Determining Optimal Caption Placement Using Eye Tracking

Andrew D. Ouzts
School of Computing
Clemson University
aouzts@cs.clemson.edu

Nicole E. Snell
Information Design and Corporate Communication
Bentley University
nsnell@bentley.edu

Prabudh Maini
School of Computing
Clemson University
pmaini@clemson.edu

Andrew T. Duchowski
School of Computing
Clemson University
andrewd@cs.clemson.edu

ABSTRACT

In this study, the effect of caption placement on information intake is examined. Eye movement data is used to quantitatively analyze the effect of four different captioning methods. Information intake (i.e. Information Assimilation (IA)) is measured via a 4-category comprehension quiz, developed by S.R Gulliver and G. Ghinea, which measures key differing aspects of captioned videos. Results indicate that caption placement can have significant effects on reading time, number of saccadic crossovers, and ratio of fixations on captions.

Categories and Subject Descriptors

D.3.2. C++ [**Programming Language**]. H2.3. SDL [**Simple DirectMedia Layer**]: *a cross-platform multimedia library*. D.4.0 Linux [**Computer Operating System**].C.2.2 TCP/IP [**Transmission Control Protocol/Internet Protocol**].

Keywords:

Closed Captioning, Information Assimilation, Eyetracking

1. INTRODUCTION

Closed captioning is a technology, as defined by US law, which ensures the civil right of an individual to have equal access to emergency information, national and local news, and public entertainment regardless of their ability to hear. In order to ensure that users of captions gain a comparable amount of information to that of viewers who have access to both the audio and video action, it is necessary to determine the most efficient and effective presentation of captions. To do so, one must determine how variables such as caption placement affect the viewer's comprehension of captioned media. We therefore investigate the impact that caption placement has on the amount of information assimilated by viewers. We hypothesize that depending on the caption placement method used, a viewer's information assimilation (IA) is either enhanced or diminished.[1]

2. METHODOLOGY

To quantitatively assess the validity of our hypothesis and each of our captioning methods, we performed a validation study. We sought to determine the captioning method(s) that minimize reading time and maximize information intake. In order to quantitatively measure these factors, we designed an eye tracking study in which four different captioning methods were examined using eye movement data, in addition to the qualitative measurements of IA.

2.1 Apparatus

Eye movements were captured by a Tobii ET-1750 eye tracker, a 17 inch (1280 x 1024) flat panel with built-in eye tracking optics. The eye tracker is binocular, sampling at 50 Hz with 0.5° accuracy. A PC with dual AMD Opteron 64 processors running Microsoft Windows XP and software provided by Tobii streams eye gaze data via TCP/IP. The display and data collection program was run on a PC with an AMD Opteron processor running Fedora Core Linux. The Linux client used TCP/IP to collect the data from the Windows server.

2.2Software

The (previously created) display and data collection application was developed with C++, OpenGL, SDL (Simple DirectMedia Layer, a cross-platform multimedia library). The ffmpeg library was used to render video so that the data collection application also served as the video player. Andrew Duchowski's Linux Tobii library[2] was used to interact with the eye tracker.

This program was further extended with the capacity to interpret and display SRT (SubRip Text - a popular file format used for storing subtitle timings and text). An additional parameter was added to the SRT files specifying the placement strategy to be used by each individual caption. Captions were displayed on an 80% transparent black box whose length is adapted to the length of the caption, using sans-serif font.

Using this program, each viewing by each participant results in a text file containing (x, y, t) data for each raw gazepoint recorded by the eye tracker. A separate text file records video events such as pauses and timestamps for each frame, and both text files are used to match gazepoints to frame numbers.

2.3 Stimulus

The video used as stimulus was a 2m34s long clip BBC News clip that aired February 16 2009 about the replacement of checks and cash with electronic payment. The clip was subtitled verbatim. The video was played in full screen on a 1280 x 1024 resolution monitor and encoded using the MPEG-1 standard, which provides VHS-quality compression.

2.4 Participants

20 participants (7 male, 13 female) took part in the study. Participants were volunteers recruited from Clemson University's Human Participation in Research student pool. Participants ages

SIGDOC'13, September 30–October 1, 2013, Greenville, North Carolina, USA.
ACM 978-1-4503-2131-0/13/09.
TTP://DX.DOI.ORG/10.1145/2507065.2507100

[1] Available at http://andrewd.ces.clemson.edu/tobii/

ranged from 17 to 51 (mean 20.85, median 19). Four participants were excluded from the final analysis due to data collection issues, bringing the total to 16 participants (6 male, 10 female).

2.5 Experimental Design

The experimental design took the form of a single factor (captioning method) with four levels. The four captioning methods examined were:

- Method A: All subtitles displayed 64 pixels from bottom
- Method B: All subtitles displayed 64 pixels from top
- Method C: Constant alternation between 64 pixels from top and 64 pixels from bottom. That is, if the n^{th} frame is displayed at the top, then the $(n+1)^{th}$ frame will be displayed at the bottom, and vice versa.
- Method D: A method in which captions were placed above a speaker to identify him/her as the source of the closed captions, and at the bottom when no speaker was visible or the visual speaker was not the source of the captions. This is similar in many respects to the methods used in [2].

All participants viewed the video using each method. Order of presentation was counterbalanced using a 4 x 4 Latin square so that approximately a quarter of all subjects saw the videos in order {A, B, C, D}, {B, C, D, A}, {C, D, A, B}, or {D, A, B, C}.

As an extension of our general hypothesis that depending on the caption placement method used, a viewer's IA is either enhanced or diminished, we in particular hypothesized the following:

1. Method A will minimize reading time. Viewers typically expect to see the captions at the bottom, and due to the unchanging nature of this method, no time is needed to search for the location of the captions.
2. Method D will maximize information intake. Due to the additional information encoded in the location of the captions, viewers should have increased scene comprehension.

2.6 Procedure

Participants were seated in front of the eye tracker at a distance of about 60 cm. The eye tracker was calibrated to the participant before each video clip was shown; that is, four times total per person. Calibration required a participant to track a moving circle with their eyes to 9 points.

Following calibration, participants were instructed to watch the video clip normally as they would at home - that is, an unguided viewing. They were informed that they would be given a quiz after playback to gauge their comprehension, but given no other specific information.

After the first video clip, participants were given a quiz based on (IA) information assimilation categories.

After the participant had viewed all 4 captioning methods, they were asked to answer some final preference-based and qualitative questions about the study. We also asked if the participant had been aware of the different captioning.

3. RESULTS

Within subjects analysis was performed for each of the three eye tracking metrics. A between subjects analysis was performed for quiz data.

A significant difference was found in the number of saccadic crossovers between each captioning method using a within-subjects ANOVA ($F(3, 45) = 4.43$, $p < 0.01$). Pairwise t-tests (no correction) showed a marginally significant difference between captioning methods C and A ($p < 0.05$).

Analysis of the reading time data also shows significance. A significant difference was found in the amount of reading time spent ($F(3,45) = 19.11$, $p < 0.01$). Pairwise t-tests (no correction) show significance between methods A and C ($p < 0.01$), B and C ($p < 0.01$), C and D ($p < 0.01$), and marginal significance between A and D ($p < 0.05$).

Differences were also found for the fixation ratio metric. Within-subjects ANOVA shows a significant difference ($F(3, 4) = 22.36$, $p < 0.01$) in the mean ratio of fixations, and pairwise t-tests (no correction) reveal significance between methods A and C ($p < 0.01$), B and C ($p < 0.01$), and D and C ($p < 0.01$).

Between subjects analysis was performed on the quiz data (based on the video the participant first saw). No significant differences were found for any of the quiz results or preferences, and the mean correct answers for methods A, B, C, and D were 52.5%, 50%, 55%, and 47%, respectively.

4. CONCLUSION

An eye tracking study was presented in which several different captioning styles were examined. Significant differences were found between eye movement metrics depending on the captioning style used, suggesting that captioning styles play an important role in viewing strategies. Participants underwent large amounts of saccadic crossovers and spent much less time reading the captions when captions changed position frequently. Future work is needed to fully examine the implications of these differences.

5. REFERENCES

1. Gulliver, S. R., and Ghinea, G. How level and type of deafness affect user perception of multimedia videoclips.

   Universal Access Information Society, 2 (2003), 374-386.
2. Vy, Q.V. and Fels, D.I. Using placement and name for speaker identification. ICCHP 2010 Conference Proceedings.