Informative or Misleading? Heatmaps Deconstructed

Agnieszka (Aga) Bojko

User Centric, Inc.
2 Trans Am Plaza Dr, Ste 100,
Oakbrook Terrace, IL 60181 USA
abojko@usercentric.com

Abstract. Eye tracking heatmaps have become very popular and easy to create over the last few years. They are very compelling and can be effective in summarizing and communicating data. However, heatmaps are often used incorrectly and for the wrong reasons. In addition, many do not include all the information that is necessary for proper interpretation. This paper describes several types of heatmaps as representations of different aspects of visual attention, and provides guidance on when to use and how to interpret heatmaps. It explains how heatmaps are created and how their appearance can be modified by manipulating different display settings. Guidelines for proper use of heatmaps are also proposed.

Keywords: Heatmaps, attention maps, eye tracking.

1 Introduction

Heatmaps are two-dimensional graphical representations of data where the values of a variable are shown as colors. Heatmaps are compelling for two reasons. First, the intuitive nature of the color scale as it relates to temperature minimizes the amount of learning necessary to understand it. From experience, we know that yellow is warmer than green, orange is warmer than yellow, and red is hot. It is not difficult to then figure out that the amount of heat is proportional to the level of the represented variable. Second, heatmaps show the data directly over the stimulus. Because the data could not be any closer to the elements to which they pertain, little mental effort is required to read a heatmap.

Heatmaps can be of great value to papers, reports, and presentations because they summarize large quantities of data that would be much more difficult to grasp if presented numerically. Heatmaps help us quickly see "the big picture" including any patterns or trends that may exist in the data.

In the user experience field, heatmaps can represent various types of data, such as usage (e.g., clicks, key presses), accuracy, or visual attention. This paper focuses exclusively on attention heatmaps, which have become popular over the past few years due to the increased usage of eye tracking technology.

Attention heatmaps can be easily generated with the help of an eye tracking software application such as Tobii's ClearView, Tobii Studio, SMI's BeGaze, NYAN, or

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2009, LNCS 5610, pp. 30–39, 2009.

EyeTools. Anyone with an eye tracker and the right software can create a heatmap. No knowledge of eye movements or of how heatmaps are created is required. As a result, heatmaps are often generated unnecessarily or are misinterpreted by those who do not understand what the visualizations are really showing or, perhaps even more importantly, not showing. Heatmaps can be deceptive because they look so intuitive that we often do not realize how much we actually do not understand.

This paper describes different heatmap types and their limitations, as well as settings used to manipulate the appearance of heatmaps. It also discusses when heatmaps should and should not be used. Proposed guidelines for using heatmaps correctly conclude this work.

2 Types of Attention Heatmaps

Heatmaps are often shown with little, if any, description of what it is they are representing. The assumption is that they are showing "attention" or "eye movements" but knowing that is certainly not enough to be able to truly understand a heatmap. There are different aspects of eye movements that heatmaps can represent. Examples include fixation count, absolute or relative gaze duration, and percentage/proportion of participants who fixated on each area of the stimulus.

Choosing the right heatmap to present depends on the study objectives and the eye movement measures that address these objectives. For example, if search efficiency was of interest to the researchers, one of the measures collected and analyzed might be the number of fixations prior to acquiring the target [1]. Therefore, assuming that the analysis would benefit from data visualization, a fixation count heatmap should be presented. A fixation count heatmap would also be appropriate if the study goal was to determine the amount of interest generated by various elements of the stimulus during a free-view task (i.e., task with no specific task instructions) [1]. However, if noticeability of a particular element was of interest, the percentage of participants who fixated on the element could be used as a measure (in addition to, for example, time to first fixation), which would warrant a participant percentage heatmap.

Because each heatmap type has different limitations that impact its interpretation, it is not only important to be aware of these limitations, but also to know the types of all heatmaps included in papers, reports, and presentations.

2.1 Fixation Count Heatmap

A visual fixation can be loosely defined as a relatively stationary eye position focused on a particular location of the stimulus (a more precise definition is discussed in section 3.1). Fixations are important events that provide insight into human cognition because during each fixation we extract visual information that we process [2].

A fixation count heatmap (see Fig. 1) shows the accumulated number of fixations across participants. Each fixation made by each participant adds a value to the color map at the location of the fixation [3]. This value is the same for each fixation regardless of its duration, so a 100 ms fixation is represented in the same way as a 900 ms fixation. Thus, when looking at a fixation count heatmap, we cannot assume that areas of the same color received similar total gaze time.

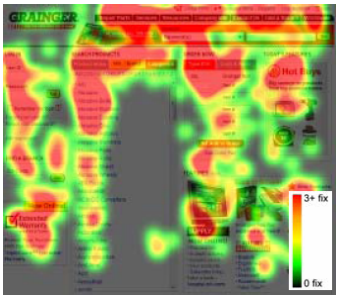

Fig. 1. Fixation count heatmap (n = 13; 0 – 12 s; free-view task)

Fixation count heatmaps can also be biased towards individuals who show high interest in elements that others do not. For example, two elements can be the same color but one attracted ten fixations from one participant only, while the other attracted attention of ten participants, one fixation from each. Therefore, we cannot assume that areas that appear similar in terms of "heat" are equivalent in terms of the number of participants who looked at them.

Another limitation of fixation count heatmaps is the fact that they can be skewed towards individuals who had a longer exposure to the stimulus and thus an opportunity to produce more fixations. For example, if participant A spent twice as much time on the stimulus than participant B, participant A's data would impact the heatmap twice as much as participant B's data. This is something to always keep in mind when viewing heatmaps created based on unequal exposure times.

2.2 Absolute Gaze Duration Heatmap

An absolute gaze duration heatmap (see Fig. 2) shows the accumulated time participants spent looking at the different areas of the stimulus. Each fixation made by each participant adds a value to the color map that is proportional to its duration [3]. For example, a 900 ms fixation will be nine times higher in color value than a 100 ms fixation. Because fixation duration is an indicator of cognitive processing [4], a heatmap that is scaled by fixation duration not only shows which areas were attended to but also represents the level of cognitive processing that the areas required.

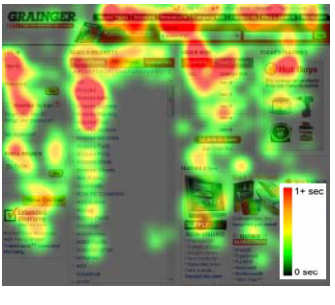

Fig. 2. Absolute gaze duration heatmap (n = 13; 0 – 12 s; free-view task)

An absolute gaze duration heatmap can be misleading because it displays different phenomena in the exact same way. For example, this type of heatmap will make one 900 ms fixation look the same as nine 100 ms fixations. A 900 ms fixation on an element indicates that one person looked at it for a while, while nine 100 ms fixations could mean, for example, that one person made nine brief fixations to the element or nine people made one brief fixation each.

In addition, similar to the fixation count heatmaps, absolute gaze duration heatmaps can be biased towards individuals who spent more time looking at the stimulus. To eliminate any bias due to unequal exposure times, the gaze duration data can be normalized to create relative gaze duration heatmaps.

2.3 Relative Gaze Duration Heatmap

A relative gaze duration heatmap shows the accumulated time each participant spent fixating at the different areas of the stimulus relative to the total time the participant spent looking at the stimulus [3]. In other words, if participant A spent 6 seconds on a web page including 2 seconds on the navigation and participant B spent 60 seconds on the same web page including 20 seconds on the navigation, this type of heatmap will make their data, as it relates to the navigation, the same weight.

Similar to the absolute gaze duration heatmap, this heatmap will also show a high gaze time by an individual (proportional to his/her total viewing time) the same way as several short gaze times by a number of individuals (proportional to their total viewing times). If the exposure time is equal across participants (e.g., all participants saw the page for 12 s), which is the case in the examples presented in this paper, the relative gaze duration heatmap and the absolute duration heatmap will be identical.

2.4 Participant Percentage Heatmap

A participant percentage heatmap (see Fig. 3) shows the percentage of participants who fixated on the different areas of the stimulus. Each participant who looked at any given location adds a value to the color map. This value is the same for each participant regardless of the number of fixations he or she made or the fixation durations. Thus, an area that was briefly fixated once by each of the participants will be presented in the same color as an area that was fixated multiple times by each of the participants and the fixations were much longer.

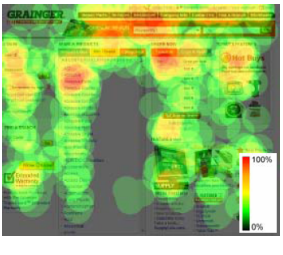

Fig. 3. Participant percentage heatmap (n = 13; 0 – 12 s; free-view task)

3 Display Settings for Creating Heatmaps

The appearance of heatmaps can be modified by manipulating various display settings. While adjusting the settings cannot change the relative distribution of attention, certain areas can be made to appear "hotter" or "colder." This can be done, for example, by changing the fixation criteria used for the analysis, showing raw data instead of fixation data, changing the upper threshold definition of the color scale, or modifying the time segment for the presented data. Since heatmap display settings can have a great impact on the appearance of heatmaps, they must be properly selected and communicated to ensure accurate interpretation.

3.1 Changing Fixation Criteria

There are several algorithms that can be used to define a fixation. A common algorithm used in commercial eye tracking software is based on duration and dispersion threshold identification. To define a fixation using this algorithm, two parameters need to be specified: minimum fixation duration (e.g., 80 ms) and maximum dispersion threshold (e.g., 0.5 degree of visual angle) [5]. A fixation defined by 80 ms and 0.5° will encompass all consecutive eye movements that occurred within 0.5° from each other for at least 80 ms. Unless noted otherwise, the heatmaps in this paper were created using these settings.

Manipulating the duration and dispersion thresholds will change the number of fixations in the data. For example, increasing the minimum fixation duration from 80 ms to 200 ms will decrease the number of fixations because it will exclude all the fixations between 80 ms and 200 ms. Increasing the maximum dispersion threshold from 0.5 degree to 1 degrees of visual angle will also decrease the number of fixations because some of the fixations that are closer together will be combined. Conversely, reducing the minimum fixation duration and maximum dispersion threshold will increase the number of fixations. More fixations in the data will increase the amount of "heat" in the heatmap, as shown in Figure 4.

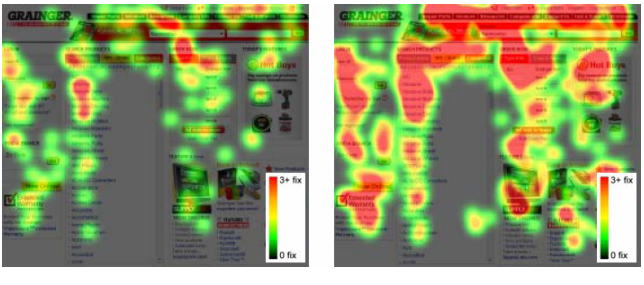

Fig. 4. Fixation count heatmaps based on the same data (n = 13; 0 – 12 s; free-view task). On the left: minimum fixation duration = 200 ms; on the right: minimum fixation duration = 80 ms.

The lack of explicit fixation definition is an issue that does not pertain specifically to heatmaps but to entire papers and reports. Many user experience studies that analyze fixation data never mention how these fixations were defined. However, this information is very important for two reasons. First, the results of different studies are

not comparable unless it is clear that fixations were defined in the same way. Second, if the definition is not provided, it cannot be verified whether or not the fixation criteria were appropriate for the stimuli used in the study. For example, fixation duration threshold for image viewing should be higher than for reading because image viewing tends to produce longer fixations than reading due to the fact that more information is being processed in a single fixation [6].

The fixation definition used to create heatmaps should match the definition used for data analysis. Changing the fixation duration to obtain a visualization of a particular intensity is not a good practice.

3.2 Displaying Raw Data Instead of Fixation Data

One step further from decreasing the fixation duration and dispersion thresholds is presenting raw data instead of fixation data. Raw data consists of meaningful eye movements (raw fixation points) and "noise" – eye movements that have little meaning in most user experience research. The noise includes eye movements that take place during saccades (rapid eye movements between fixations) as well as drifts, tremors, and flicks that occur during fixations [5]. Adding all the noise intensifies the heatmap, increasing the area covered in red (see Fig. 5).

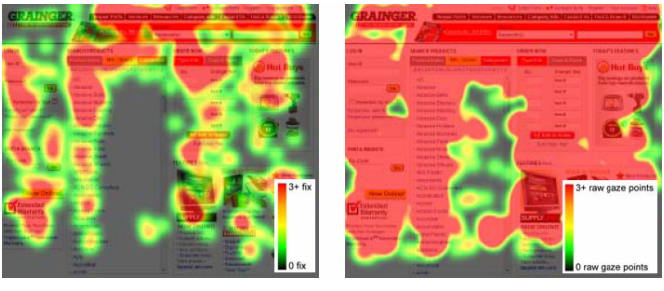

Fig. 5. Fixation count heatmaps based on the same data (n = 13; 0 – 12 s; free-view task). On the left: fixation data; on the right: raw data.

As a general rule, fixation data rather than raw data should be used when creating visualizations unless the stimulus has moving elements and the heatmap has to show smooth pursuit eye movements. We can assume that fixation data was used if the paper or report specifies how the researchers defined a fixation.

3.3 Changing the Definition of the Color Scale Upper Threshold

Another way to manipulate the amount of heat in heatmaps is by changing the definition of the upper threshold on the scale, which is usually indicated by the red color. If the requirements for an area to be red are lowered, the amount of red in the heatmap will increase. Lowering of the upper threshold of the scale can be achieved by decreasing the minimum number of fixations in fixation count heatmaps (see Fig. 6) or by decreasing the minimum gaze length in absolute gaze duration heatmaps.

There is no set process for choosing the right upper threshold. The rule of thumb is to make sure that the heatmap properly captures the range of values that are of interest to the study. Threshold selection can be compared to setting the maximum value on the Y axis in a graph, where the Y axis indicates values of the independent variable. If the Y axis is too short, the data points that exceed the maximum on the axis are cut off. As a result, we only know that these data points are higher than the maximum but we do not know what they are exactly and how they differ from one another. On the other hand, if the Y axis is too high, the graph data will appear compressed and the differences between the data points will look smaller. Similarly, if a heatmap's upper threshold is set too low, many areas will be covered in red with no differentiation between the amount of attention each attracted. Conversely, setting a heatmap's upper threshold too high will limit the range of colors (e.g., constricting it to yellow or orange as the maximum) and no areas will be covered in red.

Regardless of what criteria have been selected, they should be explicitly communicated in the figure legend or caption (e.g., "red = 10+ fixations" or "red color indicates areas that accumulated 10 s or more of gaze time"). It is also useful to put this value in context by providing the average number of fixations each participant made on the stimulus or the average time each participant spent looking at it. The heatmaps presented in this paper were created based on data with the average of 42 fixations on the page per participant and consistent exposure time of 12 seconds.

If heatmaps are generated for different experimental conditions or participant groups, their upper threshold definition should be identical, so the heatmaps can be compared. If the display settings are not the same, the differences between the heatmaps may be due to factors other than the data itself.

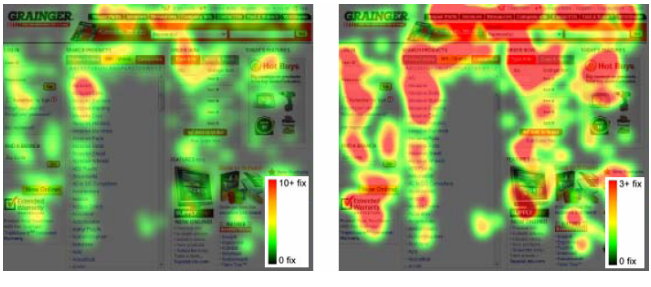

Fig. 6. Fixation count heatmaps based on the same data (n = 13; 0 – 12s; free-view task). On the left: red = 10+ fixations; on the right: red = 3+ fixations.

3.4 Modifying the Time Segment

Sometimes it may be appropriate to present data based on a shorter time segment than the total time during which a stimulus was shown to the participants. This will obviously decrease the amount of data, thus reducing the size of red areas in all heatmap types mentioned in this paper except for the relative gaze duration heatmap (see Fig. 7). Therefore, if a heatmap presents data from a time segment that is shorter than the total viewing time, this needs to be specified in the figure legend or caption.

In addition to the time segment, it should also be clear what participants were trying to do when the data presented in the heatmap was being collected. All too often heatmaps are shown without any context of the task. Eye movements are very task-dependent [7] and participants trying to log in to a website will produce very different attention distribution than if they were trying to find a product. Even if there was no specific task, it should still be noted that the data was collected in a free-view situation, which is how the heatmaps included in this paper were obtained.

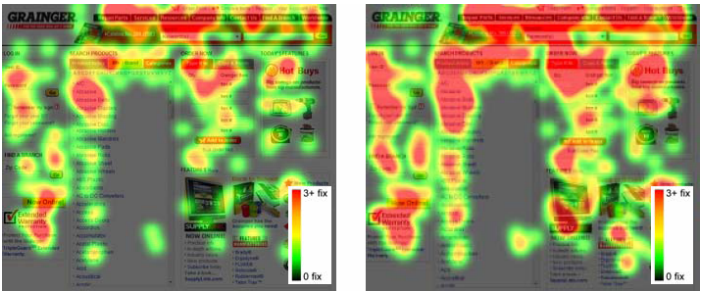

Fig. 7. Fixation count heatmaps of the same free-view task (n = 13). On the left: data from 0 to 6 seconds; on the right: data from 0 to 12 seconds.

4 Heatmap Usage

4.1 Common Mistakes

There are a few common mistakes when it comes to using heatmaps. The biggest of them is a belief that heatmaps are appropriate for just about any user experience study and for any research question. Sometimes creating heatmaps even becomes the objective of the study (e.g., "we just wanted to see where people look").

This popular "let's-track-and-see-what-happens" approach has a limited value because of its lack of focus (i.e., the study design does not target specific questions) and frequent lack of proper data analysis. Researchers often draw conclusions or make recommendations based on the results of those studies, which is inappropriate. For example, we cannot say that the reason why an element did not get much attention was because of its sub-optimal placement or insufficient size unless we have tested other conditions (i.e., alternative placements or sizes) and compared the data using appropriate statistics.

Even if different conditions were tested, sometimes conclusions regarding differences between conditions are made just by looking at the heatmaps. However, heatmaps do not lend themselves to any systematic comparison. Without any data analysis, it is impossible to tell if there are real differences between heatmaps, even if the heatmaps appear to be different.

4.2 Proper Usage Examples

Heatmaps should be used in a purposeful way and only when they add value. They can serve as illustrations of participants' viewing behavior and distribution of attention.

While they can communicate data, they cannot explain it or help analyze it. Therefore, heatmaps can rarely stand on their own. To maximize their usefulness and reduce ambiguity, heatmaps should accompany a quantitative analysis.

One of our studies investigated a new standardized label template for prescription drug labels [8]. The goal was to determine the impact of the template on pharmacists' drug selection speed and accuracy as compared to the existing label designs. The eye tracking measures included the number of fixations prior to target selection as an indicator of search efficiency, average fixation duration as a measure of information processing difficulty, and pupil diameter as a measure of cognitive workload. The results were presented in the form of statistical analyses but no heatmaps were included in the report. The study was a quantitative assessment of the effectiveness of the new labels, and heatmaps showing attention distribution were simply of no value.

In another study, we evaluated a new homepage design for a professional organization against the original homepage [9]. Our objective was to identify which design was better and why based on a series of tasks during which participants attempted to locate the correct entry point on the homepage. Measures of search efficiency such as the number of fixations and the number of eye visits to the target prior to target selection were analyzed. The analysis was supplemented with heatmaps to show the distribution of attention on the page and to help account for any inefficiencies that occurred. For example, several tasks were more efficient using the new design which had a more centralized navigation. The heatmaps of the original design showed scattered fixations covering multiple navigation areas, while heatmaps of the new design revealed fixations focused mostly around the targets.

5 Guidelines for Using Heatmaps

Even though heatmaps are very compelling and seemingly easy to understand, they should be used with caution and according to the following guidelines, summarized based on the discussion from the previous sections of this paper:

A. Generate heatmaps only if they add value to the research.
B. Use heatmaps for data visualization instead of data analysis.
C. Use heatmaps to support quantitative analysis rather than on their own.
D. Understand the different heatmap types and only use the ones that represent measures which address your study objectives (e.g., when analyzing gaze time, use a gaze duration heatmap).
E. Specify the type of data the heatmap is representing (e.g., fixation count or absolute fixation duration).
F. Know the limitations of each heatmap type to avoid incorrect interpretation.
G. When creating heatmaps, use fixation data rather than raw data.
H. Provide fixation definition and keep it consistent for analyses and visualizations within a study (e.g., min fixation duration = 100 ms and max dispersion threshold = 0.5°).
I. Provide the definition for the upper threshold of the heatmap color scale (e.g., red = 10+ fixations).
J. Put the upper threshold value in context (e.g., average number of fixations on the stimulus per participant).

K. Specify the time segment based on which the heatmap was created (e.g., the first 10 seconds of exposure).
L. Provide task context for each heatmap (e.g., data obtained from participants during the checkout task).
M. Use the same heatmap settings (e.g., upper threshold and time segment) for conditions that you are comparing.
N. If a paper or report does not provide important information about its heatmaps (e.g., type, fixation definition, upper threshold definition, time segment), ask the authors for clarification before making any assumptions.

6 Conclusion

Blinded by the attractiveness and apparent intuitiveness of heatmaps, we often do not realize how much information in addition to the visualization itself is necessary to fully understand a heatmap and properly interpret the data it represents. In other words, the biggest danger involved in creating and reading heatmaps is that we are often unaware of what we do not know, and thus we do not look or ask for the missing information. This paper has exposed some of these gaps in our meta-knowledge in the hope to encourage more critical thinking about the usage of heatmaps.

References

1. Jacob, R.J.K., Karn, K.S.: Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In: Hyona, J., Radach, R., Deubel, H. (eds.) The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research, pp. 573–605. Elsevier Science, Amsterdam (2003)
2. Liversedge, S.P., Findlay, J.M.: Saccadic Eye Movements and Cognition. Trends in Cognitive Sciences 4, 6–14 (2000)
3. Tobii.: Tobii Studio 1.X User Manual (2008)
4. Duchowski, A.: Eye Tracking Methodology: Theory and Practice. Springer, Heidelberg (2003)
5. Salvucci, D.D., Goldberg, J.H.: Identifying Fixations and Saccades in Eye-Tracking Protocols. In: Proceedings of Eye Tracking Research and Applications Symposium (2000)
6. Castelhano, M.S., Rayner, K.: Eye Movements During Reading, Visual Search, Scene Perception: An Overview. In: Rayner, K., Shem, D., Bai, X., Yan, G. (eds.) Cognitive and Cultural Influences on Eye Movements, pp. 3–33. Psychology Press (2008)
7. Yarbus, A.L.: Eye Movements and Vision. Plenum Press (1967)
8. Bojko, A.: Measuring the Effects of Drug Label Design and Similarity on Pharmacists' Performance. In: Tullis, T., Albert, W. (eds.) Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics, pp. 271–280. Morgan Kaufmann, San Francisco (2008)
9. Bojko, A.: Using Eye Tracking to Compare Web Page Designs: A Case Study. Journal of Usability Studies 1, 112–120 (2006)