A View on the Viewer: Gaze-Adaptive Captions for Videos

Kuno Kurzhals
ETH Zurich
Zurich, Switzerland
kunok@ethz.ch

Fabian Göbel
ETH Zurich
Zurich, Switzerland
goebelf@ethz.ch

Katrin Angerbauer
University of Stuttgart
Stuttgart, Germany
katrin.angerbauer@visus.uni-stuttgart.de

Michael Sedlmair
University of Stuttgart
Stuttgart, Germany
michael.sedlmair@visus.uni-stuttgart.de

Martin Raubal
ETH Zurich
Zurich, Switzerland
mraubal@ethz.ch

ABSTRACT

Subtitles play a crucial role in cross-lingual distribution of multimedia content and help communicate information where auditory content is not feasible (loud environments, hearing impairments, unknown languages). Established methods utilize text at the bottom of the screen, which may distract from the video. Alternative techniques place captions closer to related content (e.g., faces) but are not applicable to arbitrary videos such as documentations. Hence, we propose to leverage live gaze as indirect input method to adapt captions to individual viewing behavior. We implemented two gaze-adaptive methods and compared them in a user study (n=54) to traditional captions and audio-only videos. The results show that viewers with less experience with captions prefer our gaze-adaptive methods as they assist them in reading. Furthermore, gaze distributions resulting from our methods are closer to natural viewing behavior compared to the traditional approach. Based on these results, we provide design implications for gaze-adaptive captions.

Author Keywords

Eye tracking; gaze input; gaze-responsive display; multimedia; video captions

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI);** User studies;

INTRODUCTION

Multimedia content such as feature films or online videos often apply captions as a visual aid to help people with hearing impairments understand the content, or as an affordable means to display translated content without dubbing. However, displaying text in a video is a severe intervention that influences the viewing behavior of the audience [7, 24].

Figure 1: Setup on a regular desktop PC with a low-cost eye tracker. The viewer watches a video while gaze input is used to adapt the position of the captions. The gaze position is included for illustration purposes and is not visible in the application.[1]

Traditional subtitles appear at the bottom of the screen, often distant to the current content of interest, and require viewers to focus on this area for reading, before moving their eyes back on the relevant visual content. This might result in missing important details. To compensate for this issue, different alternative layout techniques for captions have been proposed in recent years. For example, Hu et al. [17] developed an algorithm that places captions close to the respective speaker in dialogues, similar to speech bubbles in comic books. User-based evaluation showed that such a representation results in a gaze distribution in favor of the relevant image content [24] and closer to natural viewing behavior [7]. Although the alternative techniques were positively perceived by participants [16, 17], such approaches face one issue that traditional captions do not have: The position of new captions is dynamic, so the viewer cannot predict where new text will appear. This is a drawback in cases where the viewer expects to continue reading (e.g., split captions for one sentence) but has to search for the new caption first [24]. Simultaneously, eye tracking has become more ubiquitous and convenient to use as an input

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.3376266

[1] Video source: "Why New Zealand needs predator control" CC BY the New Zealand Department of Conservation

device in desktop and mobile scenarios. As a consequence, interactive captions that adjust to a person's viewing behavior could respect individual differences between people, reduce search effort, and provide textual information close to the current focus of attention.

In this work, we present *gaze-adaptive captions* that apply different placement strategies in response to the viewer's gaze behavior (Figure 1). We introduce two approaches: Direct Captions (**DC**) that appear close to the viewer's last gaze position on the screen and Salience-Sensitive Captions (**SSC**), which, in addition, avoid occlusions of potentially interesting image content. Both approaches are independent of manually annotated areas of interest (AOIs) and can be applied to arbitrary videos, an advantage over approaches that rely on speaker detection. Furthermore, we evaluate how these two methods affect the viewing experience and their influence on gaze distributions. Our contributions can be summarized as follows: (1) We introduce two gaze-adaptive techniques to display captions on videos of arbitrary content. The implemented prototype adjusts caption positions with respect to the current gaze position and salient regions in the video. (2) We conducted a user study (n=54) to evaluate how the gaze-adaptive methods influence user experience and gaze distributions on the video. The new techniques are compared with traditional subtitles and a baseline without captions. (3) From our results, we derive a set of implications for future design and application of gaze-adaptive captions.

RELATED WORK

Over the years, numerous guidelines[2] and studies [6, 11, 12] have been published, which describe a large design space for captions. Among other aspects, these studies consider font size [42], text and content editing [3, 28, 38], animations [34], shot changes [22], and text segmentation [13, 32]. While eye tracking was mainly applied as a tool to evaluate the influence of captions on viewing behavior and user experience [9, 23], it also allows for gaze-responsive design [10]. As eye gaze is a good indicator for visual attention [20] and the current point of regard [26], it has been used not only as an explicit input modality [18, 43], but also as an implicit and subtle mode of interaction in attentive user interfaces [40, 41]. We apply eye tracking to adapt the position of captions according to the live measured point of regard.

In recent years, an increasing number of alternative techniques for captions were presented. They can be separated in speaker-based and salience-based captions.

Hong et al. [16], Hu et al. [17], and Tapu et al. [39] presented speaker-based techniques that mainly focus on the detection of faces, especially speaking persons, to place new text close to the speaker. Brown et al. [7] and Kurzhals et al. [24] evaluated the user experience and influence on gaze behavior with such alternative methods. Both studies showed that the alternatives were rated positively in terms of user experience and viewers could better focus on the video content. The main shortcoming of these methods is that they are restricted to specific dialog scenes where a speaking person is visible. Different scenes with action, documentations, or generally with an off-screen speaker are typically not supported.

Alternatively, approaches exist that calculate visually salient regions in the video and perform caption placement based on this information. This can be realized by generating salience maps offline from recorded gaze data [1, 2] or with live gaze data [21]. Brooks and Armstrong [4] and Kurzhals et al. [24] discussed the idea of dynamic subtitle placement with gaze and feature avoidance. Jiang et al. [19] propose a method based on rough gaze estimation and salience detection. They divided the screen into four regions and estimated where the viewer was looking before placing a caption. However, besides the low precision, the authors did not provide an evaluation of how an interactive approach influences the viewing experience.

We expand on these ideas by deploying eye tracking hardware for precise gaze estimation to implement two new placement techniques and evaluate them with quantitative and qualitative measures. With gaze-responsive design, we can adapt the viewing experience to individual differences between participants, which is not possible with pre-calculated caption layouts. Our first technique is solely based on live gaze data. For the second one, we calculate a displacement map for salient regions of the video that is deployed to adjust the position of a caption relative to the gaze position. The displacement map is based on a pre-recorded baseline condition of participants solving a specific task. This way, we can consider task-driven, top-down salience, which is still an ongoing research topic [33].

GAZE-ADAPTIVE PLACEMENT

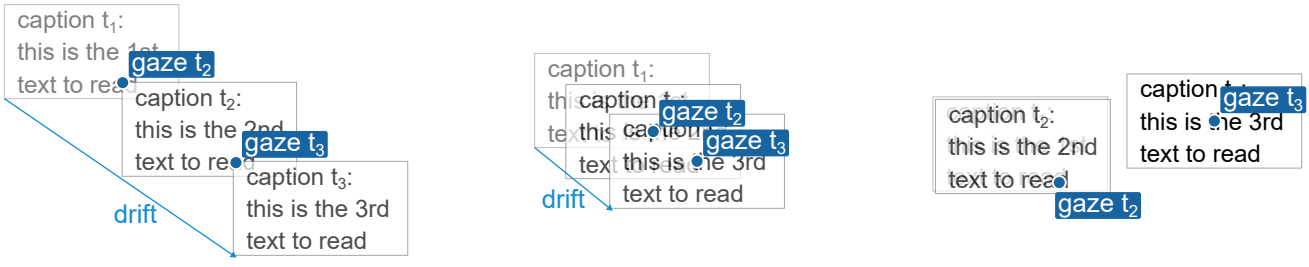
We present two techniques, i.e., direct captions (**DC**) and salience-sensitive captions (**SSC**), which focus on a text placement primarily guided by the viewer's gaze. The first technique ignores visual content and focuses on consistent placement during reading. The second technique considers overlaps of text with important regions of the stimulus and optimizes placement close to the gaze position. The presented techniques adapt to the viewer's implicit gaze input. Consequently, a live stream of eye tracking data is necessary to proceed with the algorithms. In cases where the eyes are not detected (e.g., due to blinks), the next caption will appear at the previous position. We tested our techniques on high-end eye tracking hardware, as well as on a low-cost eye tracker[3], designed for gaze-based interactions in gaming (Figure 1). In both cases, the techniques worked reliably, extending their applicability to an audience outside of a lab environment.

Direct Captions (DC)

Direct placement of text at the viewer's current gaze position potentially reduces the search effort for new captions. For this approach, the alignment of the bounding box of the caption plays an important role. One option is to align a corner point of the bounding box with the respective gaze position. Depending on the reading direction of the applied language, one could always show new captions with the top-left or top-right corner aligned to the gaze. This would enable the viewer to start reading immediately, without additional visual search (Figure 2a). However, as early experiments and related work

[2] https://bbc.github.io/subtitle-guidelines, visited: Dec. 16, 2019

[3] Tobii EyeX – https://gaming.tobii.com, visited: Dec. 16, 2019



(a) top left **(b) centered** **(c) location consistent**

Figure 2: (a) New captions that align the top-left corner of text boxes with the gaze position can cause a position drift while reading. (b) This effect can be reduced by positioning new captions centered on the gaze position. (c) A location-consistent layout where new caption positions are only set when the gaze leaves the area of the previous caption prevents the drift.

on gaze-adaptive map legends [14] showed, new captions that appear while the viewer is still reading the previous text will cause a drift that shifts new captions with each step closer to the screen borders. The drift can be reduced by placing new captions centered around the gaze point (Figure 2b). However, according to previous studies [1, 2] and also noticed during pilot testing, we identified that participants preferred a consistent position of the captions while reading. As a consequence, we further tested a location-consistent approach (Figure 2c) where the center point of the bounding boxes is kept for new placements in case the viewer is still reading. In cases where the viewer is focusing on other areas, this issue does not occur and the captions can be placed at the new gaze position.

We identify caption reading based on the geometrical properties of the bounding box of a caption. This means, if the viewer's gaze enters the bounding box, the reading process is assumed to be initiated. If the gaze leaves the bounding box, we assume the reading process to be finished. We base our hit detection for the bounding box on a circle with an approximate radius of 2.5° (100 px at a distance of 65 cm) at the gaze position to emulate the foveal area and compensate for possible imprecision of the hardware.

Salience-Sensitive Captions (SSC)

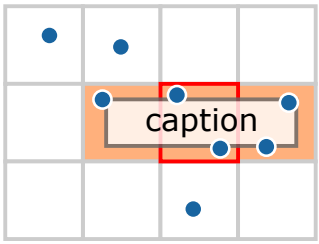
The **DC** approach is unaware of image content that could possibly be occluded. The main advantage of context-sensitive captions (e.g., Hu et al. [17]) is that text is placed in areas where irrelevant content is visible. However, applying such approaches to video content requires often tedious, manual annotation effort. Alternatively, automatic approaches are applied, which are often very specific (e.g., face detection) and hard to apply for arbitrary videos where other objects or areas might be important. Over the last years, a multitude of different salience models has been introduced, focusing on bottom-up and top-down salience in static pictures and dynamic scenes. Among other aspects, these models are based on gaze data collected from human viewers. We base our salience-sensitive approach on such gaze data, which provide an overview of common gaze distributions that indicate where potentially interesting areas are located in the stimulus. From these gaze distributions, we create displacement maps to reduce the overlap with important regions. Our approach considers gaze information from previously recorded persons and combines it with the viewer's gaze input. Before a new caption is displayed, the current gaze position is used live to look up if and where the caption should be shifted to minimize the overlap with important areas.

Data: gaze distribution grid DIST, caption CAP
Result: displacement grid DISP

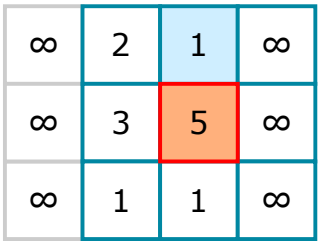
initialize coverage grid COV;
foreach *cell in DIST* **do**
  take cell as center for CAP;
  if *CAP is out of bounds* **then**
    VAL = ∞;
  else
    VAL = # gaze points in cells covered by CAP;
  end
  insert VAL in COV;
end
foreach *cell in COV* **do**
  MIN = cell with lowest value in Moore neighborhood;
  VEC = displacement vector from cell to MIN;
  insert VEC in DISP;
end
return DISP;

Algorithm 1: Calculating the displacement for one caption based on the gaze distribution for the respective time span.

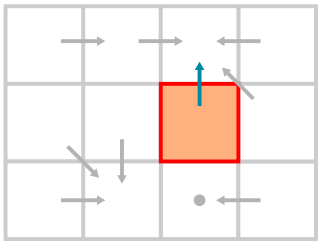
Algorithm 1 describes the procedure to calculate the displacement map for one caption element. The algorithm splits the stimulus into a grid that is processed cell-wise. Figure 3 illustrates the algorithm for one cell and displays how the displacement grid is derived. The gaze distribution (DIST) considers all gaze points within the life span of a caption to minimize current and future occlusions while the text is displayed. For each cell in the coverage grid (COV), we investigate the Moore neighborhood and set a displacement vector to the cell with the fewest gaze points. In case of equal values, the minimal Euclidean distance and the visiting order in the grid determine the displacement. This approach can be seen as the first step of a gradient descent. Generally, a smaller grid size will require more steps for the descent with an increasing degree of uncertainty for the viewer where the next caption will appear. In our study, we applied a 15×15 grid with one step because pilot testing showed that with larger cells, the caption position became unpredictable and with smaller cells, overlaps where often not sufficiently reduced. The displacement map (DISP)



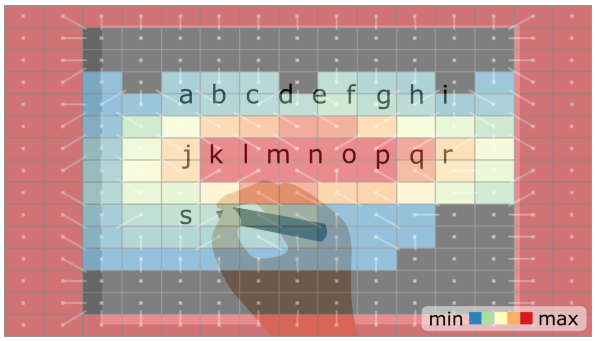
(a) gaze distribution (DIST)



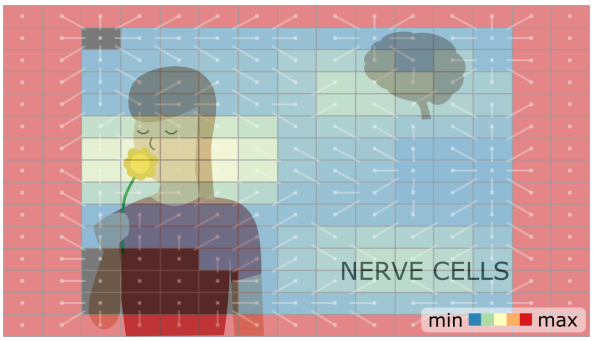
(b) coverage (COV)



(c) displacement (DISP)



(d) displacement for stimulus 1



(e) displacement for stimulus 2

Figure 3: Illustration of Algorithm 1 on how to calculate the displacement map for a caption. (a) Each cell in the gaze distribution grid is considered as a potential center for the caption. (b) The gaze coverage of the caption at this position is calculated. (c) The displacement of a cell is finally determined by the neighbor with the lowest coverage. (d, e) The resulting displacement maps reduce the overlap of captions with important areas. The stimuli in the background show illustrations of the respective videos.

is pre-calculated and deployed as look-up table during live processing of the data. Please note that we apply this approach for two main reasons:

(1) This paper does not focus on the development or comparison of generalizable salience models. Hence, we recorded gaze data for each stimulus under natural viewing conditions in a pre-study. From these data, we compute a viewer-based salience map, tailored to the investigated stimuli and task. (2) The recorded pre-study data serves as baseline for subsequent comparisons of viewing behavior with the presented approaches. However, the applied saliency map can be replaced by alternative approaches, potentially by future generalizable models that are applicable to arbitrary video content.

BASELINE RECORDING

In a pre-study experiment, we recorded 8 participants (2 female, 5 male, 1 other) watching 10 video stimuli under natural viewing conditions. The average age was 30.13 years (sd=4.68, range: 25–40) and all were good or fluent English speakers. No captions were presented during playback, only English audio. The participants' task was to pay attention to answer four questions after each video.

Stimuli

We chose ten videos from TED-Ed[4], listed in Table 1. All videos briefly explain different education topics, e.g., how cheese was discovered. In all videos, an off-camera speaker explains the content while animations are used for illustration. These videos were selected because they provide multiple regions of interest where speaker-based approaches cannot be applied. We removed the introduction and shortened the endings in order to keep playback times similar.

[4]https://ed.ted.com, visited: Dec 16, 2019

Table 1: Applied stimuli for the user study.

ID	Video	Length (min)
0	*A brie(f) history of cheese* – P. S. Kindstedt	02:05
1	*How Braille was invented* – J. Oreck	01:16
2	*How do pain relievers work?* – G. Zaidan	01:34
3	*What are those floaty things in your eye?* – M. Mauser	01:49
4	*How the Band-Aid was invented* – J. Oreck	01:10
5	*How the sandwich was invented* – J. Oreck	01:13
6	*How to visualize one part per million* – K. Preshoff	02:05
7	*The surprising link between stress and memory* – E. Cox	02:07
8	*The science of spiciness* – R. Eveleth	02:07
9	*How to unboil an egg* – E. Nelsen	01:45

Task

After watching a video, the task of the participants was to answer three questions: (1) One counting question (*How many goats were visible?*), (2) a question about colors in the video (*What is the color of the girl's hair?*), and (3) one question about the specific visual content (*What are the ingredients of the sandwich?*). Only the type of questions was known before the videos. Furthermore, we asked the participants to summarize the story of the video. All questions were answered verbally and noted down by the study supervisor. With these questions, we intended to stimulate attention on the visual content. We kept the task consistent for the main user study, described in the following section, in order to (1) collect comparable gaze distributions for the evaluation and (2) to use the data as input for the **SSC** algorithm.

USER STUDY

We conducted a user study (n=54) with a within-subject design to compare traditional captions (**TC**) with direct (**DC**) and salience-sensitive (**SSC**) captions. The same stimuli and tasks as described in the baseline recording were applied. The videos were muted to focus on the visual aspects of the stimuli and the conditions. Participants had to read captions in order to answer all questions.

Study Design

The stimuli were presented in three blocks with each block containing three videos of the same condition. *Video 0* was used for training. After each video, participants answered the content questions. After a block, the participants filled out a usability questionnaire for the viewed condition. The assignment of videos to blocks and their order was counter-balanced based on Latin Squares. The order of conditions was also systematically alternated. For the captions we applied a sans serif font (22 pt) on a screen with 96 ppi in the native resolution of the videos. We displayed text with a white font color and a semi-transparent (50%) black background (Figure 1). The captions were kept consistent between conditions.

The *independent variables* are: Direct (**DC**), Salience-Sensitive (**SSC**), and Traditional captions (**TC**). During the experiment, we collected *gaze data* at a rate of 60 Hz in order to compute fixations, approximate saccades, calculate dwell times and gaze distributions as *dependent variables*. *Task performance* was assessed by the number of correct answers given after each video. In terms of summarizing of content, we only evaluated if participants were able to give a short summary. The usability of the three conditions was measured by self-reported *cognitive load* after each condition trough the raw NASA Task Load Index [15]. Furthermore, participants were asked to rate their *experience* with the condition by filling in the User Experience Questionnaire (UEQ) [27]. In a post-study questionnaire, participants compared the conditions and ranked them based on preference. In addition, they were asked to justify their decision as free text.

Hypotheses

We state four main hypotheses on gaze behavior, based on knowledge from previous work and design considerations:

H_1 **The average saccade length for traditional captions is higher than with gaze-adaptive conditions.** The increased visual angle between text and video will cause participants to perform larger saccades to switch between both areas. This could be perceived as higher effort and more exhausting.

H_2 **The average relative dwell time on captions is higher with traditional captions.** With alternative techniques, participants read captions more efficiently, spend less time on captions, and focus more on the video.

H_1 and H_2 are derived from previous findings [24] for videos with dialogues and pre-calculated positions. We hypothesize that gaze-adaptive captions will cause similar behavior. H_3 and H_4 concern visual search and gaze distributions:

H_3 **The time to first fixation (TTFF) on a new text will be significantly lower for gaze-adaptive captions.** Due to the proximity between gaze and text, search times for captions will be reduced and participants begin to read earlier.

H_4 **Gaze distributions of DC and SSC will be closer to the baseline than TC**. Gaze-adaptive captions appear close to the viewer's gaze. Reading and viewing will be less separated than with traditional captions, leading to a gaze distribution closer to natural viewing behavior. Consequently, participants will direct their gaze more often to the same areas as in natural viewing.

Furthermore, we expect that cognitive load will decrease and the user experience will be better when participants watch videos with gaze-adaptive captions. This should be reflected in the usability questionnaires and participants' comments.

Procedure

After reading an information sheet on the experiment procedure, participants filled in a consent form and the demographics questionnaire. Then, we performed a 6-point calibration of the eye tracker. The calibration was checked between blocks and re-calibrated if necessary. Before each video, a cross in the center of the screen was displayed and participants were asked to focus on it. Following a training video, participants watched the block with the first condition, answered the questions after each video and the usability questionnaire after the block. This procedure was repeated until all three conditions were tested. Finally, a post-study questionnaire asked to compare the conditions. The participants also ranked the conditions by preference (1–3), explained their decision, and stated additional comments in a free-text form.

Participants and Apparatus

The study was conducted at two locations simultaneously: at ETH Zürich and at the University of Stuttgart. For both locations, the participant population is similar (same official language, not English) and consisted mainly of students from the respective universities. Most of the participants had a central European cultural background: 5 (9%) participants were native English speakers, 38 (70%) fluently (proficiency level C1, C2) and 11 (21%) indicated a good knowledge (proficiency level B1, B2). In total, we recruited 54 participants (29 female) with normal or corrected-to-normal vision. Additionally, we conducted an Ishihara color perception test to ensure that all questions can be answered. The average age was 25 (sd=3.9, range: 20–38) and the experiment took about 60 minutes. All participants were compensated for performing the experiment based on local standards. In both locations, we used similar hardware to keep the study procedure consistent. The eye trackers (Tobii TX 300 and Tobii T60XL) were set to a rate of 60Hz and we used the integrated displays (23 and 24 inches). Videos were displayed with a 16:9 aspect ratio and a native resolution of 1920×1080 pixels. A chin rest stabilized the participants' head position at an approximate distance of 60 cm from the display.

RESULTS

Our results are structured as follows: (1) We provide an analysis of the recorded gaze data for the evaluation of hypotheses

Table 2: Summary of pairwise comparisons for H_1–H_4. (T) t-test, (W) Wilcoxon-Signed-Rank

	Measure	TC–DC	TC–SSC	DC–SSC
H_1	Saccade Length	(T): t=10.27 p<.001	(T): t=7.38 p<.001	(T): t=-2.89 p=.010
H_2	Dwel Time	(W): Z=-4.51 p<.001	(W): Z=-3.57 p=.001	(W): Z=2.78 p=.016
H_3	TTFF	(W): Z=4.92 p<.001	(W): Z=4.55 p<.001	(W): Z=-3.01 p=.008
H_4	MSE	(W): Z=2.67 p=.012	(W): Z=2.67 p=.012	(W): Z=-1.84 p=.223

H_1–H_4. (2) We investigate how the conditions influenced the viewing experience. For statistical analysis, we used R and SPSS[5]. Based on previous work [24] and comments from the participants, we expected that the experience and familiarity with reading captions will play an important role when ranking the conditions. Hence, we divided the participants in two groups, based on the frequency of general caption usage. Inexperienced participants stated to watch few (less than 2h per week (20%)) to no subtitle content at all (30%), while experienced participants watch more than two hours a week (2–4h (26%), 4–6h (11%) and more than 6h (13%)). With the respective threshold, both groups consist of 27 participants.

Eye Tracking Data

For the recorded gaze data, we applied an identification by velocity threshold (IV-T) fixation filter [37] with a $20^\circ/s$ threshold and a minimum fixation duration of 100 ms. Approximately 4% of the stimulus recordings were discarded due to insufficient quality (< 70% valid samples). One participant was removed completely due to insufficient data for one condition. Our results are summarized in Figure 4 and Table 2. First, we tested for normality (Shapiro Wilk) and chose the respective tests. For normally distributed data, we conducted a RM-ANOVA with post-hoc Bonferroni corrected p-values for pairwise comparisons. If the data did not satisfy a normal distribution, we used Friedman tests followed by Wilcoxon tests for post-hoc testing, also with Bonferroni corrected p-values. The alpha level was 0.05. We also looked at interaction effects between the subtitle experience and the effects of the condition. If the data were non-normally distributed, we applied an aligned rank transform [44]. We found interactions for H_1 only, which leads us to the assumption that experience did not affect eye movements.

H_1 – The average saccade length for traditional captions is higher than with gaze-adaptive conditions (Figure 4a). The RM-ANOVA shows significant differences between conditions on saccade length (F=56.08, p<.001). Post-hoc testing revealed significant pair-wise differences between all conditions (Table 2). We identified an interaction effect of condition and experience with subtitles for saccade length (F=3.85, p=.020). Saccade lengths on **TC** were significantly (t=2.66, p=.010) longer for inexperienced (mean=145.81, sd=31.43, median=147.48) compared to experienced participants (mean=167.55, sd=37.68, median=165.20). When looking at experienced and inexperienced participants separately, the gaze adaptive methods had significantly shorter saccade lengths compared to **TC** for both groups (experienced: **TC**–**DC** (t=-8.85, p<.001), **TC**–**SSC** (t=-7.19, p<.001); inexperienced: **TC**–**DC** (t=-6.09, p<.001), **TC**–**SSC** (t=-3.56, p=.002)). Inexperienced participants showed a significant difference (t=-2.53, p=.040) in saccade length between **DC** (mean=122.47, sd=23.5, median=123.34) and **SSC** (mean=132.17, sd=22.3, median=132.43). (H_1 **supported**)

H_2 – The average relative dwell time on captions is higher with traditional captions (Figure 4b). There was a significant difference between conditions considering time people spent reading the captions (χ^2=24.04, p<.001). Contrary to our hypothesis, pairwise comparisons showed that participants spent a shorter amount of time looking at the **TC**, compared to other conditions (Table 2). Further, participants focused significantly longer on **DC** compared to **SSC** (p=.016). The dwell time on other content was also influenced by the condition (χ^2= 59.58, p<.001). Here, significant differences were found between all conditions (p<.001), the dwell times were reciprocal to the time on captions (Figure 4b). Hence, H_0 can be rejected, but the differences were contrary to our assumption. (H_2 **not supported**)

H_3 – The time to first fixation (TTFF) on a new text will be significantly lower for gaze-adaptive captions (Figure 4c). The condition influenced the TTFF significantly (χ^2=47.25, p<.001). The measured TTFF was significantly (Table 2) longer for **TC** compared to both gaze-adaptive techniques. (H_3 **supported**)

H_4 – Gaze distributions of DC and SSC will be closer to the baseline than TC (Figure 4d). To evaluate H_4, we calculated the gaze distributions of each stimulus and each condition based on a 100×100 grid, normalized by the number of participants watching the same combination (Figure 4e). We calculated the mean square error (MSE) between each condition and the baseline. A Friedman rank sum test on the MSE showed significant differences between conditions (χ^2=14, p<.001). Post-hoc testing showed significant differences between **TC**–**DC** and **TC**–**SSC**. (H_4 **supported**)

Task Performance

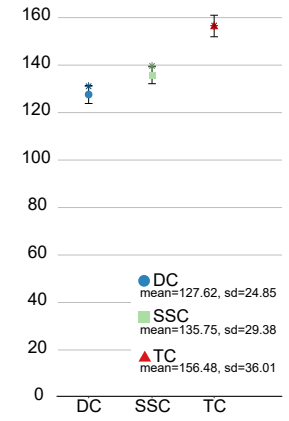
For the baseline condition, participants answered 59% of the questions correctly, 64% with **TC**, 60% with **DC**, and 59% with **SSC**. A Wilcoxon signed-rank between caption conditions and a Mann-Whitney U test for baseline comparison showed no significant differences. Since we assume that the baseline condition represents the optimal condition to answer the questions (audio explanation, no visual distraction by text), the different caption positions neither improved nor impaired the task performance. Similar findings for different line segmentation in captions can be found in the literature [31].

User Experience and Perceived Task Load

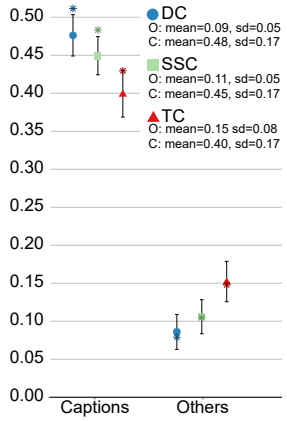
Participants rated their experience (UEQ) and their perceived task load (NASA TLX) after each condition.

The UEQ measures the user experience of a system and gives insights into the following criteria: *Attractiveness*, *Perspicuity*,

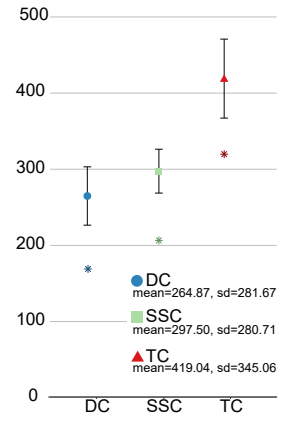
[5]IBM SPSS Statistics version 25, R version 3.6.1



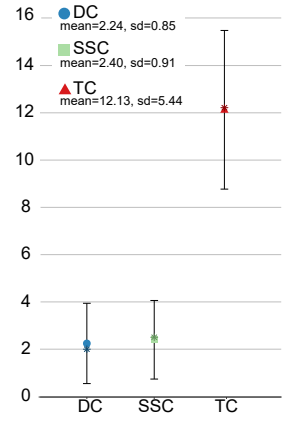
(a) Average saccade length per participant (px).



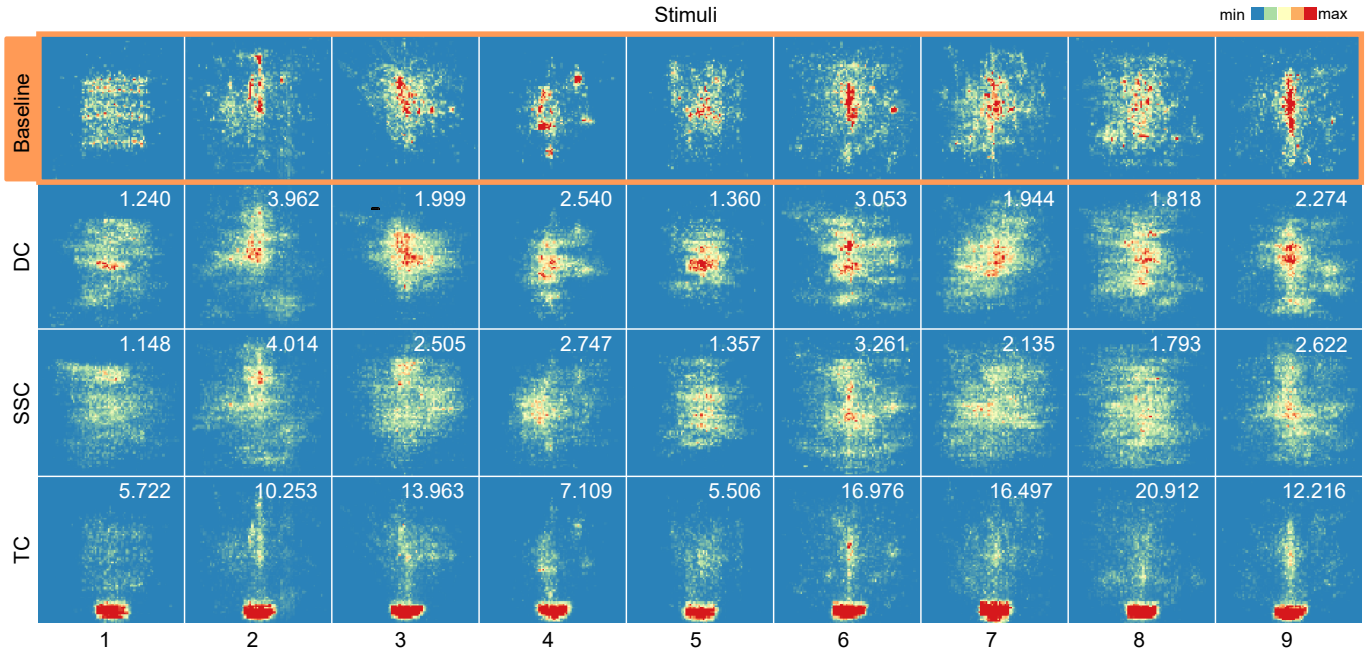
(b) Average relative dwell time on captions and other regions (%).



(c) Average time to first fixation per participant (ms).



(d) Mean square error to baseline of gaze distributions.



(e) Gaze distributions for all stimuli, normalized by the number of participants. DC and SSC are closer to natural viewing behavior than TC.

Figure 4: Summary of results for measured gaze data. (a)–(d) Results of the measures for H_1–H_4 (symbols = mean, * = median). Whiskers indicate the 95% confidence interval. (e) Visualization of the gaze distributions for the baseline and all three conditions.

Efficiency, *Dependability*, *Stimulation*, and *Novelty*. There were significant differences between the ratings of experienced and inexperienced participants in the UEQ (Figure 5). Both groups rated the *Novelty* of the gaze-adaptive methods higher than with the traditional method (both pairwise comparisons p<.001, Table 3). However, in terms of *Perspicuity* and *Dependability*, a Friedman test confirmed significant higher ratings for the traditional method. Inexperienced participants rated both, *Attractiveness* and *Efficiency* of **TC** significantly lower than experienced participants (Z=2.547, p=.011 and Z=2.977, p=.003). Furthermore, inexperienced rated the *Dependability* of **DC** significantly higher than the experienced ones (Z=-2.074, p=.038). On the other hand, experienced participants rated **TC** significantly higher in terms of *Perspicuity* and *Dependability*. Inexperienced participants found the *Stimulation* of gaze-adaptive techniques significantly higher than with the traditional method. They rated the *Dependability* of **TC** significantly higher than **SSC**.

The NASA TLX assesses the subjective difficulty of a task. Figure 6 shows the scores for each of the three conditions among the six dimensions: *Mental*, *Physical* and *Temporal Demand* as well as self-reported *Performance*, *Effort*, and *Frustration*. Despite the novelty of the gaze-adaptive conditions, a Friedman Test could not confirm significant differences between conditions.

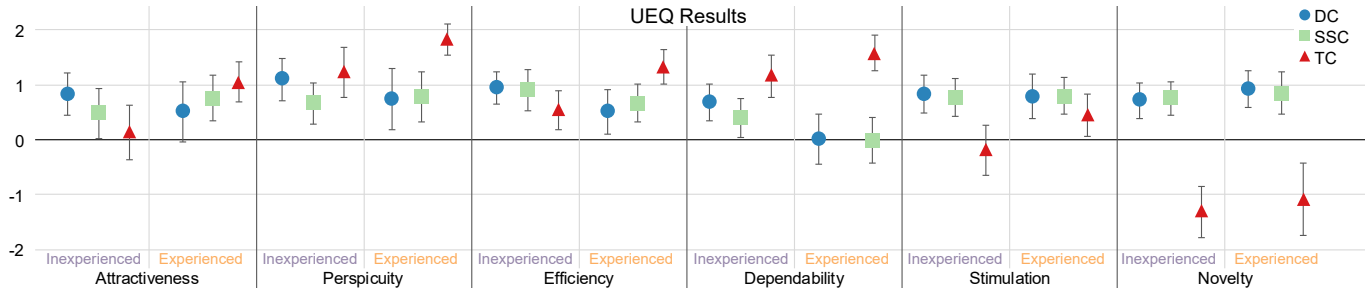


Figure 5: Results of the UEQ. Whiskers indicate the 95% confidence interval, symbols display mean values.

Table 3: Pair-wise comparison of conditions for the UEQ.

Aspect	All	Experienced	Inexperienced
Attractiveness	-	-	-
Perspicuity	**TC–SSC**: Z=-3.811, p<.001	**TC–DC**: Z=-3.263, p=.001 **TC–SSC**: Z=-3.554, p<.001	-
Efficiency	-	-	-
Dependability	**TC–DC**: Z=-4.483. p<.001 **TC–SSC**: Z=-5.004. p<.001	**TC–DC** Z=4.089, p<.001 **TC–SSC** Z=4.171, p<.001	**TC–SSC**: Z=-2.518, p=.012
Stimulation	-	-	**TC–DC** Z=3.155, p=.002 **TC–SSC**: Z=2.844, p<.004
Novelty	**TC–DC**: Z=5.528. p<.001 **TC–SSC**: Z=5.515. p<.001	**TC–DC**: Z=3.728, p<.001 **TC–SSC**: Z=3.571, p<.001	**TC–DC**: Z=4.195, p<.001 **TC–SSC**: Z=4.261, p<.001

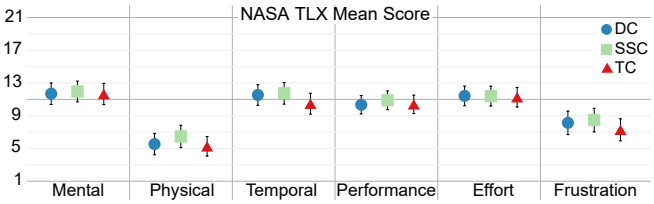


Figure 6: Result of the NASA TLX. Whiskers indicate 95% confidence intervals, symbols display mean values.

Post-Study Questionnaire

After watching all three conditions, we asked the participants to rate four questions/statements on a 6-Point Likert Scale to compare the techniques according to aspects of *Visibility*, *Content*, *Readability*, and *Visual Search* [24]. Table 4 gives an overview of the questions. We again used Friedman tests followed by Wilcoxon tests for testing differences between inexperienced and experienced participants while the effects of the condition were tested with a Mann-Whitney U test. While we could find significant differences between the conditions in general and for experienced participants only, we could not find any for inexperienced participants.

In terms of *Visibility*, there was no significant difference between the conditions (see Figure 7). However, we found that experienced participants (not shown in Figure 7) rated **TC** (mean=1.96, sd=1.16) significantly lower than the inexperienced (mean=3.52, sd=1.99). Furthermore, experienced participants gave significantly lower ratings for **TC** compared to both **DC** and **SSC**. While the type of caption seems not to affect how good participants could follow the *Content* of the video, the experienced participant's rating (mean=3.96, sd=1.22) for **TC** is significantly higher than that of inexperienced participants (mean=2.85, sd=1.46). The *Readability* with **TC** was rated higher than with **SSC**. Especially, experienced participants rated **TC** significantly higher compared to the gaze-adaptive methods. Also, significantly less *Search* is needed to find **TC** compared to the gaze-adaptive methods.

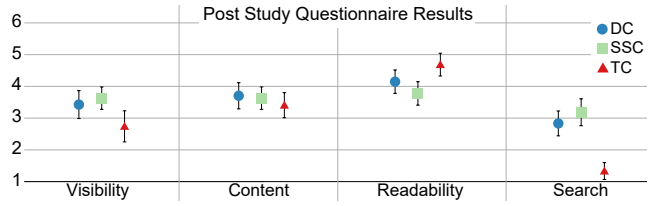


Figure 7: Result of the post study questionnaire in which participants rated each condition in terms of four characteristics.

Table 4: Post-study questions for comparison considering Visibility, Content, Readability, and Search.

Aspect	Question	Scale (1–6)
Visibility	How much did the subtitle impair your view on the video?	*few–much*
Content	How well could you follow the events in the video?	*bad–good*
Readability	The subtitles were easy to read.	*disagree–agree*
Search	I had to search for the subtitles before I could read them.	*disagree–agree*

Ranking

Conclusively, we asked the participants to rank the techniques according to their preference and comment on their decision. Figure 8 shows that almost half of the experienced participants (48,15%) rated the traditional method (**TC**) as their first choice for subtitle, followed by direct captions (**DC**, 37.04%). Salience-sensitive (**SSC**) were only preferred by 14.81 %. In contrast, inexperienced participants had a different opinion. The first rank was evenly distributed, with a slight preference for **DC** (37.04%), followed by **SSC** (33.33%) and **TC** (29.63%). Together with the preference by habit, participants mainly reported the stability of caption positions, occlusions, and tracking effort as the main reasons for their rankings.

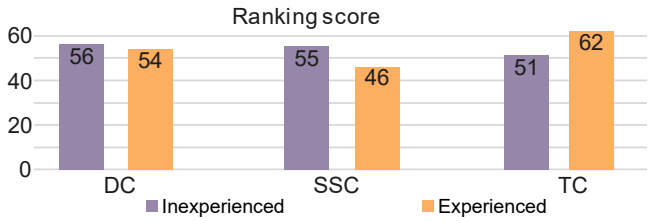


Figure 8: Ranking of the conditions based on the participants' preference. A condition gains 3 points when voted as first, 2 points for second, and 1 point for third place.

DISCUSSION

Our results support the assumption that gaze-adaptive design can improve the viewing experience, especially for viewers with less practice in reading traditional captions.

Shorter saccades indicate that less exhausting eye movement was necessary to switch between text and video. Considering dwell times, we found that with **TC**, participants focused more on the video. This finding is particularly interesting because it disagrees with the findings from other studies on subtitle layouts with fixed positions in the video [8, 24]. We interpret this difference partially as a result of the different stimuli and the task in our study. While the other studies contained videos that primarily focused on faces, our stimuli often included multiple areas that were important. Participants had to switch between all areas to answer the questions. For **TC**, some participants reported that they focused more on the video than on the text in order to answer the questions, this was not reported for the other conditions. A layering effect might also have influenced the result: The dwell time includes all fixations on a caption. However, since the captions were transparent, our measure did not differentiate if participants were focusing on the text or the content behind it.

Since the time to first fixation was significantly reduced with **DC** and **SSC**, and the text was often close to the gaze, more fixations were counted on the text than on other areas, consequently leading to higher dwell times. However, the effect of the reduced time to first fixation was not fully supported in the questionnaire in terms of the search aspect. Participants still rated the search effort for **TC** lower than for gaze-adaptive methods. This could be an indicator that the objectively reduced search time did not influence the subjective dependability of traditional captions. All ratings were in the low–mid range of the scale, indicating that for many participants this was a minor issue. In the comments, participants stated that gaze-adaptive methods helped them read the text better (*"[...] SSC was interactive and it was easy to read the subtitle along with the video."*; *"[...] it was really hard to read the traditional subtitles and also look at the video at the same time, whereas the DC and SSC type subtitles were usually in the center of the screen or near the focus point in the video, so it was relatively easy to read and watch the video at the same time."*).

As measured by the MSE values, gaze distributions for **DC** and **SSC** show more resemblance to natural viewing behavior than traditional captions. This is supported by the subjective feedback, where especially inexperienced viewers stated they could follow the content more easily and more effectively. Investigating the individual gaze distributions of the stimuli (Figure 4e), we can see that **TC** typically generates a hotspot at the bottom of the screen, while **DC** and **SSC** show similar hotspots as the baseline. Since **SSC** includes displacement from the gaze position, the data are slightly more distributed than with **DC**. Both conditions also contain horizontal gaze patterns caused by reading.

The results of the questionnaires can be summarized as follows: Although the perceived task load did not differ between traditional subtitles and gaze-adaptive methods, we could find differences in the user experience. Obviously, both gaze-adaptive captions were rated more *novel* and for inexperienced participants also more exciting and motivating (*Stimulation*) to use. Experienced participants, however, spending more than two hours per week with **TC**, rated this technique more *Perspicuous*, *Efficient* and *Dependable* than the gaze-adaptive methods. This indicates that for inexperienced people, gaze-adaptive techniques are a useful alternative that improves their viewing experience. The ranking results are in line with this hypothesis: Inexperienced participants rated **TC** to be their least favored, while the experienced preferred **TC** in 50% of the cases. Furthermore, inexperienced participants could follow the events of the video significantly worse with **TC** than the experienced viewers. They stated that the technique impaired their view on the video content significantly more.

Design Implications

Based on the participants' comments, our measurements, and the experience gathered during the development, we derived a set of design considerations for gaze-adaptive captions with respect to the target group, location consistency, displacement distance, occlusion, and perceived pace:

Target Group

As our results show, experienced viewers prefer traditional captions. However, gaze-adaptive captions were not completely neglected by experienced participants (37% preferred **DC**), as one of them stated: *"I am used to the traditional condition that is why I could perform best [...] If I would be used to them, they would probably be more comfortable to watch a movie/clip and you would notice more of the pictures in the background."* The high acceptance of the new techniques in the group of inexperienced viewers indicates that gaze-adaptive techniques will be more useful for people who only started to watch videos with captions, or watch them occasionally. At this point, individual preferences will probably determine which technique a person prefers.

Location Consistency

During reading, the location of captions should be consistent. This was also suggested by previous studies [1]. We considered this aspect in **DC** by adjusting caption positions when reading was finished. However, in **SSC** this guideline contradicts the idea of dynamic displacement for individual captions. In some cases, participants reported that they still had to search the text in **SSC**: *"[...] TC was very easy to follow and gave the time to follow the video too. DC was also on the easy side because it moved a little but not too much so that it became hectic. SSC on the other hand was so hectic and very unpleasing to follow."* To improve this aspect, an extension of

the algorithm is feasible, considering consecutive captions for optimizations with respect to temporal coherence.

Displacement Distance

The displacement relative to the gaze has to be picked carefully, because too large displacement will cause the viewer to believe that gaze has no influence on the placement and leads to searching of the captions. Although we adjusted the parameters by pilot testing, some participants were still affected by this effect: *"[...] SSC was moving more randomly on the screen, so I would lose some time trying to find the subtitles."* In our presented algorithm, the size of the grid and the number of displacement steps can be varied. We plan to further investigate these parameters by conducting experiments on just-noticeable differences (JNDs) to identify thresholds.

Occlusion

By design, **TC** will occlude few important areas, **SSC** tries to minimize the occlusions, and **DC** does not consider this aspect. Participants who were bothered by **DC** stated: *"DC often overlaid relevant images, thus the traditional method is preferred [...]"*. While another participant mentioned: *"[...] I liked SSC best, because it was easier to keep an eye on the video and the subtitles bothered me least in that case."* One participant suggested presenting just text without a bounding box to reduce the occlusion. We plan to investigate how alternative presentation methods will influence this aspect.

Perceived Pace

We noticed that some comments on **DC** concern a difference in perceived pace: *"I prefer DC over SSC, because the pace of the subtitles was slower."*;*"Subtitles were easier to read in DC and Traditional and they were slowly paced, whereas in SSC, it was an either or situation between subtitles and the content of the video."*; *"(about DC) The fact that the subtitles were moving across the screen but with a steady pace (not fast and not moving too much contrarily to SSC) helped me see more of the video content while reading at the same time."* The reduced time to first fixation might influence these perceived differences, as the timing of captions was consistent between conditions. Further research will be necessary to investigate this effect.

Other Application Fields

Our experiments were conducted on a screen with one person watching a video. With the current development in eye tracking, new applications for mobile devices become feasible.

Mobile Devices

On small screens, the use of traditional captions seems reasonable as the foveal area covers more content than on a big screen and viewers can switch more easily between reading and viewing. Hence, future work has to investigate for which screen sizes gaze-adaptive captions will become beneficial. An application of the presented techniques on small screens poses new challenges due to the increasing screen space needed to display text in a readable size.

Mixed/Virtual Reality

Captions in mixed and virtual reality require further research, including where to place captions within the field of view [5, 29, 36]. Brown et al. [8] and Rothe et al. [35] investigated speaker-based captions in 360° videos. Similarly, Peng et al. [30] applied dynamic captions with a Microsoft HoloLens, showing speech bubbles next to a speaker in a conversation. With eye tracking in head-mounted displays, gaze-adaptive captions can be applied, for instance, to translate live conversations [30] or narrative text without visible speakers (e.g., interactive tourist guides [25]). We hypothesize that with appropriate hardware, gaze-responsive placement would improve this technology. For a displacement-based approach, it must be evaluated if existing salience models are applicable to live 3D content, or if new models will be necessary.

CONCLUSION

We presented two techniques to adapt the position of captions with respect to the viewer's gaze. Our results show that casual, inexperienced viewers prefer gaze-adaptive methods as they found them easier to read and follow a video simultaneously. Gaze distributions are more similar to natural viewing behavior with gaze-adaptive techniques than with traditional captions. Our results refer to participants between the age of 20–38. However, we see potential for gaze-adaptive captions as an accessible assistance for elderly people where hearing impairments are more frequent than in our sample group. On purpose, we did not focus on multiple viewers, which requires different technology to foster individual gaze tracking and caption positioning. In general, we consider our approach applicable to video content with and without on-screen speakers. Speaker-following captions [17] could be combined with our gaze-adaptive approach by calculating the displacement map based on object detection algorithms (e.g., faces).

The design space of gaze-adaptive captions is not limited to positioning. We plan to extend our studies on additional modalities to improve the viewing experience. As a first step, we developed a prototype for inexperienced viewers that adjusts the playback speed or stops the video during the reading process until the viewer starts focusing on the video content again. Further research will be necessary to evaluate how this technique is best combined with the presented approaches. Additionally, our current results exclude audio which also influences gaze behavior. Future work will consider the influence of audio on gaze-adaptive methods and also in-the-wild studies under everyday life conditions.

In summary, gaze-adaptive captions are a promising technique especially for inexperienced viewers who become more distracted by traditional captions. Moreover, this approach is not limited to video captions but can also be applied for dynamic label placement in general, on monitor-based systems and in mobile applications with mixed reality.

ACKNOWLEDGMENTS

We would like to thank all participants for contributing to this work and Tiffany Kwok for supporting the making of the video and the figures. This work was partially funded by the German Research Foundation (DFG) – Project-ID 251654672 within the SFB/Transregio 161.

REFERENCES

[1] Wataru Akahori, Tatsunori Hirai, Shunya Kawamura, and Shigeo Morishima. 2016. Region-of-interest-based subtitle placement using eye-tracking data of multiple viewers. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*. 123–128.

[2] Wataru Akahori, Tatsunori Hirai, and Shigeo Morishima. 2017. Dynamic subtitle placement considering the region of interest and speaker location. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 102–109.

[3] Katrin Angerbauer, Heike Adel, and Ngoc T Vu. 2019. Automatic compression of subtitles with neural networks and its effect on user experience. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*. 594–598.

[4] Mike Armstrong and Matthew Brooks. 2014. Enhancing subtitles. In *Adjunct Publication of the ACM International Conference on Interactive Experiences for Television and Online Video*. 1–2.

[5] Ronald Azuma and Chris Furmanski. 2003. Evaluating label placement for augmented reality view management. In *Proceedings of the Second IEEE and ACM International Symposium on Mixed and Augmented Reality*. 66–75.

[6] Marie-Josée Bisson, Walter J. B. Van Heuven, Kathy Conklin, and Richard J. Tunney. 2014. Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics* 35, 2 (2014), 399–418.

[7] Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic subtitles: The user experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*. 103–112.

[8] Andy Brown, Jayson Turner, Jake Patterson, Anastasia Schmitz, Mike Armstrong, and Maxine Glancy. 2017. Subtitles in 360-degree video. In *Adjunct Publication of the ACM International Conference on Interactive Experiences for TV and Online Video*. 3–8.

[9] Stephen Doherty and Jan-Louis Kruger. 2018. The development of eye tracking in empirical research on subtitling and captioning. In *Seeing into screens – Eye tracking and the moving image*, Tessa Dwyer, Claire Perkins, Sean Remond, and Jodi Sita (Eds.). Bloomsbury, London, 46–64.

[10] Andrew T. Duchowski. 2017. Serious gaze. In *Proceedings of the 9th International Conference on Virtual Worlds and Games for Serious Applications*. 276–283.

[11] Géry d'Ydewalle and Wim De Bruycker. 2007. Eye movements of children and adults while reading television subtitles. *European Psychologist* 12, 3 (2007), 196–205.

[12] Géry d'Ydewalle, Caroline Praet, Karl Verfaillie, and Johan Van Rensbergen. 1991. Watching subtitled television automatic reading behavior. *Communication Research* 18, 5 (1991), 650–666.

[13] Olivia Gerber-Morón, Agnieszka Szarkowska, and Bencie Woll. 2018. The impact of text segmentation on subtitle reading. *Journal of Eye Movement Research* 6, 5 (2018), 1–12.

[14] Fabian Göbel, Peter Kiefer, Ioannis Giannopoulos, Andrew T. Duchowski, and Martin Raubal. 2018. Improving map reading with gaze-adaptive legends. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*. 29:1–29:9.

[15] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908.

[16] Richang Hong, Meng Wang, Xiao-Tong Yuan, Mengdi Xu, Jianguo Jiang, Shuicheng Yan, and Tat-Seng Chua. 2011. Video accessibility enhancement for hearing-impaired users. *ACM Transactions on Multimedia Computing, Communications, and Applications* 7S, 1 (2011), 24:1–24:19.

[17] Yongtao Hu, Jan Kautz, Yizhou Yu, and Wenping Wang. 2014. Speaker-following video subtitles. *ACM Transactions on Multimedia Computing, Communications, and Applications* 11, 2 (2014), 32:1–32:17.

[18] Robert J. K. Jacob. 1990. What you look at is what you get: Eye movement-based interaction techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 11–18.

[19] Bo Jiang, Sijiang Liu, Liping He, Weimin Wu, Hongli Chen, and Yunfei Shen. 2017. Subtitle positioning for e-learning videos based on rough gaze estimation and saliency detection. In *SIGGRAPH Asia Posters*. 15–16.

[20] Marcel A. Just and Patricia A. Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive Psychology* 8, 4 (1976), 441–480.

[21] Harish Katti, Anoop Kolar Rajagopal, Mohan Kankanhalli, and Ramakrishnan Kalpathi. 2014. Online estimation of evolving human visual interest. *ACM Transactions on Multimedia Computing Communication Applications* 11, 1 (2014), 8:1–8:21.

[22] Izabela Krejtz, Agnieszka Szarkowska, and Krzysztof Krejtz. 2013. The effects of shot changes on eye movements in subtitling. *Journal of Eye Movement Research* 6, 5 (2013), 1–12.

[23] Jan-Louis Kruger, Agnieszka Szarkowska, and Izabela Krejtz. 2015. Subtitles on the moving image: An overview of eye tracking studies. *Refractory : A Journal of Entertainment Media* 25 (2015), 1–14.

[24] Kuno Kurzhals, Emine Cetinkaya, Yongtao Hu, Wenping Wang, and Daniel Weiskopf. 2017. Close to the action: Eye-tracking evaluation of speaker-following subtitles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 6559–6568.

[25] Tiffany C. K. Kwok, Peter Kiefer, Victor R. Schinazi, Benjamin Adams, and Martin Raubal. 2019. Gaze-guided narratives: Adapting audio guide content to gaze in virtual and real environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 491:1–491:12.

[26] Michael Land, Neil Mennie, and Jennifer Rusted. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 11 (1999), 1311–1328.

[27] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work*, Andreas Holzinger (Ed.). Berlin, Heidelberg, 63–76.

[28] Maryam Sadat Mirzaei, Kourosh Meshgi, Yuya Akita, and Tatsuya Kawahara. 2017. Partial and synchronized captioning: A new tool to assist learners in developing second language listening skill. *ReCALL - The Journal of the European Association for Computer Assisted Language Learning* 29, 2 (2017), 178–199.

[29] Jason Orlosky, Kiyoshi Kiyokawa, and Haruo Takemura. 2014. Managing mobile text in head mounted displays: Studies on visual preference and text placement. *ACM SIGMOBILE Mobile Computing and Communications* 18, 2 (2014), 20–31.

[30] Yi-Hao Peng, Ming-Wei Hsi, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and others. 2018. Speechbubbles: Enhancing captioning experiences for deaf and hard-of-hearing people in group conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 293–302.

[31] Elisa Perego, Fabio Del Missier, Marco Porta, and Mauro Mosconi. 2010. The cognitive effectiveness of subtitle processing. *Media Psychology* 13, 3 (2010), 243–272.

[32] Dhevi J. Rajendran, Andrew T. Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives* 21, 1 (2013), 5–21.

[33] Vasili Ramanishka, Abir Das, Jianming Zhang, and Kate Saenko. 2017. Top-down visual saliency guided by captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7206–7215.

[34] Raisa Rashid, Quoc Vy, Richard Hunt, and Deborah I. Fels. 2008. Dancing with words: Using animated text for captioning. *International Journal of Human–Computer Interaction* 24, 5 (2008), 505–519.

[35] Sylvia Rothe, Kim Tran, and Heinrich Hußmann. 2018. Dynamic subtitles in cinematic virtual reality. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*. 209–214.

[36] Rufat Rzayev, Paweł W. Woźniak, Tilman Dingler, and Niels Henze. 2018. Reading on smart glasses: The Effect of text position, presentation type and walking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 45:1–45:9.

[37] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*. 71–78.

[38] Agnieszka Szarkowska, Izabela Krejtz, Zuzanna Klyszejko, and Anna Wieczorek. 2011. Verbatim, standard, or edited?: Reading patterns of different captioning styles among deaf, hard of hearing, and hearing viewers. *American Annals of the Deaf* 156, 4 (2011), 363–378.

[39] Ruxandra Tapu, Bogdan Mocanu, and Titus Zaharia. 2019. DEEP-HEAR: A multimodal subtitle positioning system dedicated to deaf and hearing-impaired people. *IEEE Access* 7 (2019), 88:150–88:162.

[40] Roel Vertegaal. 2002. Designing attentive interfaces. In *Proceedings of the ACM Symposium on Eye tracking Research & Applications*. 23–30.

[41] Roel Vertegaal. 2003. Attentive user interfaces. *Communications of the ACM* 46, 3 (2003), 30–33.

[42] Toinon Vigier, Yoann Baveye, Josselin Rousseau, and Patrick Le Callet. 2016. Visual attention as a dimension of QoE: Subtitles in UHD videos. In *Proceedings of the Eighth International Conference on Quality of Multimedia Experience*. 1–6.

[43] Colin Ware and Harutune H. Mikaelian. 1986. An evaluation of an eye tracker as a device for computer input. *ACM SIGCHI Bulletin* 17, SI (1986), 183–188.

[44] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 143–146.