UEyes: Understanding Visual Saliency across User Interface Types (Supplementary Materials)

Yue Jiang
yue.jiang@aalto.fi
Aalto University
Finland

Luis A. Leiva
name.surname@uni.lu
University of Luxembourg
Luxembourg

Paul R. B. Houssel
name.surname@uni.lu
University of Luxembourg
Luxembourg

Hamed R. Tavakoli
hamed.rezazadegan_tavakoli@nokia.com
Nokia Technologies
Finland

Julia Kylmälä
julia.kylmala@aalto.fi
Aalto University
Finland

Antti Oulasvirta
antti.oulasvirta@aalto.fi
Aalto University
Finland

1 UEYES DATASET

In the following we describe the key features of our dataset. It includes:

- **Design images:** 495 images of each UI category, 1980 images in total. Images are divided into 55 blocks of 36 images each.
- **Eye-tracking logs:** 554 raw logs from the eye tracker in CSV format. Each log includes eye movement data for one participant and one image block.
- **Image types:** Categorization of each image in a separate CSV file. Each image can belong to only one category.
- **Multi-duration saliency maps:** Saliency maps for 1 s, 3 s, and 7 s of free-viewing. Each fixation is weighted by the time duration of each fixation.
- **Scanpaths:** Sequences of fixations for every participant. Figure 1 shows the distributions of time duration and saccade length.
- **Segmentation information:** A JSON file for each UI with bounding boxes of detected images, texts, and faces. For this, we modified the UIED model [17] by (1) solving a model limitation that ignored text detection and (2) integrating face detection with OpenCV face detection approach using Haar feature-based cascade classifiers [4] into the model.

1.1 Saliency Maps and Scanpaths Examples

Figure 2 shows some examples of saliency maps and scanpaths.

2 VISUAL SALIENCY VS. VISUAL IMPORTANCE

We should note that visual saliency accounts for information about eye movements, while visual importance is captured by proxy data, like mouse movements or manual annotations. Visual importance results are generated from the UMSI model [3]. Although a cursor-based interface can be seen as a proxy for eye-tracking [1, 5, 8, 9], deciding where to move a cursor reflects different cognitive processes from eye movements [16]. As can be observed in Figure 3, visual importance often covers more extensive areas of the UI than a user may have had time to inspect in a relatively short time (up to 7 s in our study). Visual importance captures informative areas such as images and texts, but most of them do not really attract the user's attention.

3 SALIENCY MAP MODEL: UMSI++

The main module of the original UMSI model was trained with KL-divergence [6] and Cross-Correlation [12] losses with coefficients 10 and -3. The output of the UMSI model is the flipped saliency maps requiring postprocessing of black-to-white invert. Our UMSI++ model contains two steps during the training process instead of the flipped ones.

We first trained the model on the Mean Squared Error (MSE) loss between predicted and ground-truth saliency maps for 10 epochs, to encourage the model to approach the ground-truth saliency maps. Next, in addition to the original KL-divergence and Cross-Correlation [12] loss terms used in UMSI, we further incorporated two additional loss terms: Normalized Scanpath Saliency (NSS) and Similarity.

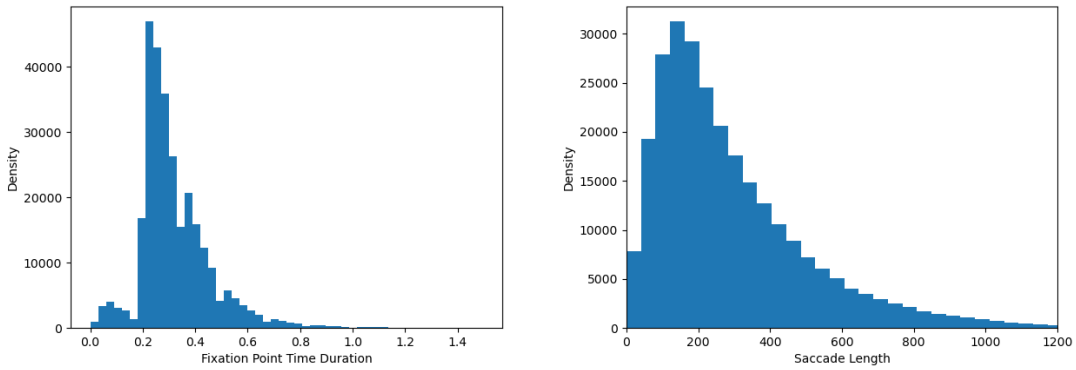
Given the predicted saliency map $\hat{\mathbf{H}}$ and the ground-truth binary map of fixation locations $\mathbf{F}$, the Normalized Scanpath Saliency (NSS) loss term $\mathcal{L}_{\text{NSS}}$ is defined as

$$\mathcal{L}_{\text{NSS}}(\hat{\mathbf{H}}, \mathbf{F}) = \frac{1}{|\mathbf{F}|} \sum_{p} \left(W(\hat{\mathbf{H}}) \circ \mathbf{F}\right)_p, \text{ where } W(\hat{\mathbf{H}}) = \frac{\hat{\mathbf{H}} - \mu(\hat{\mathbf{H}})}{\sigma(\hat{\mathbf{H}}) + \epsilon} \tag{1}$$

where p is the index of pixels and $|\mathbf{F}|$ is the total number of fixation points on the ground-truth binary fixation map $\mathbf{F}$. The symbol $\circ$ is the Hadamard product, which is the element-wise multiplication. The function $W(\cdot)$ is a whitening transformation performing a center-surround operation. ϵ is a regularization term.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) Fixation duration (s) **(b) Fixation saccade length (px)**

Figure 1: Distributions of fixation duration and saccade length.

Furthermore, given the predicted saliency map $\hat{\mathbf{H}}$ and the ground-truth saliency map $\mathbf{H}$, we define the Similarity loss term $\mathcal{L}_{\text{SIM}}$ as

$$\mathcal{L}_{\text{SIM}}(\hat{\mathbf{H}}, \mathbf{H}) = \sum_p \left(min_e(N(\hat{\mathbf{H}}), N(\mathbf{H}))\right)_p$$
$$\text{where } N(\hat{\mathbf{H}}) = \frac{R(\hat{\mathbf{H}})}{\sum_p \hat{\mathbf{H}}_p + \epsilon},\ R(\hat{\mathbf{H}}) = \frac{\hat{\mathbf{H}} - min(\hat{\mathbf{H}})}{max(\hat{\mathbf{H}}) - min(\hat{\mathbf{H}}) + \epsilon} \tag{2}$$

where p is the index of the p-th pixel on the saliency map. The function $min_e(\cdot)$ computes the element-wise minimum values between two saliency maps, while $min(\cdot)$ and $max(\cdot)$ compute the minimum and maximum values on the given saliency map respectively. $N(\cdot)$ is a function normalizing saliency map values in the range of $[0, 1]$, and $R(\cdot)$ rescales the saliency map with the min-max normalization algorithm.

Eventually, we train the model on the loss $\mathcal{L} = 10 \cdot \mathcal{L}_{\text{KL}} - 3 \cdot \mathcal{L}_{\text{CC}} - \mathcal{L}_{\text{SIM}} - 0.5 \cdot \mathcal{L}_{\text{NSS}}$. After training for 10 epochs, the coefficient of the NSS loss term $\mathcal{L}_{\text{NSS}}$ decreases to -5, and we further train the model for another 10 epochs. The coefficients of $\mathcal{L}_{\text{CC}}$, $\mathcal{L}_{\text{SIM}}$, and $\mathcal{L}_{\text{NSS}}$ are negative since they are measurements for saliency map similarity, i.e., larger values mean better saliency map results. For comparison, we also apply the same training pipeline and loss terms to the SAM architecture to get the result of model SAM++. For both models, training each epoch takes about 2 minutes on one NVIDIA GeForce RTX 2080Ti GPU. Thus, the entire training process takes about 1 hour on our UEyes dataset.

4 EVALUATION METRICS FOR SALIENCY MAP MODELS

In the following, we explain six widely used saliency evaluation metrics.

4.0.1 Area under ROC Curve (AUC). Area under ROC Curve (AUC) is the most commonly used evaluation metric for measuring saliency map performance. It evaluates saliency as a binary classifier of fixation points at various thresholds. The Receiver Operating Characteristic Curve (ROC Curve) is a curve showing the rates of the true positives and false positives at different discrimination thresholds. The Area Under the ROC curve (AUC) measures the true and false positive rates under such a binary classifier. AUC-Judd [2, 7] is a variation of AUC where the true positive rate is defined as the ratio of the true positive points in relation to the number of ground-truth fixation points above various threshold values, and the false positive rate is the ratio of false positive points in relation to the total number of non-fixated pixels. Therefore, given the predicted saliency heatmap map $\hat{\mathbf{H}}$ and the ground-truth fixation map $\mathbf{F}$, the AUC-Judd evaluation metric is defined as

$$\mathcal{L}_{\text{AUC-Judd}}(\hat{\mathbf{H}}, \mathbf{F}) = \int_t \text{ROC}\left(TP_t(\hat{\mathbf{H}}, \mathbf{F}), FP_t(\hat{\mathbf{H}}, \mathbf{F})\right)$$
$$\text{where } TP_t(\hat{\mathbf{H}}, \mathbf{H}) = \frac{|(\hat{\mathbf{H}} \circ F) \geq t|}{|\mathbf{F}|},\ FP_t(\hat{\mathbf{H}}, \mathbf{H}) = \frac{|(\hat{\mathbf{H}} \circ F) < t|}{|\mathbf{1} - \mathbf{F}|} \tag{3}$$

where $TF_t(\cdot)$ and $FP_t(\cdot)$ represent the true positive rate and false positive rate at the threshold value t respectively. $|\mathbf{F}|$ is the number of points with the value 1, which is the number of fixation points, while $|\mathbf{1} - \mathbf{F}|$ is the number of non-fixation points. The symbol $\circ$ is the Hadamard product, which is the element-wise multiplication. Thus, $|(\hat{\mathbf{H}} \circ F) \geq t|$ shows the number of positive points on the predicted heatmap $\hat{\mathbf{H}}$ above the threshold value t. In practice, the threshold value t can be selected as the set of unique saliency values on the predicted heatmap $\hat{\mathbf{H}}$.

4.0.2 Normalized Scanpath Saliency (NSS). Normalized Scanpath Saliency (NSS) [13] is the average normalized saliency at fixation points. The detailed definition has been shown in Equation 1. NSS is more sensitive to detecting false positive points than the AUC (although it may still be high with many false positives given a large number of true positives, since a small number of false positives does not affect the AUC value. However, all the false positives decrease the NSS (in other words, NSS penalizes false positives).

4.0.3 Information Gain (IG). Information Gain (IG) [10, 11] is used for measuring saliency results beyond systematic bias. Given the predicted saliency heatmap $\hat{\mathbf{H}}$ and the ground-truth heatmap $\mathbf{H}$ and

Figure 2: Examples of saliency maps and scanpaths in the UEyes dataset.

fixation map **F**, IG is defined as

$$\mathcal{L}_{\text{IG}}(\hat{\mathbf{H}}, \mathbf{H}) = \frac{1}{|F|} \sum_{p} \left(\log_2(N(\hat{\mathbf{H}}) \circ \mathbf{F} + \epsilon) - \log_2(N(\mathbf{B}) \circ \mathbf{F} + \epsilon) \right)_p$$

$$\text{where } N(\hat{\mathbf{H}}) = \frac{R(\hat{\mathbf{H}})}{\sum_p \hat{\mathbf{H}}_p + \epsilon}, \; R(\hat{\mathbf{H}}) = \frac{\hat{\mathbf{H}} - min(\hat{\mathbf{H}})}{max(\hat{\mathbf{H}}) - min(\hat{\mathbf{H}}) + \epsilon} \tag{4}$$

where p is the index of the p-th pixel on the heatmap and $|\mathbf{F}|$ is the total number of fixation points on the ground-truth binary fixation map **F**. The baseline map **B** is the systematic bias. In practice, it can

Figure 3: Examples of visual importance and visual saliency.

be e.g. a zero map or a center bias prior map [10, 11]. The function $N(\cdot)$ is a function normalizing values in the range of $[0, 1]$. ϵ is for regularization. The information gain evaluation metric demonstrates the average information gain of the normalized saliency heatmap over the normalized baseline bias map at the ground-truth fixation points.

4.0.4 Similarity (SIM). The similarity (SIM) [14, 15] metric measures the intersection between the predicted heatmap and the ground-truth heatmap, indicating the overlapping of the two heatmaps. It is defined as the sum of the minimum value of the normalized predicted heatmap and the normalized ground-truth heatmap. The detailed definition has been shown in Equation 2. SIM is lower for sparse heatmap maps and very sensitive to failed detection of saliency points, since missing saliency values would lead to zero similarity, thus reducing the similarity score.

4.0.5 Pearson's Correlation Coefficient (CC). Pearson's Correlation Coefficient (CC) [12] is a measurement for evaluating the correlation or dependence between the predicted saliency heatmap and the ground-truth heatmap. Given the predicted saliency heatmap $\hat{\mathbf{H}}$ and the ground-truth heatmap $\mathbf{H}$, CC is defined as

$$\mathcal{L}_{\text{CC}}(\hat{\mathbf{H}}, \mathbf{H}) = \sigma(W(\hat{\mathbf{H}}), W(\mathbf{H})), \text{ where } W(\hat{\mathbf{H}}) = \frac{\hat{\mathbf{H}} - \mu(\hat{\mathbf{H}})}{\sigma(\hat{\mathbf{H}}) + \epsilon} \quad (5)$$

Metric	Sensitive to FN	Sensitive to FP	Measurement	Metric Category
AUC-Judd	+	-	Similarity	Location-based
NSS	+	+	Similarity	Location-based
InfoGain	++		Similarity	Location-based
Similarity	++		Similarity	Distribution-based
CC	+	+	Similarity	Distribution-based
KL	++		Dissimilarity	Distribution-based

Table 1: Different saliency prediction evaluation metrics are sensitive to different errors, e.g, false positives (FP) or false negatives (FN). Here '+' means that the metric is sensitive to the error, and '++' means significant sensitivity, while '-' shows the metric ignores the error. Metrics measuring similarity have higher values, while the one measuring dissimilarity has lower values for better prediction models. Location-based evaluation metrics focus on fixation points, while distribution-based ones focus on continuous distributions of the saliency heatmaps.

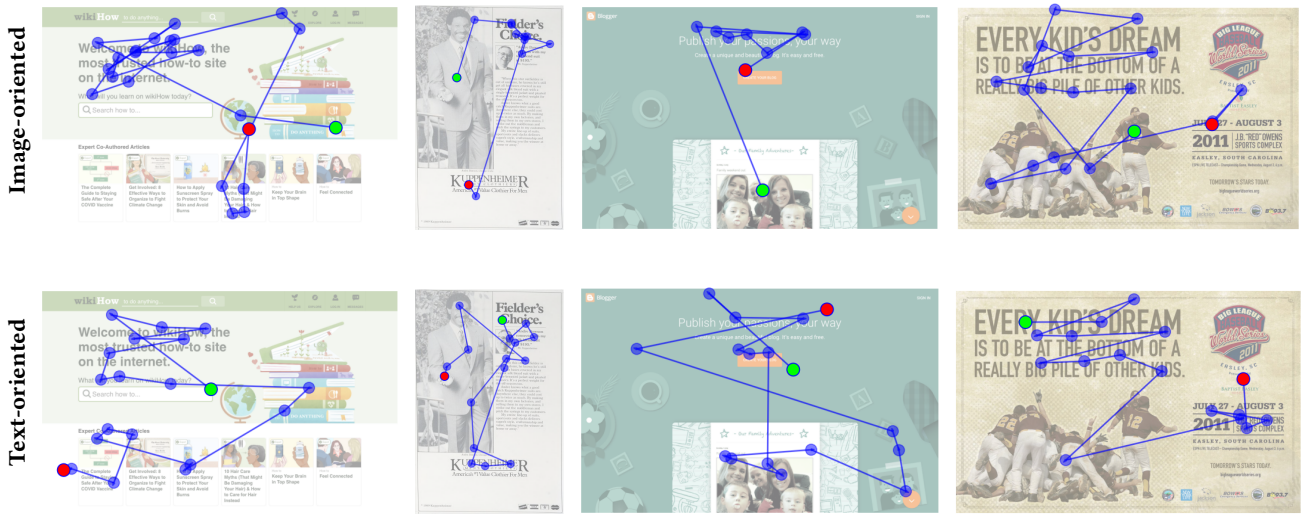


Figure 4: Different users have different viewing strategies on user interfaces. Image-oriented users often look at images before text, while text-oriented users have the opposite preference.

where the function $\sigma(\cdot,\cdot)$ is the computation of covariance and $W(\cdot)$ is a whitening transformation performing a a center-surround operation. ϵ is for regularization. The CC metric is sensitive to both false positive and false negative points. It is symmetric and thus cannot distinguish between false positives and false negatives.

4.0.6 Kullback-Leibler Divergence (KL). Unlike the above-mentioned saliency evaluation metrics, which evaluate how close they are to some ground-truth value, the Kullback-Leibler Divergence (KL) [6] metric measures the difference between the distributions of the saliency prediction and the ground-truth. A lower KL score indicates a better estimation of the saliency map. The KL metric significantly penalizes false negatives, especially when the prediction is close to zero for salient areas.

4.0.7 Metric Properties. We describe the properties of the saliency evaluation metrics regarding their sensitivity to false positives or false negatives, similarity (or dissimilarity) measurement, and metric categories (Table 1).

(1) Sensitivity: All metrics are sensitive to false negatives. KL, IG, and SIM penalize significantly false negatives, especially when the predicted values are close to zero. The normalization step of NSS increases the penalty for detecting false positives and thus makes it more sensitive to false positives than other metrics. CC is a symmetric metric according to its definition, so it has equal sensitivity to false positives and false negatives. The AUC score is insensitive to false positives.
(2) Measurement: KL is dissimilarity measure, while the other metrics similarity measures. Thus, better saliency prediction models have lower scores for KL and higher for any of the other metrics.
(3) Metric Category: Location-based metrics (AUC, NSS, InfoGain) evaluate models based on fixation points, whereas

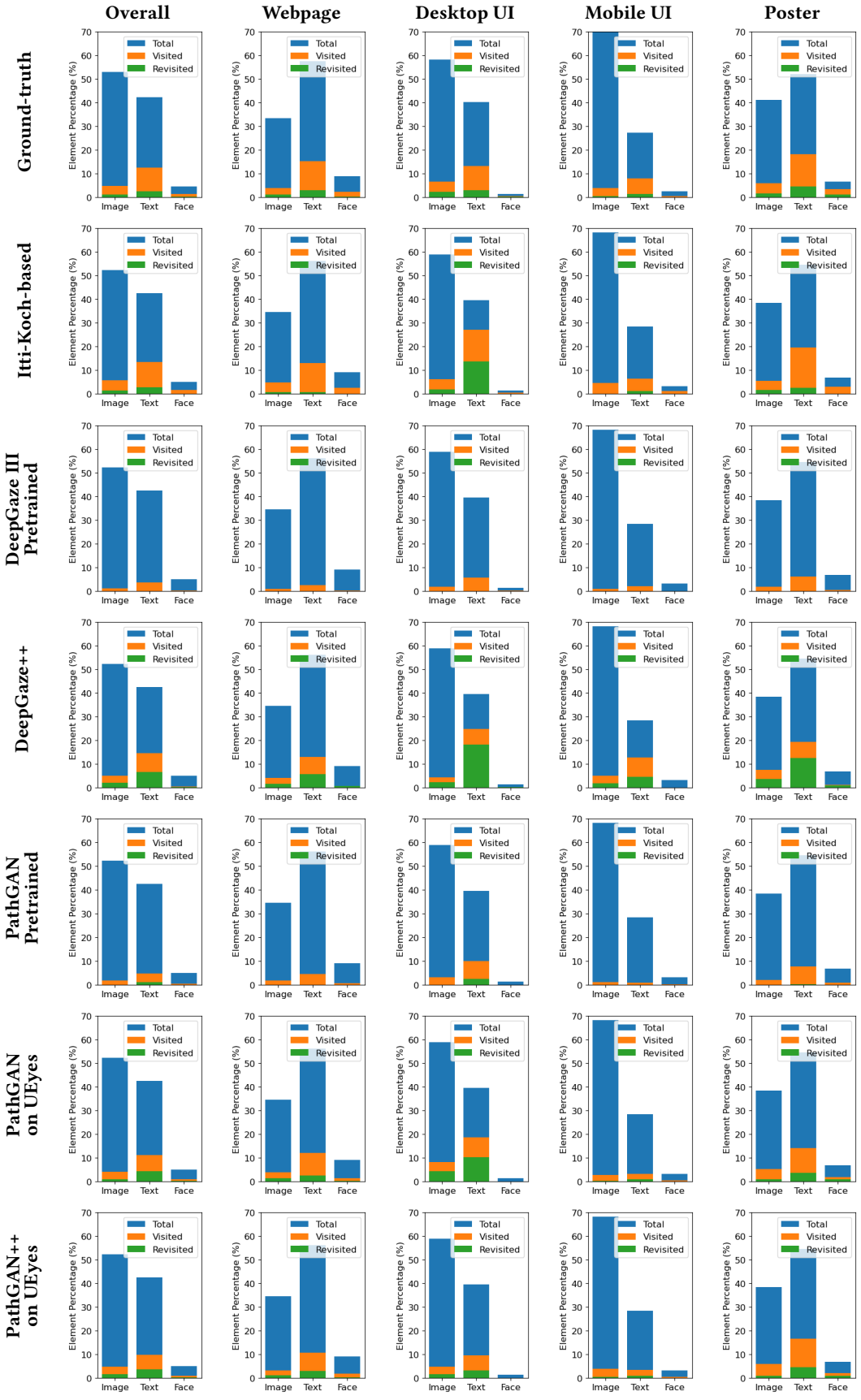


Figure 5: Visited and revisited elements comparison for the scanpath predictive models.

distribution-based metrics (SIM, CC, KL) evaluate models based on saliency heatmaps, as a continuous distribution.

5 INDIVIDUAL VIEWING STRATEGIES

We asked participants to self-report their viewing strategies after the study. Some participants prefer to start by looking at images, while others prefer to look at text or titles first, as shown in Figure 4. When asked to describe the strategies they used to look at these UIs, some participants indicated that they prefer to see "colorful images and images of people or animals, which caught my attention". In contrast, some others "tried to look at images rather than read the information".

To get the overall idea of the UI, some participants indicated that they first checked "the whole picture" and then focused on "the most interesting parts". Some participants looked for titles "to grab the general ideas" and utilized pictures "to understand the content better". Others focused on the center of the page first "if the layout is center-aligned." and then scanned from top to bottom and from left to right. These observations can guide future efforts in understanding individual strategies across UI types and ultimately may generate more personalized predictive models.

6 VISITED VS. REVISITED ELEMENTS OF SCANPATH MODELS

Figure 5 shows the visited and revisited element ratio for the scanpath predictive models.

REFERENCES

[1] Roman Bednarik and Markku Tukiainen. 2007. Validating the restricted focus viewer: A study using eye-movement tracking. *Behavior research methods* 39, 2 (2007).

[2] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. 2015. MIT saliency benchmark.

[3] Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii. 2020. Predicting visual importance across graphic design types. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 249–260.

[4] Kruti Goyal, Kartikey Agarwal, and Rishi Kumar. 2017. Face detection and tracking: Using OpenCV. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Vol. 1. IEEE, 474–478.

[5] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. SALICON: Saliency in Context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1072–1080. https://doi.org/10.1109/CVPR.2015.7298710

[6] James M Joyce. 2011. Kullback-leibler divergence. In *International encyclopedia of statistical science*. Springer, 720–722.

[7] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*. IEEE, 2106–2113.

[8] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Frẽdo Durand, and Hanspeter Pfister. 2017. BubbleView: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017). https://doi.org/10.1145/3131275

[9] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Aude Oliva, Krzysztof Z Gajos, and Hanspeter Pfister. 2015. A crowdsourced alternative to eye-tracking for visualization understanding. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 1349–1354.

[10] Matthias Kümmerer, Thomas Wallis, and Matthias Bethge. 2014. How close are we to understanding image-based saliency? *arXiv preprint arXiv:1409.7686* (2014).

[11] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. 2015. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences* 112, 52 (2015), 16054–16059.

[12] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. 2007. Predicting visual fixations on video based on low-level visual features. *Vision research* 47, 19 (2007), 2483–2498.

[13] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. 2005. Components of bottom-up gaze allocation in natural images. *Vision research* 45, 18 (2005), 2397–2416.

[14] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40, 2 (2000), 99–121.

[15] Michael J Swain and Dana H Ballard. 1991. Color indexing. *International journal of computer vision* 7, 1 (1991), 11–32.

[16] Hamed R. Tavakoli, Fawad Ahmed, Ali Borji, and Jorma Laaksonen. 2017. Saliency Revisited: Analysis of Mouse Movements Versus Fixations. In *Proc. CVPR*.

[17] Mulong Xie, Sidong Feng, Zhenchang Xing, Jieshan Chen, and Chunyang Chen. 2020. UIED: A Hybrid Tool for GUI Element Detection. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Virtual Event, USA) *(ESEC/FSE 2020)*. Association for Computing Machinery, New York, NY, USA, 1655–1659. https://doi.org/10.1145/3368089.3417940