Eye movements while viewing narrated, captioned, and silent videos

Nicholas M. Ross

Department of Psychology, Rutgers University, Piscataway, NJ, USA

Eileen Kowler

Department of Psychology, Rutgers University, Piscataway, NJ, USA

Videos are often accompanied by narration delivered either by an audio stream or by captions, yet little is known about saccadic patterns while viewing narrated video displays. Eye movements were recorded while viewing video clips with (a) audio narration, (b) captions, (c) no narration, or (d) concurrent captions and audio. A surprisingly large proportion of time (>40%) was spent reading captions even in the presence of a redundant audio stream. Redundant audio did not affect the saccadic reading patterns but did lead to skipping of some portions of the captions and to delays of saccades made into the caption region. In the absence of captions, fixations were drawn to regions with a high density of information, such as the central region of the display, and to regions with high levels of temporal change (actions and events), regardless of the presence of narration. The strong attraction to captions, with or without redundant audio, raises the question of what determines how time is apportioned between captions and video regions so as to minimize information loss. The strategies of apportioning time may be based on several factors, including the inherent attraction of the line of sight to any available text, the moment by moment impressions of the relative importance of the information in the caption and the video, and the drive to integrate visual text accompanied by audio into a single narrative stream.

Introduction

Saccadic eye movements take the line of sight to areas of interest in the visual scene in an effortless but purposeful way. They are indispensable for coping with the wealth of information that is distributed throughout the visual world. Decisions about how to plan saccades in space and time thus play a crucial role in apprehending the content of natural scenes.

A great deal of prior work has focused on identifying the factors that drive saccadic decisions while inspecting static scenes or performing visual or visuomotor tasks (e.g., Epelboim et al., 1995; Epelboim & Suppes, 2001; Hayhoe & Ballard, 2005; Johansson, Westling, Bäckström, & Flanagan, 2001; Kibbe & Kowler, 2011; Kowler, 2011; Land & Hayhoe, 2001; Land, Mennie, & Rusted, 1999; Malcolm & Henderson, 2010; Motter & Belky, 1998; Najemnik & Geisler, 2005; Pelz & Canosa, 2001; Steinman, Menezes, & Herst, 2006; Torralba, Oliva, Castelhano, & Henderson, 2006; Turano, Geruschat, & Baker, 2003; Wilder, Kowler, Schnitzer, Gersch, & Dosher, 2009; Yarbus, 1967). Much of the discussion has surrounded the relative role played by *bottom-up* versus *top-down* factors in controlling saccadic decisions. Bottom-up factors refer to the properties of the visual stimulus itself, typically, the contrast of visual features of the display (Koch & Ullman, 1985). Top-down factors encompass everything else, including voluntary attention, the judged importance or relevance of different locations, the constraints imposed by limitations of memory, and (in the case of visuomotor tasks) the coordination of eye and arm. Tatler, Hayhoe, Land, and Ballard (2011) concluded on the basis of a recent review that top-down factors are more important than bottom-up factors in driving saccadic decisions but that an understanding of the nature and operation of the relevant top-down factors is a complex endeavor that is still at a relatively early stage.

The debate about the factors that control saccadic decisions has been recently extended to the characteristics of eye movements made while watching movies or videos (Berg, Boehnke, Marino, Munoz, & Itti, 2009; Carmi & Itti, 2006; Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Itti, 2005; Le Meur, Le Callet, & Barbra, 2007; Tseng, Carmi, Cameron, Munoz, & Itti, 2009; Vig, Dorr, & Barth, 2009). Videos are interesting stimuli, more representative of natural visual arrays than static pictures. Their content changes over time and includes motion as well as a top-down component that originates from the attempts to understand and

Citation: Ross, N. M., & Kowler, E. (2013). Eye movements while viewing narrated, captioned, and silent videos. *Journal of Vision*, *13*(4):1, 1–19, http://www.journalofvision.org/content/13/4/1, doi:10.1167/13.4.1.

doi: 10.1167/13.4.1 Received July 25, 2011; published March 1, 2013 ISSN 1534-7362

interpret the depicted events (Itti, 2005; Pantelis et al., 2011; Zacks & Tversky, 2001). In contrast to studies in which the changes to the visual stimulus are produced by observers' actions (e.g., Epelboim et al., 1995; Johansson et al., 2001; Land & Hayhoe, 2001; Pelz & Canosa, 2001; Steinman et al., 2006), videos allow comparisons of performance when content remains the same across all observers. Thus, the study of eye movements while watching videos can provide a useful addition to the array of approaches being used to identify the factors that drive saccadic decisions.

A few previous studies have described eye movements while watching videos. One question dominating this prior work was the role of physical salience in predicting fixated locations. Analyses showed that measures of salience based on either flicker or motion were better predictors of fixated locations than measures based on either intensity or color (Carmi & Itti, 2006; Le Meur et al., 2007). The results also showed preferences to maintain gaze near the center of the display (Berg et al., 2009; Dorr et al., 2010; Le Meur et al., 2007; Tseng et al., 2009), analogous to what has been found for viewing static pictures (Tatler, 2007). Centering preferences may reflect strategies of looking at the most important or vivid objects, which are often placed near the center of the image (Dorr et al., 2010; Tseng et al., 2009),or strategies of positioning gaze at the location that may be best for resolving the greatest number of details across the screen (Tatler, 2007).

Studies of eye movements while watching videos have also examined the same global characteristics of saccadic patterns that have been traditionally studied in both simple and complex visual tasks (Findlay & Gilchrist, 2003; Rayner, 1998), namely, the distributions of sizes of saccades, the durations of fixation pauses, and the scatter of fixated locations. These characteristics are often used to define typical viewing patterns and provide measures that have been used to infer scanning strategies. For example, Dorr et al. (2010) found longer fixation durations, smaller saccades, and less scatter of landing positions while viewing videos than static pictures. They concluded that these differences reflect preferences to maintain gaze near the center of the video images or, in the case of what they called "natural" movies, preferences for occasional large shifts of gaze between clusters of interesting regions. Berg et al. (2009) compared eye movements of monkey and human subjects watching the same videos. They found that monkeys made larger and more frequent saccades and were less likely to confine gaze to the center of the screen than were humans. They attributed this species difference to top down factors, in particular, to the inability of the monkeys to follow the events or to understand the importance of the main actors (who were often located in the center of the images). They assumed that the inability to fully interpret the sequence of depicted events encouraged the monkeys to explore over wider regions of the displays.

One limitation in the prior work on eye movements while viewing videos has been the absence of sound or narration. Narration is often found in videos and provides additional information that guides the interpretation of events (Carmi & Itti, 2006). Narration could change the scanning strategies or scanning characteristics due to the contribution of top-down factors. There have been prior studies of eye movements while viewing static pictures that incorporated narration in the form of spoken sentences or captions. These studies, however, were concerned with using eye movements to infer properties of real-time language processing and were not concerned with characterizing the saccadic strategies used to inspect visual scenes (Andersson, Ferreira, & Henderson, 2011; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Trueswell, Sekerina, Hill, & Logrip, 1999). There also have been studies of multimedia learning that included auditory information with videos, but these studies were concerned with how the choice of fixated locations contributed to the understanding and retention of information (Hyönä, 2010; Schmidt-Weigand, Kohnert, & Glowalla, 2010).

There are two main goals of the present study. The first is to study the strategies of reading captioned videos. Captions present challenges to viewers because gaze has to shift continually between video and text. Strategies of saccadic guidance should, ideally, be configured so as to minimize loss of information from either the captions or the video portion of the display. However, prior results using static pictures suggest that observers have strong preferences to read text regardless of its utility. For example, viewers show preferences to read text present in static pictures even when the text is neither vivid nor important (Cerf, Frady, & Koch, 2009). Preferences to fixate text persist when text is scrambled into nonsense words, turned upside down, or presented in an unfamiliar language (Wang & Pomplun, 2012). The preferences to look at text even when text is uninformative or redundant are interesting because such preferences appear to lead to no important gain of information, in contrast to tasks such as search, where the optimal nature of saccadic strategies has been emphasized (Najemnik & Geisler, 2005). There are also some prior reports of preferences to read captions while watching videos with redundant audio (Bisson, van Heusen, Conklin, & Tunney, 2011; d'Ydewalle, Praet, Verfaillie, & Van Rensbergen, 1991). These prior studies were limited in that they used stimuli consisting of conversations or "talking heads," in which the information conveyed by the narration was critical to interpretation. In the present study, no conversations or talking heads were present. The

narration provided background information or explanations of the depicted visual events. We hypothesize that viewers will spend little time reading captions in the presence of redundant audio because this would take attention away from the video.

The second goal of the present study is to determine effects of audio narration on major characteristics of the eye movement patterns that have been studied in the past (see above) to infer scanning strategies. These characteristics are: (a) the distributions of saccade sizes and pause durations, (b) centering tendencies (i.e., scatter of landing positions), and (c) the influence of physical salience. Any differences between these characteristics when the video is viewed with and without narration would not be due to physical (visual) salience but rather provides evidence for a role of top-down factors. On the basis of prior work on centering tendencies while viewing videos, we would hypothesize that the added information provided by narration should increase attention to the flow of events and that the increased attention to the events would be reflected by increased centering tendencies (Berg et al., 2009) and a reduced influence of physical salience. On the other hand, if centering is due to purely visual factors (Dorr et al., 2010; Tseng et al., 2009), no effect of narration on centering would be expected. In addition, finding an influence of narration on the size of saccades or on the intersaccadic pause durations, analogous to previous studies comparing these characteristics in videos and static pictures (see above), would point to effects of narration on global aspects of viewing strategies, including processes used to apprehend the events or the time allocated to processing fixated material.

Two experiments were conducted. Experiment 1 used long (2 min) videos, accompanied by audio narration, captions, both, or neither. Experiment 2 used shorter duration videos (15 s) with or without concurrent audio. The main findings were that audio narration had little effect on saccadic patterns with the exception of a slight increase in the distributions of landing locations in the absence of any narration to guide interpretation of events, while captions had large effects with strong and surprising preferences to spend a lot of time reading captions even with redundant audio.

Experiment 1: Effect of captions and audio narration on eye movements while viewing videos

Methods

Eye-movement recording

Eye movements were recorded using the Eyelink 1000 (SR Research, Osgoode, Canada) tower mounted version, sampling at 1000 Hz. Stimuli were presented on a Viewsonic G90fb CRT monitor, 1024 × 768 resolution, 60 Hz refresh rate, located at a viewing distance of 119 cm. The display area subtended 16.2° horizontally by 12.3° vertically. A chin rest was used to stabilize the head. Eye movements were recorded from the right eye. View of the left eye was occluded by a patch. Viewing was limited to the recorded eye because studies of binocular eye movements have shown that, when binocular view is permitted, the two eyes do not necessarily fixate the same location (Kowler et al., 1992; Steinman et al., 2006). Thus, recording from one eye during a binocular view may not necessarily provide an accurate measure of the intended fixated location.

Stimuli

Sixteen 2-min video clips were tested. Four clips were cut from each of the following four source videos, all documentaries: *Meerkat Manor: The Story Begins* (Discovery Communications, LLC, 2008), *March of the Penguins* (Warner Home Video, 2005), *Destiny in Space* (Warner Home Video, 2005), and *Earth the Biography: Volcanoes* (BBC Worldwide Ltd. Program, 2008). The long duration of the clips (about as long as typical movie trailers) was used because it seemed to allow sufficient time for viewers to understand and follow the sequence of developing events. Clips were chosen with the constraint that the narrator was not depicted on screen, that is, there were no talking heads or conversations. Clips also contained enough information to allow a brief post-trial test of memory for the contents. Clips were edited to remove instances of long ($>\sim$3 s) uneventful pauses. Captions, when present, contained an average of 7.67 words ($SD = 3.15$).

Procedure

Each subject was tested in a single experimental session. Before testing began subjects were told that they would view two-minute video clips with each followed by six four-alternative multiple choice questions testing memory for the content. These questions were important for providing a motivation for the video watching. Subjects were also told that videos would contain captions, audio, captions and audio, or neither captions nor audio.

An experimental session consisted of 16 trials, organized in blocks of four. Within each block, each of the four viewing conditions was tested once: captions only, audio only, captions + audio, and no audio/no captions. The content of the captions was identical to the content of the audio.

In any given block, each viewing condition was paired with a clip from a different source video so that each of the four source videos was represented once in

each block. As a result, by the end of the 16 trials, clips from each source video were seen an equal number of times in each of the four viewing conditions. No clip was seen more than once. The order of the conditions and clips within a block was haphazard with the constraint that the same viewing condition was never tested in consecutive trials across blocks. The memory test given after each trial consisted of six multiple choice questions. Questions were equally divided among those that tested memory for content presented in the narration only, the video only, or both. Performance was 74% correct over all subjects when some form of narration was provided (captions and/or audio) and 54% when no narration was given. The additional errors in the condition without narration were due to those questions that were primarily drawn from the content of the narration.

The calibration routine built into the Eyelink software was run before the start of the experiment and again before each trial. After the Eyelink calibration subjects fixated a central cross and started the trial by button press when ready. This was followed by a presentation of five crosses, one in the center and one in each corner of the display to serve as an additional check on calibration. Calibration scale factors were adjusted for each trial depending on the outcome of these additional calibrations. Adjustments in scale factors were typically <10%. After the video ended, subjects removed their head from the chin rest and answered six multiple choice questions with pencil and paper.

Subjects

Six subjects (paid volunteers and Rutgers University students) were tested. All had normal or corrected to normal (soft contact lenses) vision and were naïve to the experimental design and hypothesis. Results from the six individual subjects will be identified by an arbitrary two letter code (SA, SC, SJ, ST, SL, SN). All subjects except one (SA) were native English speakers (SA learned English as a child). The project was approved by the Rutgers University IRB for the protection of human subjects.

Analysis

The beginning and ending positions of saccades were detected offline by means of a computer algorithm employing a velocity criterion to find saccade onset and offset. The value of the criterion was determined empirically for individual observers by examining a large sample of analog recordings of eye positions. Portions of data containing blinks or episodes where tracker lock was lost were eliminated (SA 26%, SC 27%, SJ 11%, ST 3%, SL 2%, SN 4%). These proportions were about the same across conditions (audio 11%; captions 11%; neither audio or caption narration 13%; both audio and captions 13%). Data reported are based on the analysis of 1454 s for SA, 1434 s for SC, 1748 s for SJ, 1906 s for ST, 1925 s for SL, and 1886 s for SN. Note that the larger portion of time in which lock was lost for two of the subjects (SA and SC) is not surprising given that the long durations of the trials meant that frequent blinks were likely as well as episodes in which lock was lost due to the eyelid obscuring portions of the pupil. The data available for all subjects (more than 24 minutes of recording for each) was sufficient to obtain a reliable estimate of performance.

Video clips containing captions were examined to determine the frame numbers of the onset and offset of episodes in which captions were present. Caption onset and offset times for each trial were then adjusted for any frames that were dropped during the presentations using the record of dropped frames maintained in the Eyelink software. Any pair of captions that occurred consecutively with a gap of less than 210 ms between them was considered to be part of a single caption episode.

Results

Spatial distributions of eye movements

Figure 1 illustrates a typical eye trace for a representative subject viewing several seconds of the clip *Earth the Biography: Volcanoes*. Saccades (about 1°–3°) can be seen occurring about once or twice per second. Brief episodes of smooth pursuit (~90–120°/s) can be seen at second 98 and again at 101.

Figure 2 shows the spatial distribution of eye positions (resolution 1 ms with samples during saccades omitted) for the same subject for each of the four narration conditions: audio only; captions only; audio + captions; and neither audio or captions. The eye positions were pooled across the four clips tested. For the two conditions without captions, namely, audio only, and neither captions or audio (left panels), the line of sight almost always remained within the central 5° × 5° region of the 16° × 12° display, i.e., about 13% of the total display area (see also Berg et al., 2009; Dorr et al., 2010; Le Meur et al., 2007; Tseng et al., 2009). When captions were present, the patterns changed in that a large proportion of eye samples also fell in the lower portion of the display where the caption was located. Distributions of eye positions were similar for all subjects (see Supplemental Figures S1–S5).

Saccades made within the caption and video regions

Figure 3 compares the preferences to fixate within the caption region to preferences to fixate within the video region of the display. The functions show distributions of the position of the line of sight along

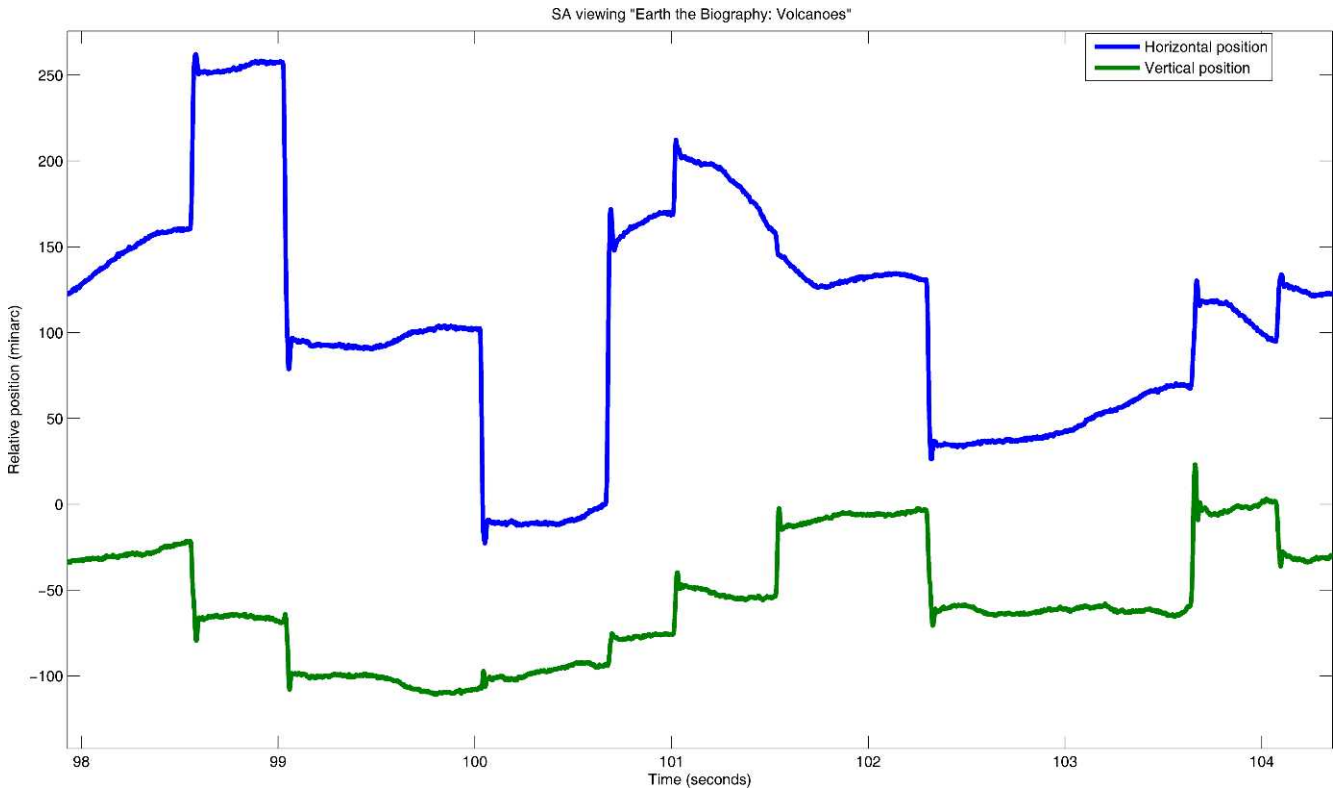

Figure 1. Sample eye trace from subject SA while viewing *Earth the Biography: Volcanoes* over a period of about 6 s.

the vertical meridian at the onset time of saccades for each condition.

Without captions (the audio only and no captions/no audio conditions), the distributions peaked near the center of the display and seldom fell outside the central 2° by 2° region, regardless of the presence of audio. When captions were present, and regardless of whether redundant audio narration was also available (the captions-only and captions + audio conditions), the distributions peaked in the lower region of the display containing the caption with a secondary peak near the display center. The presence of audio was influential in that more saccades shifted over from the caption to the video region in the captions + audio condition than in the captions-only condition. The effect of audio will be analyzed in greater detail in the following section.

A large proportion of time was spent reading captions

The analyses above suggest that there were strong preferences to read captions and that concurrent audio narration reduced these preferences. To examine the effects of the redundant audio narration on reading of the captions more closely, eye movements were examined during time intervals when a caption appeared on the screen. A *caption episode* was defined as the interval between the onset and offset of a caption. In the event that two or more captions were presented with intervening intervals shorter than 210 ms, the captions were considered to constitute a single caption episode. Caption episodes lasted 6 s on average ($SD = 3.5$ s) and there were an average of 12.6 caption episodes ($SD = 3.5$) per video. Given a video duration of 2 min, this works out to caption episodes taking up about 63% of the time the video was presented.

Figure 4 shows in detail how subjects apportioned their time during the caption episodes in both the captions-only and captions + audio conditions. Time was divided into the following categories: (a) time spent within the caption area, including the pause durations between successive saccades and the in-flight time of saccades; (b) time spent either within the video area or traveling between caption and video areas including the pause durations between successive saccades and the in-flight time of saccades; (c) the latency of the first saccade made into the caption area in response to the onset of a caption episode; and (d) intervals in which tracker lock was lost.

Consider first the captions-only condition. When captions were on screen, a large proportion of time—ranging from 48%–78% (average 63%)—was spent in the caption area. In the captions + audio condition, the proportions were still high: 32%–66% (average 44%) of time was spent reading the captions. The subjects who spent the largest proportion of time reading in the captions-only condition also spent the largest proportion of time reading when audio was available (captions + audio condition).

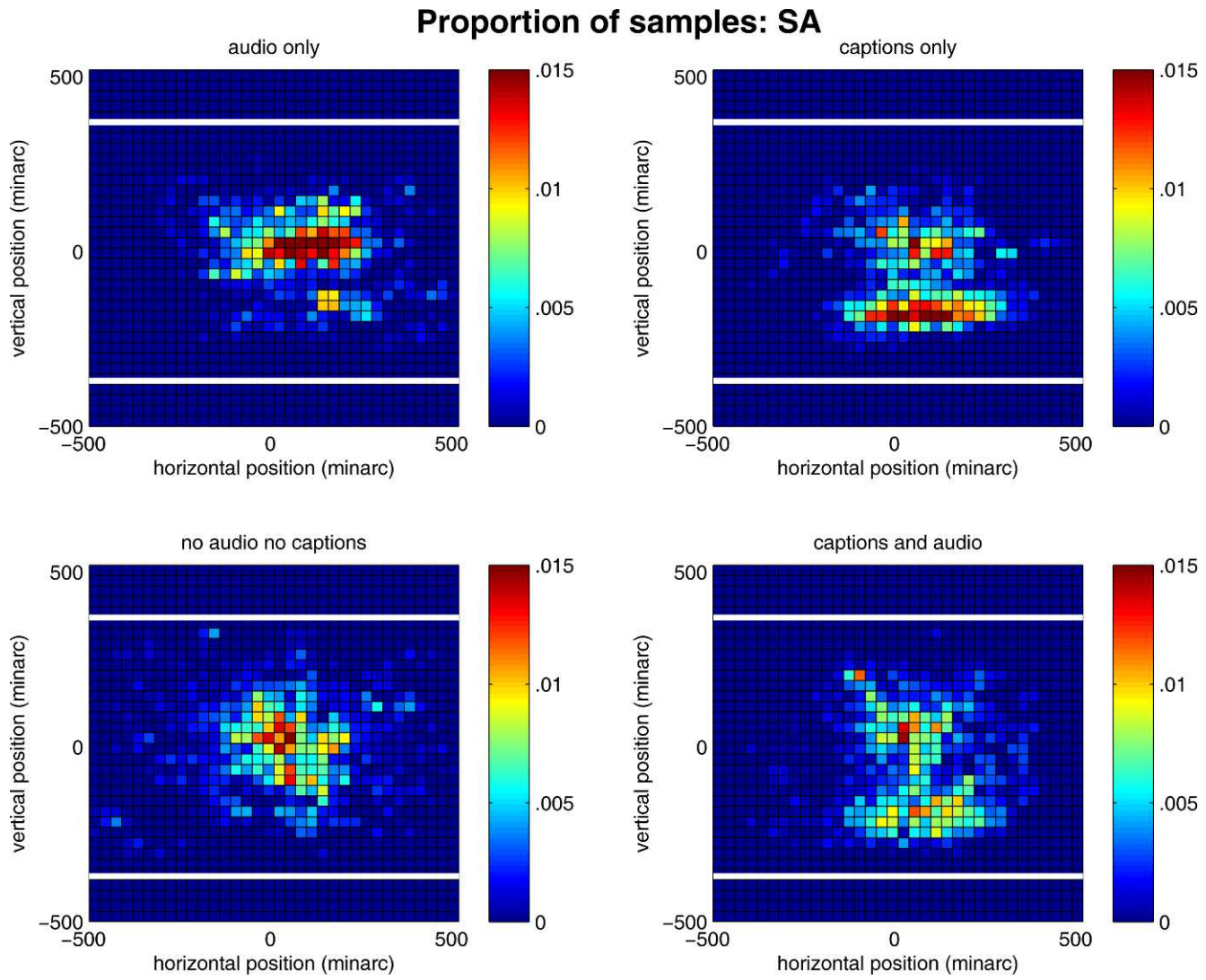

Figure 2. Three-dimensional plot of sampled eye positions during a trial (excluding samples during saccades) for subject SA for each viewing condition (audio only, captions only, neither, and both captions and audio). Data for each condition are pooled across the four clips viewed in that condition. The *x* and *y* axes represent horizontal and vertical position, respectively, in minutes of arc. Color represents the proportion of samples. The horizontal white lines indicate the vertical display boundaries. The horizontal boundaries of the display coincided with the horizontal boundaries of the plot.

These values are surely underestimates. This is because using long trials (2 min), while advantageous for presenting a coherent and developing narrative, restricted the opportunity for intertrial blinks and thus had the expected consequence of frequent intervals in which tracker lock was lost. The total amount of time in which lock was lost varied among the observers, but within an observer, the totals were the same for the two conditions compared in Figure 4 (captions only; captions + audio). If we recompute the percentage of time spent in the caption area, eliminating from consideration the intervals in which lock was lost, the percentage of time spent reading increases to an average of 76% in the captions-only condition and 55% in the captions + audio condition. Note that the magnitude of the difference between the two conditions is unchanged.

Each subject spent less time in the caption area when the audio was present. The difference between the time spent in the caption area in the captions-only condition and in the captions + audio condition (see Figure 4) was significant, paired t test: $t(5) = 3.34$, $p = 0.02$. Further statistical confirmation of the differences between these two conditions was provided by a repeated measures analysis of variance performed on the proportion of time during fixation pauses in which the line of sight was in the caption area while captions were present (arcsine-square-root transform was used

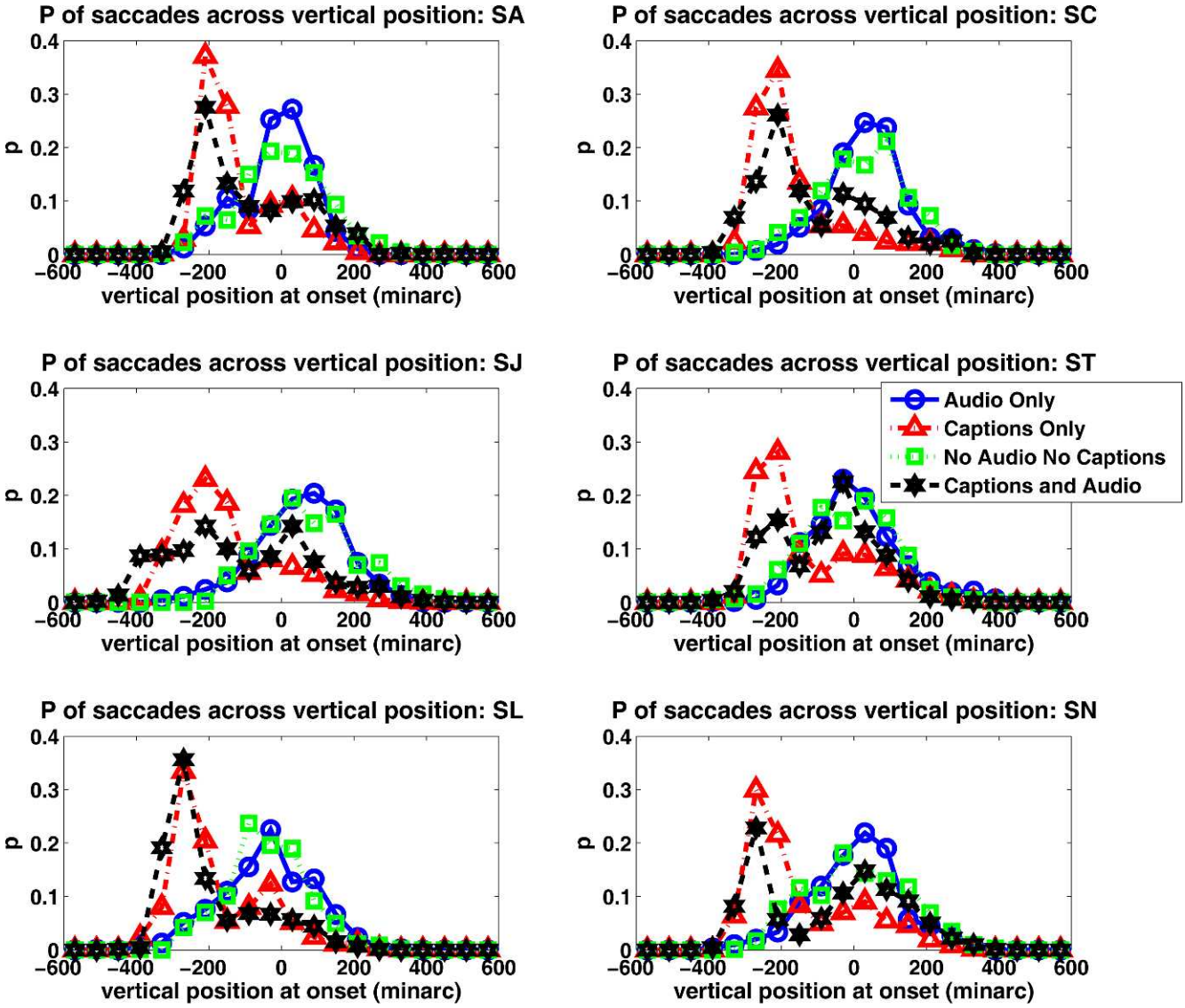

Figure 3. Histograms showing the proportions of saccades originating from different vertical positions for each viewing condition (audio only; captions only; no audio, no captions; both captions and audio). Data from six subjects. Each histogram is based on approximately 2094 to 3554 saccades. Differences across the conditions were significant, $F(3, 22293) = 1885$, $p < 10^{-5}$.

on the proportions). This analysis confirmed the significant differences between the time spent processing the captions between the captions-only and captions + audio conditions, $F(1,586) = 73.45$, $p < 0.0001$.

In summary, a large proportion of time was devoted to reading captions, even with concurrent audio, with less time reading captions when the audio was present.

It is not surprising that captions were read. The surprising finding is that so much time, indeed, any time, was spent reading captions in the presence of redundant audio when the alternative—watching the video while listening to the narration—would seem to ensure no information loss.

Audio narration reduced the duration of visits to the captions

Figure 4 showed that less time was spent on the captions when the audio was present. Was this because audio led to some captions being ignored entirely, read faster, or read only in part?

The mean frequency of visits to the caption area from the video area was the same with or without audio (1.8 visits per caption in the captions-only condition and 1.7 visits per caption in the captions + audio condition). Thus, audio did not lead to selected captions being ignored entirely. In addition, the pattern of saccades made while reading the captions was the same in both conditions. The mean size of forward

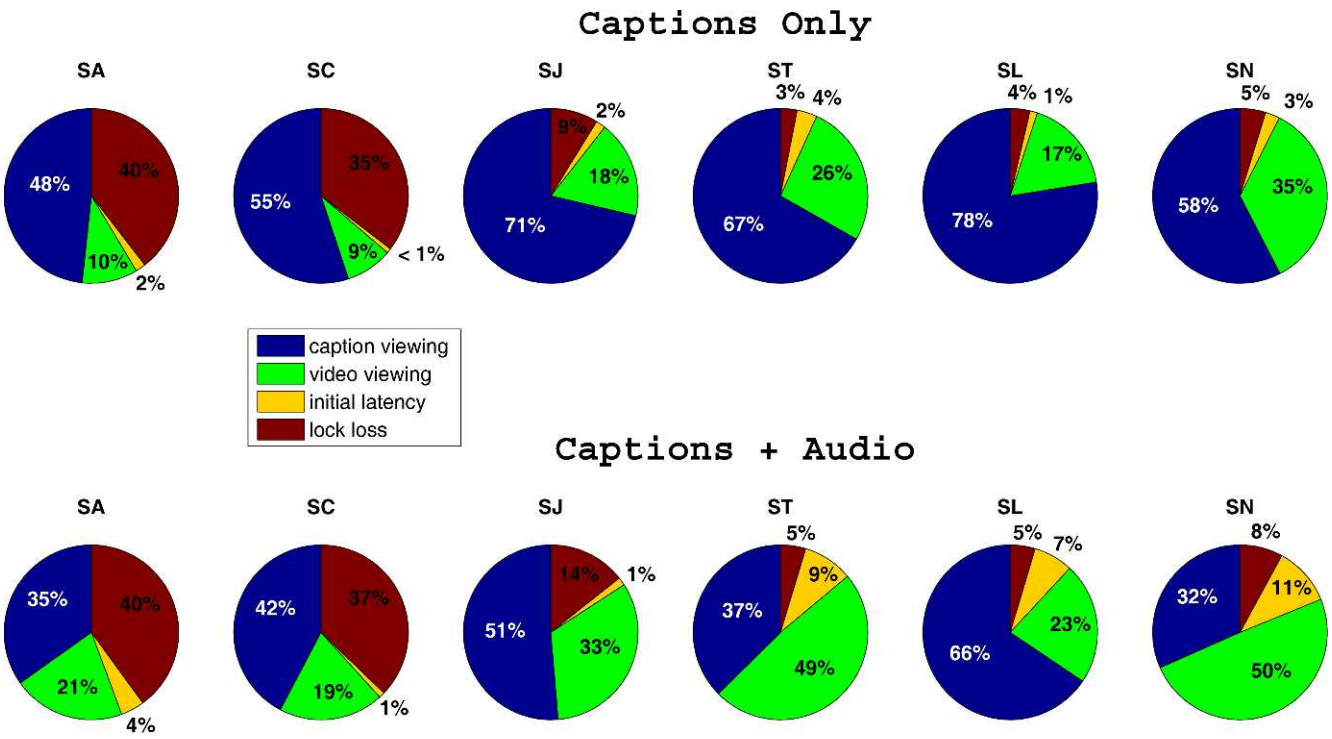

Figure 4. The proportion of time during all caption episodes for a given subject and condition (captions only, top; captions + audio, bottom) that was allocated to: (a) viewing within the caption area, including pause durations between successive saccades and in-flight time of saccades; (b) viewing within the video area or traveling between caption and video areas, including pause durations between successive saccades and in-flight time of saccades; (c) the latency of the first saccade made into the caption area in response to the onset of a caption episode; (d) intervals in which tracker lock was lost.

(rightward) saccades and the mean duration of intersaccadic pauses were each nearly identical with or without audio (Table 1); thus, the information provided by the audio did not speed up the reading nor did reading slow down in an attempt to keep time with the audio stream. The proportion of leftward saccades during reading (that is, regressions and the resetting saccades made to bring the line of sight to a new line of text) were also the same across the two conditions (mean = 36% with audio and 38% without audio), i.e., there is no evidence that the absence of audio encouraged more rereading (Schnitzer & Kowler, 2006).

The major consequence of having concurrent audio was to reduce the duration of the visits to the caption area. The duration of each visit within a given caption episode was defined as the time between a saccade into the caption region and a subsequent saccade out of the caption region. This definition allows for multiple visits during the same caption episode. The average duration of visits was 2 s ($SD = 1.5$) when audio was present and significantly longer, 2.6 s/visit ($SD = 2$), when audio was absent, $t(709) = 4.48$, $p < 0.00001$. The latency of the first saccade into the caption following its initial appearance also was longer when audio was present

	Sizes of saccades within the caption area (minarc)		Intersaccadic pause durations within the caption area (ms)	
Subject	Captions only mean (*SD*) N	Captions and audio mean (*SD*) N	Captions only mean (*SD*) N	Captions and audio mean (*SD*) N
SA	104(58.3) 495	107(51.9) 342	224(84.4) 506	221(103.7) 366
SC	114(71.9) 599	110(69.9) 367	238(119.1) 625	265(144.4) 415
SJ	116(77.8) 749	118(81.1) 505	245(131.1) 808	243(126.3) 557
ST	118(71.7) 663	117(81.8) 308	224(137.5) 722	213(110.4) 363
SL	115(78.3) 595	112(75.8) 586	257(183.2) 624	283(164.6) 624
SN	148(85.0) 544	149(81.3) 266	254(103.7) 640	246(137.8) 346
Mean	119(14.8) 6	119(15.1) 6	240(14.3) 6	245(26.1) 6

Table 1. Characteristics of saccades made within the caption area: Experiment 1. *Notes*: The *caption area* was defined as the portion of the screen >∼140 min arc below screen center.

	Audio only mean (*SD*) N	Captions only mean (*SD*) N	Neither mean (*SD*) N	Both mean (*SD*) N
Vector size (min arc)	96(80) 2132	96(75) 728	102(89) 1972	100.8(83) 1250
Intersaccadic pause duration (ms)	429(310) 2511	424(285) 1091	411(297) 2278	419(280) 1654

Table 2. Characteristics of saccades in the video area: Experiment 1. *Notes*: The *video area* was defined as the top of the display down to ~140 min arc below screen center.

than when the captions were presented without audio, $F(1,391) = 14.87$, $p = 0.0001$ (see Figure 4).

In summary, the concurrent audio did not alter the reading pattern and did not prevent gaze from being attracted to the caption region. Concurrent audio instead led to a portion of the caption being skipped and increased the latency of saccades to the caption.

Differences between saccadic patterns in the caption and video areas

The presence of both captions and video within the same trial provided an opportunity to compare saccadic patterns when reading versus when examining the pictorial material in the video portion of the display. In general, saccades when reading (Table 1) were made at a faster rate than saccades when viewing the video (Table 2). The average pause durations for inspection of the video (~420 ms; see Table 2) were longer than those reported by Dorr et al. (2010) for movie trailers (mean ~ 340 ms). Average sizes and pause durations of saccades in the video area were not affected by the presence of narration (see Table 1).

Effect of audio narration on inspection of the video

Eye fixations clustered near the center of the display (Figures 2 and S1–S5). To compare centering across the four conditions, the two-dimensional scatter of saccadic offset positions for saccades landing within the video area was determined. Two-dimensional scatter was summarized by the bivariate contour ellipse area, BCEA, which represents the size of the area in which saccadic landing positions were found 68% of the time (see Steinman, 1965; Vishwanath & Kowler, 2004; Wu, Kwon, & Kowler, 2010; for other examples of the use of the BCEA for describing the two-dimensional scatter of eye positions in different oculomotor tasks).

Figure 5 (top) shows that BCEA's (averaged across subjects) were about 20 deg^2–25 deg^2, which is about 10%–13% of the total area of the display. BCEA's were largest in the absence of any narration, although the differences were not significant, $F(3, 15) = 1.82$, $p = 0.19$. Inspection of the scatter of landing positions for the different video clips suggested that the absence of any narration (no captions; no audio) may have encouraged somewhat larger scatter for one of the clips (*Earth the Biography: Volcanoes*) which contained frequent scene cuts (Figure 5; middle; bottom graphs). The possibility that narration affects the scatter of landing positions in videos containing many scene cuts will be examined in Experiment 2.

Discussion

The presence of captions had large effects on the eye movements made while viewing videos. The main finding was that a high proportion of the viewing time was devoted to reading the captions even when the captions were redundant with the audio. Characteristics of the saccades made to read the captions, namely, the number of visits to the caption area, the size and frequency of the saccades, and the frequency of leftward saccades, were unaffected by the presence of redundant audio. The redundant audio, however, was not ignored in that it reduced the average duration of visits to the captions, as well as increased the latency of the initial saccades to the captions.

It would seem that the best strategy to use to avoid information loss when both captions and audio are available would be to process both video and audio streams concurrently, that is, keep the line of sight within the video portion of the display and use the audio stream to listen to the narration, ignoring the captions. Instead, all subjects chose to read the captions and spend somewhat less time reading in the presence of audio. Interestingly, once the decision was made to read the captions, the reading saccades themselves were unaffected by the audio. Thus, the effect of narration on strategies was found at a higher level in the decision tree, involving the choice of which region to look at (captions or video), rather than how to look within these regions. A more detailed examination of the possible reasons for the preferences to read captions in the presence of redundant audio will be taken up in the General discussion.

The suggestion that the most prominent effects of narration are found at a higher level in the decision tree is consistent with the observations made about saccades within the video portion of the display. Just as the pattern of reading saccades was unaffected by narration, the pattern of saccades made while examining the video was largely unaffected by the presence of narration. The average size of saccades and average intersaccadic pause durations were the same across the four conditions

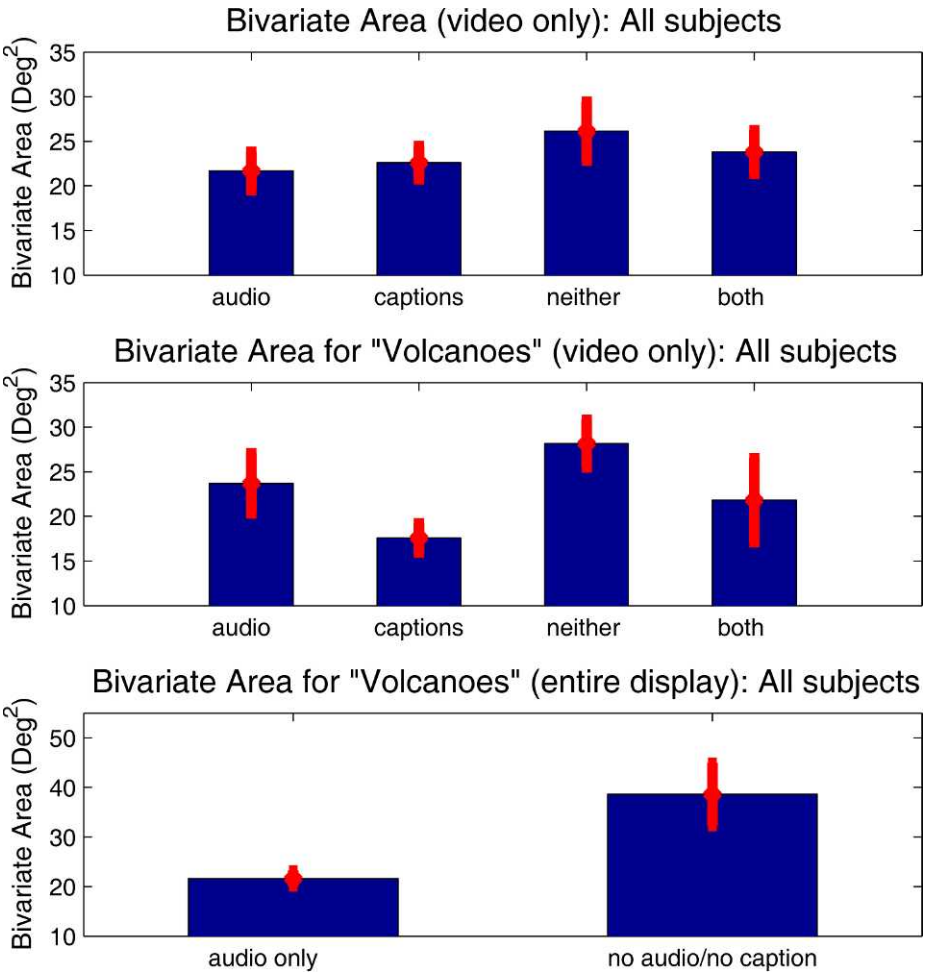

Figure 5. Average scatter of landing locations (bivariate area in degrees squared) across subjects and trials for each viewing condition (audio only, captions only, neither captions nor audio, and captions + audio) based on saccades in the video area of the display (top of display down to about 140 min arc below screen center) for all videos (top panel), for data from *Earth the Biography: Volcanoes* (middle panel), and for data from the entire display while subjects viewed *Earth the Biography: Volcanoes* (bottom panel). Error bars represent ±1 standard error. Each mean in the top panel is based on 24 trials and each mean in the middle and bottom panel is based on six trials.

(Table 2). There was a small tendency for the scatter of fixated positions to increase in the absence of any narration for some video clips, perhaps because of decisions to search across a larger region of the display in an attempt to better interpret the events. Experiment 2 further investigates the role of audio narration on saccades using a larger number of different video clips, shorter duration clips, and no captions.

Experiment 2: Effect of audio narration on eye movements

The purpose of the second experiment was to further investigate the role of audio narration when viewing videos. Given the extensive discussion of salience in prior work on eye movements while viewing videos (see Introduction), the normalized salience levels of fixated locations was also analyzed in addition to saccades sizes, pause durations, and the scatter of saccadic landing positions.

Experiment 2 used shorter duration video clips (15 s). Three conditions were tested: audio narration during the video, audio narration prior to the video, and no narration. A condition with narration prior to the video was included to examine whether any influence of narration is different when information is drawn from memory.

Methods

Stimuli

Nine video clips were tested, each with a duration of 15 s. Clips were taken from each of the following:

NHL: Greatest Moments (Warner Home Video, 2005), *Magnetic Storm* (WGBH, 2003), *Destiny in Space* (Warner Home Video, 2005), *Earth the Biography: Volcanoes* (BBC Worldwide Ltd. Program, 2008), and *Indiana Jones and the Kingdom of the Crystal Skull* (Lucasfilm Ltd., 2008). These videos were chosen because they were more eventful (i.e., contained more scene cuts) than those used in the first experiment allowing us to examine viewing behavior in videos where scenes were changing more frequently, the condition that seemed to encourage scanning over a wider region in Experiment 1. Clips were once again chosen such that no talking heads would be in the frame and so that each clip contained enough information to test for memory of content. Clips were chosen such that none started in the middle of what seemed to be an ongoing event. All clips were edited to remove instances of long ($>\sim$2 s) uneventful pauses. The text of the audio narration was written by the experimenters and consisted of simple sentences describing the events (no conversations or monologues) and lasting as long as the video clip.

Procedure

Three conditions were tested: audio narration prior to video, audio narration concurrent with video, and no narration. Each subject ran in three trials per condition for a total of nine trials. One multiple choice question was asked after each clip in order to provide motivation for paying attention to the video's content. Experiment 2 was run with the same procedures as Experiment 1 except that the multiple choice question in Experiment 2 was presented and answered on the computer, allowing the subject's head to remain in the chinrest between trials. A small proportion of data was lost due to blinks and or loss of tracker lock (audio during, 0.6%, audio prior, 1.4%, and no audio, 0.5%).

Subjects

Subjects were 18 undergraduate and graduate students at Rutgers University serving for either course credit or payment. All had normal hearing and normal or corrected-to-normal (soft contact lenses) vision. All subjects were naïve to the experimental design and hypothesis. The project was approved by the Rutgers University IRB for the protection of human subjects.

Results

Influence of narration on distributions of saccadic landing positions, pause durations, and saccade sizes

Figure 6 compares distributions of saccade sizes, pause durations, and two-dimensional scatter of saccadic landing positions (BCEA) for all three narration conditions: audio narration during the video, audio before the video, and no narration. These characteristics were largely the same across the narration conditions; saccades size: $F(2, 34) = 0.42$, $p = 0.66$; pause duration: $F(1.44, 24.88) = 0.83$, $p = 0.41$; scatter of landing positions: $F(2, 108) = 1.27$, $p = 0.28$. Note that small saccades ($<0.5°$), a category that includes microsaccades, were rare ($<5\%$), similar to what has been found during active visual tasks (Collewijn & Kowler, 2008; Malinov, Epelboim, Herst, & Steinman, 2000).

The effects of narration on saccadic performance over time

Saccadic patterns might change over time as the events evolve or as viewers become more familiar with the overall structure and theme of the depicted events (see also Itti, 2005). To determine the influence of narration over time, the same three measures of saccadic performance shown in Figure 6 were analyzed over separate 5 s epochs of the 15 s video.

Time was quite influential. Sizes of saccades increased over time, $F(2, 4566) = 5.32$, $p = 0.005$, pause durations decreased over time, $F(2, 4566) = 19.09$, $p < 0.0001$, and scatter increased over time for all conditions, $F(2, 477) = 8.92$, $p = 0.0002$ (Figure 7). By contrast, there were no significant main effects of the presence of narration on either saccade size or pause duration and no significant interactions. The largest scatter was found with no narration and the smallest with concurrent narration; however, differences did not reach significance, $F(2, 385) = 2.52$, $p = 0.08$.

Taking into account all three measures of saccadic performance, Figure 7 reveals a pattern of scanning that changed over time in that saccades became larger and more frequent, and landing positions were scattered over a larger region of the display. There was a small tendency for larger scatter in the absence of narration.

Comparison of salience levels of fixated locations

Previous studies of eye movements during the viewing of videos found a correspondence between levels of physical salience and the choice of fixated locations (Berg et al., 2009; Carmi & Itti, 2006; Le Meur et al., 2007). These results showed that physical salience, motion and flicker in particular, can be of some value in predicting which locations are fixated, regardless of whether the correlations between salience and saccades are due to some intrinsic attraction to salient features of the image or because locations with higher salience often correspond to locations of important events (Carmi & Itti, 2006). We compared the salience levels of fixated locations (as in Berg et al., 2009) when viewing the videos with and without narration.

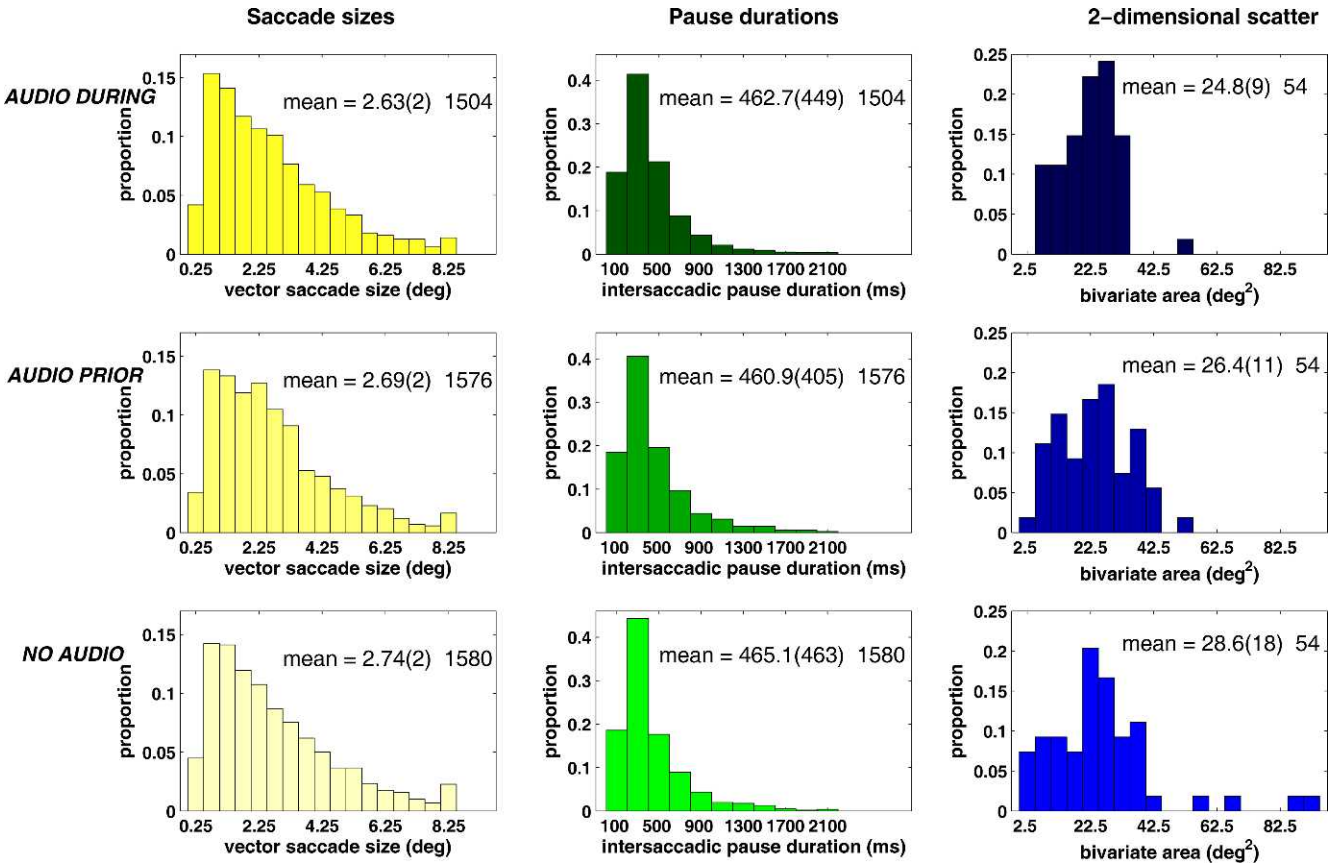

Figure 6. Left: distributions of the proportions of saccade sizes with half-degree bins. Middle: distributions of proportions of pause durations (plots were truncated so that pause durations >2200 ms [<1% of data] are not shown). Right: distributions of the proportions of bivariate contour ellipse areas with 5 deg^2 bins. Each histogram contains data from all subjects and trials.

Salience levels in the video frame corresponding in time to the offset of each saccade were computed using the algorithms available at http://ilab.usc.edu/toolkit/ using default parameter values (center and surround scales, c = {2, 3, 4}, and center-surround scale differences, $\delta = \{3, 4\}$) (Itti, 2005).

Figure 8 shows the average normalized salience values at fixated locations for five different dimensions (flicker, motion, intensity, orientation, and color), and for the combined salience across the dimensions. The results are similar to prior reports (Berg et al., 2009; Carmi & Itti, 2006; Le Meur et al., 2007) in two

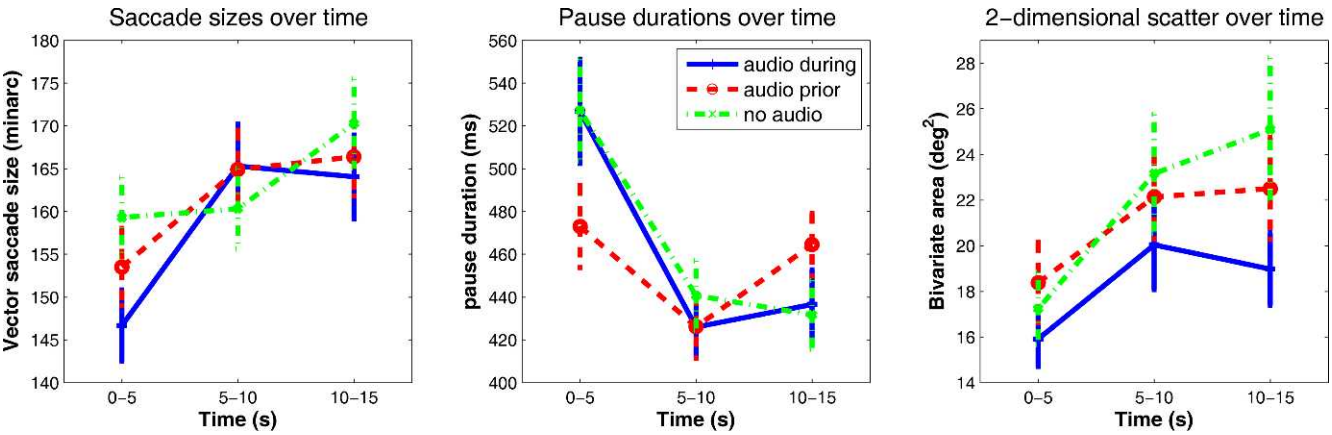

Figure 7. Average saccade sizes, pause durations, and bivariate area in three sequential 5 s epochs spanning the length of the trial (15 s). Each mean is based on 54 trials (18 Ss × 3 trials/subject). Error bars represent ± 1 standard error.

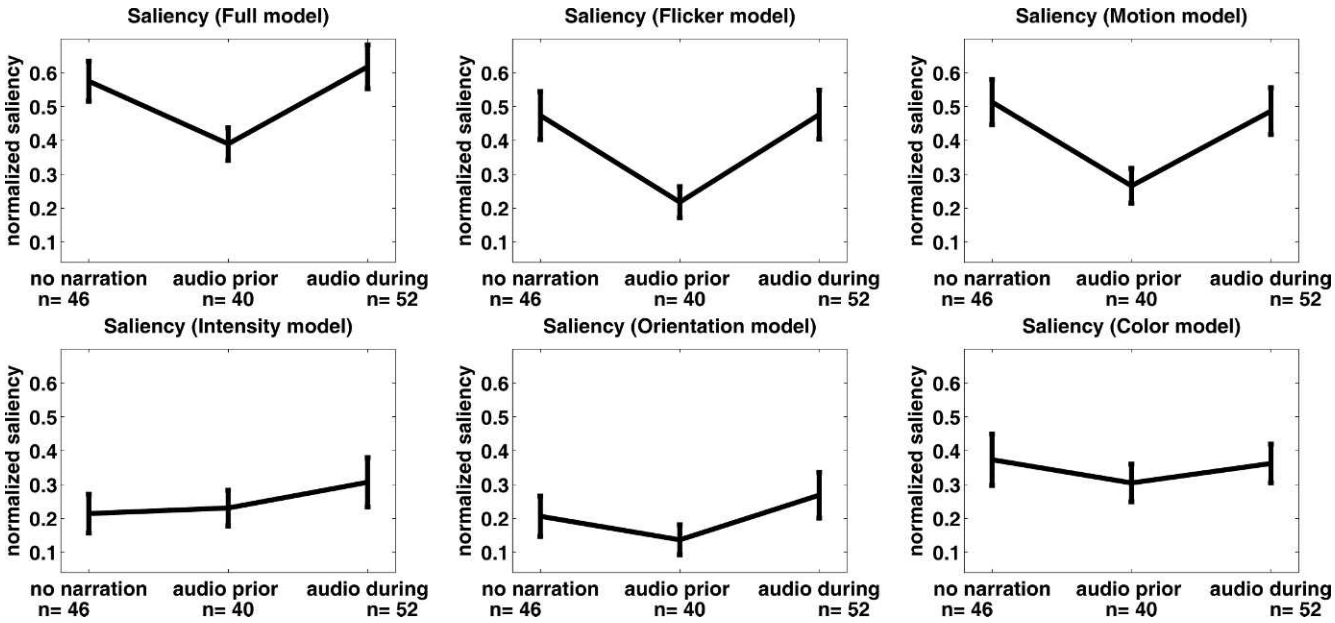

Figure 8. Average normalized saliency values at fixated locations in the three narration conditions. Different panels represent the different feature dimensions and the composite saliency model. Error bars represent ±1 standard error. The number of trials used to compute each mean can be found below the *x* axis.

respects. First, the normalized salience levels were all above zero, which shows a preference to look at locations of higher than average salience. Second, the normalized salience levels at fixated locations were higher for the two dynamic dimensions, flicker and motion, than for intensity, color, and orientation.

Figure 8 also shows that salience levels at fixated locations were lowest when the narration preceded the video for motion, $F(2, 135) = 4.05$, $p = 0.02$, flicker, $F(2, 135) = 4.59$, $p = 0.01$, and composite salience, $F(2, 135) = 3.93$, $p = 0.02$. Given that the salience levels at fixated locations did not differ between the no-narration and concurrent-narration conditions, the results when narration preceded the video may have resulted from influences not directly related to narration, such as distractions due to the attempts to retrieve details of the narration from memory, or reduced levels of interest over time (Carmi & Itti, 2006). Further investigation will be needed to distinguish among these possibilities.

Salience and the scatter of saccadic landing positions

Informal inspection of the eye movements while viewing the videos suggested that the line of sight tended to remain near a selected object for a few seconds while the location of maximum salience shifted much more frequently. Sustained interest in a single object reflects the operation of a top-down strategy that overrides salience (see Koch & Ullman, 1985, for discussion of attempts to build the sustained interest in a single object or location into the predictions of fixation positions).

To test these informal observations, we compared the two-dimensional scatter (BCEAs) of the observers' saccades (Figure 7) to the BCEAs of the locations of highest salience (Figure 9). The locations of highest salience were taken only from the frames corresponding to the offset time of each saccade in each of the three narration conditions (narration during the video, prior to the video, or no narration). Thus, the same number of samples and the same portions of the videos contributed to the computed BCEAs of the subjects (Figure 7) and to the locations of highest salience (Figure 9). We expected the scatter of fixated locations of the observers to be smaller than the scatter of the locations of maximum salience based on informal observations noted above.

Comparing Figures 7 and 9 shows that BCEAs were, as expected, larger for the salience model than for the saccadic landing positions for all feature dimensions, as well as for the composite salience. One unexpected finding was the increase in the scatter of the locations of peak salience increased over time. Effects of time on the model's BCEA were significant ($10^{-9} < p < 10^{-3}$).

The increase in scatter of the locations of peak salience over time (Figure 9) may explain in part the increase in scatter of the saccadic landing positions over time (Figure 7). Some of the increase in the scatter of the saccadic landing positions over time may have been connected to visual factors, perhaps originating from

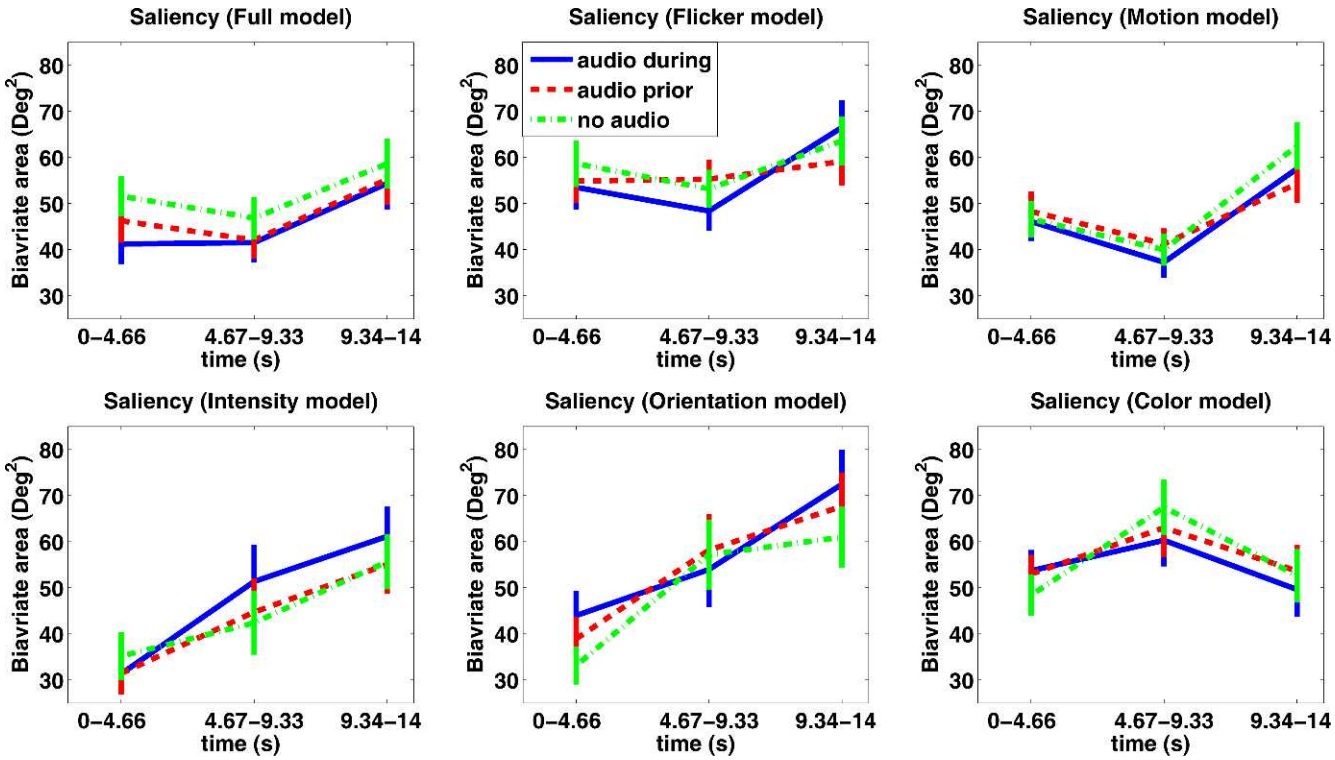

Figure 9. Average bivariate contour ellipse area (deg^2) of location of highest salience in three sequential 5 s epochs spanning the length of the trial (15 s). Each mean is based on 54 trials (18 Ss $\times$ 3 trials/subject). Different panels represent the different feature dimensions and the composite saliency model. Error bars represent $\pm$1 standard error.

the progression of events in the video clip, rather than exclusively to an evolving cognitive understanding of the depicted events.

Influence of scene cuts on the probability of a saccade occurring

A final analysis examined effects of abrupt scene cuts on saccadic production. Abrupt transitions and abrupt onsets have been associated with brief inhibition in the production of saccades (Henderson & Smith, 2009; Reingold & Stampe, 2002; 2004; van Dam & van Ee, 2005). Abrupt transitions between static scenes or between different videos have also been associated with changes in saccadic patterns, such as increased centering (Dorr et al., 2010) and a greater influence of visual salience on landing positions (Carmi & Itti, 2006). We examined whether reports of brief disruptions to the production of saccades following transient events would also apply to the transients that were part of the flow of events in a professional video.

In order to examine the influence of scene cuts on saccades, the probability of a saccade was computed for the intervals (duration = 660 ms) preceding and following cuts. (Intervals containing blinks or loss of eye-tracker lock were omitted.) Analyses were restricted to situations where the video frame immediately preceding the cut was free of saccades and in which the time between any pair of cuts was at least 1 s. Analyses were also restricted to the first saccade after each cut in order to prevent any confounding influences from saccade-produced transients.

Figure 10 shows that, consistent with prior results cited above, scene cuts tended to delay saccades. The probability of a saccade occurring during frames prior to a cut was about 0.09. Saccade rates decreased to about 0.07 following the cut and steadily increased to about 0.12 over the next several hundred milliseconds. The slope of the best fit line describing the probability of a saccade prior to a cut was not significantly different from zero, $t(7) = -0.04$, $p = 0.97$, whereas the slope of the best fit line describing the probability of a saccade following a cut was significantly greater than zero, $t(8) = 5.38$, $p = 0.001$. This pattern suggests that the scene cut, like other transient events, produced a brief inhibition or delay of saccades.

Discussion

Experiment 2 showed that over time saccades became larger, intersaccadic pause durations became

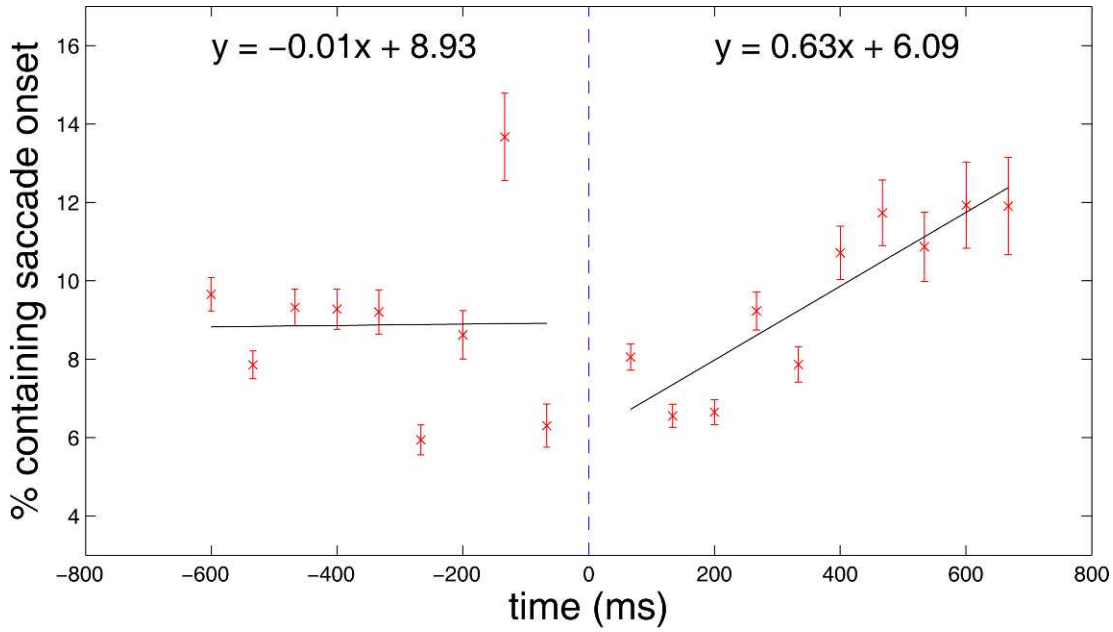

Figure 10. Percentage of abrupt changes containing saccade onsets in each of nine sequential 66 ms epochs prior to an abrupt scene cut and 10 epochs following a cut. The best fit line for each set of data (pre and post cut) are also shown. Error bars represent ±1 standard error. The vertical dashed blue line represents the onset of the frame immediately following a cut. After the first saccade onset occurred in a given sequence of frames, the frames immediately following were not analyzed (see text).

shorter, and saccadic landing positions were scattered across a larger spatial region of the display. Narration did not have prominent effects on these global measures of saccadic performance. At most, there were small increases in scatter in the absence of narration that did not reach significance.

Physical salience, particularly motion and flicker, had some value in predicting landing positions under all narration conditions, but effects were not large. Interestingly, the scatter of locations of maximum salience increased over time along with the scatter of saccadic landing positions. This relationship suggests that visual factors within the video, and not solely changes in the cognitive understanding of the events, could have accounted for some of the increase in the scatter of the saccadic landing positions over time.

General discussion

The present study examined how the presence of narration, either audio or captions, influenced strategies of saccadic planning while viewing videos. The two main questions were: (a) How do people apportion viewing time when they have to decide between viewing the video and reading the caption, and (b) How does the presence of narration (audio or caption) influence the pattern of saccades?

How do people apportion viewing time when they have to decide between viewing the video and reading the caption?

A surprising and consistent trend in Experiment 1 was the tendency to spend a large proportion of the time reading the captions, regardless of whether the redundant audio was present.

On the face of it, this strategy makes little sense. The captions provided no new information that was not already in the audio. Moreover, reading captions took the line of sight away from the video, leading to the potential and completely unnecessary risk of losing information. Preferences to read captions rather than listening to audio narration may be optimal when there are only two sources of information to monitor (text and audio narration) because reading text leads to faster processing (Levy-Schoen, 1981). With three sources of information to monitor, text, audio narration, and video, the strategies would have to be more complicated. Assuming that saccadic behavior has some rational basis (an assumption that is questioned from time to time, e.g., Viviani, 1990), we suggest below three factors that could have contributed to preferences to read the captions when redundant audio was present.

One possible reason for the preference to read the captions when audio was present is an attempt to integrate visual (captions) and audio (narration) information into a single processing stream. People often

have the subjective impression when viewing movies with subtitles that the two sources are integrated and that they are reading the text in synchrony with the audio stream. True temporal correspondence of captions and audio at all processing stages seems unlikely, however, given that processing of spoken language is slower than reading (Levy-Schoen, 1981), and reading of the captions did not slow down to keep time with the audio (Table 1). But true temporal correspondence at all processing stages is not necessary. Mechanisms of multisensory integration can create the impression of temporal correspondence even with asynchronous visual and auditory signals (for review, see Recanzone, 2009). The fact that the first saccade made into the caption area when the audio narration was present was delayed relative to the first saccade made without audio is consistent with the idea that saccades were drawn to the captions after auditory processing of the narration reached some criterion level, perhaps as part of an attempt to integrate the visual (text) with the auditory narration. Note we are not proposing a reason why people might want to integrate these signals but rather making the point that integration across modalities is an important characteristic of sensory systems (Ghazanfar & Schroeder, 2006), and eye movements may, in some cases, be called upon to facilitate such integration.

Another possible reason for why redundant captions were read is habit. Prior work has shown that the line of sight is drawn to text present in static pictures even when the text is not particularly useful (Cerf et al., 2009) and even when the text shows nonsense words or words in an unfamiliar language (Wang & Pomplun, 2012). These investigators suggested that people have a habit of reading any available text because our prior experiences have taught us that text often conveys important information (Cerf et al., 2009; Wang & Pomplun, 2012). Habits may be difficult to reverse or alter. Devoting effort or resources to altering a habitual strategy could be more of a hindrance to performance than allowing a proportion of relatively useless saccades (Araujo et al., 2001; Hooge & Erkelens, 1998).

Finally, captions may have been read as part of a strategy of directing the line of sight to the region (caption or video) judged to have the greatest momentary value or importance. This view implies that people are continually judging the value of information coming from all available sources, in this case, the video, the caption, and the audio stream. It may have been, for example, that the risk of losing some information from the video was judged to be minimal relative to the advantages of spending some time reading the informationally-dense and reliable visual text. However, the finding that visits to the captions were shorter in duration when the audio was present (Figure 4) suggests that the strategies used to apportion viewing time did not uniformly give preference to the captions but instead may have been influenced by the processing of the incoming information. For example, the appearance of new or potentially interesting visual details, gleaned from eccentric vision, could have prompted a shift of the line of sight out of the captions and into the video region. Alternatively, subjects may have strategically traded off reading and listening within a given caption episode, perhaps spending more time on the captions as part of selected attempts to confirm, accelerate, or supplement the information obtained from the audio narration. The suggestion that saccadic decisions are based on the judged momentary importance or value of different regions has been made previously in studies of the saccades made to guide motor behavior (Ballard, Hayhoe, & Pelz, 1995; Epelboim et al., 1995; Flanagan & Johansson, 2003; Hayhoe & Ballard, 2005; Johansson et al., 2001).

Any of the above processes, sensory integration, habit, or active selection, operating separately or together, may have contributed to decisions about when to read the captions and when to concentrate on the video.

How does the presence of narration influence the patterns of saccades?

The presence of narration had little influence on the sizes of saccades, intersaccadic pause durations, or the scatter of landing positions made to inspect the non-captioned portions of the video. The only suggestion of an effect of narration was a small increase in the scatter of landing positions in the absence of any narration (Figures 5 and 7), as if the absence of the cues about the depicted events prompted search over a wider region.

Saccadic patterns were dominated by visual factors. For example, except for reading the captions (Experiment 1), the line of sight rarely strayed outside the central 10%–13% of the display. In addition, saccades were attracted to regions with high levels of motion and flicker. Since such regions are generally associated with important or interesting events, including the motion of living things, the attraction to such regions could be due to intrinsic motivation to follow the events depicted in the video (Elazary & Itti, 2008). Saccades were also affected by the passage of time, with saccades over time becoming larger, more frequent, and more scattered over the display (Figure 7). At least some of these temporal changes could have been due to visual factors, namely, the increased scatter of regions of high levels of temporal change over time (Figure 9).

Under some conditions, narration might be expected to override centering tendencies or any attraction to regions of high temporal changes. Instructional videos (Daniel & Tversky, 2012) are a good example. Instructions could prompt a search for selected objects

or details and thus would change saccadic patterns and strategies to some extent. Our results show that the presence of narration, by itself, did not induce large changes in the global patterns of saccades in contrast to the differences that have been observed between saccades made to inspect pictures versus videos (Dorr et al., 2010). Visual factors, including those that highlight important display regions, not narration, determined the global characteristics of saccades.

Conclusion

The present results showed that when watching videos with narration, the line of sight was drawn to captions, even in the presence of redundant audio narration. The presence of narration, by itself, did not alter global characteristics of saccades to inspect the video (sizes, pause duration, scatter of landing positions), except for small increases in scatter of saccadic landing positions in the absence of any narration. Saccades were drawn to the display center, to regions containing temporal change (motion or flicker), and to captions.

These results suggest that strategies of saccadic planning while viewing narrated videos are affected by several factors, including the motivation to look at regions with the highest information value (actions or events), an inherent attraction to text (a region of densely-packed important information content), and attempts to facilitate integration of visual text with auditory accompaniment. Some of the factors influencing saccadic decisions may reflect the active continual evaluation of incoming information in order to direct the line of sight to regions where the most interesting or useful details will be found at any given moment. Other factors, however, such as the attraction to text or to regions with high levels of temporal change, may reflect operating principles that are followed, not necessarily because of their judged immediate value but because of built-in preferences or learned habits. Although reliance on any built-in preferences or habits can lead to a suboptimal use of time, the advantage is that the momentary cognitive load attached to decision-making is reduced.

Keywords: eye movements, saccadic eye movements, saccades, cognition, videos, movies, narration, captions, salience models, event perception, multi-sensory integration, reading

Acknowledgments

Supported by NSF DGE 0549115 and NIH EY015522. We thank Edinah Gnang, Jacob Feldman, Brian Schnitzer, Elizabeth Torres, Richard Knowles, Gaurav Kharkwal, and Karin Stromswold for valuable discussions and for comments on earlier drafts of the manuscript and Alexander Diaz for assistance with Experiment 1.

Commercial relationships: none.
Corresponding author: Nicholas M. Ross.
Email: nickross@rci.rutgers.edu.
Address: Department of Psychology, Rutgers University, Piscataway, NJ, USA.

References

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: The integration of complex speech and scenes during language comprehension. *Acta Psychologica, 137*(2), 208–216.

Araujo, C., Kowler, E., & Pavel, M. (2001). Eye movements during visual search: The costs of choosing the optimal path. *Vision Research, 41,* 3613–3625.

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Cognitive Neuroscience, 7,* 66–80.

Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., & Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision, 9*(5):19, 1–15, http://www.journalofvision.org/content/9/5/19, doi:10.1167/9.5.19. [PubMed] [Article]

Bisson, M. J., van Heusen, W., Conklin, K., & Tunney, R. (2011, August). *Processing of foreign language films with subtitles: An eye-tracking study*. Poster session presented at the European Conference on Eye Movements, Marseille, France.

Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research, 46,* 4333–4345.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12):10, 1–15, http://www.journalofvision.org/content/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article]

Collewijn, H., & Kowler, E. (2008). The significance of microsaccades for vision and oculomotor control. *Journal of Vision, 8*(14):10, 1–21, http://www.journalofvision.org/content/8/14/20, doi:10.1167/8.14.20. [PubMed] [Article]

Daniel, M. P., & Tversky, B. (2012). How to put things

together. *Cognitive Processes, 13*(4), 303–319, doi: 10.1007/s10339-012-0521-5.

Dorr, M., Martinetz, T., Gegenfurtner, K., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10*(10): 28, 1–17, http://www.journalofvision.org/content/10/10/28, doi:10.1167/10.10.28. [PubMed] [Article]

d'Ydewalle, G., Praet, C., Verfaillie, K., & Van Rensbergen, J. (1991). Watching subtitled television. Automatic reading behavior. *Communication Research, 18,* 650–666.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3):3, 1–15, http://www.journalofvision.org/content/8/3/3, doi: 10.1167/8.3.3. [PubMed] [Article]

Epelboim, J. L., Steinman, R. M., Kowler, E., Edwards, M., Pizlo, Z., Erkelens, C. J., et al. (1995). The function of visual search and memory in sequential looking tasks. *Vision Research, 35*(23-24), 3401–3422.

Epelboim, J., & Suppes, P. (2001). A model of eye movements and visual working memory during problem solving in geometry. *Vision Research*, *41*(12), 1561–1574.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision*. New York: Oxford University Press.

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769–771.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences, 10,* 278–285.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

Henderson, J. M., & Smith, T. J. (2009). How are eye fixation durations controlled during scene viewing? Further evidence from a scene onset delay paradigm. *Visual Cognition, 17,* 1055–1082.

Hooge, I., & Erkelens, C. J. (1998). Adjustment of fixation duration in visual search. *Vision Research, 38,* 1295–1302.

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction, 20*(2), 172–176.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*(6), 1093–1123.

Johansson, R. S., Westling, G., Bäckström, A., & Flanagan, J. R. (2001). Eye-hand coordination in object manipulation. *Journal of Neuroscience, 21*(17), 6917–6932.

Kibbe, M. M., & Kowler, E. (2011). Visual search for category sets: Tradeoffs between exploration and memory. *Journal of Vision, 11*(3):14, 1–21, http://www.journalofvision.org/content/11/3/14, doi:10.1167/11.3.14. [PubMed] [Article]

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*, 219–227.

Kowler, E. (2011). Eye movements: The past 25 years. *Vision Research, 51*(13), 1457–1483.

Kowler, E., Pizlo, Z., Zhu, G.-L., Erkelens, C. J., Steinman, R. M., & Collewijn, H. (1992). Coordination of head and eyes during the performance of natural (and unnatural) visual tasks. In A. Berthoz, W. Graz, & P. P. Vidal (Eds.), *The head neck sensory motor system* (pp. 419–426). Oxford: Oxford University Press.

Land, M., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41,* 3559–3566.

Land, M. F., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception, 28,* 1311–1328.

Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research, 47*(19), 2483–2498.

Levy-Schoen, A. (1981). Flexible and/or rigid control of oculomotor scanning behavior. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception* (pp. 299–316). Hillsdale, NJ: Erlbaum.

Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision, 10*(2):4, 1–11, http://www.journalofvision.org/content/10/2/4, doi:10.1167/10.2.4. [PubMed] [Article]

Malinov, I. V., Epelboim, J., Herst, A. N., & Steinman, R. M. (2000). Characteristics of saccades and vergence in two kinds of sequential looking tasks. *Vision Research, 40*(16), 2083–2090.

Motter, B. C., & Belky, E. J. (1998). The guidance of eye movements during active visual search. *Vision Research, 38,* 1805–1815.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*(7031), 387–391.

Pantelis, J. B., Cholewiak, S. A., Ringstad, P., Sanik, K., Weinstein, A., Wu, C., & Feldman, J. (2011). Perception of mental states in autonomous virtual agents. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference*

of the Cognitive Science Society (pp. 1990–1195). Austin, TX: Cognitive Science Society.

Pelz, J. B., & Canosa, R. L. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research, 41,* 3587–3596.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 191–201.

Recanzone, G. (2009). Interactions of auditory and visual stimuli in space and time. *Hearing Research, 258,* 89–99.

Reingold, E. M., & Stampe, D. M. (2002). Saccadic inhibition in voluntary and reflexive saccades. *Journal of Cognitive Neuroscience, 14*(3), 371–388.

Reingold, E. M., & Stampe, D. M. (2004). Saccadic inhibition in reading. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(1), 194–211.

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). Explaining the modality and contiguity effects: New insights from investigating students' viewing behavior. *Applied Cognitive Psychology, 24*(2), 226–237.

Schnitzer, B. S., & Kowler, E. (2006). Eye movements during multiple readings of the same text. *Vision Research, 46,* 1611–1632.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45*(4), 447–481.

Steinman, R. M. (1965). Effect of target size, luminance, and color on monocular fixation. *Journal of the Optical Society America, 55,* 1158–1165.

Steinman, R. M., Menezes, W., & Herst, A. N. (2006). Handling real forms in real life. In M. R. M. Jenkin & L. R. Harris (Eds.), *Seeing spatial form* (pp. 187–212). New York: Oxford University Press.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*(5):5, 1–23, http://www.journalofvision.org/content/11/5/5, doi:10.1167/11.5.5. [PubMed] [Article]

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766–786.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition, 73,* 89–134.

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7):4, 1–16, http://www.journalofvision.org/content/9/7/4, doi:10.1167/9.7.4. [PubMed] [Article]

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research, 43*(3), 333–346.

van Dam, L. C., & van Ee, R. (2005). The role of (micro)saccades and blinks in perceptual bi-stability from slant rivalry. *Vision Research, 45*(18), 2417–2435.

Vig, E., Dorr, M., & Barth, E. (2009). Efficient visual coding and the predictability of eye movements on natural movies. *Spatial Vision, 22*(5), 397–408.

Vishwanath, D., & Kowler, E. (2004). Saccadic localization in the presence of cues to three-dimensional shape. *Journal of Vision*, *4*(6):4, 445–458, http://www.journalofvision.org/content/4/6/4, doi:10.1167/4.6.4. [PubMed] [Article]

Viviani, P. (1990). Eye movements in visual search: Cognitive, perceptual and motor control aspects. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes* (pp. 353–393). Amsterdam: Elsevier.

Wang, H.-C., & Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *Journal of Vision, 12*(6):26, 1–17, http://www.journalofvision.org/content/12/6/26, doi:10.1167/12.6.26. [PubMed] [Article]

Wilder, J. D., Kowler, E., Schnitzer, B. S., Gersch, T. M., & Dosher, B. A. (2009). Attention during active visual tasks: Counting, pointing and simply looking. *Vision Research, 49,* 1017–1031.

Wu, C.-C., Kwon, O.-S., & Kowler, E. (2010). Fitts's Law and speed/accuracy trade-offs during sequences of saccades: Implications for strategies of saccadic planning. *Vision Research, 50*(21), 2142–2157.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127,* 3–21.