Exploring Gaze-Based Prediction Strategies for Preference Detection in Dynamic Interface Elements

Melanie Heck
University of Mannheim

Mannheim, Germany
melanie.heck@uni-mannheim.de

Jonathan Bünemann
University of Mannheim

Mannheim, Germany
jonathanbuenemann@gmail.com

Janick Edinger
University of Hamburg

Hamburg, Germany
edinger@informatik.uni-hamburg.de

Christian Becker
University of Mannheim

Mannheim, Germany
christian.becker@uni-mannheim.de

ABSTRACT

Digitization is currently infiltrating all daily processes, forcing casual computer users to become acquainted with unfamiliar tools. In order to avoid overstraining these users, simplified interfaces that are reduced to the functionality and content which are relevant to the individual user are imperative. Gaze-contingent systems thus monitor viewing behavior during natural system interactions to predict relevant interface elements. The prediction performance is highly dependent on the underlying features and algorithm, especially when the interface consists of dynamic elements such as videos. In this paper, we conduct two studies with a total of 233 subjects in which we record the viewers' gaze while watching videos. We then compare the quality of preference predictions for video elements of majority voting to the performance of machine learning. Our results indicate that (1) majority voting can predict preferences with an accuracy of up to 73% (66%) for two (four) elements, (2) machine learning improves the performance to 82% (74%), (3) prediction accuracy depends on the strength of the user's preference for an element, and (4) we can rank preferences for individual elements.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Interaction design process and methods*.

KEYWORDS

eye tracking; adaptive user interfaces; preference prediction

ACM Reference Format:

Melanie Heck, Janick Edinger, Jonathan Bünemann, and Christian Becker. 2021. Exploring Gaze-Based Prediction Strategies for Preference Detection in Dynamic Interface Elements. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21)*, March 14–19, 2021, Canberra, ACT, Australia. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3406522.3446013>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHIIR '21, March 14–19, 2021, Canberra, ACT, Australia

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8055-3/21/03...\$15.00

<https://doi.org/10.1145/3406522.3446013>

1 INTRODUCTION

The global spread of the Covid-19 pandemic and the concomitant necessity for 'social distancing' has more than ever increased the importance of digital channels. While this has enabled a massive boost for digitization, it has also revealed the shortcomings of current solutions. As technology is integrated in more and more daily activities, targeted presentation of selected information is key. Unfortunately, conventional interfaces are often overloaded. Since human attention is a scarce resource [2], only a small fraction of the information on cluttered displays can be processed. Efficient user interfaces must therefore identify what is relevant to the user and dynamically personalize the presented content. Personal devices can establish comprehensive user profiles based on usage history. In contrast, shared devices such as public terminals cannot gather, store and use preferences if users do not log into an account. Moreover, content relevance is dynamic and can thus not necessarily be derived from past interactions. A user's conception of what is relevant today may not be the same as yesterday.

Gaze-contingent systems therefore use eye tracking to infer object relevance on the fly. The common approach is to extract gaze features and then apply a prediction algorithm to derive user preferences for interface elements. Especially in the context of dynamic content such as videos, the selected gaze features and algorithm have a significant impact on the performance of preference predictions. We thus conduct two large-scale user studies ($N = 233$) with a total of eight videos and make the following contributions:

- In the first user study, we define a standard set of gaze features. We use videos with relatively neutral elements to demonstrate the feasibility of using gaze data to predict the viewer's preferred element in a video.
- In the second user study, we collect gaze data while subjects watch videos of varying design. We compare the predictive performance of majority voting to machine learning models.
- We evaluate how the video design affects the predictions.

2 RELATED WORK

Gaze-contingent systems presuppose that preferences can be inferred from gaze. Evidence for a correlation between gaze and preference was found as early as in the 1960s. Mackworth and Morandi [38] report that important aspects of pictures receive more fixations. Yet, visual attention does not necessarily imply preference

for an element. Attentional processes can be classified as *top-down* or *bottom-up* [15]. Top-down processes are goal-driven, i.e., they are controlled by the person [7]. Bottom-up attention, in contrast, is a shift of the gaze towards salient elements (e.g., bright colors) and is not driven by preferences. Nevertheless, citing the results of several psychological studies, Bednarik [3] asserts that a link between cognitive processes and gaze data can be assumed.

Market research traditionally analyzes gaze data offline to learn about the determinants of customers' decision-making [13, 70]. In contrast, gaze-contingent systems infer preferences from gaze data to dynamically adapt the system. Prior research with static interface elements has shown that relevant display areas [16, 62] and loss of interest in an element [28, 33, 58] can be inferred from users' gaze while browsing websites. By quantifying the strength of a user's preference for an item, gaze-based recommender systems dynamically adapt the displayed items [12, 21, 23, 36, 55, 73].

Xu et al. [72] measure visual attention to key frames of a scene for personalized video recommendations. In contrast to our approach, attention is measured on the granularity of entire videos, instead of comparing gaze allocation to multiple elements in a scene. In [4], looking at one out of several muted video stream for multiple seconds triggers the playback of its soundtrack. The video communication system in [67] only plays the speech output of the person the user is looking at. The adaptive movies in [45] and [68] use eye tracking to dynamically adapt the storyline. The viewer's preferred continuation of a scene is inferred from logical rules, but no assessment of the prediction performance is presented. Kandemir et al. [31] collect gaze data while subjects watch a video and rank the relevance of the elements in a video frame based on logistic regression. They find that incorporating optical information about the video elements improves predictions. Yet, their findings are based on the evaluation of a single classifier using a fixed feature set. In this paper, we thus compare multiple machine learning classifiers. We analyze whether their performance improvement over majority voting justifies the use of more complex models.

3 STUDY 1: DEFINING GAZE FEATURES

Even though eye tracking has evolved into an established research technology during the past decades, no standard set of gaze features has emerged yet for analyzing a subject's visual attention. We therefore designed a user study to evaluate the predictive performance of 17 gaze features. The objective of the study was to identify and define a set of standard gaze features that are suitable for predicting a user's preferred element in a video. We analyze two questions:

- **Q1:** *Can a standard gaze feature set for preference predictions in videos be defined?* We compute all 17 gaze features and exclude those that show low correlation with the reported preference, or are highly correlated with another feature.
- **Q2:** *Do the features accurately predict the preferred element?* We apply majority voting to each feature and compare the prediction of the preferred AOI to the reported preference.

3.1 Participants

101 subjects (36 female, aged between 18 and 52 years) participated in the study. We recruited students and employees from both the business and informatics department of our research institution. All

participants had normal or corrected to normal vision. Participation was voluntary and no monetary incentive was given. No subject was fully aware of the purpose of the experiment, but they were informed that their gaze was being tracked.

3.2 Apparatus

The employed hardware consisted of a 2.8 GHz Intel i7-4900 MQ HP Envy 15 notebook. The PC drove a 22.5 inch Samsung 225BW LCD monitor (1680 x 1050 pixels) which displayed the videos. The monitor was mounted on a table at 45 cm off the edge. For gaze data collection, an infrared screen-based Tobii Pro X3-120 eye tracker with 120 Hz sampling frequency was attached to the display. Calibration and gaze feature extraction were performed with the Tobii Pro Studio software. The experiment was performed in our research lab under controlled conditions. We asked participants to sit in a chair in front of the display. They were free to adjust the distance to the display for their own comfort. During the experiment, participants were wearing headphones in order to minimize distractions.

3.3 Procedure

The study took about 10 minutes for each participant. A nine-point calibration was performed for each subject. We verified that, after calibration, gaze accuracy was sufficient to unambiguously assign gaze data to an AOI. Subjects were shown three videos in fullscreen mode (1280 x 720 pixels), each featuring two AOIs (see Figure 1): Video A shows two objects, namely a violin and an oboe turning slowly around their own axis. Video B features humans (two cellists playing chamber music) and introduces the potentially biasing element of sound. In Video C, two musicians take turns describing their instruments, thus eliciting preferences for human elements. The theme for the videos was chosen based on the assumption that most people have rather weak preferences for a particular instrument. We thus expect the standard feature set resulting from our evaluations to perform at least as well for other videos, independent of strength of the viewer's preferences. The videos were edited to minimize visual distractions such as human hands, shadows, and text elements.

Figure 1: Video sequences for feature definition. Participants watched each scene while their gaze was recorded, and subsequently indicated their preferred element in the scene.

Immediately after watching a video sequence, a screen based selection box asked participants which element they wished to learn more about. Since the AOI in Video C encompasses the instrument (Video C.1) and the person describing it (Video C.2), subjects were asked to state their preference for both elements and the person.

3.4 Data Collection

We extracted 14 features ($f_1 - f_{14}$) which are commonly used in gaze-contingent systems, and included 3 additional features ($f_{15} - f_{17}$) from Tobii Studio [66] (see Table 1):

- (1) **gazeCount**: Total # gaze points
- (2) **decayGaze**: Declining weight of gaze points over time
- (3) **firstFix**: Whether the first fixation was in the AOI
- (4) **lastFix**: Whether the last fixation was in the AOI
- (5) **fixCount**: Total # fixations
- (6) **totalFixDur**: Total duration of all fixations
- (7) **meanFixDur**: Mean duration of fixations
- (8) **stdvFixDur**: Standard deviation of fixations
- (9) **maxContGazeDur**: Maximum continuous gaze time
- (10) **meanContGazeDur**: Mean continuous gaze time
- (11) **pupilSize**: Mean pupil size while looking at the AOI
- (12) **endGazeCount**: Gaze points during the last 400 ms
- (13) **revisitCount**: # times the gaze returned to the AOI
- (14) **maxFix**: Duration of the longest fixation
- (15) **longFix**: # fixations longer than 400 ms
- (16) **maxContFixCount**: Maximum # continuous fixations
- (17) **meanContFixCount**: Mean # continuous fixations

All features are AOI specific, i.e., they are defined with regard to their occurrence within an AOI.

Table 1: Gaze features used in gaze-contingent systems

	FEATURES																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
[5]		•															
[10]	•																
[11]			•		•												
[12]					•												
[17]					•	•	•	•						•	•		
[24]					•	•	•	•									
[28]					•	•	•	•						•	•		
[30]								•		•							
[31]					•	•	•										
[35]					•	•	•							•			
[36]					•	•	•		•	•				•			
[37]					•	•	•										
[38]					•	•	•										
[39]					•	•	•	•						•	•		
[45]										•							
[47]					•	•											
[54]				•		•											
[55]				•		•											
[56]					•	•	•			•							
[59]	•													•	•		
[63]	•	•															
[64]					•	•	•	•						•	•		
[68]	•	•			•												
[66]															•	•	•

3.5 Results

Due to technical issues with the eye tracker, six experimental runs failed to produce a valid gaze output. The following analyses were thus conducted with the remaining 95 samples.

3.5.1 Q1: Can a standard gaze feature set for preference predictions in videos be defined? The first objective of the study was to define a set of standard gaze features. We reduced the original feature set

based on Spearman rank correlations. In order to increase generalizability, we aggregated the reported preference and features from all three videos. We used participants' reported preference for the instrument (instead of the person) in Video C, as this metric resulted in higher correlations between reported and predicted preference.

y	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
y	1.00	0.30	0.31	-0.20	0.19	0.20	0.30	0.07	0.03	0.18	0.27	0.00	0.20	0.05	0.06	0.21	0.23
(1)	0.30	1.00	0.83	-0.20	0.39	0.62	0.99	0.33	0.40	0.67	0.75	0.09	0.32	0.32	0.44	0.61	0.50
(2)	0.31	0.83	1.00	-0.19	0.55	0.47	0.85	0.36	0.42	0.66	0.69	0.13	0.43	0.20	0.44	0.45	0.42
(3)	-0.20	-0.20	-0.19	1.00	-0.12	-0.18	-0.20	-0.05	0.00	-0.10	-0.14	-0.07	-0.14	-0.23	0.01	-0.17	-0.14
(4)	0.19	0.39	0.55	-0.12	1.00	0.28	0.44	0.15	0.17	0.26	0.27	-0.02	0.88	0.35	0.23	0.24	0.23
(5)	0.20	0.62	0.47	-0.18	0.28	1.00	0.62	-0.44	-0.24	0.27	0.26	0.06	0.21	0.41	-0.07	0.38	0.76
(6)	0.30	0.99	0.83	-0.20	0.44	0.62	1.00	0.33	0.40	0.66	0.74	0.08	0.37	0.35	0.45	0.62	0.50
(7)	0.07	0.33	0.36	-0.05	0.15	-0.44	0.33	1.00	0.82	0.46	0.56	0.05	0.14	-0.11	0.65	0.27	-0.32
(8)	0.03	0.40	0.42	0.00	0.17	-0.24	0.40	0.82	1.00	0.56	0.56	0.04	0.16	-0.05	0.92	0.20	-0.19
(9)	0.18	0.67	0.66	-0.10	0.26	0.27	0.66	0.46	0.56	1.00	0.75	0.09	0.18	0.04	0.60	0.39	0.35
(10)	0.27	0.75	0.69	-0.14	0.27	0.26	0.74	0.56	0.56	0.75	1.00	0.09	0.22	-0.17	0.53	0.55	0.36
(11)	0.00	0.09	0.13	-0.07	-0.02	0.06	0.08	0.05	0.04	0.09	0.09	1.00	-0.05	0.05	0.01	0.09	0.09
(12)	0.20	0.32	0.43	-0.14	0.21	0.37	0.14	0.16	0.18	0.22	-0.05	-0.09	0.32	0.20	0.17	0.18	0.10
(13)	0.05	0.32	0.20	-0.23	0.35	0.41	0.35	-0.11	-0.05	0.04	-0.17	0.05	0.32	1.00	0.04	0.13	0.10
(14)	0.06	0.44	0.44	0.01	0.23	-0.07	0.45	0.65	0.92	0.60	0.53	0.01	0.20	0.04	1.00	0.21	-0.06
(15)	0.21	0.61	0.45	-0.17	0.24	0.38	0.62	0.27	0.20	0.39	0.55	0.09	0.17	0.13	0.21	1.00	0.34
(16)	0.23	0.50	0.42	-0.14	0.23	0.76	0.50	-0.32	-0.19	0.35	0.36	0.09	0.18	0.10	0.06	0.34	1.00
(17)	0.19	0.53	0.42	-0.10	0.15	0.89	0.52	-0.43	-0.25	0.28	0.37	0.02	0.10	0.01	-0.11	0.37	0.80

Figure 2: Correlation of features and reported preference (y).

Figure 2 shows the correlations between the features and with reported preference (y). We excluded four features (decayGaze, lastFix, totalFixDur, meanContFixCount) due to high multicollinearity with another feature (corr. $\geq .8$). Four features (firstFix, stdvFixDur, pupilSize, revisitCount) were excluded that had no significant correlation with participants' reported preference (corr. $\leq .05$). The remaining nine features were defined as standard features.

3.5.2 Q2: Do the features accurately predict the preferred element? The second objective of the study was to demonstrate the ability of the extracted gaze features to infer a viewer's preferences for specific video elements. We defined a *Majority Voter* which predicts the viewer's preferred element by determining the AOI with the highest value of an individual gaze feature (see Algorithm 1).

Algorithm 1 Majority Voter

Input : featureValues of sample x_i for all AOI r_i of set R
Output : r_i with max{featureValue}

```

max = 0
foreach value( $r_i$ ) in featureValues do
    if value( $r_i$ ) > max then
        max ← value( $r_i$ )
    end
end

```

We applied the Majority Voter to the extracted features to predict the viewer's preferred element in each of the three videos. We intentionally used videos for which participants were expected to show very weak preferences. This allowed us to identify those features which predict preferences even under fairly inconclusive conditions. Prediction accuracies ranged from 50.71% (maxFix) to 65.96% (gazeCount). Table 2 summarizes the performance of the features. a one-sided z-test for proportions showed that predictions are significantly more accurate than the random baseline for gazeCount (p-value = .0129) and meanContGazeDur (p-value = .0575).

Table 2: Overview of gaze feature performance from Study 1

GAZE FEATURE	EXCLUSION		ACCURACY(MAJORITY VOTER)					p-value
	(y)-corr	coll	Video A	Video B	Video C.1	Video C.2	TOTAL	
(1) gazeCount			71.28%	61.70%	64.89%	61.70%	65.96%	.0129**
(2) decayGaze		x	59.57%	62.77%	71.28%	76.60%	64.54%	—
(3) firstFix	x		35.11%	52.13%	70.21%	64.89%	52.48%	—
(4) lastFix		x	56.38%	52.13%	61.70%	52.13%	56.74%	—
(5) fixCount			69.15%	60.64%	44.68%	51.06%	58.16%	(.1295)
(6) totalFixDur		x	68.09%	59.57%	65.96%	60.64%	64.54%	—
(7) meanFixDur			45.74%	48.94%	65.96%	60.64%	53.55%	(.3122)
(8) stdvFixDur	x		45.74%	55.32%	59.57%	56.38%	53.55%	—
(9) maxContGazeDur			57.45%	56.38%	59.57%	58.51%	57.80%	(.1404)
(10) meanContGazeDur			60.64%	57.45%	65.96%	69.15%	61.35%	.0575*
(11) pupilSize	x		50.00%	50.50%	43.62%	48.94%	47.87%	—
(12) endGazeCount			50.00%	52.13%	63.83%	60.64%	55.32%	(.2314)
(13) revisitCount	x		52.13%	36.17%	23.40%	23.40%	37.23%	—
(14) maxFix			44.68%	45.26%	53.19%	52.13%	50.71%	(.4610)
(15) longFix			63.83%	50.00%	47.87%	53.19%	53.90%	(.2953)
(16) maxContFixCount			58.51%	55.32%	41.89%	48.94%	51.77%	(.4036)
(17) meanContFixCount		x	61.70%	63.83%	51.06%	55.32%	58.87%	—

Note: (y)-corr = low correlation with the reported preference; coll = multicollinearity with another feature.

Features which were excluded based on correlations are greyed out. Total accuracy is the percentage of correct predictions using the gaze feature across videos A, B, and C.1. p-values are calculated with a one-sided z-test for the combined accuracy of all videos (significance level: *10%, **5%).

We further found that the prediction performance varies considerably between the videos. It has been suggested that audio disparity influences bottom-up attention [26]. However, since accuracy scores for Video C (audio disparity: people talking in turns) were higher than for Video B (audio uniformity: musicians playing in unison), we found no evidence for a biasing effect of sound.

Instead, we found indications that predictions are more accurate for objects that the viewer is well familiar with. In Video C, feature correlation was stronger with subjects' preference for the instrument than for the person. The high accuracy for Video A, where preferences were elicited for an instrument (instead of a person) support this assumption. We suspect that the effect is caused by the strength of the viewer's preference, assuming that preferences are stronger for familiar objects than for a random person. In a second study, we thus evaluate different algorithms on our feature set, and test whether their performance depends on the video design.

4 STUDY 2: PREDICTING PREFERENCES

We ran a second study (N = 123) to quantify how accurately preferences can be predicted. We recorded participants' gaze while watching five videos. The continuation of a video was determined by their explicit selection of an element. We examine five questions:

- **Q3: Can we predict viewers' preferred element by looking at their absolute gaze allocation to each video element?**
We apply majority voting to individual gaze features and compare the predicted preference to the viewer's choice.
- **Q4: Does relative gaze allocation suggest a visual bias?**
Absolute gaze might be biased by bottom-up attention. We thus compute relative gaze allocation, i.e., we multiply the gaze data recorded inside each AOI with its in the sample.

- **Q5: Can machine learning improve prediction accuracy?**
Machine learning might be able to capture interdependencies between video elements or systematic biases towards a salient element. We thus examine the prediction performance of different machine learning classifiers.
- **Q6: Does the video design systematically influence gaze?**
Design properties such as color, motion, or size of elements might bias a viewer's gaze. We analyze the effect of multiple design properties on the accuracy of preference predictions.
- **Q7: Can we quantify the strength of the viewer's preference?**
So far, we predicted which element a viewer prefers. Here, we test whether we can infer the strength of this preference.

4.1 Participants

We recruited 132 participants (73 male, mean age = 26.61, $\sigma = 9.22$) with normal or corrected to normal vision. Our sample consisted of students and employees of our research institution, as well as visitors of a science fair. Participation was voluntary and no monetary incentive was given. Subjects were told that their gaze was being tracked, but were not fully aware of the purpose of the experiment.

4.2 Apparatus

The software ran on a 2.5GHz Intel Core i7-6500 HP PC. The PC drove a 23.6 inch iiyama touch monitor (1920 x 1080 pixels) on which the videos were displayed. Gaze data was collected with the same Tobii Pro X3-120 eye tracker that we used in the first study.

The experiment was performed in our research lab and a science fair. Participants were seated in front of the display, maintaining a distance between their eyes and the eye tracker of 20-30 inches.

4.3 Experimental Material

The experiment consisted of five videos reflecting the genres commonly available on multimedia platforms.

Table 3: Design properties of videos I-V

VIDEO	FACTORS INFLUENCING BOTTOM-UP ATTENTION												AOI	TILES	LENGTH
	COLOR		MOTION		TEXTURE		SIZE		DETAIL		PREFERENCE				
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ			
(I)	.50	.02	.29	.01	.16	.01	.42	-	.10	.03	.76	.08	2	2	42
(II)	.35	.06	.45	.16	.38	.11	.24	.12	.16	.02	.71	.11	4	1	36
(III)	.19	.01	.01	.00	.18	.01	.42	-	.02	.01	.72	.11	2	2	18
(IV)	.94	.08	.48	.03	.26	.07	.32	-	.21	.04	.71	.12	4	4	53
(V)	.15	.04	.53	.03	.12	.01	.61	.09	.30	.00	.70	.09	2	1	25

Note: Values of factors influencing bottom-up attention are normalized over all videos. μ = mean value of all AOI in the video; σ = standard deviation of all AOI in the video.

We used videos that vary with regard to the factors that have been identified in previous literature to influence bottom-up attention (see Table 3). This allows us to investigate which video design properties bias the user’s viewing patterns in dynamic video scenes. Specifically, we evaluated the following *inherent* (i.e., attributable to the video) design properties: *color* [6, 8, 14, 22, 27, 50, 51, 60, 65] (calculated according to the visual saliency model proposed by Itti and Koch [27]), the degree of *motion* [6, 26, 65] measured by the number of changed pixels per frame, *texture* [6, 8, 51] (defined as the number of sharp edges in a frame), *size* [8, 14, 65], and *detail* [6, 8, 38].

We additionally considered one *external* (i.e., attributable to the viewer) metric, namely the strength of *preference* that an element generally elicits. A strong influence of this metric on the prediction accuracy would indicate that adaptive systems only work well when preferences for an adaptation target differ significantly across users. We conducted a prestudy with 227 subjects (86 female, mean age = 31.76, σ = 7.23) on MTurk, in which participants were shown the five videos. After watching a video, they were asked to quantify for each AOI (using a 7-point Likert scale) their desire to watch a scene featuring exclusively the respective element.

- (I) **Geography:** Parallel streams of a bird’s eye view on the Namibian landscape and attractions of New York.
→ *tiled screen, 2 AOIs*

- (II) **Instrumental:** Performance of a jazz band featuring a pianist, a saxophone player, a bassist, and a drummer.
→ *fullscreen, 4 AOIs*

- (III) **Advertising:** Parallel slideshows of a fashion show and a car presentation. Images are switched simultaneously after 3 seconds.
→ *tiled screen, 2 AOIs*

- (IV) **Sports:** Parallel streams of the last few seconds from a basketball game, a rowing race, a biathlon, and a 800m final.
→ *tiled screen, 4 AOIs*

- (V) **Infotainment:** Opening scene of an infotainment show featuring two show hosts engaged in silent conversation. Overlaid with opening music.
→ *fullscreen, 2 AOIs*

Figure 3: Experimental material consisting of 5 videos from different genres: Geography [19],[61], Instrumental, Advertising [9],[18], Sports [40],[42],[43],[69], Infotainment.

We used two fullscreen videos (II, V) and three tiled videos (I, III, IV) (see Figure 3). The latter show two (or four) video streams in parallel. Such a design allows, e.g., to simultaneously present multiple trailer previews. An adaptive system may subsequently display a teaser scene of the movie that received the most attention.

4.4 Procedure

The experiment took approximately 15 minutes for each participant. At the beginning of the experiment, a 5-point calibration was performed for each subject. Completing the procedure required successful eye detection. This ensured that gaze data was collected for all participants. Participants were instructed to sit in a comfortable position and watch a sequence of videos. By granting them freedom of movement, we receive more realistic estimates of the system’s predictive performance in a real-world scenario. Subjects were then shown the first video while the eye tracker recorded their gaze. The video halted at a decisive point and participants were asked to select the stream they would like to watch until the end. They were informed that they would not be able to watch any of the other possible continuations afterwards. Subsequently, the chosen stream resumed, but no further gaze data was collected. After watching all videos, participants were presented a questionnaire to rate the strength of their preference for each element in the videos. Upon finishing the experiment, they were asked to fixate five circles which consecutively appeared in each corner and in the center of the display to verify the accuracy of the recorded gaze data.

4.5 Results

29 samples were excluded due to insufficient quality of gaze recordings. The remaining 103 samples allow determining gaze allocation to an AOI (mean error < .9 degrees of visual angle). Fixations were calculated based on the fixation filters proposed by Tobii [41].

4.5.1 Q3: Can we predict viewers’ preferred element by looking at their absolute gaze allocation to each video element? In a first step, we evaluated the performance of majority voting for each feature separately. Absolute features select the AOI with the highest feature value computed for an individual participant. Their main advantage is that no further knowledge is required (in contrast to relative features and machine learning that both require prior data collection). The prediction accuracies are summarized in Table 4. Seven features (fixCount, meanFixDur, maxContGazeDur, meanContGazeDur, endGazeCount, maxFix, maxContFixCount) were dominated by at least one other feature. The performance of gazeCount and longFix differed across the videos. Consistent with our findings

Table 4: Accuracy of preference predictions for videos I-V using the Majority Voter on each feature

Feature	ACCURACY									
	Video I		Video II		Video III		Video IV		Video V	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
(1) gazeCount	.462	.712	.500	.452	.731	.712	.657	.626	.598	.686
(5) fixCount	.644	.731	.279	.365	.462	.587	.556	.485	.578	.598
(7) meanFixDur	.615	.721	.298	.317	.423	.500	.232	.232	.588	.588
(9) maxContGazeDur	.606	.731	.346	.433	.317	.481	.182	.182	.471	.480
(10) meanContGazeDur	.558	.712	.337	.433	.279	.519	.192	.202	.480	.510
(12) endGazeCount	.385	.721	.442	.452	.731	.731	.576	.566		
(14) maxFix	.635	.731	.308	.327	.433	.500	.394	.364	.598	.598
(15) longFix	.731	.731	.240	.260	.404	.490	.343	.333	.608	.608
(16) maxContFixCount	.731	.731	.135	.135	.356	.356	.283	.283	.598	.598

Note: Features which dominate all others in at least one video are highlighted. Absolute features select the AOI with highest feature value. Relative features multiply absolute features with the proportion by which the element was preferred in the sample.

from the first study, gazeCount overall delivered the best predictions. This suggests that preference for elements in a video can be best predicted when the entire gaze trajectory (including both fixations and saccades) is taken into account. Aggregating several gaze points into a gaze feature negatively impacts the prediction.

When applying majority voting to gazeCount, the largest improvement was achieved in Video IV (*Sports*). While random guessing only delivers 25% accurate predictions, we correctly predicted the preferred element in 65.7% of the samples. A one-sided z-test for proportions confirmed a significant improvement over the random baseline with 99% confidence for Video II (p-value = .000), Video III (p-value = .000), and Video IV (p-value = .000). The effect was significant on a 90% level for Video V (p-value = .079). In Video I (p-value = .290), majority voting did not perform significantly better.

Figure 4: Gaze of a viewer while watching Video IV. AOI 1 receives the most attention and is also selected by the viewer (indicated in yellow).

An exemplary gaze output for Video IV is presented in Figure 4: Gaze allocation (defined by gazeCount) is strongly correlated with the viewer's preference ratings. The stream that received the most visual attention was also selected by the viewer to be continued.

4.5.2 Q4: Does relative gaze allocation suggest a visual bias? In a second step, we evaluated the prediction performance of relative gaze features. These take into account the distribution of preferences in the entire sample and thus account for the visual bias caused by bottom-up attention.

We normalized each gaze feature to account for varying units and gaze data quality across subjects. The value of feature f in AOI r_i of test subject n_j was determined in proportion to the sum

of the values over all AOI $r_i \in R$:

$$abs_f(r_i, n_j) = \frac{f(r_i, n_j)}{\sum_{i=1}^R f(r_i, n_j)} \quad (1)$$

We then introduced a popularity index pop , which is calculated as the number of times the AOI r_i represented a subject's preferred element $pref(r_i, n_j)$, divided by the sample size N . The feature value is then multiplied with the popularity index:

$$pop(r_i) = \frac{\sum_{j=1}^N pref(r_i, n_j)}{N}, pref(r_i, n_j) \in [0, 1] \quad (2a)$$

$$rel_f(r_i, n_j) = abs_f(r_i, n_j) * pop(r_i) \quad (2b)$$

Based on the resulting relative features, we again evaluated the performance of gaze-based preference predictions with the Majority Voter. Table 4 contrasts the resulting accuracy values with the performance of absolute gaze features. Similar to the results from Q3, most accurate predictions were achieved with the feature gazeCount. A one-sided z-test evinced that the prediction accuracy of the Majority Voter when applied to the relative gazeCount improved significantly compared to the absolute gazeCount for videos I (p-value = .000***) and V (p-value = .093*). Coincidentally, these videos had also resulted in below-average prediction performance when applying the Majority Voter to absolute gaze features. Similar to machine learning models, relative features take into account empirical preference distributions. A priori collected scene-specific data is considered in the calculation the popularity index pop . This confirms our assumption that, in sequences with especially low accuracy, the absolute features were biased by bottom-up attention.

We therefore proceeded to evaluate the prediction performance of the gaze features when applied to machine learning models. We investigate whether the a priori gathering of information about the general viewing and preference patterns of viewers for a specific scene can improve the accuracy of preference predictions.

4.5.3 Q5: Can machine learning improve prediction accuracy? We evaluated whether machine learning models can deliver more accurate preference predictions than majority voting. The models are trained with gaze data from subjects who have previously been shown the same scenes. This accounts for possible biases if the gaze data is influenced by bottom-up attention [15].

Table 5: Performance of machine learning algorithms for best feature subsets

		BEST FEATURE SUBSET								ACCURACY
		[1]	[5]	[7]	[9]	[10]	[12]	[14]	[15]	
Video I	Majority Voter (absolute)	•								46.15%
	Majority Voter (relative)	•								71.15%
	Logistic Regression	•								73.05%
	SVC	•								73.05%
	Decision Tree (Tree)	•								73.05%
	Random Forest (RF)	•					•		•	74.00%
	MLP	•			•		•		•	75.90%
	AdaBoost	•								73.05%
	Passive Aggressive			•	•					74.00%
	Gradient Boosting	•				•				74.00%
	k-Nearest Neighbor (k-NN)		•		•				•	74.10%
Video II	Majority Voter (absolute)	•								50.00%
	Majority Voter (relative)	•								45.19%
	Logistic Regression	•	•	•	•		•			55.81%
	SVC	•	•							49.05%
	Decision Tree (Tree)		•			•		•	•	47.10%
	Random Forest (RF)	•	•	•	•	•	•	•		52.90%
	MLP	•								52.86%
	AdaBoost	•	•			•	•	•		50.90%
	Passive Aggressive	•				•	•		•	50.59%
	Gradient Boosting	•	•		•		•		•	54.86%
	k-Nearest Neighbor (k-NN)	•		•		•				47.14%
Video III	Majority Voter (absolute)	•								73.08%
	Majority Voter (relative)	•								71.15%
	Logistic Regression	•							•	76.90%
	SVC	•	•						•	80.71%
	Decision Tree (Tree)	•								79.76%
	Random Forest (RF)	•	•	•	•	•	•		•	78.86%
	MLP	•	•							81.76%
	AdaBoost	•								79.76%
	Passive Aggressive		•	•		•	•	•	•	74.81%
	Gradient Boosting	•								78.76%
	k-Nearest Neighbor (k-NN)	•		•	•				•	75.95%
Video IV	Majority Voter (absolute)	•								65.66%
	Majority Voter (relative)	•								62.63%
	Logistic Regression	•	•		•					70.74%
	SVC	•	•							68.63%
	Decision Tree (Tree)	•		•					•	44.26%
	Random Forest (RF)	•			•			•	•	69.79%
	MLP	•			•				•	67.68%
	AdaBoost	•						•		63.63%
	Passive Aggressive	•		•				•		67.79%
	Gradient Boosting	•	•		•	•	•	•	•	73.74%
	k-Nearest Neighbor (k-NN)	•								65.63%
Video V	Majority Voter (absolute)	•								59.80%
	Majority Voter (relative)	•								68.63%
	Logistic Regression	•	•	•						67.67%
	SVC	•						•	•	69.62%
	Decision Tree (Tree)	•	•		•			•		63.81%
	Random Forest (RF)	•		•	•				•	69.52%
	MLP	•	•	•	•		•		•	69.76%
	AdaBoost	•								61.90%
	Passive Aggressive	•		•						65.62%
	Gradient Boosting	•		•					•	63.76%
	k-Nearest Neighbor (k-NN)	•	•		•		•	•		69.57%

Nine classifiers from the Python scikit-learn library [46] were tested using 5-fold cross-validation. The classifiers were selected based on their usage in related systems which infer user preferences or intentions from implicit user feedback:

- Logistic regression [31],[36],[62]
- SVC [20],[25],[29],[49],[55],[62]
- Decision Tree [34],[53]
- Random Forest [59]
- Multi-Layer Perceptron (MLP) [53]
- AdaBoost [59],[62]
- Passive Aggressive Algorithm [62]
- Gradient Boosting [57],[62]
- k-Nearest Neighbor [52],[53],[32]

We performed best subset selection by training each classifier with all possible combinations of gaze features (see Table 5). For videos with four AOIs, accuracy of the best performing classifier ranged from 55.81% (Video II) to 73.74% (Video IV). In the binary case, we even achieved values ranging from 69.76% (Video V) to 81.76% (Video III). Figure 5 shows that machine learning outperformed not only the random baseline, but also the Majority Voter with both absolute and relative features throughout the entire set of tested videos. The improvement over majority voting with absolute features was significant with 95% confidence for videos I ($\Delta acc = 33.56\%$, $p\text{-value} = .000$) and V ($\Delta acc = 12.63\%$, $p\text{-value} = .027$), where relative features also returned better results. Overall, the MLP classifier delivered the most reliable predictions, with a mean accuracy of 75.81% (64.78%) over all videos with two (four) AOIs.

Figure 5: Accuracy of preference predictions. MLP outperforms both the random baseline and majority voting.

4.5.4 Q6: Does the video design systematically influence gaze? While filtering out confounding factors, machine learning requires collecting training data for each scene. Even though our results show that machine learning overall produces more accurate predictions, majority voting does deliver comparable results for some videos (i.e., videos II, III, and IV). In order to determine whether it can be anticipated for which videos the Majority Voter returns largely unbiased predictions, we investigated the effect of video design properties on the *visual bias* which is induced by bottom-up attention. We defined visual bias as the difference between the prediction accuracy of the Majority Voter with absolute gazeCount and MLP (i.e., our overall most reliable classifier). This leverages the fact that absolute gaze features are direct measures of visual attention. They thus encompass both top-down attention (i.e., preference) and bottom-up attention (i.e., visual bias). Machine learning models, in contrast, filter out the part in visual attention which is not caused by preferences. This is due to bottom-up attention being similar across subjects and therefore being reflected in the training data.

Figure 6 illustrates the relationship between the visual bias of a video and the degree of difference between its AOIs, measured by their standard deviation. We analyzed the difference measure for each design property. It shows that none of the inherent video design properties is positively correlated with visual bias. A weak relationship does exist though for the external metric *preference*, which measures the strength of anticipated preference for a video element. Thus, rather than by the video design, the gaze appears to be mostly guided by the strength of the viewer's own preferences.

Figure 6: Effect of salience metrics on visual bias, normalized over all videos. PREFERENCE is plotted in reversed order to account for its expected negative effect on accuracy.

4.5.5 Q7: Can we quantify the strength of the viewer's preference?

Determining user preferences in longer videos where elements can appear in different constellations or leave the plot may require establishing not only the viewer's preferred element in the video, but also the strength of their preference for each element.

Table 6: Multiple regression of gaze features on reported preference ratings

STATISTIC	(I)	(II)	(III)	(IV)	(V)
<i>Adjusted R²</i>	.050	.132	.406	.430	.044
<i>F-statistic</i>	11.93	64.25	36.31	100.39	5.68
<i>Prob(F-statistic)</i>	.00***	.00***	.00***	.00***	.00***
(1) <i>beta</i>	—	.209	.525	.455	.146
<i>p-value</i>	—	.00***	.00***	.00***	.02**
(5) <i>beta</i>	.108	—	-.285	-.079	—
<i>p-value</i>	.00***	—	.00***	.00***	—
(10) <i>beta</i>	—	—	-.171	—	—
<i>p-value</i>	—	—	.01**	—	—
(12) <i>beta</i>	—	—	.201	—	—
<i>p-value</i>	—	—	.00***	—	—

Note: Not significant coefficients are omitted (significance level: *10%, ** 5%, *** 1%).

We therefore ran a multiple regression analysis of gaze features on the strength of participants' preference ratings for each element in the videos (as expressed in the follow-up questionnaire). The regression output reveals that we can indeed rank the importance of each element (see Table 6). Depending on the video, the variation

Figure 7: Scatterplots of the relationship between absolute gazeCount and reported preference ratings. A clear trend shows for videos with significant correlations (p-value < .01).

in the strength of preference for an element is to a statistically significant part explained by gazeCount (II, III, IV, V), fixCount (I, III, IV), meanContGazeDur (III), and endGazeCount (III).

Figure 7 illustrates the distribution of gaze (measured by gazeCount) relative to viewers' reported preference ratings for video elements. The scatterplots for videos II-IV show that data points are clustered around the regression line for videos with highly significant correlations (p-value < .01). In contrast, no patterns exist for videos I and V where gazeCount is not significantly correlated with the strength of reported preferences. The results indicate that gaze data does not only allow to predict the viewer's preferred element, but on top reveals information about the preference strength.

5 DISCUSSION

In two independent user studies, we demonstrated that it is possible to spontaneously tailor a video to the viewer's preferences without any active user involvement. Our findings may guide developers of gaze-contingent systems in choosing the optimal features and algorithm for inferring a person's preference for video elements. In the following, we discuss the significance of our findings, as well as their generalizability with regard to the video material and the environmental conditions.

5.1 Significance of the Results

Using machine learning classifiers, we were able to predict which of two elements in a video a person would like to continue watching with an accuracy ranging from 70% to 82%. When there is a choice between four videos, accuracies still range from 56% to 74%. We further showed the ability of gaze data to elicit latent preferences, providing a more detailed view on the strength of the viewer's preference for different video elements. Although the evaluated strategies already provide good predictions, we believe that extending the feature set by smooth pursuits might lead to even better results. The current feature set was derived from previous work focusing on static images. In dynamic scenes, smooth pursuits occur when the gaze follows a moving object [71]. This being a deliberate focus on the object, we expect a high correlation with interest.

5.2 Generalizability Across Videos

Absolute gaze distributions only accurately predict the viewer's preferred element if the video does not evoke bottom-up attention. Machine learning models filter out confounding factors, but require prior collection of scene-specific training data. In order to avoid repeating the time-consuming data collection for each new scene, Qvarfodt and Zhai [48] propose filtering out bottom-up attention

by assigning parameters for confounding factors to each object. However, our results indicate that, rather than the inherent design properties of a video, the strength of the viewer's own preferences are the most important determinant for prediction accuracy. Visual bias caused by bottom-up attention, in contrast, showed to have a negligible effect. Cheng et al. [11] propose a generalized two-step prediction model: First, the new scene is compared to a database of scenes for which training data has been collected. Second, machine learning is applied to the training data of comparable scenes. The fitted model is then used to predict preferences for the new scene.

5.3 Influence of Environmental Factors

The reported accuracy values of our preference predictions have been achieved with gaze data that was collected under controlled laboratory conditions. In a real-world setting, numerous confounding factors (changing lighting conditions, users frequently diverting their gaze away from the screen, etc.) will lower the quality of the recorded gaze data and make preference predictions less accurate. Further, infrared eye tracking systems like the *Tobii Pro X3-120* that we used in our studies require special hardware which is impractical and too expensive for a common user. Open-source solutions that use built-in device cameras [1, 44, 74, 75] are still less accurate than specialized hardware. However, they are rapidly catching up in their performance and are becoming more resilient to changing environmental conditions.

6 CONCLUSION

We evaluated the performance of different gaze features and prediction strategies for inferring a viewer's preferred element in a video. In a first user study ($N = 101$), defined a standard feature set for gaze-based preference prediction in videos encompassing nine gaze features. In a second user study ($N = 132$), we showed that absolute features can predict user preferences with an accuracy of up to 73% for two elements, and 66% for four elements. Using the MLP classifier, prediction accuracy can be increased to 82% (74%) for videos with two (four) elements. We also provided evidence that gaze data reveals insights into the strength of the viewers' preference. We further found that prediction performance is largely dependent on the user's intrinsic strength of preference for the evaluated video elements, rather than by extrinsic video design properties.

In future work, we plan to evaluate how personalized content delivery based on gaze data affects user engagement. We further plan to study how additional features derived from physiological sensors can be used to improve the system.

REFERENCES

- [1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: An Open Source Facial Behavior Analysis Toolkit. In *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision* (Lake Placid, NY, USA) (WACV '16). IEEE, 1–10. <https://doi.org/10.1109/WACV.2016.7477553>
- [2] Hyejin Bang and Bartosz W. Wojdowski. 2016. Tracking Users' Visual Attention and Responses to Personalized Advertising Based on Task Cognitive Demand. *Computers in Human Behavior* 55, PB (2016), 867–876. <https://doi.org/10.1016/j.chb.2015.10.025>
- [3] Roman Bednarik. 2005. Potentials of Eye-Movement Tracking in Adaptive Systems. In *Proceedings of the 4th Workshop on Empirical Evaluation of Adaptive Systems* (UM '05). Edinburgh, Scotland, 1–8.
- [4] Richard A. Bolt. 1981. Gaze-Orchestrated Dynamic Windows. *SIGGRAPH Computer Graphics* 15, 3 (aug 1981), 109–119. <https://doi.org/10.1145/965161.806796>
- [5] Boris Brandherm, Helmut Prendinger, and Mitsuru Ishizuka. 2007. Interest Estimation Based on Dynamic Bayesian Networks for Visual Attentive Presentation Agents. In *Proceedings of the 9th International Conference on Multimodal Interfaces* (Nagoya, Aichi, Japan) (ICMI '07). ACM, New York, NY, USA, 346–349. <https://doi.org/10.1145/1322192.1322253>
- [6] Neil D. B. Bruce, Calden Wloka, Nick Frosst, Shafin Rahman, and John K. Tsotsos. 2015. On Computational Modeling of Visual Saliency: Examining What's Right, and What's Left. *Vision Research* 116, B (nov 2015), 95–112. <https://doi.org/10.1016/j.visres.2015.01.010>
- [7] Raymond R. Burke and Alex Leykin. 2014. Identifying the Drivers of Shopper Attention, Engagement, and Purchase. *Review of Marketing Research* 11 (2014), 147–187. <https://doi.org/10.1108/S1548-643520140000011006>
- [8] Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. 2009. What Do You See when You're Surfing?: Using Eye Tracking to Predict Salient Regions of Web Pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) (CHI '09). ACM, New York, NY, USA, 21–30. <https://doi.org/10.1145/1518701.1518705>
- [9] Chanel. 2019. Herbst-Winter 2019/20. https://www.chanel.com/de/_JDE/mode/kollektion/herbst-winter-2019-20/die-looks.html
- [10] Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. 2011. Eye Activity as a Measure of Human Mental Effort in HCI. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (Palo Alto, CA, USA) (IUI '11). ACM, New York, NY, USA, 315–318. <https://doi.org/10.1145/1943403.1943454>
- [11] Shiwei Cheng, Xiaojian Liu, Pengyi Yan, Jianbo Zhou, and Shouqian Sun. 2010. Adaptive User Interface of Product Recommendation Based on Eye-tracking. In *Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction* (Hong Kong, China) (EGIHMI '10). ACM, New York, NY, USA, 94–101. <https://doi.org/10.1145/2002333.2002348>
- [12] Shiwei Cheng and Ying Liu. 2012. Eye-Tracking Based Adaptive User Interface: Implicit Human-Computer Interaction for Preference Indication. *Journal on Multimodal User Interfaces* 5, 1-2 (2012), 77–84. <https://doi.org/10.1007/s12193-011-0064-6>
- [13] Andrew T. Duchowski. 2002. A Breadth-First Survey of Eye-Tracking Applications. *Behavior Research Methods, Instruments, and Computers* 34, 4 (2002), 455–470. <https://doi.org/10.3758/BF03195475>
- [14] F. L. Engel. 1974. Visual Conspicuity and Selective Background Interference in Eccentric Vision. *Vision Research* 14, 7 (jul 1974), 459–471. [https://doi.org/10.1016/0042-6989\(74\)90034-0](https://doi.org/10.1016/0042-6989(74)90034-0)
- [15] Ulrich Engelke and Patrick Le Callet. 2015. Perceived Interest and Overt Visual Attention in Natural Images. *Signal Processing: Image Communication* 39, B (2015), 386–404. <https://doi.org/10.1016/j.image.2015.03.004>
- [16] Andrea Finke, Kai Essig, Giuseppe Marchioro, and Helge Ritter. 2016. Toward FRP-based Brain-Machine Interfaces - Single-trial Classification of Fixation-related Potentials. *PLoS ONE* 11, 1 (2016), 1–19. <https://doi.org/10.1371/journal.pone.0146848>
- [17] Matthew Gingerich and Cristina Conati. 2015. Constructing Models of User and Task Characteristics from Eye Gaze Data for User-Adaptive Information Highlighting. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (Austin, Texas) (AAAI '15). AAAI Press, 1728–1734.
- [18] Wolfgang Gomoll. 2018. Mercedes C300 de: Spürbare Elektro-Vitaminkur. Retrieved April 16, 2020 from <https://www.automobil-produktion.de/hersteller/neue-modelle/mercedes-c-300-de-spurbare-elektro-vitaminkur-353.html>
- [19] Gondwana Collection Namibia. 2019. The Best of Namibia - A Bird's Eye View. Retrieved April 16, 2020 from <https://www.youtube.com/watch?v=eQvosFjxK2M&list=PLB2F63866773CB519>
- [20] David R. Hardoon, John Shawe-Taylor, Antti Ajanki, Kai Puolamäki, and Samuel Kaski. 2007. Information Retrieval by Inferring Implicit Queries from Eye Movements. *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics* 2 (mar 2007), 179–186. <http://proceedings.mlr.press/v2/hardoon07a.html>
- [21] David R. Hardoon, John Shawe-Taylor, Antti Ajanki, Kai Puolamäki, and Samuel Kaski. 2007. Information Retrieval by Inferring Implicit Queries From Eye Movements. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics* (Procedings of Machine Learning Research), Marina Meila and Xiaotong Shen (Eds.), Vol. 2. PMLR, San Juan, Puerto Rico, 179–186. <http://proceedings.mlr.press/v2/hardoon07a.html>
- [22] Jonathan Harel, Christof Koch, and Pietro Perona. 2007. Graph-based Visual Saliency. In *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, 545–552. <https://doi.org/10.7551/mitpress/7503.003.0073>
- [23] Melanie Heck, Janick Edinger, and Christian Becker. 2019. Gaze-based Product Filtering: A System for Creating Adaptive User Interfaces to Personalize Stateless Point-of-sale Machines. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (UIST '19). ACM, New Orleans, LA, USA, 75–77. <https://doi.org/10.1145/3332167.3357120>
- [24] John M. Henderson, Svetlana V. Shinkareva, Jing Wang, Steven G. Luke, and Jenn Olejarczyk. 2013. Predicting Cognitive State From Eye Movements. *PLoS ONE* 8, 5 (may 2013), e64937. <https://doi.org/10.1371/journal.pone.0064937>
- [25] Eric Horvitz. 1999. Principles of Mixed-initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) (CHI '99). ACM, New York, NY, USA, 159–166. <https://doi.org/10.1145/302979.303030>
- [26] Laurent Itti and Christof Koch. 2001. Computational Modelling of Visual Attention. *Nature Reviews Neuroscience* 2, 3 (mar 2001), 194–203. <https://doi.org/10.1038/35058500>
- [27] Laurent Itti, Christof Koch, and Ernst Niebur. 1988. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (nov 1988), 1254–1259. <https://doi.org/10.1017/S1042771600010292>
- [28] Natasha Jaques, Cristina Conati, Jason M. Harley, and Roger Azevedo. 2014. Predicting Affect From Gaze Data During Interaction With an Intelligent Tutoring System. In *Proceedings of the International Conference on Intelligent Tutoring Systems* (ITS '14). Springer, 29–38. https://doi.org/10.1007/978-3-319-07221-0_4
- [29] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately Interpreting Clickthrough Data as Implicit Feedback. *SIGIR Forum* 51, 1 (aug 2017), 4–11. <https://doi.org/10.1145/3130332.3130334>
- [30] Melih Kandemir and Samuel Kaski. 2012. Learning Relevance from Natural Eye Movements in Pervasive Interfaces. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (Santa Monica, California, USA) (ICMI '12). ACM, New York, NY, USA, 85–92. <https://doi.org/10.1145/2388676.2388700>
- [31] Melih Kandemir, Veli-Matti Saarinen, and Samuel Kaski. 2010. Inferring Object Relevance from Gaze in Dynamic Scenes. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (Austin, Texas) (ETRA '10). ACM, New York, NY, USA, 105–108. <https://doi.org/10.1145/1743666.1743692>
- [32] Przemysław Kazienko and Michał Adamski. 2007. AdROSA - Adaptive Personalization of Web Advertising. *Information Sciences* 177, 11 (jun 2007), 2269–2295. <https://doi.org/10.1016/j.ins.2007.01.002>
- [33] Peter Kiefer, Ioannis Giannopoulos, Dominik Kremer, Christoph Schlieder, and Martin Raubal. 2014. Starting to Get Bored: An Outdoor Eye Tracking Study of Tourists Exploring a City Panorama. In *Proceedings of the Eye Tracking Research and Applications Symposium* (ETRA '14). ACM, Safety Harbor, Florida, USA, 315–318. <https://doi.org/10.1145/2578153.2578216>
- [34] Jong Woo Kim, Byung Hun Lee, Michael J. Shaw, Hsin-Lu Chang, and Matthew Nelson. 2001. Application of Decision-Tree Induction Techniques to Personalized Advertisements on Internet Storefronts. *International Journal of Electronic Commerce* 5, 3 (dec 2001), 45–62. <https://doi.org/10.1080/10864415.2001.11044215>
- [35] Arto Klami, Craig Saunders, Teófilo E. de Campos, and Samuel Kaski. 2008. Can Relevance of Images Be Inferred from Eye Movements?. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval* (Vancouver, Canada) (MIR '08). ACM, New York, NY, USA, 134–140. <https://doi.org/10.1145/1460096.1460120>
- [36] László Kozma, Arto Klami, and Samuel Kaski. 2009. GaZIR: Gaze-based Zooming Interface for Image Retrieval. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* (Cambridge, Massachusetts, USA) (ICMI-MLMI '09). ACM, New York, NY, USA, 305–312. <https://doi.org/10.1145/1647314.1647379>
- [37] Yan Liu, Pei-Yun Hsueh, Jennifer Lai, Mirweis Sangin, Marc-Antoine Nussli, and Pierre Dillenbourg. 2009. Who is the Expert? Analyzing Gaze Data to Predict Expertise Level in Collaborative Applications. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo* (New York, NY, USA). IEEE, 898–901. <https://doi.org/10.1109/ICME.2009.5202640>
- [38] Norman H. Mackworth and Anthony J. Morandi. 1967. The Gaze Selects Information Details Within Pictures. *Perception & Psychophysics* 2, 11 (nov 1967), 547–552. <https://doi.org/10.3758/BF03210264>
- [39] Pascual Martínez-Gómez and Akiko Aizawa. 2014. Recognition of Understanding Level and Language Skill Using Measurements of Reading Behavior. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (Haifa, Israel) (IUI '14). ACM, New York, NY, USA, 95–104. <https://doi.org/10.1145/2557500.2557546>
- [40] NBA. 2012. Top 10 Plays of the 2011–2012 NBA Regular Season. Retrieved April 16, 2020 from <https://www.youtube.com/watch?v=SsC4RrcKeR8&list=PLB2F63866773CB519>

- [41] Anneli Olsen and Ricardo Matos. 2012. Identifying Parameter Values for an I-VT Fixation Filter Suitable for Handling Data Sampled with Various Sampling Frequencies. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) (ETRA '12). ACM, 317–320. <https://doi.org/10.1145/2168556.2168625>
- [42] Olympic. 2012. Men's Eight Rowing Heats Highlights - London 2012 Olympics. Retrieved April 16, 2020 from <https://www.youtube.com/watch?v=DniKM5SK6c&t=22s>
- [43] OlympicTorino2006. 2010. Biathlon - Men's 15Km Mass Start - Turin 2006 Winter Olympic Games. Retrieved April 16, 2020 from <https://www.youtube.com/watch?v=yMp9ICLKYl8>
- [44] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. 2018. Learning to Find Eye Region Landmarks for Remote Gaze Estimation in Unconstrained Settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research and Applications* (Warsaw, Poland) (ETRA '18). ACM, New York, NY, USA. <https://doi.org/10.1145/3204493.3204545>
- [45] John Paulin Hansen, Allan W. Andersen, and Peter Roed. 1995. Eye-gaze Control of Multimedia Systems. In *Symbiosis of Human and Artifact*, Yuichiro Anzai, Katsuhiko Ogawa, and Hirohiko Mori (Eds.). Advances in Human Factors/Ergonomics, Vol. 20. 37–42. [https://doi.org/10.1016/S0921-2647\(06\)80008-0](https://doi.org/10.1016/S0921-2647(06)80008-0)
- [46] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [47] Helmut Prendinger, Aulikki Hyrskykari, Minoru Nakayama, Howell Istance, Nikolaus Bee, and Yosiyuki Takahashi. 2009. Attentive Interfaces for Users with Disabilities: Eye Gaze for Intention and Uncertainty Estimation. *Universal Access in the Information Society* 8, 4 (2009), 339–354. <https://doi.org/10.1007/s10209-009-0144-5>
- [48] Pernilla Qvarfordt and Shumin Zhai. 2005. Conversing with the User Based on Eye-gaze Patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) (CHI '05). ACM, New York, NY, USA, 221–230. <https://doi.org/10.1145/1054972.1055004>
- [49] Filip Radlinski and Thorsten Joachims. 2005. Query Chains: Learning to Rank from Implicit Feedback. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (Chicago, Illinois, USA) (KDD '05). ACM, New York, NY, USA, 239–248. <https://doi.org/10.1145/1081870.1081899>
- [50] Konstantinos Rapantzikos, Nicolas Tsapatsoulis, Yannis Avrithis, and Stefanos D. Kollias. 2007. Bottom-up Spatiotemporal Visual Attention Model for Video Analysis. *IEEE Image Processing* 1, 2 (jun 2007), 237–248. <https://doi.org/10.1049/iet-ipr:20060040>
- [51] Babak Rasolzadeh, Alireza Tavakoli Targhi, and Jan Olof Eklundh. 2007. An Attentional System Combining Top-down and Bottom-up Influences. In *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint, International Workshop on Attention in Cognitive Systems (WAPCV '07)*. Springer, 123–140. https://doi.org/10.1007/978-3-540-77343-6_8
- [52] Berthier Ribeiro-Neto, Marco Cristo, Paulo B. Golgher, and Edleno Silva de Moura. 2005. Impedance Coupling in Content-targeted Advertising. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) (SIGIR '05). ACM, New York, NY, USA, 496–503. <https://doi.org/10.1145/1076034.1076119>
- [53] Ling Rothrock, Richard Koubek, Frederic Fuchs, Michael Haas, and Gavriel Salvendy. 2002. Review and Reappraisal of adaptive Interfaces: Toward Biologically Inspired Paradigms. *Theoretical Issues in Ergonomics Science* 3, 1 (2002), 47–84. <https://doi.org/10.1080/14639220110110342>
- [54] Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel Kaski. 2003. Can Relevance be Inferred from Eye Movements in Information Retrieval?. In *Proceedings of the 4th Workshop on Self-Organizing Maps* (Hibikino, Japan) (WSOM '03). 261–266. <https://doi.org/10.1145/1460096.1460120>
- [55] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. [n.d.]. . Technical Report. Helsinki University of Technology.
- [56] Jarkko Salojärvi, Kai Puolamäki, Jaana Simola, Lauri Kovanen, Ilpo Kojo, and Samuel Kaski. 2005. Inferring Relevance from Eye Movements: Feature Extraction. In *Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling* (Whistler, BC, Canada). 45–67.
- [57] Stefan Schroedl, Anand Kesari, and Leo Neumeyer. 2010. Personalized Ad Placement in Web Search. In *Proceedings of the 4th Annual International Workshop on Data Mining and Audience Intelligence for Online Advertising* (Washington D.C., USA) (ADKDD '10). ACM.
- [58] Björn Schuller, Ronald Müller, Florian Eyben, Jürgen Gast, Benedikt Hörnler, Martin Wöllmer, Gerhard Rigoll, Anja Höthker, and Hitoshi Konosu. 2009. *Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-life Application*. Vol. 27. 1760–1774 pages. <https://doi.org/10.1016/j.imavis.2009.02.013>
- [59] Christina Schweikert, Louis Gobin, Shuxiao Xie, Shinsuke Shimojo, and D. Frank Hsu. 2018. Preference Prediction Based on Eye Movement Using Multi-layer Combinatorial Fusion. In *Proceedings of the International Conference on Brain Informatics* (BI '18). Springer, 282–293. https://doi.org/10.1007/978-3-030-05587-5_27
- [60] Chengyao Shen, Xun Huang, and Qi Zhao. 2015. Predicting Eye Fixations on Webpage with an Ensemble of Early Features and High-Level Representations from Deep Network. *IEEE Transactions on Multimedia* 17, 11 (2015), 2084–2093. <https://doi.org/10.1109/TMM.2015.2483370>
- [61] Sights of Sites. 2014. New York City Bird's-Eye View. Retrieved April 16, 2020 from <https://www.youtube.com/watch?v=g1m1ShinkZ0>
- [62] Gino Slanzi, Jorge A. Balazs, and Juan D. Velásquez. 2017. Combining Eye Tracking, Pupil Dilation and EEG Analysis for Predicting Web Users Click Intention. *Information Fusion* 35 (2017), 51–57. <https://doi.org/10.1016/j.inffus.2016.09.003>
- [63] India Starker and Richard A. Bolt. 1990. A Gaze-Responsive Self-Closing Display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '90). ACM, Seattle, Washington, USA, 3–10. <https://doi.org/10.1145/97243.97245>
- [64] Ben Steichen, Giuseppe Carenini, and Cristina Conati. 2013. Adaptive Information Visualization: Using Gaze Data to Infer Visualization Types, Tasks and User Characteristics. In *Proceedings of the 18th International Conference on Intelligent User Interfaces* (Santa Monica, CA, USA) (IUI '13). ACM, 317–328.
- [65] Alistair Sutcliffe and Abdallah Namoun. 2012. Predicting User Attention in Complex Web Pages. *Behaviour & Information Technology* 31, 7 (jul 2012), 679–695. <https://doi.org/10.1080/0144929X.2012.692101>
- [66] Tobii AB. 2016. *Tobii Studio User's Manual*. Technical Report. 1–164 pages. <https://doi.org/10.1017/CBO9781107415324.004>
- [67] Roel Vertegaal, Jeffrey S. Shell, Daniel Chen, and Aadil Mamuji. 2006. Designing for Augmented Attention: Towards a Framework for Attentive User Interfaces. *Computers in Human Behavior* 22, 4 (jul 2006), 771–789. <https://doi.org/10.1016/j.chb.2005.12.012>
- [68] Tore Vesterby, Jonas C. Voss, John Paulin Hansen, Arne John Glenstrup, Dan Witzner Hansen, and Mark Rudolph. 2005. Gaze-Guided Viewing of Interactive Movies. *Digital Creativity* 16, 4 (aug 2005), 193–204. <https://doi.org/10.1080/14626260500476523>
- [69] Virginiaathletics. 2011. University of Virginia - Robby Andrews wins 800m at NCAA. Retrieved April 16, 2020 from <https://www.youtube.com/watch?v=y1vjpVWcmw>
- [70] Michel Wedel and Rik Pieters. 2006. Eye Tracking for Visual Marketing. *Foundations and Trends in Marketing* 1, 4 (2006), 231–320. <https://doi.org/10.1561/1700000011>
- [71] Dan Witzner Hansen and Qiang Ji. 2010. In the Eye of the Beholder: A Survey on Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 3 (1010), 478–500. <https://doi.org/10.1109/TPAMI.2009.30>
- [72] Songhua Xu, Hao Jiang, and Francis C.M. Lau. 2008. Personalized Online Document, Image and Video Recommendation via Commodity Eye-tracking. In *Proceedings of the 2008 ACM Conference on Recommender Systems* (Lausanne, Switzerland) (RecSys '08). ACM, New York, NY, USA, 83–90. <https://doi.org/10.1145/1454008.1454023>
- [73] Xing Yi, Liangjie Hong, Erheng Zhong, Nathan Nan Liu, and Suju Rajan. 2014. Beyond Clicks: Dwell Time for Personalization. In *Proceedings of the ACM Conference on Recommender Systems* (RecSys '14). ACM, Foster City, Silicon Valley, CA, USA, 113–120. <https://doi.org/10.1145/2645710.2645724>
- [74] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2019. Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland) (CHI '19). ACM, New York, NY, USA, Article 416, 13 pages. <https://doi.org/10.1145/3290605.3300646>
- [75] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (CVPR '17).