Behavior Research Methods (2021) 53:2650–2667
https://doi.org/10.3758/s13428-021-01576-8

Statistical modeling of dynamic eye-tracking experiments: Relative importance of visual stimulus elements for gaze behavior in the multi-group case

Mara Stadler[1,2,3] · **Philipp Doebler**[1] · **Barbara Mertins**[4] · **Renate Delucchi Danhier**[4]

Accepted: 5 March 2021 / Published online: 23 May 2021
© The Author(s) 2021

Abstract

This paper presents a model that allows group comparisons of gaze behavior while watching dynamic video stimuli. The model is based on the approach of Coutrot and Guyader (2017) and allows linear combinations of feature maps to form a master saliency map. The feature maps in the model are, for example, the dynamically salient contents of a video stimulus or predetermined areas of interest. The model takes into account temporal aspects of the stimuli, which is a crucial difference to other common models. The multi-group extension of the model introduced here allows to obtain relative importance plots, which visualize the effect of a specific feature of a stimulus on the attention and visual behavior for two or more experimental groups. These plots are interpretable summaries of data with high spatial and temporal resolution. This approach differs from many common methods for comparing gaze behavior between natural groups, which usually only include single-dimensional features such as the duration of fixation on a particular part of the stimulus. The method is illustrated by contrasting a sample of a group of persons with particularly high cognitive abilities (high achievement on IQ tests) with a control group on a psycholinguistic task on the conceptualization of motion events. In the example, we find no substantive differences in relative importance, but more exploratory gaze behavior in the highly gifted group. The code, videos, and eye-tracking data we used for this study are available online.

Keywords Eye tracking · Dynamic gaze behavior · Saliency map · Relative importance

Introduction

Exploring gaze behavior is a popular research method in many domains, since it can tell us how we are filtering information and how we might differ in our perception. Examples of between-group comparisons include differences in gaze behavior of experts and laymen (Bernal et al., 2014; Giovinco et al., 2014; Harezlak, Kasprowski, & Kasprowska, 2018), differences between elderly and younger people (Fontana et al., 2017) or differences in visual exploration due to native language (Stutterheim, Andermann, Carroll, Flecken, & Mertins, 2012). Besides comparisons of natural groups, differences in gaze behavior are studied subject to different experimental conditions like manual driving and highly automated driving (Navarro, Reynaud, & Gabaude, 2017). The data-analytic approach for group comparisons proposed in this paper is illustrated by differences in exploration between a group of people with particularly high cognitive ability and a control group.

Our visual environment is mostly characterized by dynamic processes and therefore the focus in this paper is on modeling dynamic scenes. We model the relative importance (RI) of different stimulus elements in dynamic scenes for gaze behavior for two natural groups by employing raw eye-tracking data. Therefore, we extend the approach of Coutrot and Guyader (2017) to a multi-group case. The model builds on linear combinations of feature maps to form a master saliency map while taking into account the highly dynamic nature of visual exploration, influenced by many time-dependent factors. The feature maps in the model can be, for instance, the static or dynamic salient contents of a video stimulus or predetermined areas of interest (AoIs). In addition, we reflect the individual steps in the modeling process. Before detailing the proposed approach, we review existing techniques for dynamic group comparisons and review modeling based on saliency maps.

✉ Mara Stadler
mara.stadler@helmholtz-muenchen.de

Extended author information available on the last page of the article.

Existing techniques for dynamic group comparisons

There are different approaches for comparing gaze behavior in dynamic scenes. Besides using metrics such as reaction time, dwelling time in AoIs and energy concentration ratios (Bernal et al., 2014; Fontana et al., 2017), some approaches take into account scan patterns. These approaches are based on evaluating similarity with sequence alignment scores followed by testing for statistical differences (Feusner & Lukoff, 2008) or providing a similarity score for two scanpaths based on their morphology and, optionally, duration in an AoI (Frame, Warren, & Maresca 2018). Navarro, Reynaud, and Gabaude (2017) analyze approaches based on the visual screen but without information on displayed images which take into account x and y axis variability of both groups or detect observer-based AoIs via heat maps and compare the consequent matrices by a Wilcoxon signed-ranks test. Furthermore, Navarro et al. (2017) compare techniques which do take into account the information on the visual stimulus by considering the percentage of time spent looking at a region of 5 degrees around a tangent point or by analyzing gaze positions relative to a dynamic gaze point on the stimulus but with a decomposition of gaze positions in horizontal and vertical components. Coutrot, Hsiao, and Chan (2017) introduce an approach in which hidden Markov models are learned from a group of scanpaths. This is useful to visualize and compare the gaze behavior of two different groups of observers. Other important scanpath algorithms have been introduced by Kübler, Rothe, Schiefer, Rosenstiel, and Kasneci (2017), where the scanpath comparison and classification is based on subsequence frequencies, and by Cristino, Mathot, Theeuwes and Gilchrist (2010), who present an approach for comparing saccadic eye- movement sequences based on the Needleman–Wunsch algorithm used in bioinformatics to compare DNA sequences.

Holmqvist and Andersson (2017) present different overtime calculations, like AoI over time with line graphs showing the proportion of participants gazing at a particular AoI. These methods are illustrated for static stimuli and do not involve direct group comparisons, but also provide feature importance curves that could be compared for different groups. However, this approach is not based on a statistical model, which means that the strength of the effects of the individual features on visual fixations cannot be quantified. Furthermore, this method considers each AoI individually and not in a combined manner. This also leads to multiple allocations of fixations in the case of overlapping AoIs.

Modeling with saliency maps

Some studies show that saliency maps from the computer vision field play an important role in the prediction of gaze behavior in different settings like watching videos, egocentric vision, or in computer games (Coutrot & Guyader, 2014; Sundstedt, Stavrakis, Wimmer, & Reinhard 2008; Yamada et al., 2011). On the other hand, some studies show that tasks overrule saliency when the participant takes the task very seriously (e.g., Chen & Zelinsky, 2006; Land & Hayhoe, 2001). Stimulus-driven saliency, also called bottom-up saliency, can be defined by predetermined AoIs, but also by static and dynamic saliency. Many computational models for visual attention, such as the model by Koch and Ullman (1985), are based on the Feature Integration Theory (FIT) by Treisman and Gelade (1980). A well-known approach that also focuses on features such as contrast, color or orientation, is the model by Itti, Koch, and Niebur (1998). This approach has later been extended by motion filters to obtain saliency models for video stimuli by Peters and Itti (2008). Another saliency model for video stimuli has been proposed by Le Meur, Thoreau, Le Callet, and Barba (2005). Le Meur and Baccino (2012) provide an extensive overview of computational modeling methods of visual attention and survey the strengths and weaknesses of common assessment methods based on diachronic (scanpaths or saliency maps) eye-tracking data. Marat et al. (2008) propose a spatio-temporal saliency model, which is biologically inspired and based on luminance information. In this model, high spatial frequencies are processed to extract luminance orientation and frequency contrast through a bank of Gabor filters and normalized to strengthen the spatially distributed maxima to obtain the static saliency of a frame. Under the assumption of luminance consistency between two consecutive frames, the dynamic pathway of the same model can be used to create dynamic saliency maps. Here, the moving areas are extracted by using low spatial frequencies. This model is also used in the approach of Coutrot and Guyader (2017), on which the method in this work is based.

Coutrot and Guyader (2017) combine the more popular bottom-up features with the observer-based top-down features linearly to a master saliency map. The model takes into account the dynamic aspect of the stimulus by using a statistical shrinkage method, which is a crucial difference from other common models in eye-tracking experiments. The works of Zhao and Koch (2011) and Peters and Itti (2007) for instance, are based on a similar model setup, but use a least-squares approach. Moreover, there exist methods based on deep learning networks that provide even higher correct classification rates and are state-of-the-art in terms of visual saliency prediction (Bylinskii, Isola, Bainbridge, Torralba, & Oliva 2015; Coutrot et al., 2017). However, these methods have the disadvantage that they depend on many parameters that are difficult to interpret (Lipton, 2018). In this paper, we extend the approach of Coutrot and Guyader (2017) to a multi-group model,

which allows to compare gaze behavior between two or more groups.

The remainder of this paper is organized as follows: After discussing the extension of the approach in Coutrot and Guyader (2017) in the subsequent "Methods", several worked examples are given in the "Practical application", followed by concluding remarks in the "Conclusions".

Methods

In this section, the idea of modeling eye-position density maps based on the approach of Coutrot and Guyader (2017) is described, though many details differ from the original exposition. The creation of feature maps and the estimation of eye-position density maps based on raw eye-tracking data is reviewed in detail, and the least absolute shrinkage and selection operator (LASSO) is discussed. Subsequently, we present the novel extension of the eye-position density modeling approach to the multi-group case.

Modeling eye-position density maps

The aim is to predict salient regions in complex natural scenes by linearly combining feature maps to a so-called master saliency map, which identifies regions that might lead to increased visual attention. The features in the model can refer to the stimulus (like contrast, motion, or predetermined AoIs), the so-called bottom-up features, or to the observer, the top-down features (like group membership). An often-observed behavior-based bias is the center bias (e.g., Tseng, Carmi, Cameron, Munoz, & Itti, 2009). This top-down feature describes the tendency to visually focus rather on the center than on the edges of a stimulus. The weights of the feature maps in the model vary systematically over time. The choice of feature maps also plays an important role since it has a strong impact on the predictions quality.

Let $\boldsymbol{S}$ be a master saliency map, $\boldsymbol{M}_k(t)$ the feature map for the kth feature at time t, $k \in \{1, ..., K\}$, and $\beta_k(t)$ the corresponding feature map weight at time t. The master saliency map $\boldsymbol{S}(t)$ is given by the linear combination

$$\boldsymbol{S}(t) = \sum_{k=1}^{K} \beta_k(t) \boldsymbol{M}_k(t).$$

We suppress the dependency on time in the notation for ease of exposition, i.e., $\boldsymbol{S} = \boldsymbol{S}(t)$, $\boldsymbol{M}_k = \boldsymbol{M}_k(t)$ and $\beta_k = \beta_k(t)$ for all k. The maps $\boldsymbol{S}$ and $\boldsymbol{M}_k$, $k = 1, ..., K$, can be understood as vectors with a length corresponding to the number of pixels of the stimulus frame. The vector of weights $\boldsymbol{\beta}$ is learned using eye-tracking data. Visual experiments are dynamic processes affected by many time-related factors, so the statistical shrinkage method LASSO is used to sieve out relevant feature maps.

Feature map generation

The generation of feature maps to be included in the model is described next. Note that all feature maps are represented by matrices, and all are normalized to obtain a bivariate probability density function by dividing each entry through the sum of all entries of the map.

Uniform map The uniform map is a bottom-up feature with the same value $\frac{1}{w \cdot h}$ at each entry or pixel, $\boldsymbol{M}_U = (\frac{1}{w \cdot h})_{i=1,...,w,\ j=1,..,h} \in \mathbb{R}^{w \times h}$, where w and h represent the stimulus width and the stimulus height in pixels. This feature represents a "catch-all hypothesis" for fixations, which can only be weakly explained by the other features.

Center bias map The center bias is a bottom-up feature generated by a time-independent bivariate Gaussian function $\mathcal{N}(0, \boldsymbol{\Sigma})$ with a diagonal covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_x^2, \sigma_y^2)$, which is centered at the center of the monitor. Standard deviations σ_x and σ_y are chosen proportional to frame size by dividing the stimulus height and width by 12.

Static and dynamic saliency map The static and dynamic saliency maps are top-down features that highlight areas of the stimulus that stand out statically or dynamically from the other areas of the stimulus. Saliency maps can be created using different saliency models. In this paper, saliency maps are first determined using two different saliency models and then the resulting feature map weights are compared. The Graph-Based Visual Saliency Algorithm (GBVS) and the Real-Time Three-Path Saliency Model (TVSM) are used for this purpose. The comparison of these rather different approaches is carried out to get an impression of the influence of the choice of the saliency model on the results in the eye-position density map modeling approach. The GBVS algorithm is a graph-based approach (Harel, Koch, & Perona, 2006). The spatio-temporal saliency model TVSM is a biologically inspired model based on luminance information (Marat, Rahman, Pellerin, Guyader, & Houzet, 2013) and is also applied in Coutrot and Guyader (2017). Both approaches use the same model to create dynamic saliency maps and static saliency maps. The approach to model dynamic saliency maps considers information from the previous frame. For this reason, it is not possible to create a dynamic saliency map for the first frame.

We find that for stimuli where the dynamic content is also statically very different from the remainder of the stimulus, the TVSM algorithm is more suitable for determining the saliency maps: Fig. 9 includes the comparison of the GVBS and the TVSM saliency maps for an illustrative frame as

well as some resulting estimated curves, which show that the resulting feature map weights for the video stimulus differ considerably between the approaches. The contrary course of the feature map weightings of static and dynamic saliency due to the correlation of the two features is particularly strong under the GBVS approach. For dynamic scenes where the statically salient content is more distinct from the dynamically prominent content, such as the stimulus from the freely accessible supplementary material from Coutrot and Guyader (2017), the inaccuracy of the GBVS algorithm has little effect on the estimated feature map weights.

Consequently, the saliency features are determined using the TVSM algorithm. Figure 1 shows the static and dynamic saliency map as well as the original frame for one frame of the video stimulus *car cornfield* from our experiment.

Areas of interest (AoIs) The AoIs are defined as polygons with known vertices. Based on these coordinates, we create binary matrices for each frame, which are 1 on the pixels inside the polygon and 0 otherwise. Figure 2 illustrates the dynamic and static AoI in a video stimulus.

Eye-position density map estimation

For each natural or experimental group, an eye-position density map is estimated frame by frame from raw eye-tracking data by kernel density estimation, a non-parametric approach for estimation of probability densities. Since the stimuli are two-dimensional videos, a bivariate kernel density with bivariate least squares cross-validation bandwidth matrix is appropriate (e.g., Duong, 2004). The bivariate kernel density estimator on a grid out of pixels $\boldsymbol{x}$ for a random sample out of N fixation coordinates $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ is given by

$$\hat{f}(\boldsymbol{x}; \boldsymbol{H}) = N^{-1} \sum_{i=1}^{N} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_i),$$

with two-dimensional vectors $\boldsymbol{x} = (x_1, x_2)^T$ and $\boldsymbol{X}_i = (X_{i1}, X_{i2})^T, i = 1, ..., N$. The scaled kernel is denoted by $K_{\boldsymbol{H}}$ and $\boldsymbol{H} \in \mathbb{R}^{2\times 2}$ is a non-random, symmetric, positive definite bandwidth matrix. The relation to the non-scaled kernel K is given in a general form by $K_{\boldsymbol{H}}(\boldsymbol{x}) = |\boldsymbol{H}|^{-1/2} K(\boldsymbol{H}^{-1/2}\boldsymbol{x})$. The scaled bivariate gaussian kernel, on which the calculations in this paper are based, is given by $K_{\boldsymbol{H}}(\boldsymbol{x}) = (2\pi)^{-1}|\boldsymbol{H}|^{-1/2} \exp(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{H}^{-1}\boldsymbol{x})$.

The bandwidth matrix $\hat{\boldsymbol{H}}_{\text{LSCV}}$ is the solution of the minimization problem $\text{argmin}_{\boldsymbol{H}} \text{LSCV}(\boldsymbol{H})$, with

$$\text{LSCV}(\boldsymbol{H}) = \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{x}; \boldsymbol{H})^2 d\boldsymbol{x} - 2N^{-1} \sum_{i=1}^{N} \hat{f}_{-i}(\boldsymbol{X}_i; \boldsymbol{H}),$$

where $\hat{f}_{-i}(\boldsymbol{X}_i; \boldsymbol{H}) = (N-1)^{-1} \sum_{j=1, j\neq i}^{N} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{X}_j)$ is the *leave-one-out*-estimator. This procedure differs from the kernel density estimation in the work of Coutrot and Guyader (2017), in which a bivariate Gaussian kernel with a standard deviation of 1 degree of visual angle is chosen. In particular, a new bandwidth is selected here for each frame, which allows for more variability between the densities of the individual frames (a comparison of the two approaches can be seen in Fig. 8). A smooth eye position density map $\boldsymbol{Y}$ results. With this approach, no further background information on the experimental setup (like visual angle) is required. Overall, the smoother the kernel density estimation, the less individual feature maps stand out and the more similar the resulting weights are.

Least absolute shrinkage and selection operator algorithm

The advantage of the LASSO over other regression methods, especially least squares regression or the expectation maximization algorithm, is that it allows selecting relevant features and to reject the other features. This property can lead to a more efficient and more interpretable model (Hastie, Tibshirani, & Friedman, 2009). Here, the LASSO shrinks feature map weights $\boldsymbol{\beta}$ by imposing a L_1 penalty with irrelevant map weights shrunk to 0 and, hence, removed entirely from the master saliency map.

Fig. 1 Frame of a video stimulus (*left*) with corresponding static (*middle*) and dynamic (*right*) saliency maps calculated with TVSM. Contrasts and luminance influence the static map, while the moving truck in the otherwise steady scene dominates the dynamic salience

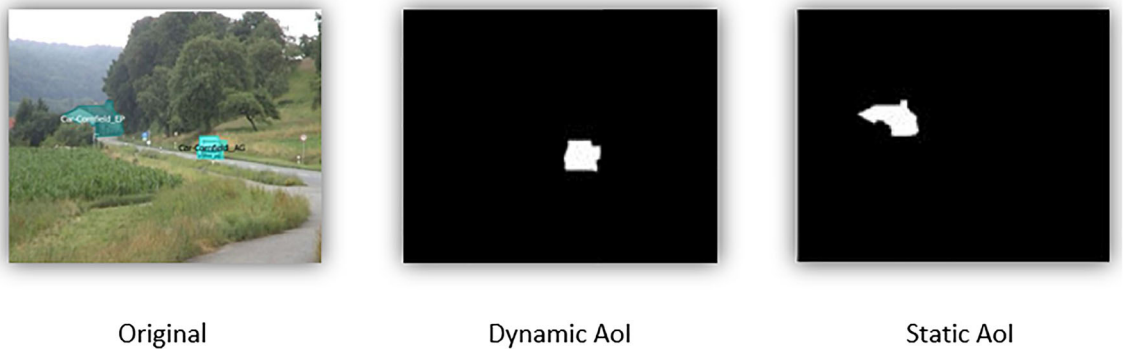

Fig. 2 Frame of a video stimulus with highlighted AoIs (*left*) and dynamic (*middle*) and static AoI maps (*right*) corresponding to binary matrices. The dynamic AoI map differs from frame to frame while the static AoI map is the same for each frame of the video stimulus since it does not move and there is no camera motion

Recall that the eye-position density map is denoted by $\boldsymbol{Y}$, and that the master saliency is equal to $\sum_{k=1}^{K} \beta_k \boldsymbol{M}_k$. The parameter $\lambda > 0$ is a tuning parameter, which controls the amount of shrinkage. The LASSO estimate solves the minimization problem

$$\boldsymbol{\beta}^{LASSO}(\lambda) = \text{argmin}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \sum_{k=1}^{K} \beta_k \boldsymbol{M}_k)^2 + \lambda \sum_{k=1}^{K} |\beta_k| \right\}.$$

For $\lambda = 0$ the LASSO algorithm corresponds to the least squares estimate. For $\lambda \to \infty$ the weights β_k, $k = 1, ..., K$, are shrunk towards zero. For an increasing λ, the variance decreases and the bias increases (James, Witten, Hastie, & Tibshirani, 2013).

The R package *glmnet* (Friedman, Hastie & Tibshirani, 2010) finds $\boldsymbol{\beta}^{LASSO}(\lambda)$ values for a regularization path, i.e., for a sequence of λ values. Following Coutrot and Guyader (2017), λ is chosen so that $\boldsymbol{\beta}^{LASSO}(\lambda)$ is optimal in terms of the Bayesian Information Criterion (BIC), given here by

$$\text{BIC} = \text{BIC}(\boldsymbol{S}|\boldsymbol{Y}) = -2 \log L(\boldsymbol{S}|\boldsymbol{Y}) + K \log n,$$

where L is Gaussian likelihood of $\boldsymbol{S} = \boldsymbol{S}(\lambda)$, K is the number of feature maps in the model (equal to the number of nonzero β_k) and $n = w \cdot h$ denotes the number of pixels in $\boldsymbol{Y}$. Zou, Hastie, and Tibshirani (2007) show that the number of non-zero coefficients provides an unbiased estimate of the degrees of freedom in LASSO, which does not require further assumptions on the predictors. In addition, it is shown that the unbiased estimator is asymptotically consistent and thus model selection criteria, such as the BIC, are acceptable.

Extension to multi-group case

The extension of the approach of Coutrot and Guyader (2017) is based on the method for modeling interactions between qualitative and quantitative predictors in general linear models, e.g., Kutner, Nachtsheim, and Neter (2005). To extend the model for the two-group case, a binary dummy variable $\tilde{\boldsymbol{M}}_G$ is introduced, which denotes whether the information on the jth pixel refers to the treatment or the control group and is given by

$$\tilde{\boldsymbol{M}}_{Gj} = \begin{cases} 1, & \text{if } j \text{ refers to the treatment group} \\ 0, & \text{if } j \text{ refers to the control group,} \end{cases}$$

where $j = 1, ..., 2 \cdot w \cdot h$, with w the number of pixels in width and h the number of pixels in height. In the following model, extension of the first $w \cdot h$ entries refer to the treatment group and the second $w \cdot h$ entries refer to the control group. Thus, $\tilde{\boldsymbol{M}}_G$ is given by $\tilde{\boldsymbol{M}}_G := \left[\boldsymbol{1}_{w \cdot h}^T \boldsymbol{0}_{w \cdot h}^T\right]^T \in \mathbb{R}^{2 \cdot w \cdot h}$.

The feature maps are given in form of vectors in the model and the variable $\tilde{\boldsymbol{M}}_G$ is interacted with each feature map in the model. This is done by elementwise vector multiplication, denoted by "$\circ$". The model with K feature maps $\boldsymbol{M}_1, ..., \boldsymbol{M}_K$ is given by

$$\tilde{\boldsymbol{Y}} = \boldsymbol{\beta} \tilde{\boldsymbol{M}} + \epsilon, \tag{1}$$

where $\tilde{\boldsymbol{M}} = \left[\tilde{\boldsymbol{M}}_1 \, ... \, \tilde{\boldsymbol{M}}_K \;\; \tilde{\boldsymbol{M}}_1 \circ \tilde{\boldsymbol{M}}_G \, ... \, \tilde{\boldsymbol{M}}_K \circ \tilde{\boldsymbol{M}}_G\right]$ denotes the design matrix with $\tilde{\boldsymbol{M}}_i := \left[\boldsymbol{M}_i^T \boldsymbol{M}_i^T\right]^T \in \mathbb{R}^{2 \cdot w \cdot h}$, for $i = 1, ..., K$, $\tilde{\boldsymbol{Y}} = \left[\boldsymbol{Y}_T^T \boldsymbol{Y}_C^T\right]^T \in \mathbb{R}^{2 \cdot w \cdot h}$ with $\boldsymbol{Y}_T$ and $\boldsymbol{Y}_C$ the eye-position density maps of the treatment and control group in form of a vector and $\boldsymbol{\beta} = (\beta_1, ..., \beta_K, \beta_{1,G}, ..., \beta_{K,G})^T \in \mathbb{R}^{2 \cdot K}$ the regression coefficient vector. The density on the jth pixel is therefore given by

$$\begin{aligned} \tilde{\boldsymbol{Y}}_j &= \beta_1 \tilde{\boldsymbol{M}}_{1j} + ... + \beta_K \tilde{\boldsymbol{M}}_{Kj} + \beta_{1,G}(\tilde{\boldsymbol{M}}_{1j} \cdot \tilde{\boldsymbol{M}}_{Gj}) + ... \\ &\quad + \beta_{K,G}(\tilde{\boldsymbol{M}}_{Kj} \cdot \tilde{\boldsymbol{M}}_{Gj}) + \epsilon_j \\ &= \begin{cases} (\beta_1 + \beta_{1,G})\tilde{\boldsymbol{M}}_{1j} + ... + (\beta_K + \beta_{K,G})\tilde{\boldsymbol{M}}_{Kj} + \\ \quad ... + \epsilon_j, \; j = 1, ..., n \\ \beta_1 \tilde{\boldsymbol{M}}_{1j} + ... + \beta_K \tilde{\boldsymbol{M}}_{Kj} + \epsilon_j, \; j = n+1, ..., 2 \cdot n, \end{cases} \end{aligned}$$

with $n = w \cdot h$. If $\beta_{1,G}, ..., \beta_{K-1,G}$ or $\beta_{K,G}$ differ significantly from zero, it can be interpreted as differences

in gaze behavior between the two groups. Confidence intervals therefore need to be estimated, which will not be further specified in this paper.

Practical application

In the following section, we describe the structure of the experiment, the stimulus material, as well as the available data material and data processing. Subsequently, the method is applied to the stimulus and data material. In addition to the analysis of the stimuli in our specific experiment, we also include the evaluation of two static stimuli in the Appendix to demonstrate that our method also works for other types of stimuli. The analyses are performed in the statistics software R (R Core Team, 2020). The code, visual material, and eye-tracking data we used for this study are available online.[1]

Material

Participants and experiment The eye-tracking experiment was carried out with two groups at different time points. The first group consists of $N_T = 33$ members of the *Mensa in Deutschland e.V.*, an association for participants with particularly high cognitive abilities. This group is denoted as the treatment group, where a particularly high cognitive ability stands for the treatment. The second group contains $N_C = 102$ participants, which are primarily members of a bachelor's program in German philology. This group is considered to be the control group. All participants are multi-lingual and speak German at a native-speaker level. The treatment group in our experiment was aware that their gaze behavior would be compared with a control group, which could influence their gaze behavior.

In the experiment, several video stimuli are presented to the participants. The task was to briefly describe orally what is happening in the video. This task (or pseudo-task) is common in many psycholinguistic gaze behavior studies since it aims to achieve greater comparability between participants, as visual behavior can vary greatly without any task (e.g., Castelhano, Mack, & Henderson, 2009).

Visual material The video stimuli are taken from a study that compares the gaze behavior of speakers of different native languages (Stutterheim, Andermann, Carroll, Flecken, & Mertins, 2012). A distinction was made between speakers of an aspect language and non-aspect language. In terms of the use of tenses, aspect languages, such as English, distinguish between an ongoing action (John was crossing the street) and a completed action in the past (John has crossed the street). Non-aspect languages, such as German, do not make such a distinction, but need time adverbs to clarify that an action is happening right now. In the study, it was shown that speakers of non-aspect languages, when considering dynamic stimuli, put a stronger focus on the expected—but not occurring—endpoint towards which an object is moving (Stutterheim et al., 2012). In the context of the current work, influence of cognitive ability on gaze behavior is studied, while keeping the variable language constant. In a study by Vigneau, Caissie, and Bors (2006) on differences in gaze behavior when solving the Advanced Progressive Matrices Test, a speech-free multiple-choice intelligence test, it could be shown, for example, that subjects with high test scores consider all elements of the matrix to be completed. In contrast, subjects with low test scores only considered the elements in the row and column of the element to be completed in the matrix.

The video stimuli contain one moving object, the dynamic AoI, and we have defined a fixed end point, the static AoI. The stimuli end before the end point is reached by the moving object. The stimuli have no camera pan and no sound. The procedure is exemplified on several stimuli in this paper and the detailed procedure is described using the stimulus *car cornfield* as an example.

The refresh rate is 25 Hz and the resolution of the stimulus is $w \cdot h = 720 \times 576$ pixels. This video stimulus shows a car, the dynamic AoI, driving in the direction of a house, representing the static AoI, see Fig. 3. The duration of the video stimulus is approximately 7 s and therefore the stimulus consists out of $N_F = 174$ frames.

Eye-tracking data The eye-tracking data are given as x and y coordinates of fixations and saccades on the monitor. Following Coutrot and Guyader (2017), only the coordinates of the right eye are considered. The data were recorded with an SMI RED 60 device. The distance of a participant to the monitor was between 55 and 65 cm. The resolution of the monitor is 1920×1080 pixels and the stimulus was enlarged to full monitor height and proportionally adjusted in width. Therefore, the video area has a resolution of 1350×1080 pixels with black areas on the sides with a width of 258 pixels each. The fixations and saccades were recorded at a sampling rate of 60 Hz. The upper left corner of the monitor represents the coordinate $(0, 0)$, which is also recorded if there is a loss of vision or if the respondent blinks.

Data processing The number of recorded fixations or saccades varies slightly between 407 and 410 data points per participant in the treatment group and between 407 and 419 in the control group due to eye-tracker inaccuracies. The dataset does not provide any information about the points in time at which the gaze coordinates are lost, which is why the number of gaze coordinates is shortened by discarding

[1] https://github.com/marastadler/Lasso_eyeposition

Fig. 3 Excerpt of the video stimulus at the beginning, in the middle, and at the end (1st, 87th, and 174th frame)

the last gaze coordinates to the minimum available number of gaze coordinates per respondent. Thus, inaccuracies of up to 12/60 s can be assumed. The gaze coordinates (0, 0) are removed, as they represent that the gaze could not be tracked. In addition, the coordinates tracked outside the stimulus area on the monitor are removed. The tracking rate of 60 Hz and the refresh rate of 25 Hz result in 2.4 coordinates per person and frame. In order to consider one coordinate per person per frame, the first viewing coordinate that remains completely on the respective frame is selected. With 2.4 view coordinates per frame, the 1st, 4th, 6th, 9th, 11th, 13th etc. are thus selected. The remaining view coordinates are not included in the analysis.

When estimating causal effects in observational data, a randomized experiment should be replicated as accurately as possible to ensure that the distribution of covariates in the treatment and control groups is as similar as possible. To ensure this, a matching is carried out. Since the groups have rather small overlaps in their covariates, a propensity score matching (PSM) with an optimal matching algorithm and subsequent balance diagnostics (Zhang, Kim, Lonjon, & Zhu, 2019) on the covariates gender and age is performed using the R package *MatchIt* (Ho, Imai, King, & Stuart, 2011). PSM can be helpful if there is a high level of imbalance in the covariates (King & Nielsen, 2019). By using a caliper of 0.1 the matching algorithm selects only 25 participants from the control group and ten participants from the extreme group with high cognitive abilities and thus rejects 77 participants from the control group and 23 participants from the group of participants with high cognitive abilities. Although the result does not provide satisfactory group sizes, the two-group model is illustrated on the basis of these matched groups.

Results

Figure 4 illustrates the single-group model for one frame of the video stimulus. The two-dimensional maps can be understood as matrices $\boldsymbol{M}$ of the dimension $w \times h$, where w stands for the number of pixels in width and h for the number of pixels in height of the stimulus. Each pixel corresponds to a number on the grayscale, where 0 stands for black and 1 for white. The uniform (U), center bias (CB), static saliency (S), and dynamic saliency (D) maps, as well as the dynamic AoI (AOI1) and the static AoI (AOI2) are all included in the model. The matrices are treated as vectors in the model definition, so that the value of the kernel density estimate on a pixel corresponds to an observation in the model. Therefore, $\boldsymbol{M}_i := \vec{\boldsymbol{M}}_i$, $i \in \{\text{U, CB, S, D, AOI1, AOI2}\}$, applies. Mathematically, the model has the following form,

$$\begin{aligned}\boldsymbol{M} = {} & \beta_U \boldsymbol{M}_U + \beta_{CB} \boldsymbol{M}_{CB} + \beta_S \boldsymbol{M}_S + \beta_D \boldsymbol{M}_D \\ & + \beta_{AOI1} \boldsymbol{M}_{AOI1} + \beta_{AOI2} \boldsymbol{M}_{AOI2} + \epsilon,\end{aligned}$$

where $\boldsymbol{Y}, \boldsymbol{M}_i \in \mathbb{R}^{w \cdot h}$, $i \in \{\text{U, CB, S, D, AOI1, AOI2}\}$, and $w \cdot h = 576 \cdot 720 = 414720$. The residuals ϵ ignore any remaining spatial dependencies and framewise homoscedasticity is assumed. Each feature map vector is divided by the sum of all entries of the vector to obtain probability density functions. The eye-position density map and feature maps are centered and standardized in the model. Thus, the units on the y-axis are standard deviations.

First, the initial model is adapted separately for both unmatched groups of participants. The following Fig. 5 shows the estimated relative importance (RI) curves and the adjusted coefficient of determination R^2 for each frame. The term 'relative importance' refers to the effect of each feature map on the prediction of the fixation density on the corresponding frame in the stimulus compared to the effect of the other feature maps in the model. The RI of the feature maps can be compared between different feature maps on one frame (at the same time) or between several frames (throughout the stimulus duration).

In both groups, the feature maps do not predict the fixations on the first frames very well. After about the 10th frame, the R^2 values increase. Since it is assumed that there are latent influences on human gaze behavior, even lower R^2 values can be considered acceptable. The curve of the

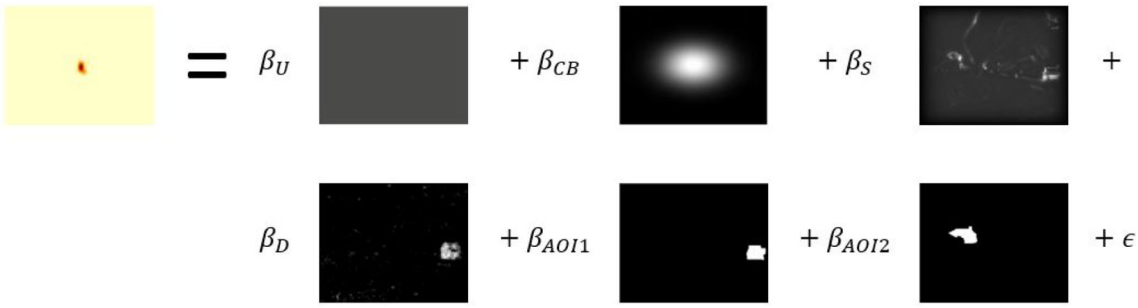

Fig. 4 Model illustration for the second frame of the video stimulus (U = uniform, CB = center bias, S = static, D = dynamic)

coefficient of determination R^2 of the treatment group is similar to the curve of the control group, but on a noticeably lower level. It can be concluded that the participants of the treatment group exhibit a more explorative behavior during this stimulus and thus the feature maps predict the coordinates of the fixations less accurately. This finding is also obtained when drawing a random sample from the control group that corresponds to the sample size of the treatment group.

The extension to a two-group model for the jth pixel, $j = 1, ..., 2 \cdot w \cdot h = 2 \cdot 576 \cdot 720 = 829440$, is given by

$$
\begin{aligned}
\tilde{\boldsymbol{Y}}_j \;=\; & \beta_U \tilde{\boldsymbol{M}}_{Uj} + \beta_{CB} \tilde{\boldsymbol{M}}_{CBj} + \beta_S \tilde{\boldsymbol{M}}_{Sj} + \beta_D \tilde{\boldsymbol{M}}_{Dj} \\
& + \beta_{AOI1} \tilde{\boldsymbol{M}}_{AOI1j} + \beta_{AOI2} \tilde{\boldsymbol{M}}_{AOI2j} \\
& + \beta_{U,G} (\tilde{\boldsymbol{M}}_{Uj} \cdot \tilde{\boldsymbol{M}}_{Gj}) \\
& + \beta_{CB,G} (\tilde{\boldsymbol{M}}_{CBj} \cdot \tilde{\boldsymbol{M}}_{Gj}) + \beta_{S,G} (\tilde{\boldsymbol{M}}_{Sj} \cdot \tilde{\boldsymbol{M}}_{Gj}) \\
& + \beta_{D,G} (\tilde{\boldsymbol{M}}_{Dj} \cdot \tilde{\boldsymbol{M}}_{Gj}) + \beta_{AOI1,G} (\tilde{\boldsymbol{M}}_{AOI1j} \cdot \tilde{\boldsymbol{M}}_{Gj}) \\
& + \beta_{AOI2,G} (\tilde{\boldsymbol{M}}_{AOI2j} \cdot \tilde{\boldsymbol{M}}_{Gj}) + \epsilon_j, \\
= \; & \begin{cases} (\beta_U + \beta_{U,G}) \tilde{\boldsymbol{M}}_{Uj} + (\beta_{CB} + \beta_{CB,G}) \tilde{\boldsymbol{M}}_{CBj} + ... \\ + \epsilon_j, \text{ for } j = 1, ..., n \\ \beta_U \tilde{\boldsymbol{M}}_{Uj} + \beta_{CB} \tilde{\boldsymbol{M}}_{CBj} + ... + \epsilon_j, \text{ for } j = n+1, \\ ..., 2 \cdot n, \end{cases}
\end{aligned}
$$

with $\tilde{\boldsymbol{M}}_i := \left[\boldsymbol{M}_i^T \boldsymbol{M}_i^T\right]^T \in \mathbb{R}^{2 \cdot w \cdot h}$ for $i \in \{\text{U, CB, S, D, AOI1, AOI2}\}$, $\tilde{\boldsymbol{M}}_G := \left[\boldsymbol{M}_{G,Tre}^T \boldsymbol{M}_{G,Con}^T\right]^T = \left[\boldsymbol{1}_{w \cdot h}^T \boldsymbol{0}_{w \cdot h}^T\right]^T \in \mathbb{R}^{2 \cdot w \cdot h}$ and $n = w \cdot h = 414720$.

The bivariate kernel density estimate $\tilde{\boldsymbol{Y}} = \left[\boldsymbol{Y}_{Tre}^T \boldsymbol{Y}_{Con}^T\right]^T \in \mathbb{R}^{2 \cdot w \cdot h}$ and the feature maps are centered and standardized separately for each group. The kernel density estimation in the groups is carried out separately for both groups and is therefore based on different bandwidths. A standardization across both groups could lead to a group having a strong peak if there are strong mean differences in the densities of the groups.

The R^2 values in the group model for the stimulus vary over the entire duration of the stimulus between values close to zero and 0.5, with most frames showing R^2 values between 0.1 and 0.4. Apart from the very low R^2 values of the models of the first frames, no temporal influence on the R^2 values can be seen. The comparison of the results of the model with a LASSO penalty to the results of a least squares approach shows that there are no notable differences in the results (see Fig. 10) which means that all features in the model play an essential role in explaining the gaze behavior. Figure 6 shows the non-normalized RI curves or feature map weights $\beta_U, ..., \beta_{AOI2}$ for the control group in transparent colors and $\beta_U + \beta_{U,G}, ..., \beta_{AOI2} + \beta_{AOI2,G}$ for the treatment group. The weights here are illustrated in a non-normalized form, since the densities of the two groups were not scaled equally and differ in particular in their maximum. When interpreting such results, it should always be taken into account that inaccuracies in the eye tracker may lead to a fixation being incorrectly assigned to a feature. For the AoIs and the center bias as well as for the dynamic saliency this problem should be rather negligible, since these features cover comparatively large and dense areas of the stimulus. The influence of these inaccuracies can be greater for the static saliency, which in some cases highlights very fine contours (see Fig. 1). The curves are descriptive in nature and do not indicate significant influences of some features on the fixations or differences between the groups. In both groups, however, the dynamic AoI (AoI1) seems to have a higher weighting than the other features, which suggests that gaze behavior is strongly driven by the stimulus content. Nevertheless, the dynamic AoI is moving in a linear way with no changes in velocity or directions, so for the participants it is very easy to predict the development of the depicted movement, which in turn frees them to explore the rest of the scene. Overall, the curves in both groups run at a similarly high level. The curves indicate that the groups do not react to image elements represented by the feature maps at exactly the same time, but with a time lag. This behavior can be seen for example in the center bias curves (green). Also, the curves of the dynamic AoI (AoI1) indicate that the groups do not always focus on the car at the same time. Since we use the L_1 penalty in our model, feature map weights, which are not relevant for the prediction of fixations, would get

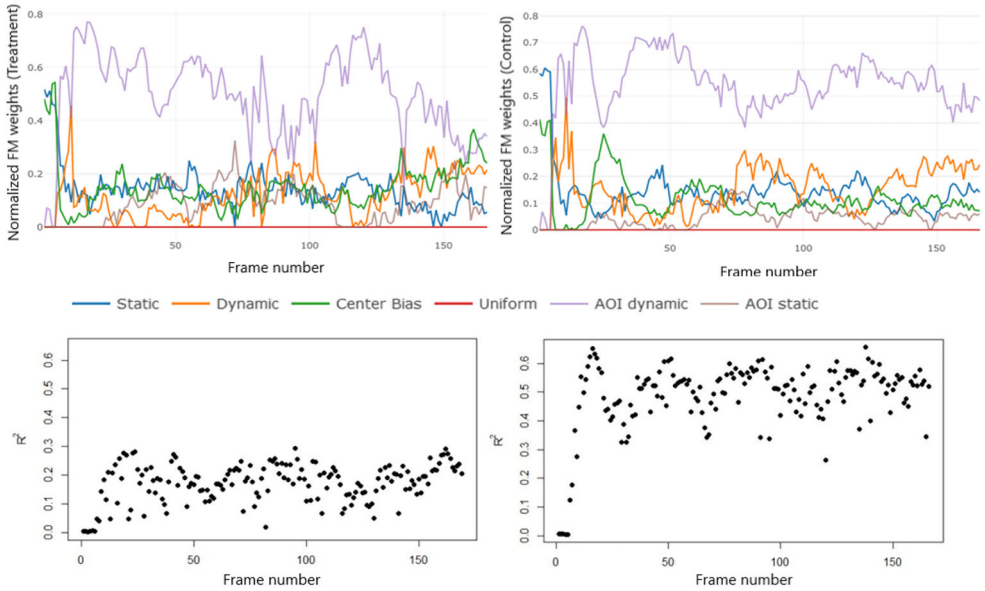

Fig. 5 RI curves of respective feature for both groups separately in the non-extended model (*top*) and R^2 values for both groups (*bottom*) (*left*: treatments, *right*: controls)

a value of zero, which is not the case here except for the uniform map, which serves as a control instance and should therefore be zero. It can be concluded that all feature maps in our model, the bottom-up feature center bias as well as the top-down features, seem to be relevant for the prediction of fixations.

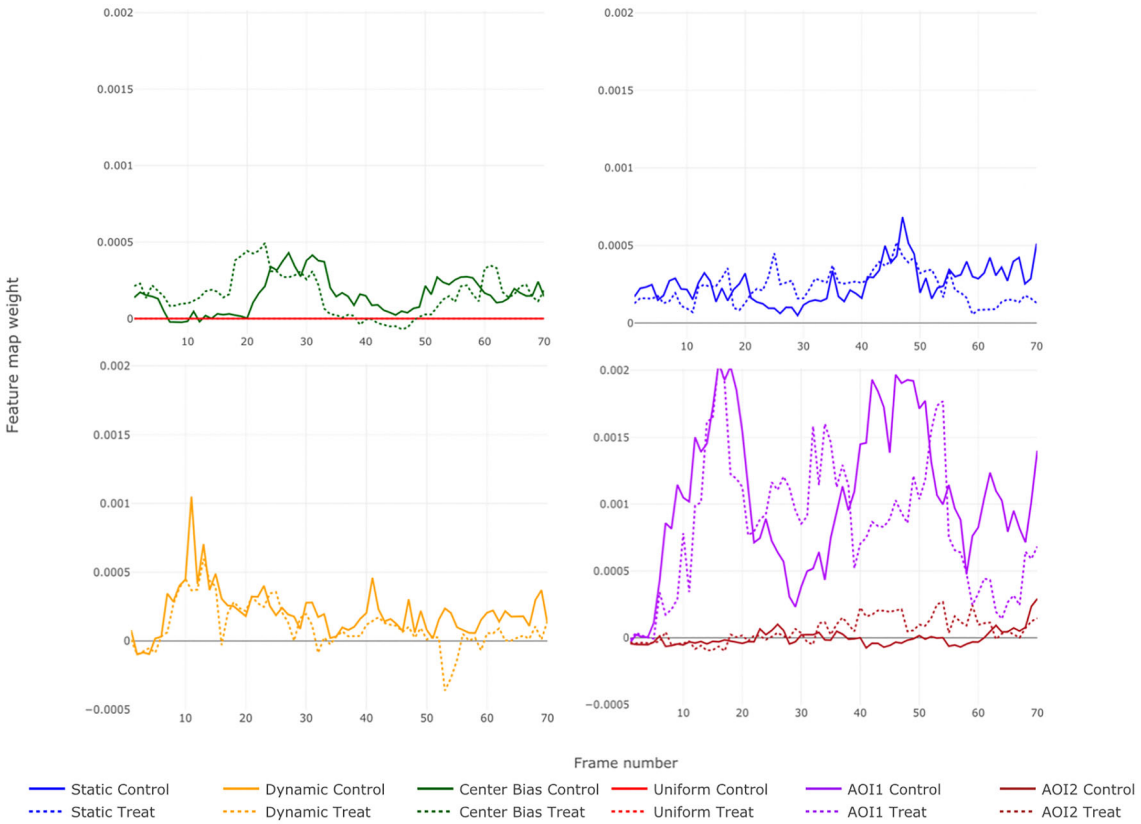

Fig. 6 Estimated feature map weights for the first 70 frames in the two-group model for the clarity in multiple plots. *Dashed lines* lines stand for feature map weights in the treatment group and solid for the control group

We have performed permutation tests based on an equidistant sequence of ten frames (for runtime reasons) to make statements about significant differences between the groups for each frame. The following Fig. 7 shows the boxplots of all observed coefficients of group differences $\hat{\beta}_{U,G}$, ..., $\hat{\beta}_{AOI2,G}$ in $P = 1000$ permutations and the corresponding p values. The observed coefficients are highlighted with a red cross. If the red cross is located inside the box, the regression coefficient hardly differs from the coefficients of a random group assignment and this indicates that there exists no difference in gaze behavior between the two groups in this frame. Conversely, observations located outside of the box represent significant group differences. The results we obtain reflect what the relative importance curves indicate.

For example, in frame 42 for both the static AoI (AoI2) and the dynamic AoI (AoI1) and also for the center bias the red crosses are outside of the box as one would expect when looking at the RI curves. For frame 21, on the other hand, no significant difference can be detected in both AoIs, which is also indicated by the curves. The fact that the AoI1 boxes are not exactly centered around zero and are also very large overall shows that in the dynamic AoI, the gaze behavior also varies more among individuals in general. Nevertheless, the difference between the two groups examined here is particularly noticeable.

We again use model (1) and the same groups to analyze a further stimulus *walking market*, which includes a market stall as an endpoint and a lady as a moving object moving towards the market stall. This stimulus also contains other possible areas of interest such as a pigeon walking through the image, which are not included as individual features in the model, but are covered by the static and dynamic saliency. Again, we find that the dynamic AoI (AoI1), i.e., the lady, gains the highest weight in both groups. In contrast to the previous stimulus, the static AoI (AoI2) has high weights at the beginning, which can be explained by the

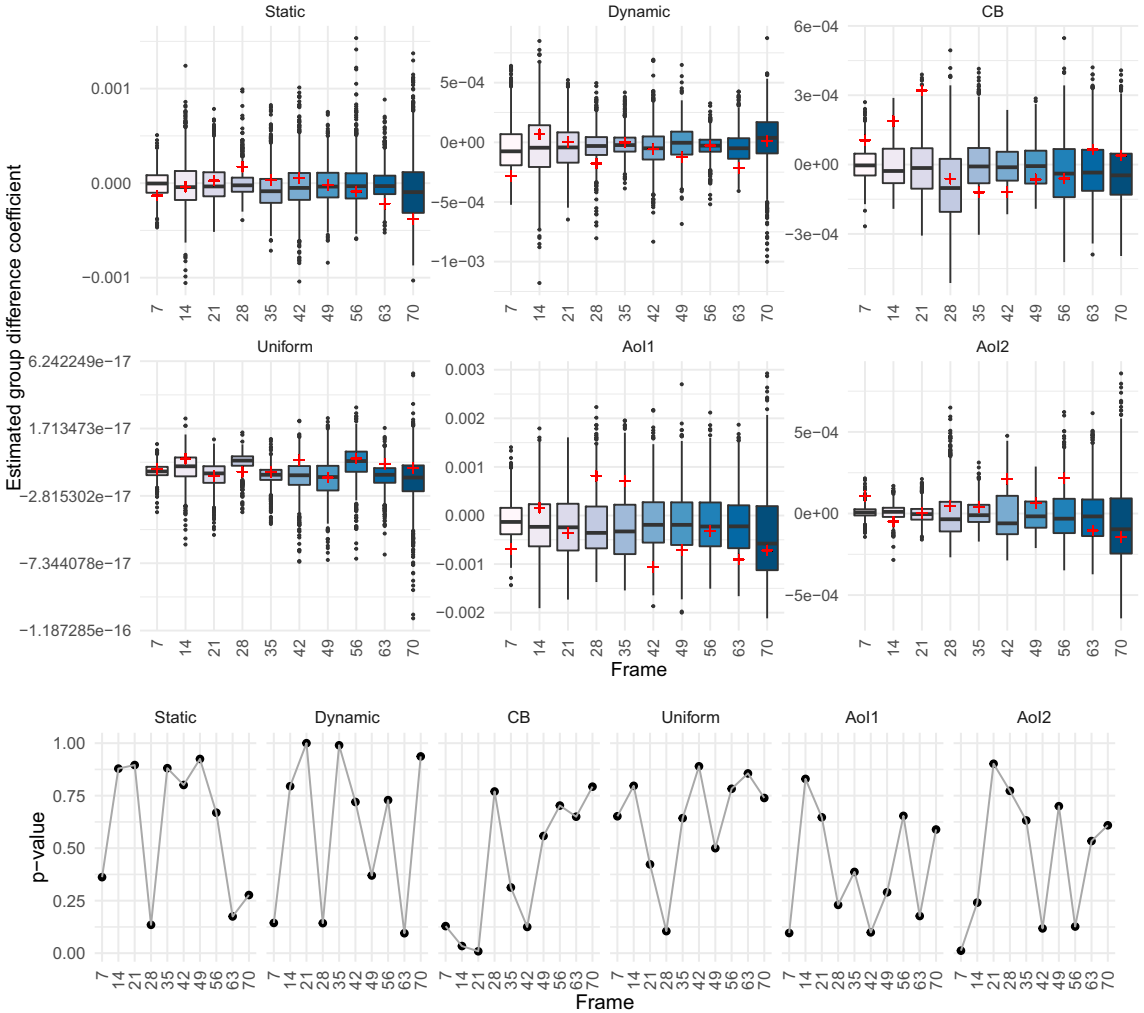

Fig. 7 Boxplots showing the estimated coefficients of the group differences from 1000 permutations in all feature maps in our proposed model (1) for a selection of equidistant frames of the stimulus *car cornfield* (*top*). *Red crosses* indicate the estimated coefficient of the true groups. Corresponding p values for each feature map and frame (*bottom*)

numerous elements in the picture. Example frames of the stimulus and the resulting RI curves are shown in Fig. 11 and results from permutation tests in Fig. 12.

Depending on the experimental context, additional AoI maps could be added, say to model objects competing for attention. Following Coutrot and Guyader (2017), we think that both saliency maps are advisable to include in calculations, but the underlying LASSO regression modeling framework continues to work when one or both saliency maps are removed from Eq. 1, changing model interpretation when doing so. To illustrate the generalizability of our approach, we include the evaluation of static stimuli in which the dynamic AoI and the dynamic saliency in model (1) are omitted (see Figs. 13 and 14). We expect that the method can also be applied to video stimuli with camera panning, since neither kernel density estimation, AoI maps nor saliency map calculations rely on static scenes.

Conclusions

This article provides a multi-group extension of a visual saliency model for dynamic stimuli by Coutrot and Guyader (2017). This allows to compare two or more natural or experimental groups in terms of the relative importance (RI) of visual features. Standardized RI plots provide an interpretable summary. The practical application of the method shows that the RI curves have similar shape in both groups, despite the more explorative gaze behavior in the treatment group. The method thus represents a group comparison tool which is robust against possible intentional changes in gaze behavior and investigates differences in highly automated and subconscious gaze behavior. In contrast to dynamic models on the level of individuals, gaze behavior is first aggregated groupwise for each frame. Hence, model coefficients and especially RI have to be interpreted as parameters of the groups gaze behavior distribution. In general, it is not possible to make predictions for individuals. In principle, the linear model framework in the background is extensible to further covariates. However, we caution that the gaze distribution needs to be estimable by kernel densities or similar approaches, which breaks down when too few individuals are available.

We demonstrate that the method also works for fewer features and for static stimuli. The method and the provided code are applicable to other natural groups and video stimuli without camera panning without any major changes. The single steps of model construction can be individually adapted and should be reflected with regard to the stimulus material.

Appendix

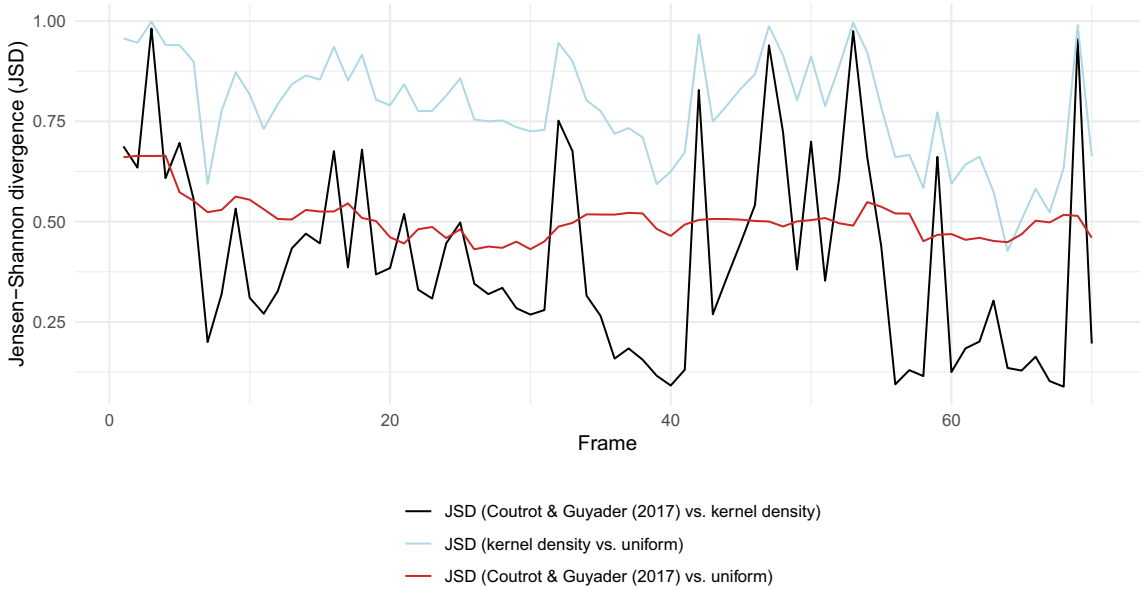

Fig. 8 Jensen–Shannon divergence (JSD) between the kernel density estimate (KDE) in Coutrot and Guyader (2017) (bandwidth 1 degree of visual angle) and the KDE with a bandwidth selection via bivariate least squares cross-validation as well as the JSD between the kernel densities and the uniform distribution for the stimulus *car cornfield*

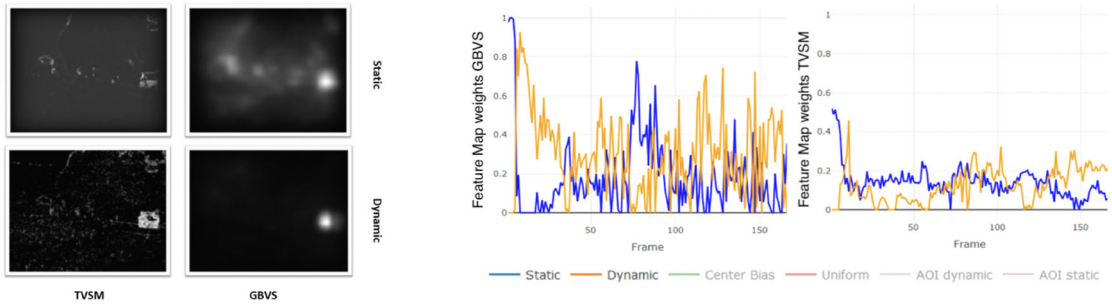

Fig. 9 Illustration TVSM and GBVS saliency maps for one frame of the stimulus *car cornfield* (*left*) and resulting feature map weights (static and dynamic saliency) for GBVS and TVSM saliency maps (*right*)

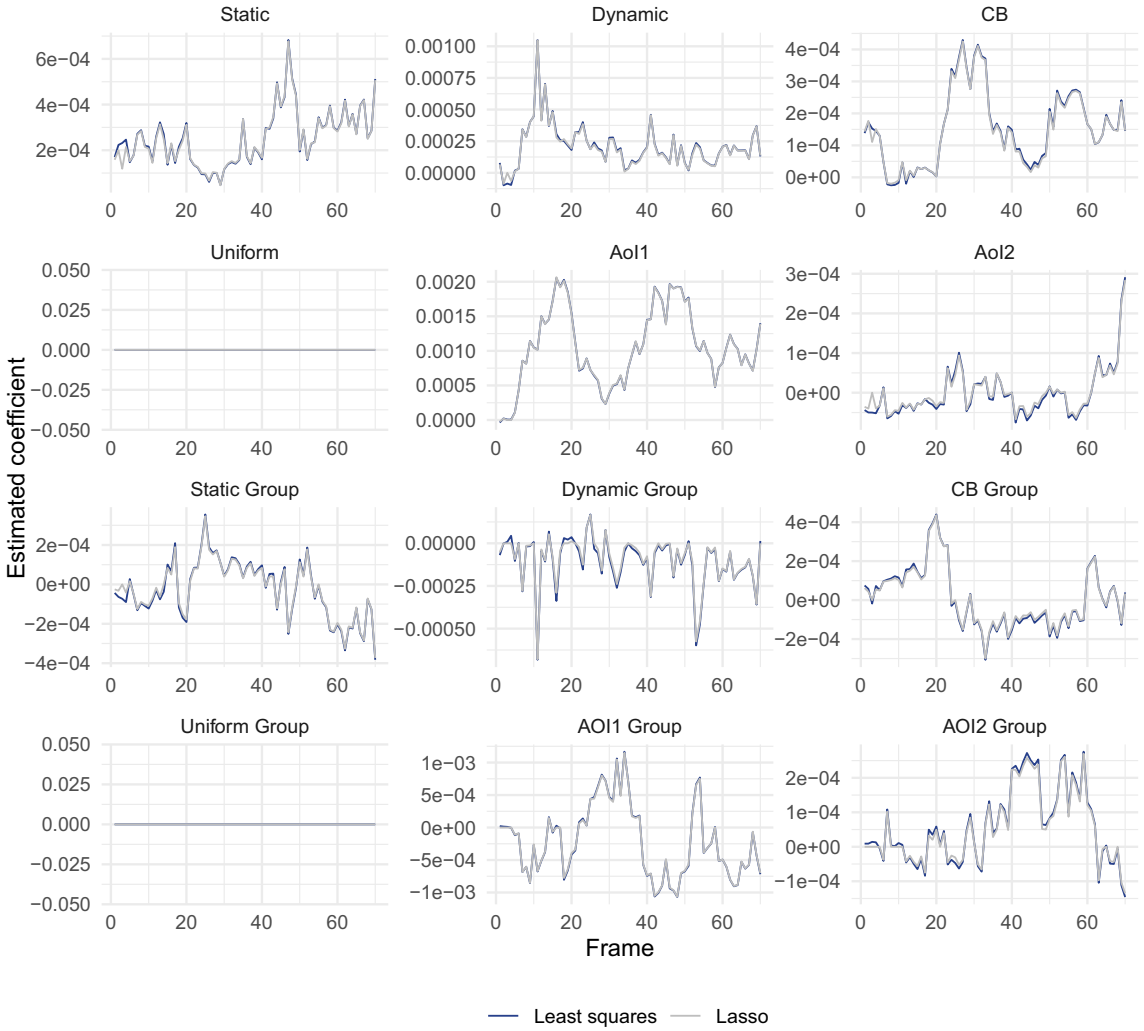

Fig. 10 RI curves for the least-squares approach (*blue*) and for the lasso penalty (*gray*)

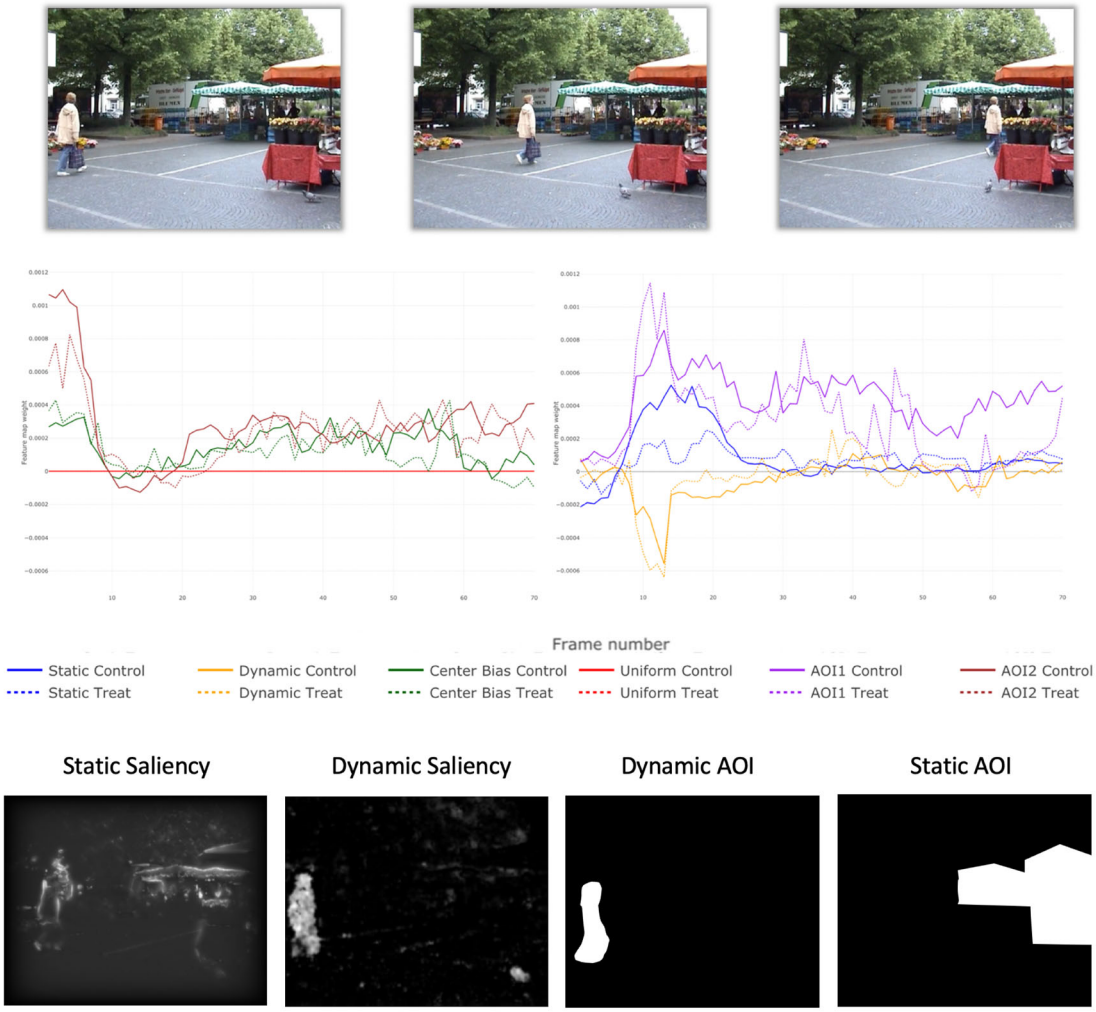

Fig. 11 Relative importance curves of the stimulus walking market with the lady as dynamic AoI (AOI1) and the market stalls as static AoI (AOI2). There is also a pigeon walking through the video (starting at the lower right corner), which is not modeled as a separate AoI

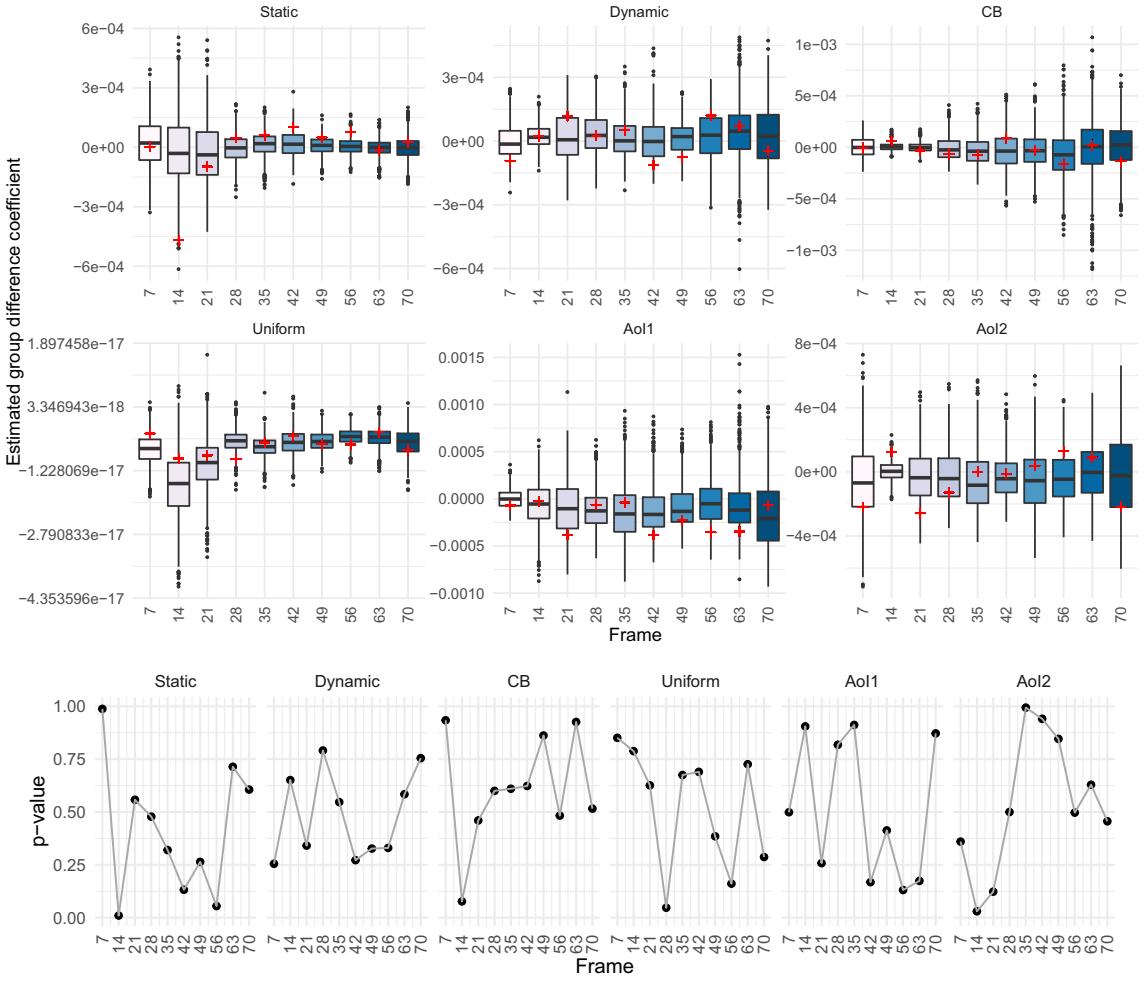

Fig. 12 Boxplots showing the estimated coefficients of the group differences from 1000 permutations in all feature maps in our proposed model (1) for a selection of equidistant frames of the stimulus *walking market (top)*. *Red crosses* indicate the estimated coefficient of the true groups. Corresponding p values for each feature map and frame (*bottom*)

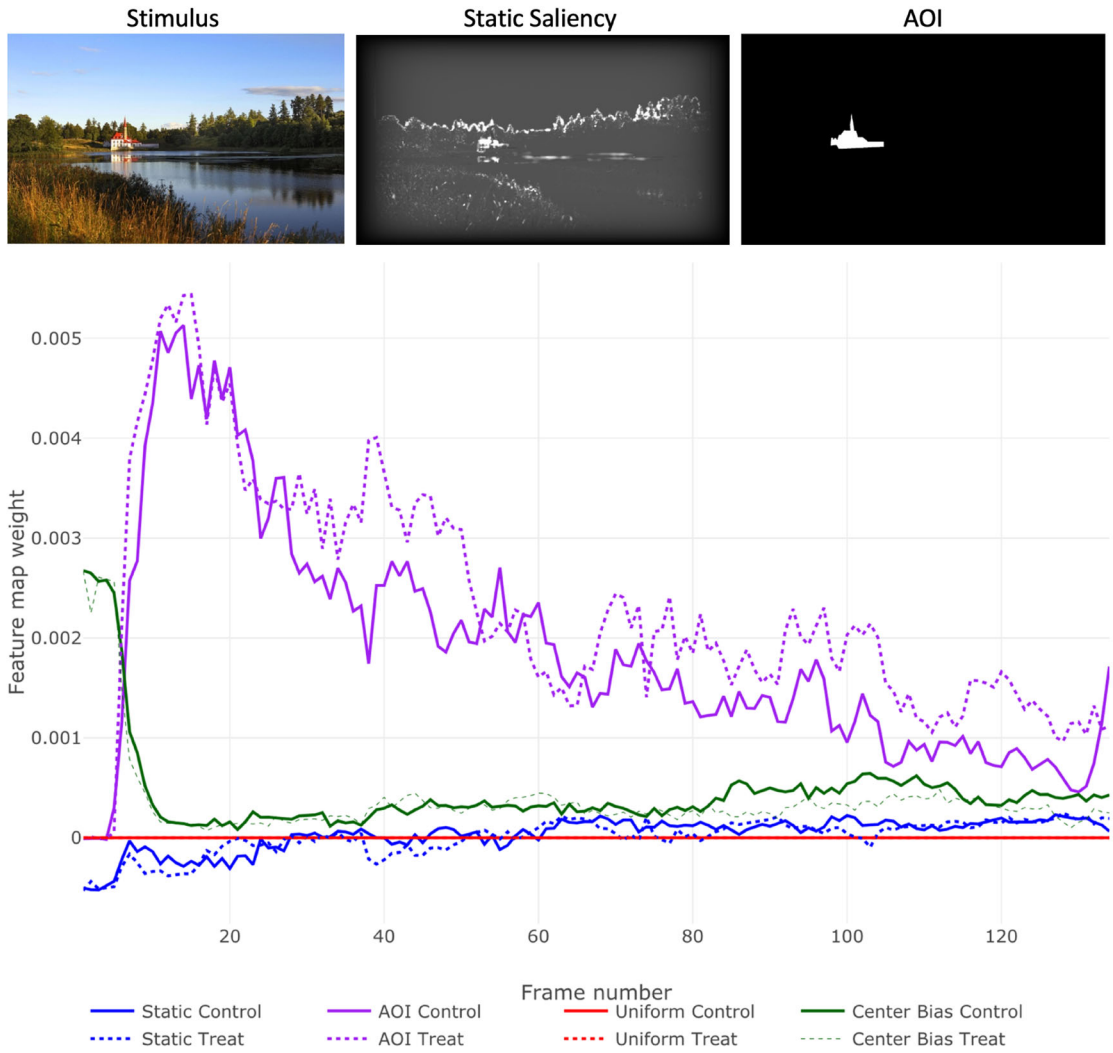

Fig. 13 Group comparison of gaze behavior between a group of $N_T = 55$ (after matching) architecture and civil engineering students (treatment) and a group of $N_C = 55$ (after matching) linguistic students (control) over time for a static stimulus and one AoI (church)

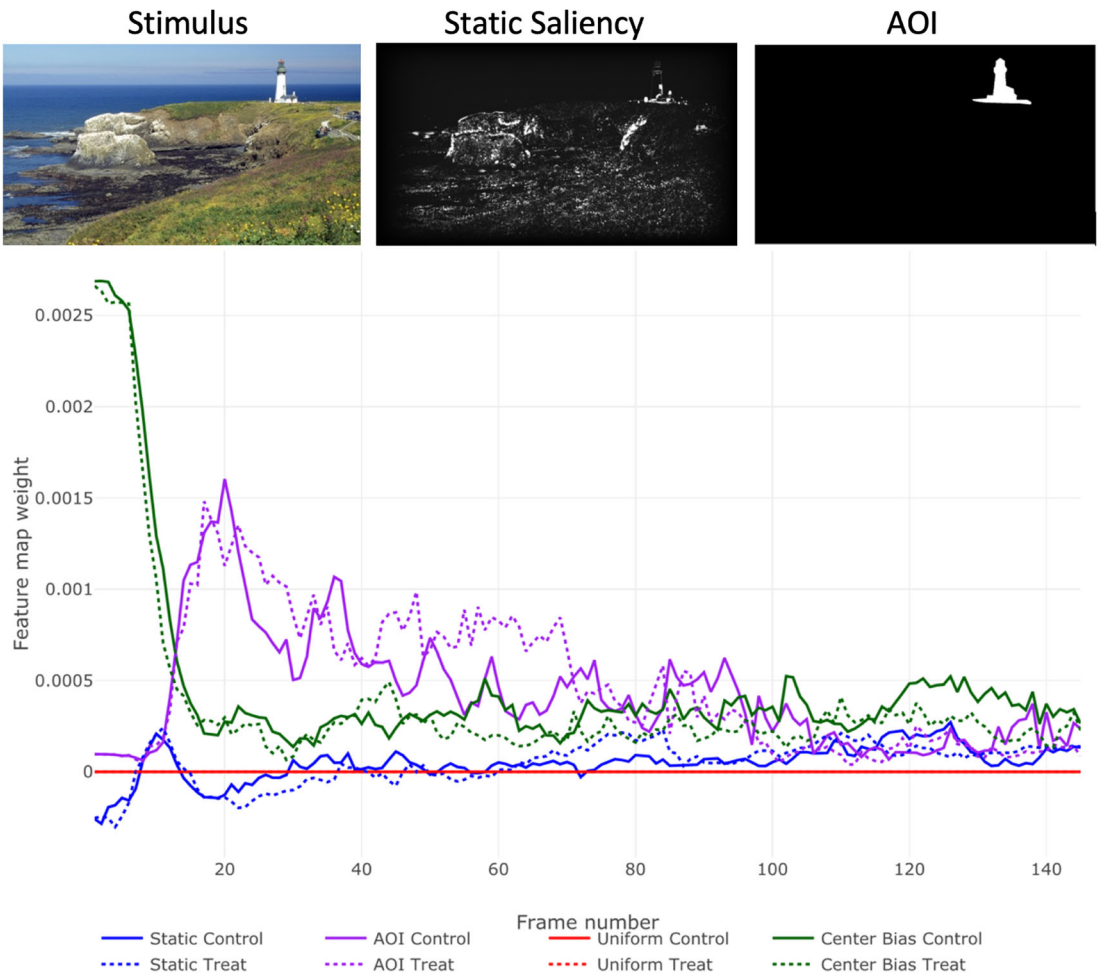

Fig. 14 Group comparison of gaze behavior between a group of $N_T = 55$ (after matching) architecture and civil engineering students (treatment) and a group of $N_C = 55$ (after matching) linguistic students (control) over time for a static stimulus and one AoI (lighthouse)

Acknowledgements We thank Timo Budszuhn for his support in the evaluation of the stimuli. M.S. is supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS".

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Practices Statements Data and analysis code for the two dynamic stimuli *car cornfield* and *walking market* are available at https://github.com/marastadler/Lasso_eyeposition. An example code can be used to replicate the analyses of these two stimuli (model, interactive plots, permutation tests) at https://github.com/marastadler/Lasso_eyeposition/blob/master/README.Rmd.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Bernal, J., Sánchez, F., Vilariño, F., Arnold, M., Ghosh, A., & Lacey, G. (2014). Experts vs. novices: Applying eye-tracking methodologies in colonoscopy video screening for polyp search. In *Eye Tracking research and applications symposium (ETRA)* (pp. 223–226). https://doi.org/10.1145/2578153.2578189

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, 116. https://doi.org/10.1016/j.visres.2015.03.005

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3), 6–6. https://doi.org/10.1167/9.3.6

Chen, X., & Zelinsky, G. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, *46*, 4118–33. https://doi.org/10.1016/j.visres.2006.08.008

Coutrot, A., & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision*, *14*. https://doi.org/10.1167/14.8.5

Coutrot, A., & Guyader, N. (2017). Learning a time-dependent master saliency map from eye-tracking data in videos. arXiv:1702.00714

Coutrot, A., Hsiao, J., & Chan, A. (2017). Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods*, *50*, 1–18. https://doi.org/10.3758/s13428-017-0876-8

Cristino, F., Mathot, S., Theeuwes, J., & Gilchrist, I. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, *42*, 692–700. https://doi.org/10.3758/BRM.42.3.692

Duong, T. (2004). Bandwidth selectors for multivariate kernel density estimation, University of Western Australia), Doctoral dissertation.

Feusner, M., & Lukoff, B. (2008). Testing for statistically significant differences between groups of scan patterns. Eye Tracking Research and Applications Symposium (ETRA), 43–46. https://doi.org/10.1145/1344471.1344481

Fontana, F., Uding, A., Cleneden, A., Cain, L., Shaddox, L., & Mack, M. (2017). A comparison of gaze behavior among elderly and younger adults during locomotor tasks. https://doi.org/10.13140/RG.2.2.16892.44165

Frame, M., Warren, R., & Maresca, A. (2018). Scanpath comparisons for complex visual search in a naturalistic environment. *Behavior Research Methods*, *51*, 1454–1470. https://doi.org/10.3758/s13428-018-1154-0

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Giovinco, N., Sutton, S., Miller, J., Rankin, T., Gonzalez, G., Najafi, B., & Armstrong, D. (2014). A passing glance? Differences in eye tracking and gaze patterns between trainees and experts reading plain film bunion radiographs. *The Journal of Foot and Ankle Surgery*, *54*. https://doi.org/10.1053/j.jfas.2014.08.013

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Proceedings of the 19th international conference on neural information processing systems*, (pp. 545–552). Canada: MIT Press.

Harezlak, K., Kasprowski, P., & Kasprowska, S. (2018). Eye movement traits in differentiating experts and laymen. In A. Gruca, T. Czachórski, K. Harezlak, S. Kozielski, & A. Piotrowska (Eds.) *Man-machine interactions, chap 5*, (pp. 82–91). Cham: Springer International Publishing.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining inference and prediction*. New York: Springer.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Holmqvist, K., & Andersson, R. (2017). *Eye-tracking: A comprehensive guide to methods, paradigms and measures*. Lund: Lund Eye-Tracking Research Institute.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 1254–1259. https://doi.org/10.1109/34.730558

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4).

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

Kübler, T., Rothe, C., Schiefer, U., Rosenstiel, W., & Kasneci, E. (2017). Subsmatch 2.0: Scanpath comparison and classification based on subsequence frequencies. *Behavior Research Methods*, *49*(3), 1048–1064. https://doi.org/10.3758/s13428-016-0765-6

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2005). Applied linear statistical models. McGraw-Hill/Irwin.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25), 3559–3565. https://doi.org/10.1016/S0042-6989(01)00102-X

Le Meur, O., & Baccino, T. (2012). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods*. https://doi.org/10.3758/s13428-012-0226-9

Le Meur, O., Thoreau, D., Le Callet, P., & Barba, D. (2005). A spatiotemporal model of the selective human visual attention. *ICIP*, *3*, 1188–1191.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. (2008). Spatio-temporal saliency model to predict eye movements in video free viewing. In *16th European signal processing conference (EUSIPCO)* (pp. 1–5). https://doi.org/10.5281/zenodo.40947

Marat, S., Rahman, A., Pellerin, D., Guyader, N., & Houzet, D. (2013). Improving visual saliency by adding 'face feature map' and 'center bias'. *Cognitive Computation*, *5*(1), 63–75. https://doi.org/10.1007/s12559-012-9146-3

Navarro, J., Reynaud, E., & Gabaude, C. (2017). Eye movement analysis in dynamic scenes: Presentation and application of different methods in bend taking during car driving. *Le Travail Humain*, *80*, 307. https://doi.org/10.3917/th.803.0307

Peters, R., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proceedings of IEEE computer vision and pattern recognition*. https://doi.org/10.1109/CVPR.2007.383337

Peters, R., & Itti, L. (2008). Applying computational tools to predict gaze direction in interactive visual environments. *TAP*, 5. https://doi.org/10.1145/1279920.1279923

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Stutterheim, C. V., Andermann, M., Carroll, M., Flecken, M., & Mertins, B. (2012). How grammaticized concepts shape event conceptualization in language production: Insights from linguistic analysis, eye tracking data, and memory performance. *Linguistics*, *50*, 833–867. https://doi.org/10.1515/ling-2012-0026

Sundstedt, V., Stavrakis, E., Wimmer, M., & Reinhard, E. (2008). A psychophysical study of fixation behavior in a computer game. In *APGV 2008—Proceedings of the symposium on applied perception in graphics and visualization*. https://doi.org/10.1145/1394281.1394288

Treisman, A., & Gelade, G. A. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7), 4. https://doi.org/10.1167/9.7.4

Vigneau, F., Caissie, A., & Bors, D. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*, 261–272. https://doi.org/10.1016/j.intell.2005.11.003

Yamada, K., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A., & Hiraki, K. (2011). Can saliency map models predict human egocentric visual attention? In R. Koch, & F. Huang (Eds.) *Computer vision – ACCV 2010 workshops*, (pp. 420–429). Berlin: Springer.

Zhang, Z., Kim, H., Lonjon, G., & Zhu, Y. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine*, *7*, 16. https://doi.org/10.21037/atm.2018.12.10

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*. https://doi.org/10.1167/11.3.9

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, *35*(5), 2173–2192. https://doi.org/10.1214/009053607000000127

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Mara Stadler[1,2,3] · Philipp Doebler[1] · Barbara Mertins[4] · Renate Delucchi Danhier[4]

Philipp Doebler
doebler@statistik.tu-dortmund.de

Barbara Mertins
barbara.mertins@tu-dortmund.de

Renate Delucchi Danhier
renate.delucchi@tu-dortmund.de

1 Department of Statistics, TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany

2 Present address: Department of Statistics, Ludwig Maximilian University of Munich, Ludwigstr. 33, 80539 Munich, Germany

3 Present address: Institute of Computational Biology, Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany

4 Department of Culture Studies, TU Dortmund University, Emil-Figge-Str. 50, 44227 Dortmund, Germany