Is This Really Relevant? A Guide to Best Practice Gaze-based Relevance Prediction Research

Melanie Heck
University of Mannheim
Mannheim, Germany
melanie.heck@uni-mannheim.de

Paulina Sonntag
University of Mannheim
Mannheim, Germany
psonntag@mail.uni-mannheim.de

Christian Becker
University of Mannheim
Mannheim, Germany
christian.becker@uni-mannheim.de

ABSTRACT

As eye tracking is becoming feasible on commodity devices, it provides a powerful tool for inferring users' perceived relevance of objects. Yet the prediction quality depends on multiple parameters that have to be considered when designing the prediction model. In this paper, we review approaches to predict relevance from gaze with regard to five design issues: 1) extracting features, 2) defining the algorithm, 3) setting a prediction scope, 4) eliminating visual distractors, and 5) evaluating the system. The insights may serve as a guide to establish best practices for the design and evaluation of relevance prediction models, thus allowing for better comparability of future work. We further discuss promising fields of application that will drive future research on gaze-based relevance prediction.

CCS CONCEPTS

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

KEYWORDS

relevance prediction, eye tracking, adaptive media, gaze-contingent systems

ACM Reference Format:
Melanie Heck, Paulina Sonntag, and Christian Becker. 2021. Is This Really Relevant? A Guide to Best Practice Gaze-based Relevance Prediction Research. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3450614.3464476

1 INTRODUCTION

Customers of online shops find that whenever they visit their favorite store, some items have already been handpicked for them. A similar experience is provided by streaming platforms, where customers are recommended content that matches their interests. In order to provide such personalization, individual preferences have to be determined.

Given the continuously maturing technology, eye tracking has enjoyed increasing popularity as an indicator for users' relevance judgments. During the last years, the ease of use of eye trackers has continuously increased up to a point where they can now be easily integrated into a user's smartphone camera or webcam [29]. The major advantage of this approach is that the users don't have to explicitly judge the relevance of the items themselves. Instead, inferences are made from their behavior while naturally using the system [22]. Yet, the accuracy of the predictions is highly influenced by a multitude of parameters that have to be considered when designing the prediction model. The lack of standardized procedures for evaluating the models make a direct comparison of different approaches difficult.

Motivated by this gap, we review the state of the art of gaze-based relevance predictions. We highlight both the advantages and disadvantages of the implemented models and evaluation methods. This shall stimulate a discussion to establish best practices for the definition and evaluation of gaze-based relevance prediction models. As a starting point, we propose a general evaluation procedure that aims to ensure comparability of future research works.

Relevant literature was found by using a keyword search in Google Scholar with the search terms ("user profile modeling" OR "user modeling") AND ("eye tracking" OR "review" OR "survey"). We started with this initial set and then additionally conducted backward reference searching in the papers we collected. We included all papers concerned with the mapping of gaze data to the user's relevance judgment of objects. We coded each paper based on the design of the prediction model. Finally, we identified five main categories: 1) Extracting gaze features; 2) Defining the prediction model; 3) Setting a scope for the prediction model; 4) Accounting for visual distractors; 5) Collecting ground truth data and evaluating the system. Starting from the application domains of the reviewed systems, we further discuss the most promising fields of application that will drive future research on gaze-based relevance predictions.

2 RELATED WORK

The existence of a *link between gaze and relevance* of elements has been established in the 1960s [30]. Citing the results of several psychological studies, Bednarik [1] more recently confirms that gaze allocation is closely related to cognitive processes. *Application areas* for gaze-based relevance predictions have thus been the subject of multiple literature reviews. In market research, gaze allocation is a popular indicator for products and product features that are perceived as relevant by consumers [7, 45]. In attentive user interfaces, in contrast, the users' focus of attention triggers a direct system response with the aim to support the users in their current task [19, 31]. A comprehensive overview of *algorithmic approaches* for relevance prediction based on implicit user feedback can be found in [22].

UMAP '21 Adjunct, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8367-7/21/06.
https://doi.org/10.1145/3450614.3464476

Yet to the best of our knowledge, the state-of-the art approaches to predict relevance from gaze have not been analyzed in a systematic manner. We thus take a look at recent research works, highlighting their advantages and disadvantages, and identify areas that can benefit from applying the knowledge created by the prediction models.

3 HOW IS RELEVANCE PREDICTED?

The quality of the prediction model depends on the gaze features that are fed into the prediction algorithm, the extent to which the model should be applicable to new objects and users, and the elimination of visual distractors. The reported performance, in turn, is highly influenced by the evaluation procedure.

3.1 Extracting gaze features

Gaze patterns consist of alternations between fixations and saccades. During fixations, the user focuses on one specific point, usually to acquire information. During the saccades in between fixations, only little information is absorbed. Relevance is therefore often inferred solely from fixations [21]. Table 1 summarizes the common gaze features.

3.1.1 What gaze features contain relevance information? A human's gaze tends to linger on relevant elements [9]. Concordingly, relevance is most often inferred either from the *total gaze duration* (in terms of time [39] or number of gaze points [40, 46]), or *fixation duration* [5, 18, 38, 41] on an object. The *total number of fixations* [5, 44] places less weight on prolonged continuous observations of one object. In contrast, the *continuous gaze* [32] or *continuous fixation duration* [5, 21], considers uniquely how long the gaze dwells on an object before readjusting the focus.

By using feature vectors, the combined information content of multiple gaze features can be used. A subset of the primary features mentioned above is usually used. In addition, secondary features are included which may not be suitable to directly infer object relevance, but explain to some degree the user's gaze allocation. Since the initial focus tends to be guided by salient features instead of interest in the element, the *first fixation* [6] and *time to first focus* [27, 29] control for this effect. In contrast, the *last fixation* is most often placed on an object that is perceived as relevant [15, 36]. The intensity of the focus is measured by the *number of transitions* to an element [26, 27, 32, 36, 40], *standard deviation of the number of fixations* [27], *distance between fixations* [27], *distance of fixations from the object center* [25], and *saccade length* [13, 24, 25]. *Pupil dilation* tends to be smaller [6, 24] when focusing on relevant items.

A comprehensive image of the user's gaze patterns is represented by the *scanpath* [24, 35]. This time series collection of fixations and saccades provides the most holistic representation of the user's gaze allocation.

3.1.2 How are gaze features assigned to an object? Predictions are based on the distribution of gaze data to Areas of Interest (AOI). These are rectangle areas around the elements for which relevance is elicited. An AOI may encompass a static graphic element such as an image [5, 6, 8, 14, 15, 26, 27, 29, 32, 36, 39, 41, 46] or interaction element [21]. In text documents, it usually holds a single word [12, 13, 18, 35, 38, 46]. For dynamic targets in videos [44, 46] or real-world scenes [24, 25], die AOI has to be defined for each video frame or observation window individually.

3.2 Defining the prediction model

When selecting the relevance prediction algorithm, it is imperative to consider on what level of detail the system aims to predict relevance. Table 2 provides an overview of the proposed implementation.

3.2.1 How is relevance defined? Relevance can be predicted on multiple levels of detail: On the most basic level, it refers to a *binary* evaluation [13, 18, 21, 24, 26, 29, 32, 35, 38, 41]. The prediction model infers whether a given element is relevant or not. On a deeper level, the model identifies the *most relevant* element out of a given set [14, 15, 25, 36, 39, 40, 44]. The relevance judgment is thus based on a comparison between multiple elements. The most detailed models *quantify* the relevance that each object has to the user [5, 6, 8, 12, 27, 29, 46]. This allows to rank the elements.

3.2.2 What prediction logic is applied? Most commonly, binary evaluations are based on *threshold values* [18, 21, 32, 38]. This approach was first applied in gaze-controlled systems. In order to avoid that each fixation on an object instantly triggers an action (a.k.a. Midas Touch), objects are only activated when the dwell time exceeds an empirically defined threshold [21]. It is usually defined in terms of gaze duration [18, 21, 32, 38], but can also be applied to other gaze features such as the number of fixations [5]. Empirically defined optimal thresholds range from 150 ms [21] to 800 ms [18]. In order to rank the relevance of elements, multiple threshold values can be defined [5].

In the most naive implementation, relevance is quantified by the *cumulative gaze allocation* to each object [6, 8, 12, 46]. The user's preferred object can directly be inferred by selecting the element with the highest visual focus [14, 39, 40, 44]. Starker and Bolt [40] additionally consider the time dimension by discounting the importance of old gaze points whenever a new gaze point is recorded. When using feature vectors, weights can be assigned to manipulate the impact that each feature has on the relevance prediction. Equal vector weights assign the same importance to all features [12].

If different weights are used, their optimal values can be determined explicitly based on experience [6, 8], or implicitly through *machine learning* [8, 13, 15, 24–27, 29, 35, 36, 41, 46]. Provided that ground truth data for the relevance of the elements exists, the latter determines the optimal weights in a training phase. The features are mapped to the objects' known relevance, so that the number of correct predictions is maximized. Linear discriminant analysis (LDA) [26] is a simple model which assigns the data to linear subspaces to maximize class discrimination. Classes are defined as "relevant", or "not relevant" respectively. The often more robust logistic regression (logit) maps the reported relevance class to a linear combination of the features [15, 25, 27]. It predicts the probability of an object being relevant or not, and can therefore also be used to quantify relevance. Being capable of modeling both linear and non-linear relationships, k-Nearest Neighbor (k-NN) minimizes the difference between data points in the class [15]. For the same

Table 1: Gaze features commonly used in relevance prediction models

Reference	Dimension	dwell time						intensity						sequential			
		# gaze points	total gaze time	fixation duration	# fixations	continuous gaze duration	continuous fixation duration	# transitions	# fixations (sd)	distance between fixations	distance to object center	saccade length	pupil dilation	first fixation	time to first focus	last fixation	scanpath
Starker & Bolt, 1990 [40]	single feature	•						•									
Jacob, 1990 [21]	single feature						•										
Sibert et al., 2000 [38]	single feature			•													
Hyrskykari et al., 2003 [18]	single feature			•													
Salojärvi et al., 2004 [35]	feature vector			•	•												•
Qvarfordt & Zhai, 2005 [32]	single feature					•		•									
Vesterby et al., 2005 [44]	single feature				•												
Hardoon et al., 2007 [13]	feature vector			•	•							•					
Klami et al., 2008 [26]	feature vector			•	•		•	•									
Xu et al., 2008 [46]	single feature	•															
Kozma et al., 2009 [27]	feature vector			•		•		•	•	•					•		
Cheng et al., 2010 [6]	feature vector		•		•								•	•			
Kandemir et al., 2010 [25]	feature vector			•			•				•	•					
Kandemir & Kaski, 2012 [24]	feature vector						•						•				•
Giordano et al., 2012 [12]	feature vector				•		•										
Li et al. (2017) [29]	feature vector		•												•		
Chen et al., 2017 [5]	single feature			•	•		•										
Schweikert et al., 2018 [36]	feature vector		•		•			•								•	
Song & Moon, 2019 [39]	single feature		•														
Fahim Shahriar et al., 2020 [8]	feature vector			•	•												
Sulikowski et al., 2020 [41]	single feature			•													
Heck et al., 2019 [15]	feature vector	•		•	•		•									•	

purpose, Support Vector Machines (SVM) construct a hyperplane so that the area around class borders is maximized [13, 15, 29]. Ensemble classifiers including boosting algorithms [15, 36] and Random Forests [15, 29] combine multiple machine learning algorithms and therefore often perform better. Yet Li et al. [29] found that the SVM classifier outperforms not only Decision Trees (DT), but also Random Forests (RF). While the readily interpretable Decision Trees learn simple decision rules for data classification or regression, the ensemble classifier RF predicts relevance as the averaged result of multiple randomly constructed Decision Trees. When tested against other ensemble methods, the RF classifier in turn outperforms Mixed Group Ranks (MGR), boosting, and a multi-layer neural network (NN) [15, 36]. The Multi-layer Perceptron (MLP) deep neural network implements multiple non-linear hidden layers to map the gaze features to relevance classes[15, 41]. Sulikowski and Zdziebko [41] use the model to infer relevance of objects with varying display positions, while using a small training set. If large amounts of training data are available, Passive-Aggressive (PA) classifiers can be used [15]. Gaussian Processes (GP) [24] and Hidden Markov Models (HMM) [35] can be applied to sequential scanpath features. Based on Bayesian methodology, GPs represent states in terms of conditional probability distributions. The classification is determined by the sign of the final probability [24]. HMMs represent sequential changes in the data as probabilistic transitions between hidden states. Discriminative HMMs define transitions so that the likelihood of the data being assigned to the correct class is maximized. Salojärvi et al. [35] find that discriminative HMMs outperforms LDA and SVMs. Yet the improvement over the two models, both of which use averaged feature vectors instead of sequential data, is marginal.

3.3 Setting a scope for the prediction model

The scope of a prediction model encompasses two dimensions: The generalisability of relevance predictions with regard to other users, and to other objects (see Figure 1).

3.3.1 Can the model be applied to new users? Models based on cumulative gaze allocation tend to generalize well across different users, even if different weights are assigned to each feature [6, 8]. The same applies to activation thresholds, although Sibert et al. [38] recognize that the optimal threshold value might be different for each user.

Machine learning models have proven to deliver reliable relevance predictions when trained with data of other users [13, 15, 24, 29, 35,

Table 2: Algorithms and factored distractors

		relevance concept			model specification	distractors				
		binary (0/1)	most relevant	quantification		display time	size	complexity	location	activation
cumulative models	Starker & Bolt, 1990 [40]		•							
	Vesterby et al., 2005 [44]		•			•				
	Xu et al., 2008 [46]			•						
	Cheng et al., 2010 [6]			•						
	Giordano et al., 2012 [12]			•						
	Heck et al., 2019 [14]		•							
	Song & Moon, 2019 [39]		•			•				
	Fahim Shahriar et al., 2020 [8]			•						
threshold values	Jacob, 1990 [21]	•			150-250 ms					
	Sibert et al., 2000 [38]	•			240 -360 ms		•	•		
	Hyrskykari et al., 2003 [18]	•			800 ms			•		
	Qvarfordt & Zhai, 2005 [32]	•			450 ms		•			•
	Chen et al., 2017 [5]			•	# fixations: [3.16; 2.64; 1.42], cont. fixation time: [289ms; 340ms; 144ms] total fixation time: [1089ms; 1039ms; 448ms]					
machine learning	Salojärvi et al., 2004 [35]	•			LDA, SVM, HMM					
	Hardoon et al., 2007 [13]	•			SVM		•	•	•	
	Klami et al., 2008 [26]	•			LDA					
	Kozma et al., 2009 [27]	•			logit					
	Kandemir et al., 2010 [25]		•		logit		•			
	Kandemir & Kaski, 2012 [24]	•			GP					
	Li et al., 2017 [29]	•		•	SVM, DT, RF	•			•	
	Schweikert et al., 2018 [36]		•		NN, boost, RF, MGR					
	Sulikowski et al., 2020 [41]	•			MLP			•	•	
	Heck et al., 2021 [15]		•		logit, SVM, DT, RF, MLP, boost, PA, k-NN			•	•	

36, 41]. The predictions are slightly better when training with user-specific data [25]. Yet person-independent models carry the benefit of not having to be trained for each user individually [26, 27].

3.3.2 Can predictions be extrapolated to new objects? The primary objective of the prediction models is to infer the relevance of elements that are displayed on the user interfaces. By applying content or collaborative filtering techniques, the relevance of other objects that the user has not previously looked at can be predicted [6, 8, 12, 13, 27, 39, 46].

Content-based filtering analyzes the key properties of the objects and identifies others that are similar with regard to these features [6, 8, 12, 13, 27, 46]. Automated object retrieval with color correlograms [17] selects images based on correlations of pixel colors [46]. Using a tree-structured self-organizing map, the unsupervised neural algorithm PicSOM [28] analyzes the similarity between two images based on the three low-level features color, texture, and shape [27]. Documents can be analyzed with regard to the occurrence of keywords that the user perceives as most relevant [12]. They can be represented as a bag-of-words model that decomposes the document into its individual words. SVMs are then trained with the bag-of-words representation of documents that have been identified as relevant for the user [13]. The trained SVM model is applied to the bag-of-words representation of new documents to rank their relevance. Cheng et al. [6] manually encode products into a binary representation of five product-related features. Based on the user's visual attention to different products, a genetic algorithm constructs the optimal product for a given user. The most similar products are then presented to the user. K-means clustering groups objects so that similarity within clusters is maximized, while minimizing between-cluster similarity. Objects withing the cluster that contains the largest amount of elements that are known to be relevant to the user are predicted to also be relevant. Product recommenders usually cluster objects along product-related features such as brand, category, and price [8].

Collaborative filtering, in contrast, is based on similarity between users [8, 39]. An unseen element is considered as relevant if it has been defined as relevant by other users with similar preferences. Pearson correlations provide a simple measure of similarity [39]. K-means clustering can be applied for an automated identification of similar user groups [8].

3.4 Accounting for visual distractors

Buscher et al. [4] identify three factors that influence the gaze: 1) The task and information need determines our perception of what

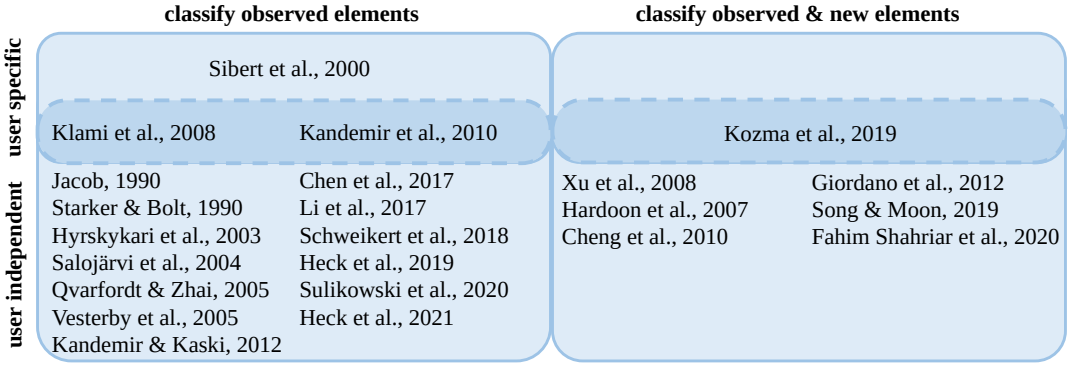

Figure 1: The scope of the prediction model is defined by its generalisability across users and objects of interest.

is relevant or not, and is therefore the target of relevance predictions. 2) Expectations about where to find relevant information is based on prior experience. 3) The gaze is primarily drawn towards text or human figures, especially faces. If no specific object is present, the gaze is directed towards the center of the screen, or towards locations with salient low-level features [23]. Seven low-level features are generally considered to determine salient regions of a scene: *orientation* [2–4, 20, 47], *luminence/ intensity* [2–4, 11, 20, 47], *color* [2–4, 20, 33, 34, 37, 42, 47], *motion* [3, 20, 42, 47], *texture* [2–4, 34, 47] (defined as the number of sharp edges in a frame), *size* [2, 4, 42], and *detail* [2–4, 30]. Unintended gaze allocation to salient regions therefore has to be accounted for. Table 2 summarizes the prevalent approaches.

3.4.1 What distractors are considered? Most low-level salient features tend to be accounted for implicitly (see Section 3.4.2). Assuming that more gaze points fall on objects that occupy proportionally more display space than others, object *size* (or equivalently word length) is often considered explicitly [13, 25, 32, 38]. In text-based relevance predictions, both the *complexity* of the reading material and the user's abilities may affect the dwell time on a word. This effect is captured by the word frequency [13, 18], type settings [38], and user age and cognitive abilities [41]. Expectations about where to find information may be accounted for by the object *location* on the page [13, 29, 41]. The *activation* history of elements reflects the logical relationships between objects without having to trace the user's scanpath [32]. Close relationship to the previously activated object indicates high relevance. Previous activation has the opposite effect, since users are less likely to be interested in objects they have already selected before.

3.4.2 How are distractors determined and integrated into the model? The magnitude of each correction parameter can be specified either explicitly, or learned implicitly. *Explicit parametrization* defines fixed values based on the empirical findings derived from experimental data [18, 32, 38]. Activation thresholds can thus be reduced to account for distractors. They may vary for each object within one application, depending on how salient the element is [18, 32, 38]. Cumulative gaze allocation models should account for varying exposure times of objects by normalizing the feature values with regard to the display time of each object [29, 39, 44].

Learned parameters are directly derived from the data by mapping the gaze to the known relevance of an object [8, 13, 14, 25, 29, 39, 41]. Machine learning models implicitly already account for visual distractors by mapping the input features to the known relevance of the objects during the training phase. Additional object-related distractors can easily be accounted for by extending the input vector [13, 25, 29, 41]. Less complex models condition object relevance on the distributions of gaze and preferences of other users [8, 14, 39]. Similar to machine learning models, this discounts the relevance of objects that generally draw much attention but are not considered as relevant by users. These data-driven approaches have a more solid empirical foundation than explicitly defined correction parameters. Yet they are only viable if gaze data and the corresponding relevance evaluation for the objects have already been collected.

3.5 Collecting ground truth data and evaluating the system

In order to determine whether the developed model accurately predicts relevance, ground truth data about the users' true relevance perceptions are required. The gaze data of users interacting with the target system thus have to be recorded in an experimental setting along with their relevance assessment of one or multiple elements.

3.5.1 How is data about the users' true relevance perception collected? The validity of the model evaluations is to a large degree influenced by the amount of collected data and the method for relevance elicitation (see Table 3).

The *amount* of collected data depends on the number of participants in the user study (ranging from 3 [35] to 132 [15]), as well as the number of tasks completed by each participant and objects per task. When participants are instructed to browse freely, the number of displayed objects is different for each person [5, 6, 8, 27, 29, 32, 41]. In experiments where users are exposed to a fixed number of objects [13, 24, 26, 35, 36], or look at a dynamic stimulus for a specified amount of time [15, 25], the total number of samples (where a sample is defined as the individual gaze record of one user while looking at one object) ranges from 70 samples [24] to 2700 samples [26].

The ideal *moment and task* for the data collection depend on the target application. Since users might not be able to recall their relevance perception for each displayed object after the experiment,

Table 3: Experimental setup and metrics for model and system evaluation

	ground truth collection				sample size		model evaluation metric						
Reference	**real-time indication**	**ex post indication**	**task-induced**	**pre-determined**	**# participants**	**# objects**	**accuracy**	**precision**	**recall**	**F1**	**AUC**	**rank metric**	**system evaluation**
Sibert et al., 2000 [38]					8	n/d							task support
Hyrskykari et al., 2003 [18]					n/d	n/d							usability
Salojärvi et al., 2004 [35]				•	3	540	•						
Qvarfordt & Zhai, 2005 [32]					12	varying							usability
Vesterby et al., 2005 [44]					11	66							manipulation awareness
Hardoon et al., 2007 [13]	•			•	6	600		•					
Klami et al., 2008 [26]	•				27	2700					•		
Kozma et al., 2009 [27]		•			6	varying		•			•		
Cheng et al., 2010 [6]		•	•		9	varying			•				usability, task support
Kandemir et al., 2010 [25]		•			4	n/d	•						
Kandemir & Kaski, 2012 [24]	•				5	70	•			•	•		
Giordano et al., 2012 [12]					30	varying		•					usability
Li et al., 2017 [29]		•			36	varying					•	•	
Chen et al., 2017 [5]			•		18	varying						•	
Schweikert et al., 2018 [36]	•				12	1440		•					
Fahim Shahriar et al., 2020 [8]		•			20	varying		•					usability
Sulikowski et al., 2020 [41]			•		52	varying	•	•	•		•		manipulation awareness
Heck et al., 2021 [15], [16]	•				132	varying	•						usability

ground truth data is often collected in real-time. This can be done by asking participants to explicitly indicate their relevance assessment of an object while looking at it [13, 15, 24, 26, 36]. Alternatively, relevance can be derived from the users' interactions with functionalities that constitute an integral part of the system [5, 6, 41]. In recommender systems, this can for example be the selection of a product for purchase [5, 6], or the submission of a product review [41]. Collecting the ground truth data after the experiment has the advantage of not disturbing the user's task flow. This is especially relevant when data is collected while the participants interact with the actual target systems [6, 8, 25, 27, 29]. In these cases, participants are thus asked to use the system under the pretext of its intended purpose, and afterwards explicitly indicate how relevant each of the displayed items had been to them. Alternatively, the relevance of objects can be pre-determined by the task itself. Typical tasks include asking participants to browse through a number of objects and find specific information. The information can be retrieved from the relevant objects [13, 35].

3.5.2 What metrics are used for the assessment? Binary classification is commonly evaluated with regard to accuracy, precision, recall, F1, or AUC scores. Accuracy indicates the percentage of correct predictions [15, 24, 25, 35, 41], but is often biased in imbalanced datasets. This is a frequent issue in applications that predict relevance, since more items tend to be classified as "irrelevant" than "relevant". Precision thus measures how many of the items that were predicted as relevant were also perceived as such [8, 12, 13, 27, 36, 41]. Recall takes into account missed items by measuring how many relevant items were also predicted to be relevant [6, 41]. The effects captured by precision and recall are combined in the F1 score [24]. The area under the receiver operator characteristics curve (AUC) is a combined measure of recall and the False Positive Rate (i.e., How many irrelevant items were falsely predicted as relevant?) [24, 26, 27, 29, 41]. It indicates of how well the classifier can distinguish between relevant and irrelevant items, independent of the dataset.

The *ranking performance* of a model can be evaluated based on the Hit-Ratio@K measure [5]. It indicates the proportion of relevant items that were also predicted to be among the top K most relevant items. On a more fine-grained level, the NDCG@k measures how precisely the top k items were ranked [29].

Since few evaluations report all relevant metrics, a comparison between the models remains difficult.

3.5.3 How are system effectiveness and usability evaluated? Subjective user feedback about the system can be collected after their interaction with the system. The assessment can either be of qualitative nature in the form of interviews [32, 44], or quantitative through standardized questionnaires and log data [6, 8, 12, 16, 18, 32, 38].

Interviews may reveal whether users are aware of the manipulative nature of the system [32, 44]. They can also elicit detailed feedback about functionalities that are appreciated by the users, and uncover those that require refinement [32].

Standardized *questionnaires* only deliver feedback about a set of pre-specified metrics. Yet they provide numeric data on fixed item scales. This allows to statistically compare the systems to a

non-adaptive implementation [6]. Metrics for evaluation include satisfaction with the system [6, 12, 18, 32], ease of use [6], interest/ engagement [6, 16], and liking [8]. Recommender systems can be evaluated with regard to their perceived influence on the user's item selections [41].

Objective feedback is provided by *log data* [6, 38]. The recorded metrics can be as diverse as the number of clicks and keystrokes [6], task completion time [6, 38], and errors while performing an application-specific task [38].

4 WHAT CAN RELEVANCE PREDICTIONS BE USED FOR?

Gaze can be used to either trigger a direct system reaction, or to create user profiles based on the user's attention to the interface elements. The latter allows to predict the relevance of new items, and thus personalize the displayed content.

4.1 Using relevance predictions for known elements

The initial interest in eye tracking was nurtured by the idea to use gaze as an *interaction technique* [21].

Diagnostic applications are used in *psychology research* to gain insights into a person's preferences or behavioral patterns. Schweikert et al. [36] thus predict a user preferences for face images.

With the emergence of *attentive user interfaces*, visual attention was used to adapt the interface to better support the user's current task. These interfaces derive the users' attention from their gaze behavior and put fixated elements into focus [43]. Starker and Bolt [40] monitor eye movements in a virtual world inspired by the story of "The little prince". Whenever the user looks at an object for a prolongued time, an animated little prince talks about the focused object. The interactive map iTourist [32] helps users plan a city trip by displaying additional information about fixated locations. By applying relevance predictions to real-world objects, Kandemir et al. [24, 25] personalize augmented reality information for objects in an art gallery. In interactive movies, the user's attention to elements in a scene determines how the movie continues [16, 44]. Attentive reading assistants [18, 38] reason that words that are fixated longer than others are unfamiliar to the user, and provide reading support. The interactive dictionary iDict [18] displays the translation for words that users reading in a foreign language are struggling with. Similarly, the GWGazer Reading Assistant [38], supports children with reading problems by reading out loud words that are difficult for the child.

4.2 Using relevance predictions for new elements

On websites, online shops, and search engines, user-adapted content has been of interest since the early days of the internet [10]. *Recommender systems* thus determine the user's preference through gaze data, and recommend only items that match these preferences. Recommendations may target search engines for documents, images, and videos [12, 13, 26, 27, 35, 46], products in an online shop [6, 8, 14], Google Play Store applications [29], or product advertisements [39]. More recently, gaze data has been used to *validate relevance predictions*. In recommender systems, validating the user's satisfaction with the displayed items allows to refine the recommendation algorithm [5, 41].

5 DISCUSSION

Our review of the various areas where relevance predictions can improve the user experience shows that a strong demand for good prediction models exists. The reported performances of the reviewed models are very encouraging. Yet, since often different metrics are used for the evaluation, comparability of different approaches remains difficult. Comparability is further decreased by the large variety of parameters that can influence the model performance, including different gaze features, algorithms, and study design. The latter is primarily influenced by the tested stimuli, target application, and participants. Few evidence thus exists for the replicability of the reported results in different settings. In order to overcome this issue, we thus propose to report the results of future research using a standardized procedure that allows a direct comparison with previous works:

(1) Clearly define the relevance concept under investigation (see Section 3.2.1)
(2) Identify the most similar previous work(s) as benchmark study and determine what gaze features (Section 3.1) and prediction logic (see Sections 3.2.2-3.4) are used
(3) Replicate the experimental setup and procedure of the benchmark study as closely as possible (see Section 3.5.1)
(4) Run multiple evaluations, altering only one model parameter (as specified in Sections 3.1-3.4) at a time, so that the predictions differ from the benchmark study in only a single aspect
(5) Report the same evaluation metrics that are used in the benchmark study. Additionally, provide unbiased metrics if these are not included in the benchmark study

6 CONCLUSION

In this paper, we reviewed the state-of-the-art approaches to predict relevance of objects from gaze data. The design consideration can be classified into: 1) gaze feature extraction, 2) algorithm selection and definition, 3) prediction scope definition, and 4) elimination of visual distractors. When evaluating the model, special attention should be paid to the procedure for collecting ground truth data, as well as to the evaluation metrics and techniques. Furthermore, we identified five fields where relevance predictions enjoy increasing importance. Of particular interest are recommender systems where relevance predictions can be used to speed up the user's search task.

A remaining challenge is the low comparability of previous work, resulting from the lack of a standardized evaluation procedure. The insight gained from our literature review shall help to establish best practices for the design and evaluation of relevance prediction models, thus allowing for better comparability of future work.

REFERENCES

[1] Roman Bednarik. 2005. Potentials of eye-Movement tracking in adaptive systems. In *4th Workshop on Empirical Evaluation of Adaptive Systems (EAS '05)*. Edinburgh, UK, 1–8.

[2] Ali Borji, Dicky N. Sihite, and Laurent Itti. 2013. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing* 22, 1 (2013), 55–69. https://doi.org/10.1109/TIP.2012.2210727

[3] 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 (2015), 95–112. https://doi.org/10.1016/j.visres.2015.01.010

[4] 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

[5] Li Chen, Feng Wang, and Pearl Pu. 2017. Investigating users' eye movement behavior in critiquing-based recommender systems. *AI Communications* 30, 3-4 (2017), 207–222. https://doi.org/10.3233/AIC-170737

[6] 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

[7] 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

[8] A. B.M. Fahim Shahriar, Mahedee Zaman Moon, Hasan Mahmud, and Kamrul Hasan. 2020. Online product recommendation system by using eye gaze data. In *Proceedings of the International Conference on Computing Advancements* (Dhaka, Bangladesh) *(ICCA '20)*. ACM, New York, NY, USA, Article 61, 7 pages. https://doi.org/10.1145/3377049.3377108

[9] Robert L. Fantz. 1961. The origin of form perception. *Scientific American* 204, 5 (1961), 66–73.

[10] Josef Fink and Alfred Kobsa. 2000. Review and analysis of commercial user modeling servers for personalization on the World Wide Web. *User Modelling and User-Adapted Interaction* 10, 2-3 (2000), 209–249. https://doi.org/10.1023/A:1026597308943

[11] Antón Garcia-Diaz, Xosé R. Fdez-Vidal, Xosé M. Pardo, and Raquel Dosil. 2012. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing* 30, 1 (2012), 51–64. https://doi.org/10.1016/j.imavis.2011.11.007

[12] Daniela Giordano, Isaak Kavasidis, Carmelo Pino, and Concetto Spampinato. 2012. Content based recommender system by using eye gaze data. In *Proceedings of the Eye Tracking Research and Applications Symposium (ETRA '12)*. ACM, Santa Barbara, CA, USA, 369–372. https://doi.org/10.1145/2168556.2168639

[13] David 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*, Marina Meila and Xiaotong Shen (Eds.), Vol. 2. PMLR, San Juan, Puerto Rico, 179–186.

[14] 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 *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

[15] 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 Conference on Human Information Interaction and Retrieval (CHIIR '21)*. ACM, Canberra, ACT, Australia, 129–139. https://doi.org/10.1145/3406522.3446013

[16] Melanie Heck, Janick Edinger, Jonathan Bünemann, and Christian Beck. 2021. The subconscious director: Dynamically personalizing videos using gaze data. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. ACM, College Station, TX, USA, 1–18. https://doi.org/10.1145/3397481.3450679

[17] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. 1997. Image indexing using color correlograms. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '97)*. IEEE, Washington D.C., USA, 762–769. https://doi.org/10.1109/CVPR.1997.609412

[18] Aulikki Hyrskykari, Päivi Majaranta, and Kari-Jouko Räihä. 2003. Proactive response to eye movements. In *Human-Computer Interaction (INTERACT '03)*. IOS Press, Zurich, Switzerland, 129–136.

[19] Aulikki Hyrskykari, Päivi Majaranta, and Kari-jouko Räihä. 2005. From gaze control to attentive interfaces. In *Proceedings of the HCI International (HCII '05)*. Springer, Las Vegas, NV, USA.

[20] Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience* 2, 3 (2001), 194–203. https://doi.org/10.1038/35058500

[21] Robert J. K. Jacob. 1990. What you look at is what you get: Eye movement-based interaction techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People* (Seattle, Washington, USA) *(CHI '90)*. ACM, New York, NY, USA, 11–18. https://doi.org/10.1145/97243.97246

[22] Dietmar Jannach, Lukas Lerche, and Markus Zanker. 2018. Recommending based on implicit feedback. In *Social Information Access*. Springer, Cham, 510–569. https://doi.org/10.1007/978-3-319-90092-6_14

[23] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. *2009 IEEE 12th International Conference on Computer Vision* (2009), 2106–2113. https://doi.org/10.1109/ICCV.2009.5459462

[24] Melih Kandemir and Samuel Kaski. 2012. Learning relevance from natural eye movements in pervasive interfaces. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '12)*. ACM, Santa Monica, CA, USA, 85–92. https://doi.org/10.1145/2388676.2388700

[25] 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 (ETRA '10)*. ACM, 105–108. https://doi.org/10.1145/1743666.1743692

[26] Arto Klami, Craig Saunders, Teófilo E. de Campos, and Samuel Kaski. 2008. Can relevance of images be inferred from eye movements?. In *Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval (MIR '08)*. ACM, Vancouver, British Columbia, Canada, 134–140. https://doi.org/10.1145/1460096.1460120

[27] 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 (ICMI-MLMI '09)*. 305–312. https://doi.org/10.1145/1647314.1647379

[28] Jorma Laaksonen, Markus Koskela, and Erkki Oja. 1999. PicSOM: Self-Organizing Maps for content-based image retrieval. In *International Joint Conference on Neural Networks (IJCNN '99)*. IEEE, Washington D.C., USA, 2470–2473. https://doi.org/10.1109/ijcnn.1999.833459

[29] Yixuan Li, Pingmei Xu, Dmitry Lagun, and Vidhya Navalpakkam. 2017. Towards measuring and inferring user interest from gaze. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 525–533.

[30] Norman H. Mackworth and Anthony J. Morandi. 1967. The gaze selects information details within pictures. *Perception & Psychophysics* 2, 11 (1967), 547–552. https://doi.org/10.3758/BF03210264

[31] Päivi Majaranta and Andreas Bulling. 2014. Eye Tracking and Eye-Based Humanâ€"Computer Interaction. In *Advances in Physiological Computing. Humanâ€"Computer Interaction Series*, S. Fairclough and K. Gilleade (Eds.). Springer, London, Chapter 3, 39–65. https://doi.org/10.1007/978-1-4471-6392-3_3

[32] 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, 221–230. https://doi.org/10.1145/1054972.1055004

[33] Konstantinos Rapantzikos, Nicolas Tsapatsoulis, Yannis Avrithis, and Stefanos D. Kollias. 2007. Bottom-up spatiotemporal visual attention model for video analysis. *IET Image Processing* 1, 2 (June 2007), 237–248. https://doi.org/10.1049/iet-ipr:20060040

[34] 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 (WAPCV '07)*. Springer, Berlin, Heidelberg, 123–140. https://doi.org/10.1007/978-3-540-77343-6_8

[35] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. 2004. *Relevance feedback from eye movements for proactive information retrieval.* Technical Report. Helsinki University of Technology. 1–6 pages. https://doi.org/10.1109/OCEANS.2007.4449348

[36] 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 *Brain Informatics (BI '18)*. Springer, Cham, 282–293. https://doi.org/10.1007/978-3-030-05587-5_27

[37] 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

[38] J. L. Sibert, M. Gokturk, and R. A. Lavine. 2000. The reading assistant: Eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the ACM Symposium on User Interface Software and Technology* (San Diego, CA, USA) *(UIST '00)*. ACM, 101–108.

[39] Hyejin Song and Nammee Moon. 2019. Eye-tracking and Social Behavior Preference-based Recommendation System. *Journal of Supercomputing* 75, 4 (2019), 1990–2006. https://doi.org/10.1007/s11227-018-2447-x

[40] India Starker and Richard A. Bolt. 1990. A gaze-responsive self-disclosing display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People (CHI '90)*. ACM, Seattle, Washington, USA, 3–10. https://doi.org/10.1145/97243.97245

[41] Piotr Sulikowski and Tomasz Zdziebko. 2020. Deep learning-enhanced framework for performance evaluation of a recommending interface with varied recommendation position and intensity based on eye-tracking equipment data processing. *Electronics* 9, 2 (2020), 1–15. https://doi.org/10.3390/electronics9020266

[42] Alistair Sutcliffe and Abdallah Namoun. 2012. Predicting user attention in complex web pages. *Behaviour and Information Technology* 31, 7 (2012), 679–695. https://doi.org/10.1080/0144929X.2012.692101

[43] 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 (2006), 771–789. https://doi.org/10.1016/j.chb.2005.12.012

[44] 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 (2005), 193–204. https://doi.org/10.1080/14626260500476523

[45] 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

[46] 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 (RecSys '08)*. ACM, Lausanne, Switzerland, 83–90. https://doi.org/10.1145/1454008.1454023

[47] Sheng Hua Zhong, Yan Liu, Feifei Ren, Jinghuan Zhang, and Tongwei Ren. 2013. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *Proceedings fo the AAAI Conference on Artificial Intelligence (AIII '13)*. 1063–1069.