Human Eye Movements After Viewpoint Shifts in Edited Dynamic Scenes are Under Cognitive Control

Raphael Seywerth[1], Christian Valuch[1,2], and Ulrich Ansorge[1]

[1] Faculty of Psychology, University of Vienna, Austria

[2] Faculty of Biology and Psychology, University of Göttingen, Germany

ABSTRACT

We tested whether viewers have cognitive control over their eye movements after cuts in videos of real-world scenes. In the critical conditions, scene cuts constituted panoramic view shifts: Half of the view following a cut matched the view on the same scene before the cut. We manipulated the viewing task between two groups of participants. The main experimental group judged whether the scene following a cut was a continuation of the scene before the cut. Results showed that following view shifts, fixations were determined by the task from 250 ms until 1.5 s: Participants made more and earlier fixations on scene regions that matched across cuts, compared to nonmatching scene regions. This was evident in comparison to a control group of participants that performed a task that did not require judging scene continuity across cuts, and did not show the preference for matching scene regions. Our results illustrate that viewing intentions can have robust and consistent effects on gaze behavior in dynamic scenes, immediately after cuts.

KEYWORDS

attention, eye tracking, fixations, editing, continuity, movies, dynamic scenes

INTRODUCTION

Edited dynamic scenes, such as films, newscasts, television shows, and all sorts of edited videos are widely prevalent in human environment. Such edited videos (as we might call them) contain frequent cuts, which are abrupt global changes of video content, occurring every few seconds and connecting different video takes. Despite the high prevalence of edited videos, only few studies systematically explored how eye movements might be affected by cuts (e.g., Carmi & Itti, 2006a; Germeys & D'Ydewalle, 2007). As outlined in the Background section below, previous studies suggested that in early time periods following scene cuts, cognitive top-down influences on eye movements are rather limited. Here, we chose an experimental approach to uncover such cognitive top-down influences on eye movements following scene cuts: We looked at how spatio-temporal eye movement patterns following scene cuts are influenced by specific task goals (Yarbus, 1967). Our study illustrates that humans exert robust cognitive control over their eye movements. Immediately after cuts, viewers can selectively fixate on scene regions that contain the most task-relevant information.

To start with, eye fixations enable humans to perceive visual information using highly accurate *foveal vision*. Yet in each moment, foveal vision captures only a small portion of the available information. To overcome this inherent selectivity, humans make several fixations per second and select or sample visual information from different spatial locations (Rayner, 2009). Thus, mechanisms of selective visual attention are tightly connected to fixation location selection (Deubel & Schneider, 1996) and ensure that relevant information is sampled and available for cognitive processing and behavior (Land & Tatler, 2009). In edited dynamic scenes, accurate and timely fixations on behaviorally relevant content are of particular importance because specific information is only transiently accessible and the video content undergoes frequent dynamic changes. A more solid understanding of the degree to which human viewers have cognitive control over their fixations in edited dynamic scenes would be an important step towards the improvement of technological applications that rely on videos and human attention. Examples are the design of viewer-acceptable video-coding

Corresponding author: Raphael Seywerth, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Wien, Austria. Phone: +43 1 4277 22 009

E-mail: office@seywerth.net

standards (Adzic, Kalva, & Furht, 2013; Salomon, 2004) or, more generally, display devices that are aware of and optimized for the viewer's attention (Ferscha, Paradiso, & Whitaker, 2014). Moreover, the principles that determine attention and eye movements in edited dynamic scenes could inspire the development of graphical user interfaces (May, Dean, & Barnard, 2003; Valuch, Ansorge, Buchinger, Patrone, & Scherzer, 2014).

Background

RECOGNITION TASKS UNCOVER COGNITIVE CONTROL

Cognitive influences on eye movements can be uncovered by manipulating the viewing task between different groups of experimental participants (Smith & Mital, 2013; Yarbus, 1967). In static images, recognition tasks have proven particularly useful for this purpose (Castelhano, Mack, & Henderson, 2009). For example, one group of participants can be asked to first memorize a series of images in a learning block and then discriminate between novel and familiar images in a transfer block (Foulsham & Kingstone, 2013; Valuch, Becker, & Ansorge, 2013). To identify how task goals modulate gaze behavior, a second group of participants can be presented with the identical series of images but with different task instructions (Valuch et al., 2013). Differences in eye movement measures between the groups of participants can then be attributed to cognitive influences (Castelhano et al., 2009).

A recent study compared fixation patterns between two groups of participants, which were both shown the same photographs of real-world scenes (Valuch et al., 2013). The main experimental group was instructed to memorize the photographs in a learning block, and then discriminate between familiar and novel scenes in a transfer block. The control group was instructed to freely view the photographs in both blocks, without the need to recognize familiar scenes. Crucially, some of the scenes that were repeated in the transfer block underwent a panoramic view shift relative to the learning block. In these shifted views, either the left or the right half of the photograph matched with the view from the learning block. In other words, the image content in the transfer block overlapped by exactly 50% with the image content that was previously presented in the learning block. The central result was that participants from the experimental group, actively recognizing familiar images, fixated significantly more often and longer on the matching scene regions than on the nonmatching, novel scene regions. In contrast, the control group, which viewed exactly the same scene photographs in both blocks but was not required to actively recognize the images, did not show this effect in their fixation locations. Other experiments have delivered a tentative explanation for the tendency to fixate on matching scene regions during recognition: In order to accurately recognize whether a scene is familiar or not, humans must direct their foveal vision to scene details that were present and fixated during learning of the scene (Foulsham & Kingstone, 2013; Valuch et al., 2013). Hence, the bias in the spatial fixation distribution in the recognition group towards overlapping scene regions reflected the degree to which viewers exerted cognitive top-down control over their eye movements.

COGNITIVE TOP-DOWN INFLUENCES IN DYNAMIC SCENES

Cognitive top-down influences are well established for static scenes, but only few studies attempted testing them in the context of dynamic scenes (e.g., Germeys & D'Ydewalle, 2007; Loschky, Larson, Magliano, & Smith, 2015; Smith & Mital, 2013). The majority of research suggests that cognitive top-down factors play a negligible to minor role for explaining eye movements in videos (Carmi & Itti, 2006a; Mital, Smith, Hill, & Henderson, 2011). Previous studies suggest that a large part of the spatial variance in fixations in videos could be explained by a strong generic viewing bias towards the center regions of a video (Tseng, Carmi, Cameron, Munoz, & Itti, 2009). In addition, fixations seem to correlate substantially with salient visual image features, such as strong motion (Mital et al., 2011), luminance and color contrasts (Carmi & Itti, 2006b), or spatio-temporal novelty (Itti & Baldi, 2009). Of particular importance for our present study, research suggests that any residual cognitive top-down influences are muted during the very first second following scene cuts in edited dynamic scenes (Carmi & Itti, 2006a; Smith & Mital, 2013): After cuts, studies reported particularly high correlations between fixation locations and salient visual characteristics (Carmi & Itti, 2006a), or generic biases towards the screen center, with cognitive top-down influences only slowly taking over gaze control as the video progresses (Smith & Mital, 2013).

One explanation for the lack of evidence for cognitive top-down influences on early fixation selection after cuts is the choice of stimuli and viewing tasks of previous studies. To start with, not all types of videos are equally suited to study differences due to cognitive top-down influences. Hollywood-like video material minimizes interobserver viewing variability and elicits particularly strong clustering of fixations in the center of the image area (Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Goldstein, Woods, & Peli, 2007). This is possibly due to tailored editing, aimed at attracting the gaze and attention of most viewers and in a similar way to the most important content in images. In contrast, more naturalistic videos of real-world scenes are known to invite higher spatial fixation variability and leave more room for detecting differences caused by cognitive top-down influences (cf. Dorr et al., 2010). Also, if participants are asked to "freely view" a series of videos, they often orient their gaze towards salient visual features (Mital et al., 2011), but this does not indicate a causal effect of visual salience on gaze control (e.g., Nuthmann & Henderson, 2010), and specific task goals could drastically change such relationships (e.g., Acik, Onat, Schumann, Einhäuser, & König, 2009; Fuchs, Ansorge, Redies, & Leder, 2011). To date, there is a general shortage of studies that would include specific task instructions, as well as suitable control conditions that could justify conclusions about causal influences of visual saliency, independent of cognitive influences, on fixation selection (Tatler, Hayhoe, Land, & Ballard, 2011).

For example, if an experiment includes only cuts between completely unrelated scenes, fixations appear to correlate more strongly with visual characteristics in the very first second following the cut (Carmi & Itti, 2006a). Notably, in the absence of an important comparison condition—cuts between related scene images—conclusions

about the general absence of cognitive top-down influences are difficult. This is problematic because edited material very often includes cuts between related scenes. A common example is viewpoint shifts, where the same scene is shown from two different camera perspectives before and after a cut. With such cuts, cognitive top-down influences can be expected, because viewers might recognize or actively search for familiar (remembered) previous visual scene content to understand how the two different scene views relate to one another (Ansorge, Buchinger, Valuch, Patrone, & Scherzer, 2014; Hochberg & Brooks, 1996). Indeed, using a novel type of recognition task, two recent eye tracking studies suggested that eye movements might be differently affected by cuts that connect visually unrelated scenes as opposed to cuts that connect two visually related views on the same scene (Valuch et al., 2014; Valuch & Ansorge, 2015). Viewers are able to faster recognize movie continuations after cuts between related scenes relative to cuts between visually unrelated scenes (Valuch et al., 2014). Moreover, if viewers do not know at which of two alternative locations a movie will continue after a cut, they make faster eye movements to the correct location after cuts between related scenes than after cuts between visually unrelated scenes (Valuch & Ansorge, 2015). While these studies looked at the temporal properties of the initial gaze orientation after cuts, they did not explore whether cuts between related scene views entail systematic cognitive top-down influences on spatio-temporal fixation distributions within the post-cut scene and how these develop over the course of the first seconds following a cut. Related, these previous studies did not manipulate the viewing task between separate groups of participants, leaving it unclear whether the observed attentional effects were due to cognitive task-dependent top-down influences or whether they could be explained by some form of task-independent stimulus-driven repetition priming effect (cf. Maljkovic & Nakayama, 1994; Theeuwes, 2013). The aim of the present study was to address these open questions.

The Present Study

We used a large set of naturalistic video recordings of real-world scenes to test if human viewers can exert cognitive control over their eye movements during the very first seconds following scene cuts. In each trial of our experiments, participants saw two video takes in succession, separated by a single cut. All takes were spatial segments cropped from originally larger wide-screen source videos and showed a city scene that did or did not continue across the cut. In the main Experiment 1, we asked participants to recognize the post-cut takes as continuations or discontinuations of the immediately preceding pre-cut takes. Among these continuous cuts, we used shifted conditions in which the view on the scene underwent a panoramic shift from the pre- to the post-cut take: The takes presented before the cut were cropped from the left or right side of the original (panoramic) source video, and the takes following the cut showed a leftward or rightward shifted view that was cropped from the same source video. In these conditions, either the left or the right 50% of the image content in the post-cut take was visually related to and, therefore, matched with the view in the pre-cut take (see Figure 1).

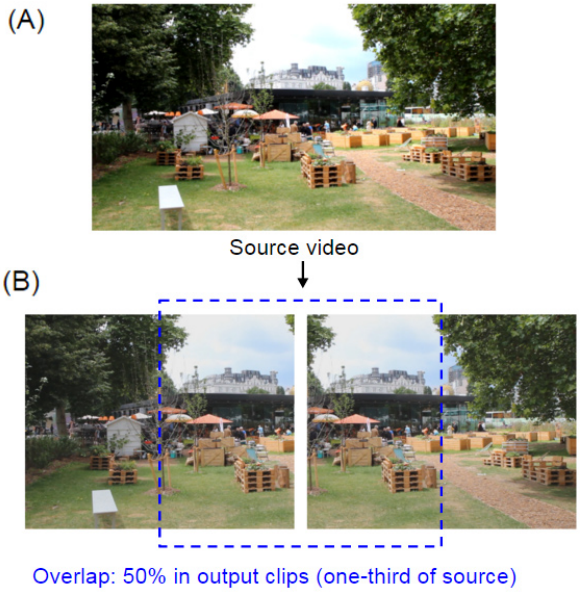

FIGURE 1.

Example images of a source video (A) and the two alternative cropped views created from this video that were used for panoramic view shifts in the shifted conditions (B). As can be seen, each cropped view corresponded to one horizontal side of the source video and, as depicted within the dotted rectangle, there was an area of 50% spatial overlap between the two different views taken from the same source video.

In the control Experiment 2, we used the same set of stimuli, but we changed the viewing task. Different from Experiment 1, we did not ask participants to recognize whether the post-cut take was a continuation of the pre-cut scene. Instead, we implemented an alternative recognition task as a control. Crucially, this control task did not require the participants to directly compare the two immediately succeeding takes within a trial. Before starting the experimental trials, participants in this control group were shown 16 videos that were also presented as to-be-recognized videos among the experimental trials. After each experimental trial, they were asked to report whether any of these 16 videos was identical to the pre-cut or the post-cut take in this trial.

In addition to these critical shifted conditions, both experiments included two further control conditions. The first control condition consisted of discontinuous cuts, where the post-cut take was completely unrelated to the pre-cut take. The second control condition consisted of full continuations, where the post-cut take was a continuation of the same scene from exactly the same view as the pre-cut take. This was only possible by inserting a blank screen (with only a central fixation cross) between all pre- and post-cut takes. In addition to allowing the inclusion of the full continuations as a control condition, this ensured that all participants started viewing the post-cut take from the same neutral central position in all conditions and trials.

We predicted that in the shifted conditions, the main experimental group (Experiment 1) would be more likely to fixate on visually related, matching scene regions compared to participants in the control Experiment 2. This is because matching scene regions contained

critical information to solve the task of recognizing whether or not the post-cut take was a continuation of the pre-cut take. Only the matching scene regions were informative about whether this was the same scene or a different, potentially similar scene, without requiring an exhaustive inspection of the whole post-cut images. In contrast, participants in Experiment 2 did not need to establish a relation between the two takes across the cut. Hence, we predicted that the control group should not show a particular preference for the matching scene regions, provided that such a preference in Experiment 1 would be solely due to the cognitive top-down demands imposed by the specific recognition task used. Across participants, we balanced the assignment of the individual video clips to the three cut conditions (see Figure 2). This allowed us to rule out any possibility that the clustering of fixations in matching scene areas of post-cut takes of Experiment 1 resulted from a higher occurrence of interesting scene content compared to the nonmatching areas. Thus, the full continuations and the discontinuous cuts served as control conditions for the shifted conditions in both experiments.

EYE TRACKING EXPERIMENTS

Methods and Materials

PARTICIPANTS

Forty-eight students took part in the experiments in exchange for partial course credit. Half of the participants (age 18–23 years, M = 19.8) took part in the main Experiment 1 and the second half (age 19–32 years, M = 24.6) took part in the control Experiment 2. All participants had normal or fully-corrected vision and gave informed consent prior to participation.

DYNAMIC SCENE STIMULI

We recorded 240 different landscape videos of street, park, or interior scenes around the city of Vienna (see Figure 3). All videos were recorded using a tripod from fixed positions, without any camera or lens movements, but movement was present within the videos at several locations in each frame. This movement was mostly caused by people walking or working, animals moving, cars passing, trees moving in the wind, or reflections on water surfaces. Videos were recorded with a wide angle lens in daylight conditions using narrow apertures to ensure high depth of field such that all image areas remained homogeneously sharp. In Experiment 1, we cut two immediately succeeding shorter video takes out of the source videos, henceforth referred to as Takes 1 and 2, each with a length of 10 s. In Experiment 2, the same Takes 1 and 2 were used, but further shortened to 5 s each (i.e., the last 5 s before each cut, and the first 5 s following each cut) because, after Experiment 1, it was clear that even 5 s are more than sufficient for understanding gaze behavior around the time of the cuts. For the creation of altogether 320 (plus a few demonstration) takes, we cropped spatially smaller frames (with a resolution of 1,280 × 1,024 pixels; 5:4 ratio) corresponding approximately to two thirds of the high definition source Takes 1 and 2 (with an original resolution of 1,920 × 1,088 pixels) (see Figure 1). The two alternative cropped views of each take depicted either the left or the right two thirds of the source takes and overlapped by precisely 50%.

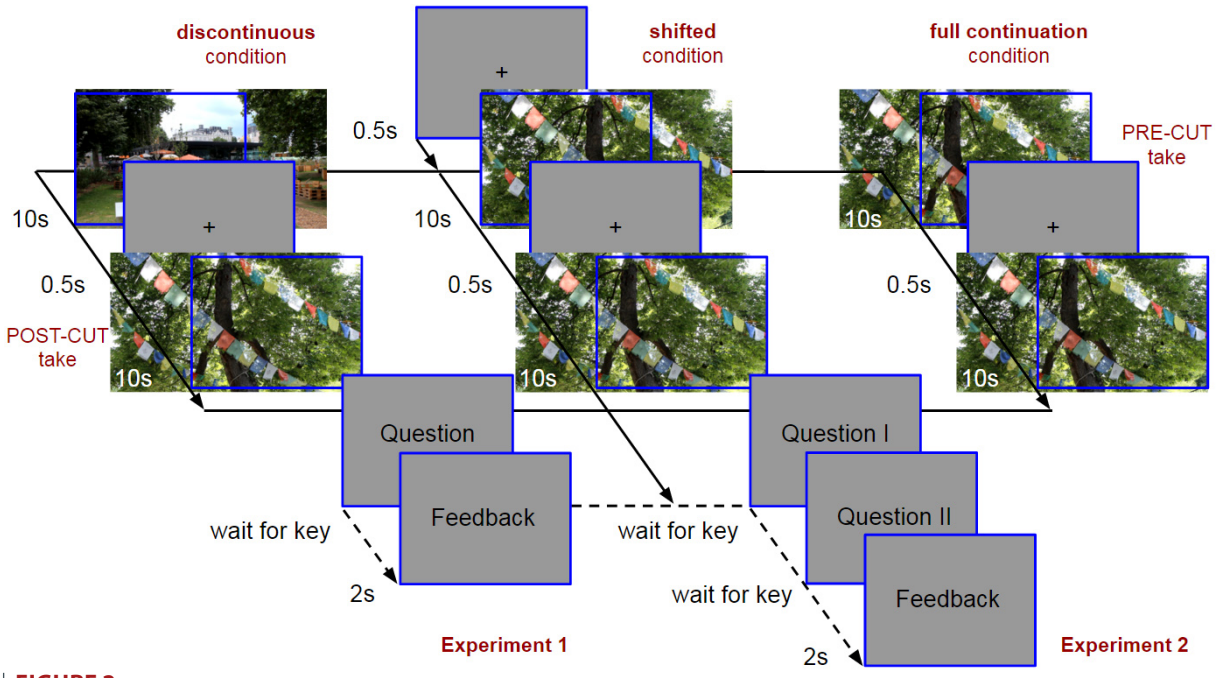

FIGURE 2.

Depicted is the same source video (in the post-cut take), assigned to the three different cut conditions: discontinuous (on the left), shifted condition (in the middle), or full continuation (on the right). In Experiment 1, each trial consisted of a 10 s pre-cut take followed by a fixation cross (cut) for 500 ms, and a 10 s post-cut take. In different versions of the experiment, the same post-cut take was used in discontinuous, shifted, or continuous cut conditions, but each participant saw only one of these versions. In Experiment 2, instead of 10, only 5 s of each take were shown.

FIGURE 3.
Example still images from the videos that were used in the current study.

APPARATUS

Eye movements were recorded using an EyeLink Desktop Mount eye tracker (SR Research Ltd.) at a sampling rate of 1,000 Hz. The system was calibrated to each participant's dominant eye using a standard 9-point calibration procedure. Every time the takes started or stopped, the exact timestamp was saved to the eye tracking data file, which allowed analyzing fixation latencies, durations, and frequencies with millisecond precision relative to the onset of each stimulus. After every tenth trial, calibration was checked using a standard drift check procedure and, if necessary, recalibrated. The videos were displayed on a 19-in. color CRT monitor (Sony Multiscan G400) at a resolution of 1,280 × 1,024 pixels and a refresh rate of 60 Hz. The experimental procedure was implemented in MATLAB (MathWorks) using the Psychophysics toolbox and the Eyelink toolbox (Kleiner et al., 2007). Viewing distance to the monitor was 64 cm, supported by chin and forehead rests, resulting in an apparent size of the full screen videos of 31 × 24.2°.

PROCEDURE AND DESIGN

Following six demo trials, every participant saw 160 experimental trials, each of them consisting of two takes—one pre-cut take of 10 s (Experiment 1) or 5 s (Experiment 2) and one post-cut take of 10 s (Experiment 1) or 5 s (Experiment 2)—and a cut between them (here, a short break of 500 ms). All takes were presented in full screen and in color. Prior to each trial and during the cut between the takes within each trial, the screen went grey for 500 ms, with the exception of a black fixation cross at screen center. Only after the post-cut take finished, participants were shown a grey response screen until they responded. In Experiment 1, the response screen contained the question: "Was the post-cut take a continuation of the scene shown in the pre-cut take?" To implement the control task in Experiment 2, participants saw and learned 16 clips for later recognition in advance of the actual experimental trials. These clips included both pre- and post-cut takes of full continuations. During the experimental trials of Experiment 2, in four instances of each of the four possible conditions (continuous cuts, discontinuous cuts, left-shifted, and right-shifted conditions), either the pre- or the post-cut clip contained a pre- or a post-cut clip of the initially learned videos, and each post-cut screen read: "Was there one of the 16 initial clips among the two takes that you just saw?" For those trials of the control task in which the participants in Experiment 2 indicated that one of the clips was part of the initially learned memory set, participants additionally had to indicate whether the first or the second take was among the initially presented clips. In this task, any "yes" (or recognition) answer was counted as correct when either the pre- or the post-cut take was from the initially learned memory set.

Throughout the experiments, participants fixated on the central fixation cross whenever it was present (i.e., before a trial started, and in between the end of the first take and the beginning of the second take). Participants pressed the *8* or *2* keys on the numerical keypad of a standard USB keyboard for their different judgments (e.g., *8* for the same scene vs. *2* for different scenes in pre- and post-cut takes of Experiment 1). Only after incorrect responses, participants saw an additional feedback screen of another 2 s that indicated that the wrong response had been given.

One half of all trials (80 trials) were discontinuous cuts in which the post-cut take showed a novel, hitherto not presented take (see Figure 2). The other half of all trials was continuations. Among the continuations, half (40) of the trials were full continuations, with the pre-and post-cut takes depicting the same view on the same scene. The other half of all continuous trials were shifted conditions. In shifted conditions, the cut constituted a panoramic view shift, with the view in the post-cut take shifted either to the right (20 trials) or to the left (20 trials) border of the original panoramic source video. To note, all of the take sequences, including the full continuations and the shifted conditions, were presented in the correct temporal order and presented 20 s (Experiment 1) or 10 s (Experiment 2) of immediately succeeding video content, without any temporal omissions, repetitions, or reversals. Also, the 5 s before the cut and the 5 s following the cut were exactly the same in Experiments 1 and 2. In Experiments 1 and 2, all different conditions were presented in a randomized order. Each trial took about 25 s (Experiment 1) or 14 s (Experiment 2), and the total test time was about 80 min (Experiment 1) or 60 min (Experiment 2).

DATA ANALYSIS

Areas of interest (AOI) for the major analyses were the post-and pre-cut takes' left and right sides. These AOIs started at two degrees eccentricity from the vertical meridian and reached until the respective image borders. Fixations were detected from the recorded gaze coordinates using the SR Research detection algorithm, as the average gaze position during periods with gaze position changes by less than 0.1°, eye movement velocity below 30°/s, and acceleration below 8000°/

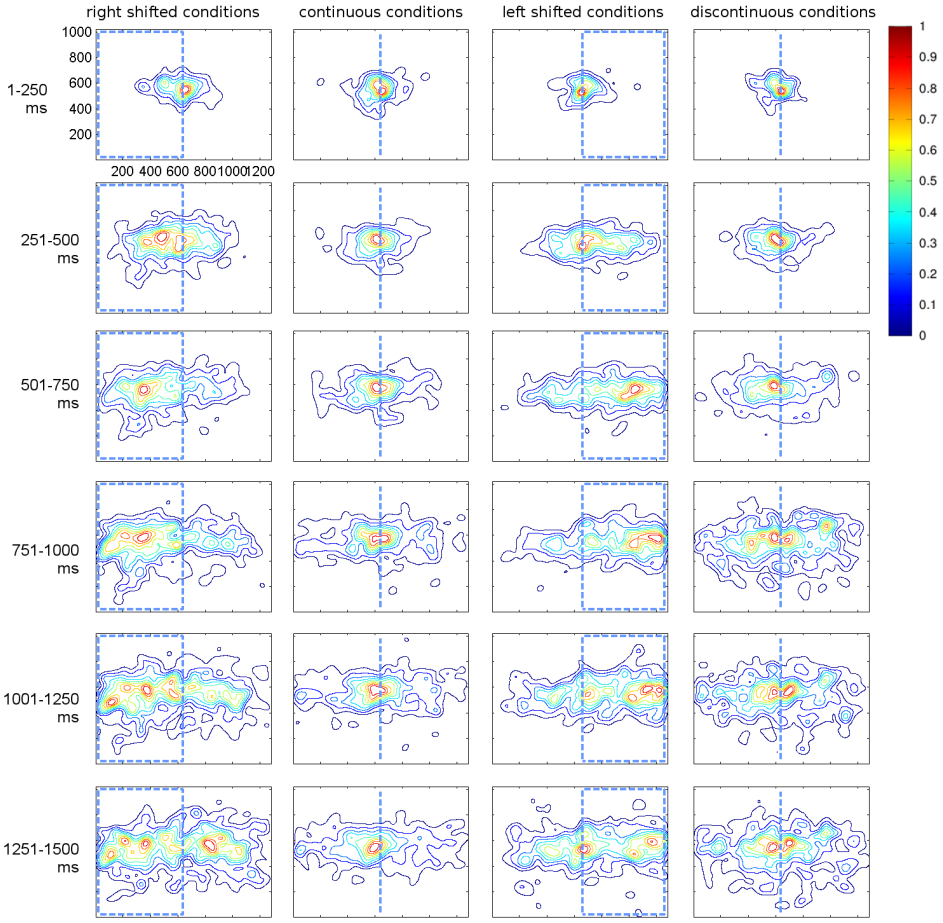

FIGURE 4.

Heat maps (across participants and images) of fixations within the first 1.5 s in the post-cut takes for the two shifted conditions (Columns 1 and 3) and for the two control conditions (Columns 2 and 4) of Experiment 1. Here, red, orange and yellow depict areas of relatively higher numbers of fixations, while green, blue, and white depict areas of lower numbers of fixations. The horizontal and vertical coordinates of each subplot correspond to the screen coordinates of the full screen post-cut takes (1,280 × 1,024 pixels). In the first column, fixation data from the right shifted conditions show more fixations on the left side, with its across-cut matching content. In Columns 2 to 4, fixations are shown for full continuations, left shifted conditions, and discontinuous cuts, respectively. The time bins into the post-cut takes are given in the rows from early at the top to further into the post-cut take at the bottom. From Rows 2 to 6, in the shifted conditions, a clustering of fixations in areas that match across the cut is evident.

s^2. Eye movement data were preprocessed in MATLAB and statistically analyzed in R (R Core Team, 2016). Statistical significance was assumed at an α level of .01 or below. (A slightly more liberal criterion of an α level of .05 would have yielded identical conclusions.) We limited our analysis of the viewing behavior in the post-cut takes to the first three seconds following the cut because after this time we observed no preferences for fixating one of the two alternative AOIs between the different conditions. All statistical tests were based on 144 out of the 160 trials because the 16 trials that contained a clip of the control task in Experiment 2 were excluded from all analyses of both experiments.

Results of Experiment 1

BEHAVIORAL TASK

Participants made 1.46% errors (*SD* = 1.18) in the scene continuity judgment task.

FIXATION FREQUENCIES

Within the first 3 s of the post-cut takes, participants made 8.21 fixations on average (*SD* = 1.31). Figure 4 gives an impression of how the spatial distribution of fixations of our participants developed across five 250 ms time bins from 0 to 1.5 s following the onset of the post-cut take. Fixations starting and ending in different bins were assigned to all bins in which they were measured. The first column of Figure 4 shows that in right shifted conditions, participants preferentially fixated on the across-cut matching left side of the post-cut take and fewer fixations were made on the nonmatching right side of the post-cut take. One can also see that this preference for one side over the other was reversed in left shifted conditions (third column of Figure 4). In contrast, no strong preferences for either side were observed in the two control conditions—that is, in the fully continuous (see second column of Figure 4) and in the discontinuous takes (see fourth column of Figure

TABLE 1.

p-Values of Pairwise Comparisons (*T*-Tests With Holm-Bonferroni Correction) Between Different Cut Conditions for Different Times Into the Post-Cut Takes of Experiment 1

	Time after the cut in ms in timespans of 250ms											
	1-250	-500	-750	-1000	-1250	-1500	-1750	-2000	-2250	-2500	-2750	-3000
right shifted												
shifted vs. discont.	1	.001*	.001*	.001*	.001*	.001*	.011	.104	.109	.877	1	1
shifted vs. fully cont.	.959	.001*	.001*	.001*	.001*	.005*	.185	1	1	1	.972	1
discont. vs. fully cont.	1	.366	.860	1	1	1	1	1	1	1	1	1
left shifted												
shifted vs. discont.	.511	.001*	.001*	.001*	.001*	.001*	.035	.138	.026	.289	.070	.022
shifted vs. fully cont.	1	.001*	.001*	.001*	.001*	.001*	.022	.536	.094	.063	.059	.001*
discont. vs. fully cont.	1	1	1	1	1	1	1	1	1	1	1	1

Note. * = significant at α $p < .01$. $df = 23$. discont. = discontinuous; cont. = continuous.

4), in which both sides of the post-cut takes were equally matching or nonmatching across the cut.

For the statistical analysis, we split the first 3 s of fixations on the post-cut takes into time bins of a length of 250 ms. We used *t* tests with Holm-Bonferroni correction to test for statistical differences between the frequencies of fixations on the left versus the right sides (see Table 1). Since each participant saw each post-cut take only as either left shifted or right shifted, a direct within-participant comparison between the two would have been beset with a difference in their visual content. To compare both of these experimental conditions separately with their respective control conditions (i.e., the full continuations and the discontinuous takes), the averages for each condition and time bin were taken for each participant and compared afterwards. The control conditions showed exactly the same post-cut takes as were used in the respective shifted conditions. As can be seen in Table 1, from at least 250 ms until about 1.5 s into the post-cut takes, fixation frequencies on across-cut matching regions of the post-cut takes in the shifted conditions differed significantly from fixation frequencies in the corresponding areas of the same post-cut takes under the control conditions. The upper three rows of Table 1 show that *t* tests confirmed that more fixations were made on the left side of the post-cut takes of the right shifted conditions than on the left side of the same post-cut takes in full continuations and discontinuous cuts. The *t* tests also showed that there were no such differences between the two control conditions (full continuations and discontinuous cuts). The lower three rows of Table 1 show that *t* tests also confirmed more fixations were made on the right

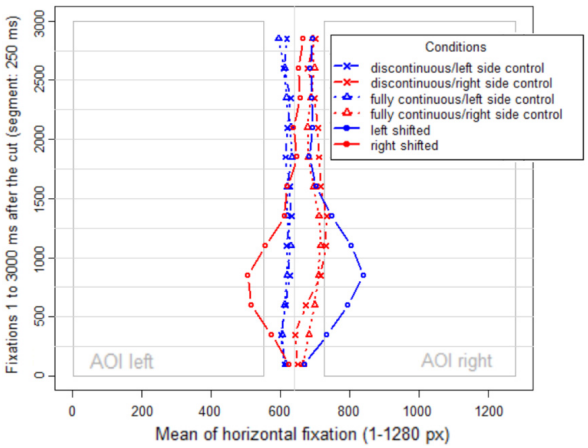

FIGURE 5.

Mean horizontal deviations of all fixations (on the abscissa) as a function of the time into the post-cut take on the ordinate, separately for different conditions of Experiment 1. One can see that in the shifted conditions (continuous lines), participants more frequently fixated locations in the across-cut matching regions of the post-cut take (left sides for right shifted, right sides for left shifted conditions) than under both control conditions (full continuations [punctuated lines] and discontinuous takes [dashed lines]).

TABLE 2.

Means and Standard Deviations (in Parentheses) of Fixation Latencies on the Left and Right Sides (or in the Respective Areas of Interest, AOI) of the Post-Cut Takes as a Function of the Different Cut Conditions in Experiment 1

	Shifted conditions	Full continuations	Discontinuous cuts
right shifted			
left AOI	*617 ms (500)*	*976 ms (759)**	*885 ms (696)**
right AOI	976 ms (703)	715 ms (621)*	730 ms (531)*
left shifted			
left AOI	1,088 (766)	853 ms (687)*	826 ms (682)
right AOI	*600 (451)*	*955 ms (714)**	*980 ms (681)**

Note. * = Wilcoxon signed-rank test significant at $\alpha < .01$ between shifted conditions and full continuations; and between shifted conditions and discontinuous cuts. Rows featuring across-cut matching AOIs of the shifted conditions are in italics.

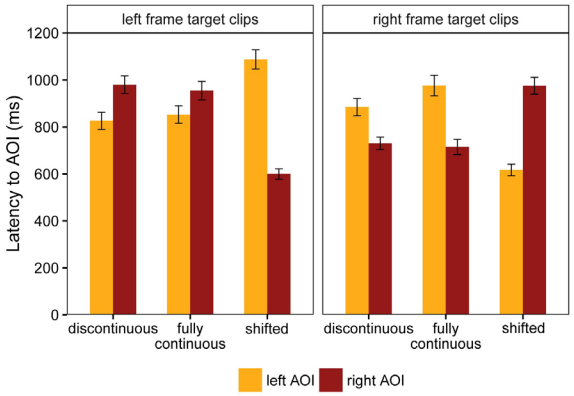

FIGURE 6.

Mean latencies of first fixations on the left side (area of interest, AOI; in yellow) and on the right side (AOI; in red) of the post-cut takes in Experiment 1. Left panel: performance in left shifted conditions and in the corresponding post-cut takes from the two control conditions (discontinuous and full continuations). Right panel: performance in right shifted conditions and in the corresponding post-cut takes from the two control conditions.

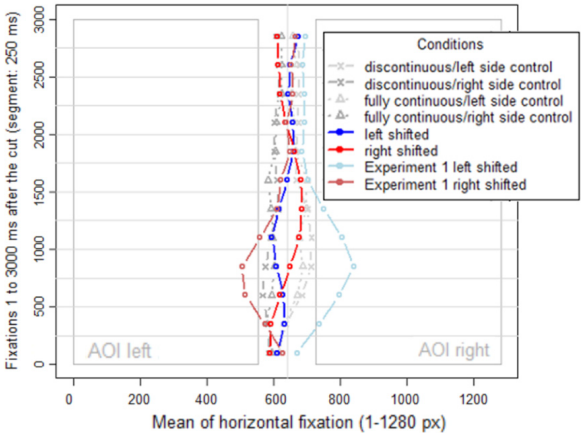

FIGURE 7.

Mean horizontal deviations of all fixations (on the abscissa) as a function of the time into the post-cut take on the ordinate, separately for different conditions of Experiments 1 and 2. One can see that in the shifted conditions of Experiment 2 (continuous lines in red and blue), participants showed a slight fixation preference on nonmatching regions (right of center for right shifted, left of center for left shifted conditions) compared to both control conditions (full continuations [punctuated lines] and discontinuous [dashed lines] cuts). This is in contrast to Experiment 1, where a strong preference for matching regions was found. For the sake of an easier comparison, the corresponding performances of Experiment 1 have also been included (pale lines).

side of the post-cut takes of the left shifted conditions than on the right side of the same post-cut takes in full continuations and discontinuous cuts. And again, the *t* tests also demonstrated that there were no such differences between the two control conditions (full continuations and discontinuous cuts).

Figure 5 shows the mean horizontal locations of the participants' fixations. From 250 ms to 1.5 s into the post-cut take, the AOI on the right side attracted more fixations in the shifted left conditions (blue line), and the AOI on the left side attracted more fixations in the shifted right condition (red line). These increased fixation frequencies were found compared to their respective control conditions: In the full continuations (dotted lines) and the discontinuous cuts (broken lines), the mean fixation locations were less lateralized and more consistently within 2° of the take center, showing spatially more balanced fixations on the left and right. To estimate the effect sizes of the fixation preference for the matching regions in the particular time bins with significant differences in the shifted conditions, we calculated Pearson's *r* correlation coefficients across participants between the horizontal axis positions in the post-cut takes. The rationale for this test is that a high preference for one side should lead to a high correlation of the horizontal fixation locations. In the shifted conditions, these correlations were of medium size ($r = 0.29$ to 0.35) for fixations from 250 ms to 750 ms, and small for the other significant time segments ($r = 0.14$ to 0.26).

TABLE 3.

p-Values of Pairwise Comparisons (*T*-Tests With Holm-Bonferroni Correction) Between Different Cut Conditions for Different Times Into the Post-Cut Takes of Experiment 2

	Time after the cut in ms in timespans of 250ms											
	1-250	-500	-750	-1000	-1250	-1500	-1750	-2000	-2250	-2500	-2750	-3000
right shifted												
shifted vs. discont.	1	1	.008*	.004*	.002*	.028	.202	.467	1	1	1	1
shifted vs. fully cont.	1	1	.891	.671	.030	.014	.016	.915	1	1	1	1
discont. vs. fully cont.	1	1	.891	1	1	1	1	1	1	1	1	1
left shifted												
shifted vs. discont.	1	1	.034	.001*	.006*	.135	1	1	1	1	1	1
shifted vs. fully cont.	1	1	.166	.129	.288	.135	.647	1	1	1	1	1
discont. vs. fully cont.	1	1	1	1	1	1	1	1	1	1	1	1

Note. * = significant at α $p < .01$. $df = 23$. discont. = discontinuous; cont. = continuous.

LATENCIES OF FIRST FIXATIONS

As a second dependent variable, the latencies of the first fixations on either AOI of the post-cut takes were analyzed. We discarded trials in which participants did not fixate inside the AOIs within 5 s following the start of the post-cut take and outliers that exceeded a criterion of 1.5 times the interquartile range of the latency distribution. As a result, 20.9% of the fixations on the post-cut takes were excluded. This resulted in 969 trials (27.3%) being excluded due to fixations outside the AOIs. Another 32 trials (0.9%) exceeding the range for outliers were excluded. On average, it took the participants 865 ms to fixate at least once on locations inside both lateral AOIs, left and right.

Table 2 shows that for post-cut takes of the shifted conditions, latencies were significantly shorter for the first fixations on across-cut matching sides than for the sides of the identical post-cut takes in the respective two control conditions (i.e., full continuations and discontinuous cuts). In addition, fixations on the across-cut matching sides were of significantly lower latency than fixations on nonmatching regions ($p < .01$). Finally, at least for the right shifted conditions, latencies of fixations on the nonmatching side were significantly higher than latencies of fixations on the same (right) side in the control conditions. Mean fixation latencies are also plotted in Figure 6. The error bars represent the standard errors of the means.

Results of Experiment 2

BEHAVIORAL TASK

The mean percentage of incorrect answers to the control task was 13% ($SD = 3.9$).

FIXATION FREQUENCIES

Participants made an average of 8.38 fixations ($SD = 0.89$) within the first 3 s of the post-cut takes. As in Experiment 1, the first 3 s were split into bins of 250 ms and t tests with Holm-Bonferroni correction were used to determine differences between conditions (see Table 3 and Figure 7).

In the upper three rows of Table 3, t tests showed that more fixations were made on the right side in right shifted conditions than on the right side of the same post-cut takes in discontinuous cuts. (Numerically, the same difference was found between the right shifted and the fully continuous conditions.) This is the opposite tendency of what has been observed in Experiment 1, where the participants made more fixations on across-cut matching regions in the shifted conditions compared to the control conditions. For the sake of an easier comparison, we plotted the data from both experiments in Figure 7 (data from Experiment 1 are rendered as pastel red and blue continuous lines). The lower three rows of t tests in Table 3 show that more fixations were made on the left side in left shifted conditions than on the left side of the same post-cut takes under discontinuous conditions. Again, these effects were in the opposite direction of the differences that we saw in Experiment 1.

LATENCIES OF FIRST FIXATIONS

In Experiment 2, there were no significant differences between the cut conditions for the latencies of the first fixations on either AOI at all.

DISCUSSION

Our study presents new evidence that viewers of edited dynamic scenes have robust cognitive top-down control over their eye movements immediately after scene cuts. This was doubtful in light of previous studies that had suggested cognitive top-down influences were minimal after cuts (Carmi & Itti, 2006a) and need more time to take effect (Smith & Mital, 2013). Here, we uncovered early task-dependent, top-down controlled gaze behavior by comparing spatio-temporal fixation distributions after standardized view shifts in a large set of videos of complex real-world scenes.

In each trial of Experiment 1, participants judged whether or not the second of two takes was a continuation of the pre-cut scene. In the critical view shifted conditions, we found robust and early task effects on eye movements: From 250 ms up to 1.5 s, the participants' fixations systematically clustered in scene regions that matched with the familiar view from before the cut (the very first time window of 0 to 250 ms did not show an effect, as it was determined by the central fixation cross that was present in the interval between the two takes). Statistical tests confirmed that participants made an overall higher number of fixations on scene regions that matched across cuts compared to nonmatching regions. This result conceptually replicates a previous study using a similar view shift manipulation with static images (Valuch et al., 2013). Moreover, in Experiment 1, fixations on matching regions had a lower mean latency compared to fixations on nonmatching regions.

We can exclude the possibility that these differences in fixations between matching and nonmatching scene regions of Experiment 1 resulted from particular visual characteristics of the videos in the shifted conditions (e.g., matching regions containing more interesting image content, hence, attracting fixations independent of the task): Assignment of the videos to the three different experimental conditions was balanced across participants, and the two control conditions of discontinuous cuts and full continuations did not result in any systematic off-center gaze clustering. Note also that in the shifted conditions, participants could not have expected the direction of the upcoming view shift because shifts occurred only in 25% of trials and with equal probability to the left and to the right. The results thus illustrate that viewers exerted immediate cognitive top-down control over their fixations in the post-cut takes, and quickly oriented their eyes towards matching scene regions, which contained the most task-relevant information.

In Experiment 2, we tested an independent group of participants with a control task. Importantly, all video stimuli and the three cut conditions were the same as in Experiment 1, but participants did not need to make judgments about scene continuity from the pre-cut take into the post-cut take. They performed a control task by indicating after each trial whether either of the two successive takes in a trial was part of a set of 16 clips that were shown before the experimental trials. Given this task, there was no utility in making fixations on scene regions that matched across view shifts because participants did not need to directly relate the pre-cut and the post-cut take to each other. In line with our prediction, in this control group, fixation latencies

were not decreased and fixation frequencies were not increased on across-cut matching scene regions. If anything, there was a slight but less pronounced tendency to fixate more frequently on the nonmatching regions of the post-cut takes. This slightly opposite tendency occurred relative to the discontinuous cut conditions of Experiment 2 and relative to the shifted conditions in Experiment 1. One possible explanation could be a tendency to visually explore scenes, in the absence of the need to identify the connecting elements. This could be similar to a preference for spatio-temporal novelty that was reported in past research using free-viewing tasks (Itti & Baldi, 2009). A tendency for novel information in Experiment 2 might have also reflected a task-specific influence because, when the post-cut clips were presented, pre-cut scenes had already been compared to the memory content, so that only the novel information in the post-cut clips had to be evaluated for its similarity to the initially learned videos. However, one should rather not over-interpret this result because it was by far weaker than the strong, task-driven effects in Experiment 1.

In any case, Experiment 2 rules out the possibility that the effects observed in Experiment 1 could be explained by an involuntary tendency to automatically look at whatever content repeats across viewpoint changes. If we would have found the same preference for matching areas in Experiment 2, this would have suggested that the effect stems from repetition priming, which is sometimes believed to influence attentional selection in a stimulus-driven way, irrespective of the task (Theeuwes, 2013). The data from Experiment 2 thus further supports the notion that the behavior observed in Experiment 1 was truly driven by the requirements imposed by the task.

The present results extend the literature in several respects. Previous studies reasoned that following scene cuts, the contribution of cognitive top-down factors to gaze behavior is minimized (Carmi & Itti, 2006a; Loschky et al., 2015). This judgment was based on increased correlations between fixation locations and salient local features (Carmi & Itti, 2006a) or an increased tendency to look at the image center following cuts (Dorr et al., 2010; Tseng et al., 2009). In Experiment 1, we clearly showed a robust and systematic off-center deviation in spatial fixation distributions towards peripheral scene regions that contained the task-relevant information very early after scene cuts. The discrepancy between our results and previous reports could partly be explained by methodological differences. Previous studies were mostly conducted in a free-viewing context, without manipulating the task between groups of participants (for an exception see, e.g., Smith & Mital, 2013). Moreover, previous studies partly relied on professionally produced video material that is known to elicit strong center biases and high inter-observer correlations in gaze direction, due to more strongly constrained visual content (Goldstein et al., 2007; Loschky et al., 2015). Maybe most importantly, studies of eye movements in edited videos sometimes used only cuts between different scenes, where the pre- and post-cut takes were visually completely unrelated (e.g., Carmi & Itti, 2006a, 2006b; Itti & Baldi, 2009). Under such conditions, it is impossible to discriminate cognitive top-down influences from stimulus-driven factors, such as visual salience (Carmi & Itti, 2006b). In contrast, our study was purposely tailored to identify the contribution of cognitive top-down control to gaze guidance early after scene cuts by comparing two groups of participants under different task instructions, and it included cuts between two related views on the same scene, such that the pre-to-post-cut view change followed a well-defined relationship in all trials of this condition. We also used videos of real-world scenes because these enabled us to implement the view shift manipulation across a large set of videos in an automated manner while retaining the potential to visually explore a complex dynamic scene.

One limitation of our present findings might be that the explicit recognition task in Experiment 1 was realized in a setting that is quite different from more everyday video viewing situations (where viewers of edited movies usually simply attentively view one edited video and follow its narrative). However, we believe that the setting we created in Experiment 1 is actually quite similar to a more implicit viewing task that viewers are usually engaged in whenever they attentively follow an edited video. Indeed, cognitive film theory suggests that following each cut, film viewers must recognize how a new take relates to what they saw before the cut (Hochberg & Brooks, 1996). One caveat here certainly is the different nature of the edited material in our study compared to professional footage (Dorr et al., 2010). Replications of our study with spatially cropped outtakes of feature films would, therefore, be informative about whether there is something particular about professionally produced feature films that mesmerizes the audience and causes all viewers to fixate on the same content (Goldstein et al., 2007) or whether such observations could reflect implicit task-goals of the viewers dedicated to understanding of how successive takes relate to each other. As such, our results could inspire further theorizing about why certain standard editing practices of film and media professionals work particularly well. In continuity editing, for example, cutters take care to facilitate the narrative connection between pre- and post-cut take (Bordwell & Thompson, 2001). At least some of the subjectively perceived smoothness of continuity editing (Shimamura, Cohn-Sheehy, & Shimamura, 2014) might be explained by the degree to which visual similarities across cuts facilitate recognition of familiar content after view changes (Valuch & Ansorge, 2015; Valuch, König, & Ansorge, 2017).

Finally, our findings have applications beyond video editing. First, the improvement of video coding standards benefits from a better understanding of the determinants of gaze behavior in edited videos (Adzic et al., 2013). Second, based on our results, computational models of human eye movements can include memory components that allow to model gaze behavior in contexts where cuts frequently occur between related scene views (Ansorge et al., 2014). Apart from edited videos, there are other situations where viewers visually explore complex dynamic scenes across view changes while being engaged in a specific task. For example, in radiographic applications and ultrasound imaging, highly complex and visually cluttered dynamic scenes have to be searched for anomalies. Recognizing the visual content that repeats across view changes is key to efficient visual orienting. In these applications, understanding the spatio-temporal limits and properties of voluntary top-down control over eye movements after cuts is of obvi-

ous relevance. More broadly, graphical user interfaces regularly include shifts between complex screens which could be conceptually similar to scene cuts in edited dynamic scenes (May et al., 2003; Valuch et al., 2014). Knowing to which degree humans possess immediate cognitive control over their gaze direction after scene cuts could facilitate overall usability and help improve user experience in computer applications in a wide variety of applications.

CONCLUSION

We showed that human eye movements in edited videos are sensitive to task-specific cognitive top-down control immediately after view changes across cuts. Different to what the existing research literature suggests, task goals can override influences of stimulus characteristics and generic viewing biases already in the very first second following a cut. We described implications for understanding common video editing practices, and for the improvement of technological applications.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF, Vienna Science and Technology Fund) grant no. CS 11-009 awarded to Ulrich Ansorge, Shelley Buchinger, and Otmar Scherzer. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Acik, A., Onat, S., Schumann, F., Einhäuser, W., & König, P. (2009). Effects of luminance contrast and its modifications on fixation behavior during free viewing of images from different categories. *Vision Research, 49*, 1541-1553. doi: 10.1016/j.visres.2009.03.011 WWW

Adzic, V., Kalva, H., & Furht, B. (2013). Exploring visual temporal masking for video compression. In T. Hirashima (Chair), *Proceedings of the IEEE International Conference on Consumer Electronics* (pp. 590-591). Las Vegas, NV: Curran Associates, Inc. doi: 10.1109/ICCE.2013.6487030

Ansorge, U., Buchinger, S., Valuch, S., Patrone, A. R., & Scherzer, O. (2014). Visual attention in edited dynamical images. In M. S. Obaidat, A. Holzinger, & E. Cabello (Eds.), *Proceedings of the 11th International Conference on Signal Processing and Multimedia Applications* (pp. 198-205). Setúbal, Portugal: SCITEPRESS Digital Library. doi: 10.5220/0005101901980205

Bordwell, D., & Thompson, K. (2001). *Film art: An introduction* (6th ed.). New York, NY: McGraw-Hill.

Carmi, R., & Itti, L. (2006a). The role of memory in guiding attention during natural vision. *Journal of Vision, 6*, 898-914. doi: 10.1167/6.9.4

Carmi, R., & Itti, L. (2006b). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research, 46*, 4333-4345. doi: 10.1016/j.visres.2006.08.019 WWW

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*, 1-15 doi: 10.1167/9.3.6 WWW

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*, 1827–1837. doi: 10.1016/0042-6989(95)00294-4 WWW

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10*, 1-17. doi: 10.1167/10.10.28 WWW

Ferscha, A., Paradiso, J., & Whitaker, R. (2014). Attention management in pervasive computing. *IEEE Pervasive Computing, 13*, 19-21. doi: 10.1109/MPRV.2014.2

Foulsham, T., & Kingstone, A. (2013). Fixation-dependent memory for natural scenes: An experimental test of scanpath theory. *Journal of Experimental Psychology: General, 142*, 41-56. doi: 10.1037/a0028227 WWW

Fuchs, I., Ansorge, U., Redies, C., & Leder, H. (2011). Salience in paintings: Bottom-up influences on eye fixations. *Cognitive Computation, 3*, 25-36. doi: 10.1007/s12559-010-9062-3

Germeys, F., & D'Ydewalle, G. (2007). The psychology of film: Perceiving beyond the cut. *Psychological Research, 71*, 458–466. doi: 10.1007/s00426-005-0025-3 WWW

Goldstein, R. B., Woods, R. L., & Peli, E. (2007). Where people look when watching movies: Do all viewers look at the same place? *Computers in Biology and Medicine, 37*, 957-964. doi: 10.1016/j.compbiomed.2006.08.018 WWW

Hochberg, J., & Brooks, V. (1996). The perception of motion pictures. In M. P. Friedman & E. C. Carterette (Eds.), *Cognitive ecology* (2nd ed., pp. 205-292). London, UK: Academic Press.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*, 1295-1306. doi: 10.1016/j.visres.2008.09.007 WWW

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., Cornelissen, F., et al. (2007). What's new in psychtoolbox-3?. *Perception, 36*, 1-89. Retrieved from: http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/attachments/ECVP2007-Kleiner-slides_5490%5b0%5d.pdf

Land, M., & Tatler, B. (2009). *Looking and acting: Vision and eye movements in natural behaviour*. Oxford, UK: Oxford University Press.

Loschky, L. C., Larson, A. M., Magliano, J. P., & Smith, T. J. (2015). What would Jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension. *PloSONE, 10*, e0142474. doi: 10.1371/journal.pone.0142474 WWW

Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory & Cognition, 22*, 657-672. doi: 10.3758/BF03209251 WWW

May, J., Dean, M. P., & Barnard, P. J. (2003). Using film cutting techniques in interface design. *Human-Computer Interaction, 18*, 325–372. doi: 10.1207/S15327051HCI1804_1

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation, 3*, 5-24. doi: 10.1007/s12559-010-9074-z

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*, 1-19. doi: 10.1167/10.8.20 WWW

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*, 1457–1506. doi: 10.1080/17470210902816461 WWW

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Salomon, D. (2004). *Data compression: The complete reference.* New York, NY: Springer.

Shimamura, A. P., Cohn-Sheehy, B. I., & Shimamura, T. A. (2014). Perceiving movement across film edits: A psychocinematic analysis. *Psychology of Aesthetics, Creativity, and the Arts, 8*, 77-80. doi: 10.1037/a0034595

Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision, 13*, 1-24. doi: 10.1167/13.8.16 WWW

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*, 1-23. doi: 10.1167/11.5.5 WWW

Theeuwes, J. (2013). Feature-based attention: It is all bottom-up priming. *Philosophical Transactions of the Royal Society B: Biological Sciences, 368*, 20130055. doi: 10.1098/rstb.2013.0055 WWW

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*, 1-16. doi: 10.1167/9.7.4 WWW

Valuch, C., & Ansorge, U. (2015). The influence of color during continuity cuts in edited movies: An eye-tracking study. *Multimedia Tools and Applications, 74*, 10161–10176. doi: 10.1007/s11042-015-2806-z

Valuch, C., Ansorge, U., Buchinger, S., Patrone, A., & Scherzer, O. (2014). The effect of cinematic cuts on human attention. In P. Oliver, P. Wright, & T. Bartindale (Chairs), *TVX '14 Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video* (pp. 119-122). New York, NY: ACM. doi: 10.1145/2602299.2602307

Valuch, C., Becker, S. I., & Ansorge, U. (2013). Priming of fixations during recognition of natural scenes. *Journal of Vision, 13*, 1-22. doi: 10.1167/13.3.3 WWW

Valuch, C., König, P., & Ansorge, U. (2017). Memory-guided attention during active viewing of edited dynamic scenes. *Journal of Vision, 17*, 1-32. doi: 10.1167/17.1.12 WWW

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

RECEIVED 14.10.2016 | ACCEPTED 09.03.2017