, Viewing patterns, Picture text integration, Human-computer Interaction

ACM Reference Format:
Chaitra Yangandul, Sachin Paryani, Madison Le, and Eakta Jain. xxxx. How many words is a picture worth? Attention allocation on thumbnails versus title text regions. In *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ETRA '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND BACKGROUND

It is often said that a picture is worth a thousand words. More formally, this phenomenon is referred to as the *picture superiority effect*, that is, pictorial information is more likely to be remembered than textual information. The implications of the picture superiority effect are greatest for education and marketing: pictures have been found to improve understanding in the educational context [Levie and Lentz 1982], and the pictorial component of an advertisement has been reported to capture more attention than text [Pieters and Wedel 2004]. As a result, there has been focused research on eye movements for integrated text and picture stimuli, such as informational and educational materials [Holsanova et al. 2009; Schmidt-Weigand et al. 2010], print advertisements [Rayner et al. 2001], and web advertisements [Simola et al. 2011]. This research has reported that viewers spend almost 50% more time looking at the picture rather than the text [Rayner et al. 2001], and that the likelihood of the users' first fixation landing on the picture was between 60 − 70% [Rayner et al. 2008].

Yet, this body of work has also shown that task has an effect on attention allocation: when participants were told that their goal was to judge the effectiveness or aesthetics of the advertisement, they spent more time looking at the picture [Rayner et al. 2008], whereas when their goal was to make a purchase, they spent more time looking at the text [Rayner et al. 2001]. In our work, participants are given a specific goal: it is Friday night and they have to find something to watch. In that sense, their goal is closer to the "buy a product" case rather than a "judge the advertisement" case. If we were studying print advertisements, we would expect that title texts would be looked at more often than thumbnails. However, the print ads considered in most advertising research involve a single product whereas browsing a video content provider service involves looking through multiple products arranged in a large grid, with high spatial contiguity between the thumbnail and the text. Layout has been shown to have an effect on attention allocation: if the picture and text have high spatial contiguity, there are a larger number of integrative saccades, i.e., shifts between the picture and the text, than if the picture and text are separated [Holsanova et al. 2009]. We might then expect that in our study both pictures and text will be looked at.

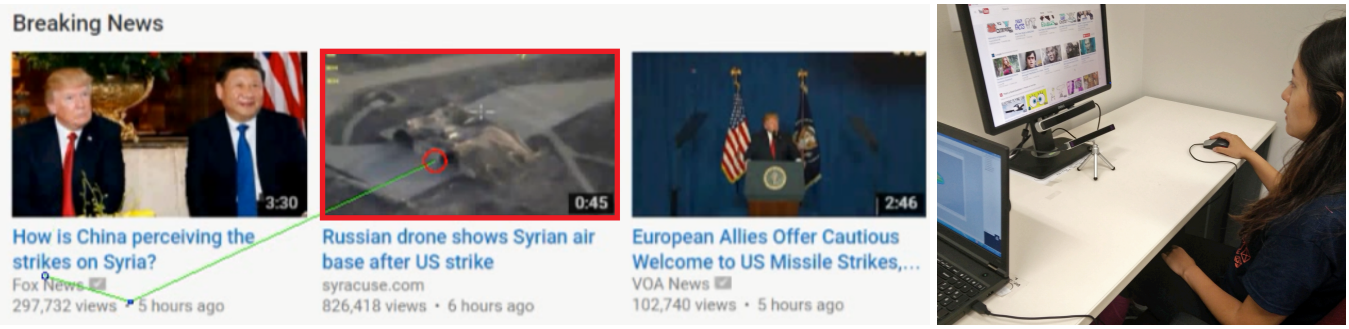

Figure 1: Left: The video thumbnail that the user fixated on is highlighted in red. Right: A remote eye-tracker is used to track fixations and saccades as users navigated the popular video content provider website YouTube. Typically on these websites, content is arranged in a grid, where each element of a grid is a thumbnail image and two to three lines of text underneath it, as seen on the computer screen above.

Resolving the attention allocation question on video content sharing websites is crucially important both for the user's experience and for the service provider's revenue. Though a large amount of data about user preferences is typically collected from the user's browsing and sharing history, it does not include gaze data because eye-tracking typically requires either a webcam or an infra-red camera to be turned on.

Researchers have proposed that mouse activity be used to infer users' interests for a variety of web browsing tasks [Goecks and Shavlik 2000; Guo and Agichtein 2010]. For example, Lagun and colleagues tracked the viewport on a mobile phone while users performed a search task, as a proxy for user attention. They reported that the view-port was a sufficient proxy in the case of tasks that did not require knowing user attention too finely [Lagun et al. 2014]. Huang and colleagues [Huang et al. 2012] examined whether mouse cursor position is a good proxy for user attention on different tasks. They reported that the cursor lags behind gaze by almost a second, likely because the eye can move much faster than a user can move the mouse. The deviation between gaze position and cursor position is also very dependent on the individual, much more so than task type. That the eye leads the hand is not surprising; it is, in fact, consistent with real-world tasks such as food preparation [Land and Hayhoe 2001] and catching a ball [Land and Tatler 2009]. As a result, the question remains open: what information does a user process from the time the web-page renders on their screen, to the first click they make?

In this study, we examined the relative importance of the pictorial region (thumbnails) versus the text region (title texts) while users browse for videos. We eye-tracked participants and recorded their screens. We annotated the screen capture with the two types of areas of interest (AOIs), and analyzed the fixations on each.

2 EXPERIMENT

A web usability task was setup on an external monitor (19"x11.8", 1680×1050 resolution) connected to a laptop (Figure 1 (Right)) as an extended screen. Participants were given the following instruction, "We are interested in users' casual browsing patterns on YouTube. Imagine that it is Friday night and you have some free time, but there is nothing good on TV. Please scroll through the videos on the YouTube homepage and click on something that catches your attention. After watching this video, please return to the home page and find something else you like. Your goal is to spend about 10 minutes."

Because our goal is to understand gaze behavior between the time the homepage is loaded and the time when the user makes their choice by clicking on a particular thumbnail, we instructed the user to only browse the homepage, and not use the search bar. To improve task compliance, we removed the keyboard entirely, and only provided the user with a mouse (Figure 1 (Right)). Additionally, we asked the user to return to the homepage by clicking on the YouTube icon on the top left corner of the screen in order to make their next selection, rather than clicking on the suggested videos based on their selection.

We define one trial as the data collected between the time stamp that the YouTube homepage was loaded, and the time stamp that the video selection was made. Instructing the participants to click on the YouTube icon on the top left corner of the page to return to the home page had the additional purpose of guiding all participants to start their browsing task from the same region of the screen, much like a fixation cross is used between trials in a traditional eye-tracking study. A few participants clicked the Back button on the browser out of habit. We did not discard these trials as the Back button is sufficiently close to the YouTube icon for the purpose of demarcating trials.

In order to keep the homepage as consistent as possible, we decided to not let the subject sign in to the service. We later realized that the homepage does change very slightly across participants, from session to session, for example, the contents of trending videos can change every day or even hourly on the same day. Because the metrics that we compute involve number of visits to either a thumbnail image or a title text region, our findings remain relevant despite the slight content differences.

While the user performed their task, their screen was recorded and their mouse movements were logged. Additionally, their gaze was recorded through the EyeTribe eye-tracker (30 Hz, infrared-based remote eye-tracking). The data capture and logging was synchronized through the open source software package OGAMA, Version 5.0 [Vosskuhler et al. 2008]. Each session began with an

informed consent procedure. Participants were calibrated using the Eyetribe 9-point calibration procedure. We validated the calibration accuracy by collecting gaze data while participants viewed a fixation cross located at the center of the four quadrants of the screen. A fixation comprised five or more consecutive gaze points.

Two instructions slides were displayed, the first slide explained the task to the participant, and the second slide showed them the location of the YouTube icon that would take them back to the homepage to start the next trial. A short questionnaire was administered to collect demographic information, questions regarding usage of the video streaming service, and their experience after their session was done. Finally, we noted the distance of the participant from the screen in inches for computing the pixels to visual angle conversion.

3 PARTICIPANTS

Twenty four participants were recruited from the university community in accordance with an IRB approved protocol. One participant did not pass the calibration procedure and did not complete the demographics and post-experiment questionnaire. Thus, we discarded this participant's data from all subsequent analysis.Of the remaining 23 participants, eleven participants were female. The age range of these 23 participants was 18-35 years, though 21 participants were less than 25 years of age. The participants' ethnicity/ racial breakdown was: Asian/Pacific Islander: 78.26%, White: 17.39,%African-American/Black: 4.35%. The academic positions of the participants were: Graduate students: 78.26%, Freshman: 0%, Sophomore: 4.35%, Junior: 4.35%, Senior: 8.69%, Other: 4.35%. Our participants were quite familiar with video browsing. More than seventy percent of the participants reported using YouTube for either an hour or more each day, or more than once a day but for less than an hour each day. More than half of the participants preferred using YouTube after signing in, possibly for content suggestions based on their browsing history (52.18% responded "Yes", 17.39% responded "Sometimes", and 30.43% responded "No").

4 QUALITATIVE ANALYSIS

For the question "What categories are you more likely to visit YouTube for", participants could select one or more of the following categories: Entertainment, Learning, Tips or Tricks, News, Politics, History, and Other (Other was a freeform text box). We found that every participant checked Entertainment. We found that the remaining categories were selected in the following preference order: Learning (13 selections), Tips or Tricks (9 selections), News (7 selections), Politics (5 selections), History (4 selections), Others (2 selections). The freeform text returned for "Others" was "Sports" and "Music".

Participants were asked "What made you click on the videos that you selected during the experiment?" The answer was given as free-form text. We found that the free-form answers ranged from 1 to 27 words (median = 6). Most participants (78.26%) gave one sentence answers, where a sentence counts as anything longer than three words. We found that 56.52% participants used some form of the word "interest" in their answer. Of the remaining answers, reasons included curiosity, mood, and familiarity with content. Only one participant mentioned the title of the video as the reason, and only one participant used the words "eye-catching" as their reason.

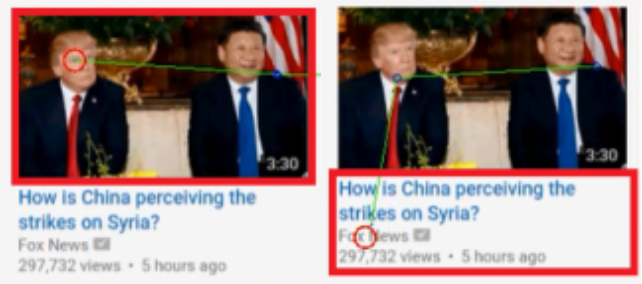

Figure 2: The areas of interest are illustrated here. Left: Thumbnail image (294 × 165 pixels), Right: Title text region (294 × 98 pixels).

Of the 23 participants, one more participant was discarded because the software hung during data collection.All gaze analysis was thus done on data from twenty two participants. Twenty participants complied with our instructions. One participant did not return to the homepage after the first video, therefore, this participant only counted towards our data as a single trial. One participant clicked on a suggested video after viewing the one they had selected. This did not impact the data in this participant's trial because we only consider gaze data from the time the YouTube icon is pressed to the time the next selection is made.

5 QUANTITATIVE ANALYSIS

Because the EyeTribe toolkit only reports levels of calibration (Perfect, Good, Moderate, Poor, Recalibrate), we allowed the participants to proceed with the experiment on getting "Perfect" calibration. Based on our numerical validation, the average calibration error was 1.09° (range 0.34° to 4.05°). Two participants were discarded because calibration error was found to be greater than 1.5°. Data from the remaining twenty participants is used for the rest of the quantitative analyses. Each data collection session lasted 114 seconds on average (time spent browsing, not counting time spent watching the selected video). In this duration, we obtained approximately four trials per participant, where a trial is defined as clicking on the YouTube icon to reach the homepage, browsing the videos on the homepage, and clicking on a thumbnail. In all but one trial, viewers watched their chosen videos for more than 30 seconds (Figure 3). The maximum number of videos watched in a session was seven, and the minimum number was two.

Figure 2 shows the thumbnail and title text box and the relative sizes of these two regions of interest. The dimensions of the AOIs that we used for analysis is 294 × 165 pixels for the thumbnails and 294 × 98 for the title text region. As a comparison, the average error in pixels is approximately 52.5. Figure 4 illustrates an instance where the thumbnail is viewed before the title text (Fixations 1,2,3,4,5) and an occurrence of the title text being viewed before the thumbnail (Fixations 6,7).

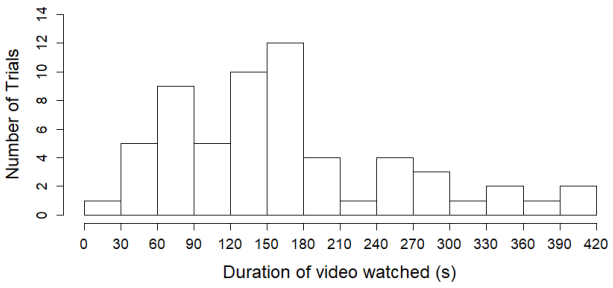

Figure 3: The distribution of watch durations suggests that viewers were interested in the videos they selected.

5.1 Amount of time spent browsing prior to a video selection

A trial is defined as the time spent on the YouTube homepage looking for a video to watch. For each trial, the trial start time is the time stamp when the participant presses the YouTube home icon. The end time is the time stamp of the first frame after the user selects a video to view. For each participant, and each trial, the time spent on the homepage for this trial is computed as the difference of these two time stamps in seconds *TimeSpentHomePage*.

The average time spent on the homepage for each participant is the mean value over all the trials for this participant (Figure 5 (a)). We found that the average time spent on the homepage across all participants is 116.9 seconds (σ = 73.5). This large variability across participants is driven by two participants who spent a lot of time looking for something to watch and five participants who were quick decision makers (they spent as little as ten seconds in selecting the video they wanted to watch). Figure 5 (b) shows that most participants are within one standard deviation of the mean.

For each participant, we also computed the *TotalTimeSpentHomePage* as the sum of the *TimeSpentHomePage* values across all trials for a given participant. The total time spent on the homepage over all trials was then converted into a percentage of the duration of the participant's entire session. We found that participants spent on average 18.8% of the duration of their session browsing for the what to watch (μ = 18.8%, σ = 10.7%).

5.2 Number of thumbnails viewed versus number of title texts viewed

We computed the total number of unique thumbnails or title text regions viewed by each participant, i.e., if a participant visited a previously visited AOI again, this visit did not increment the count. The aggregate across all trials is shown in Figure 5 (c). On average, a participant viewed approximately 99 thumbnails over the duration of the experiment (μ = 99, σ = 71.4) compared to 53.6 title texts (σ = 41.9). Participants looked at 1.84 times more thumbnails than title texts, that is, they spent nearly twice as much attention on thumbnail images as compared to title text regions. This was confirmed by a two-tailed paired samples t-test (t(19)=5.441, p<0.01).

5.3 Were thumbnails and title texts always viewed together?

In some cases the participant looks at the thumbnail and title text both, while in other cases, (s)he might look at only the thumbnail or only the title text. On average, only thumbnails were viewed much more often than both thumbnails and title texts together (μ=16.4 compared to μ=8.3). If only one AOI was visited, then thumbnail got many more visits than title text (μ = 16.4 compared to μ = 5).

5.4 Were thumbnails viewed before title texts or vice versa?

We hypothesized that if the thumbnail images were more salient than the corresponding title text, then thumbnails would be looked at first. If a participant attended both the thumbnail and the title text of a video, and if the participant looked at the thumbnail before looking at the title text of that video, then *ThumbnailsFirst* was incremented by one. Figure 4 shows an example where the participant's gaze shifted from the rightmost thumbnail to the leftmost thumbnail, and then moved to down to the title text region. In this case, *ThumbnailsFirst* was incremented by one. Alternately, if the participant had looked the title text before looking at the thumbnail of that video, then *TitleTextFirst* would have been incremented.

Figure 5 (d) compares these two metrics. On average, thumbnails were much more likely to be attended first (μ = 25, σ = 16.8) than title texts (μ = 7.8, σ = 7.4). Though there is a fair amount of variance across participants in the amount of browsing they did prior to making their selection, all participants consistently attended thumbnails before title text regions (Figure 5 (e)). This was confirmed with a two-tailed paired samples t-test (t(19)=6.264,p<0.01).

6 DISCUSSION

Our metrics showed that participants looked at almost twice as many thumbnail images as title text regions (Figure 5 (c)). We also found that participants overwhelmingly looked at thumbnails before title texts (Figure 5 (e)). Thumbnail images are typically 1.68 times the area of the title text region. We can account for this larger area as follows: if the title text region was as large as the thumbnail, it would proportionately get 1.68 times as many fixations, i.e., an average of $53.6 \times 1.68 = 90$. However, thumbnails get an average of 99 fixations, suggesting that even if title text regions were as large as thumbnails, they would still get about 10% less attention.

This finding has implications for user experience designers and content creators as viewers spent approximately 20% of their session time browsing. Based on our analyses, we make the following recommendations to UI/UX designers: (1) users look at more unique thumbnails than title text, so if you want to encourage your users to browse, use pictures; (2) users look at thumbnails before title text, so take your time picking a good picture.

In the context of previous work, even though our task is similar to the "buy a product" task of Rayner et al.[Rayner et al. 2001], there are differences in users' attention allocation on a print ad that is composed around a single product, and the shelf-like arrangement found in video content providers' websites. Print ad pictures are also different from the highly colorful, textured images that video content creators select as thumbnails. In our case, textual information is not a clear winner, and the picture dominates. From the point

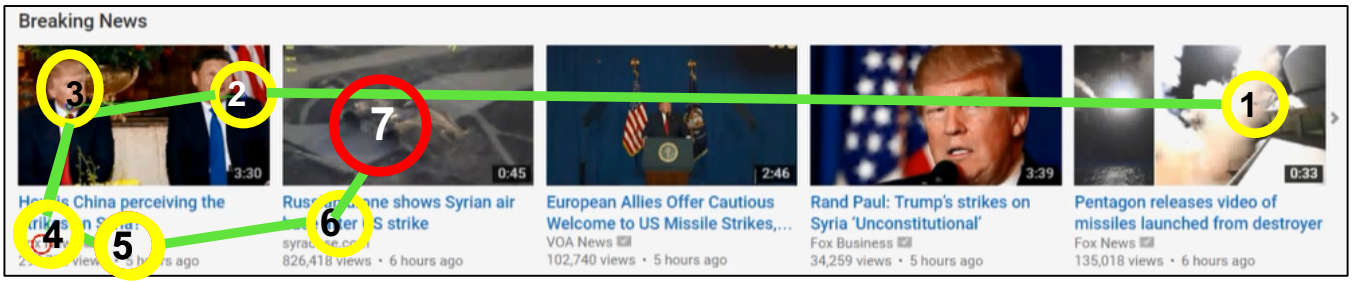

Figure 4: The scanpath for one participant is shown here. The circles represent fixations and the lines represent saccades. Fixations 2,3 are an example of "ThumbnailFirst", while fixations 6,7 are an example of "TitleTextFirst".

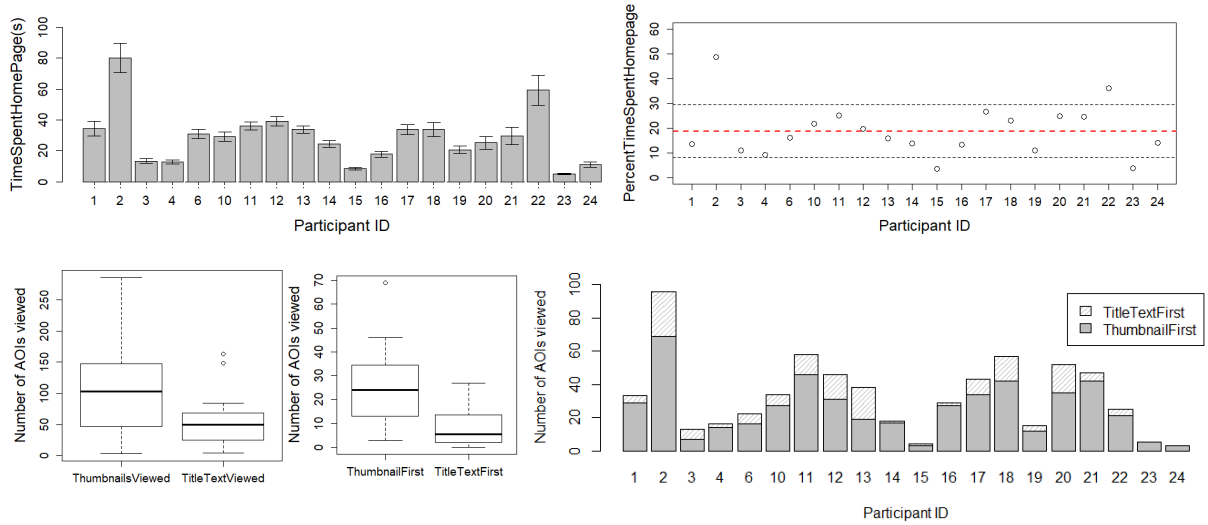

Figure 5: (a) Individual participants are fairly consistent in how they browse across trials. (b) On average, participants spent about 18% of their session time browsing for what to watch. Apart from four participants, all others were within one standard deviation of the mean. (c) (d) On average, thumbnails are viewed both more often than title texts and before title texts. (e) Participants consistently look at almost twice as many thumbnails than title texts.

of view of usability and human-computer interaction [Bergstrom and Schall 2014; Jacob and Karn 2003; Poole and Ball 2005], our experiment and data analysis illustrates the relative importance of the various design elements in the grid type layout that is the de facto standard for online product browsing.

7 FUTURE WORK

Further studies could investigate gaze patterns of users during certain portions of their browsing time, for example, analyzing for revisits and the number of revisits to the selected video's thumbnail. Additionally, this experiment could be replicated on a hand-held device such as a mobile phone, where the thumbnails are typically bigger. It is not possible to explicitly change the layout, font sizes, and other user interface factors within the YouTube website. Future work could create a custom site using interactive prototyping tools such as InVision to test each factor. It would also be interesting to test how task-specific browsing (find a video discussing the gun rights debate in the US) is different from a free browsing task.

In our study, the role of the qualitative questions was to understand the "type of user" we were studying. We found that our participant demographic was a college going student who used YouTube frequently, mostly for entertainment. This information helps with follow on studies that aim to evaluate if the same findings are replicated across different demographics. For example, younger users such as pre-teens might focus more on the pictures. Correlations between culture, ethnicity, and occupation would also be interesting to explore. However, it would also be interesting to ask users qualitative questions such as "Do you think you base your selection on thumbnails or title text information?". This might reveal the extent of user self-awareness or perhaps a surprising mismatch between beliefs and actual behavior. Eye-tracking solutions such as WebGazer [Papoutsaki et al. 2016] could be used to collect this data at a truly large scale.

REFERENCES

Jennifer Romano Bergstrom and Andrew Schall. 2014. *Eye Tracking in User Experience Design* (1st ed.). Morgan Kaufmann Publishers Inc.

Jeremy Goecks and Jude Shavlik. 2000. Learning users' interests by unobtrusively observing their normal behavior. In *International conference on Intelligent User Interfaces (IUI)*. 129–132.

Qi Guo and Eugene Agichtein. 2010. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In *ACM International conference on Research and Development in Information Retrieval (SIGIR)*. 130–137.

Jana Holsanova, Nils Holmberg, and Kenneth Holmqvist. 2009. Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology* 23, 9 (2009), 1215–1226.

Jeff Huang, Ryen White, and Georg Buscher. 2012. User see, user point: gaze and cursor alignment in web search. In *SIGCHI Conference on Human Factors in Computing Systems*. 1341–1350.

RJ 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.

Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *ACM International conference on Research and Development in Information Retrieval (SIGIR)*. 113–122.

Michael Land and Benjamin Tatler. 2009. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press.

Michael F Land and Mary Hayhoe. 2001. In what ways do eye movements contribute to everyday activities? *Vision research* 41, 25 (2001), 3559–3565.

W Howard Levie and Richard Lentz. 1982. Effects of text illustrations: A review of research. *ECTJ* 30, 4 (1982), 195–232.

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 Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016*.

Rik Pieters and Michel Wedel. 2004. Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of Marketing* 68, 2 (2004), 36–50.

Alex Poole and Linden J. Ball. 2005. Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future. (2005).

Keith Rayner, Brett Miller, and Caren M Rotello. 2008. Eye movements when looking at print advertisements: The goal of the viewer matters. *Applied Cognitive Psychology* 22, 5 (2008), 697–707.

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.

Florian Schmidt-Weigand, Alfred Kohnert, and Ulrich Glowalla. 2010. A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and instruction* 20, 2 (2010), 100–110.

Jaana Simola, Jarmo Kuisma, Anssi Öörni, Liisa Uusitalo, and Jukka Hyönä. 2011. The impact of salient advertisements on reading and attention on web pages. *Journal of Experimental Psychology: Applied* 17, 2 (2011), 174.

Adrian Vosskuhler, Volkhard Nordmeier, Lars Kuchinke, and Arthur M Jacobs. 2008. OGAMA (Open Gaze and Mouse Analyzer): open-source software designed to analyze eye and mouse movements in slideshow study designs. *Behavior research methods* 40, 4 (2008), 1150–1162.