Overlooking: The nature of gaze behavior and anomaly detection in expert dentists

Nora Castner
Perception Engineering, University of Tübingen
Tübingen, Germany
castnern@informatik.uni-tuebingen.de

Solveig Klepper
Computer Science Institute, University of Tübingen
Tübingen, Germany
solveig.klepper@student.uni-tuebingen.de

Lena Kopnarski
Computer Science Institute, University of Tübingen
Tübingen, Germany
lena.kopnarski@student.uni-tuebingen.de

Fabian Hüttig*
University Hospital Tübingen
Tübingen, Germany
fabian.huettig@med.uni-tuebingen.de

Constanze Keutel†
University Hospital Tübingen
Tübingen, Germany
constanze.keutel@med.uni-tuebingen.de

Katharina Scheiter
Leibniz-Institut für Wissensmedien
Tübingen, Germany
k.scheiter@iwm-tuebingen.de

Juliane Richter
Leibniz-Institut für Wissensmedien
Tübingen, Germany
j.richter@iwm-tuebingen.de

Thérése Eder
Leibniz-Institut für Wissensmedien
Tübingen, Germany
tf.eder@iwm-tuebingen.de

Enkelejda Kasneci
Perception Engineering, University of Tübingen
Tübingen, Germany
enkelejda.kasneci@uni-tuebingen.de

ABSTRACT

The cognitive processes that underly expert decision making in medical image interpretation are crucial to the understanding of what constitutes optimal performance. Often, if an anomaly goes undetected, the exact nature of the false negative is not fully understood. This work looks at 24 experts' performance (true positives and false negatives) during an anomaly detection task for 13 images and the corresponding gaze behavior. By using a drawing and an eye-tracking experimental paradigm, we compared expert target anomaly detection in orthopantomographs (OPTs) against their own gaze behavior. We found there was a relationship between the number of anomalies detected and the anomalies looked at. However, roughly 70% of anomalies that were not explicitly marked in the drawing paradigm were looked at. Therefore, we looked how often an anomaly was glanced at. We found that when not explicitly marked, target anomalies were more often glanced at once or twice. In contrast, when targets were marked, the number of glances was higher. Furthermore, since this behavior was not similar over all images, we attribute these differences to image complexity.

*Department of Prosthodontics
†Department of Radiology, Center of Dentistry, Oral Medicine and Maxillofacial Surgery

MCPMD'18 , October 16, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-6072-2/18/10…$15.00
https://doi.org/10.1145/3279810.3279845

CCS CONCEPTS

• **Applied computing** → **Psychology**; **Education**; • **Human-centered computing** → *Interactive systems and tools*; Visualization design and evaluation methods;

KEYWORDS

Remote Eye Tracking, Medical image interpretation, Cognitive Modelling, Expertise

ACM Reference Format:
Nora Castner, Solveig Klepper, Lena Kopnarski, Fabian Hüttig, Constanze Keutel, Katharina Scheiter, Juliane Richter, Thérése Eder, and Enkelejda Kasneci. 2018. Overlooking: The nature of gaze behavior and anomaly detection in expert dentists. In *Workshop on Modeling Cognitive Processes from Multimodal Data (MCPMD'18), October 16, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3279810.3279845

1 INTRODUCTION

Expertise in any domain is what many strive for. It is known that these skills are established through practice. Yet, there are still mechanisms that are not fully understood. Mainly, how experts process their visual input such that their domain knowledge is effectively applied.

In general, experts are not easily available due to time and work constraints. Therefore, the majority of the literature measures expertise with small samples of experts. Such small *caches* can lead to an insufficient understanding of expertise. In the literature review from Gegenfurtner et al., [4], across all expertise domains evaluated, mean expert sample sizes ranged from six to 17 experts; with the medical profession having approximately eight experts. More recently, van der Gijp et al. [10] provided a similar review that focused solely on radiology. Of the 26 studies evaluated in the meta-analysis,

only two studies were able to acquire more than 15 experts (e.g. sub-specialized experts, radiologists, or other medical specialists). Both literature reviews offer a comprehensive understanding of experts' scanning behavior in addition to performance compared to novices. However, the interplay of cognitive mechanisms that distinguish acceptable task performance is still uncertain. In medical image processing, such as radiology, an important research question is related to the reasons why an anomaly would be overlooked.

1.1 Previous Literature

As in many fields, experts in medical fields exhibit more optimal performance. However, optimal performance may or may not always be one hundred percent accurate. Often, it is a tradeoff of detecting what is most necessary with regard to a patient's health and understanding the costs. Diniz and colleagues [2] looked at the accuracy of cavity detection in OPTs for dentists with 5 to 7 years experience (10) versus students in the final semester of dental studies (10). The authors reported that the experts had a trade-off of low sensitivity to high specificity compared to the advanced students, who had high sensitivity and low specificity. They attributed their findings to the idea that more experienced dentists may overlook some cavities and focus on the more detrimental ones [2]. Employing this strategy, the more experienced dentists avoid overtreating or extensive restoration processes that are costly and may leave a patient susceptible to complications.

Filtering of non pertinent information is also crucial to effective medical image interpretation. Mallet et al., 2014 [7] measured eye movements of 65 radiologists and divided them into experienced CT colonography scan readers (27) and radiologists inexperienced in the same task (38). They found that the experienced radiologists were overall more accurate in identifying polyps in a 3D CT scan and had shorter time to first fixation on polyps. However, the time to interpret the polyps accurately was not distinguishable between the experienced and inexperienced readers [7]. Thus, experienced readers may recognize and search the polyp-prone areas more quickly, but they process and interpret the area of interest similar to radiologist inexperienced in CT scan reading.

Additionally, Drew et al., 2013 [3] had 24 expert radiologists searching 3D CT Lung scans to detect as many nodules as possible in three minutes. They were instructed to scroll through a stack of 2D image slices, and click where they found nodules. Two predominant search strategies were observed: Scanning, or searching each slice in a left to right reading fashion, and drilling, or searching multiple slices top to bottom. They found the 'drillers' had a significant increase in true positives, though no difference in false positives. Also, drillers' scanning behavior covered a larger area of the lung. When looking at the false negatives, the scanners had more search errors (not looking at the nodule areas) and drillers had higher recognition errors. Meaning, they often glanced at a nodule, but not long enough to indicate an error in their interpretation.

To our knowledge only one study has focused on radiological image interpretation (orthopantomographs, or OPTs) in the dental context. Turgeon & Lamm (2016) [9] compared 15 certified oral and maxillofacial radiologists (OMRs) to 30 fourth year dental students. Performance was not measured; however, they compared students to experts' eye gaze on subtle and non subtle anomalies in the OPTs. They found that eye movement behavior was different between experts and novices. More interesting, experts had longer total time and more fixations in areas of interest when the images had more subtle anomalies. Whether these eye movement behaviors are indicative of accurate detection is of interest to this work. We aim to look into the correlation of gaze on anomalies of interest to the actual detection of anomalies of interest. Additionally, if gaze behavior can also indicate recognition or interpretation errors is of interest.

We aim to further explore the relationship between gaze and anomaly detection in medical image interpretation. Specifically, whether accurate glances correlate to an accurate anomaly detection. Additionally, whether search or interpretation errors can be measured by the number of glances; where a higher number of glances on an anomaly that was not determined as such may be indicative of an interpretation error.

By incorporating a drawing paradigm into the current study, we are able to create comprehensive expert ground truth performance data. Then, by comparing the gaze data, we can further explore the cognitive process that underly expertise in this domain.

2 METHODOLOGY

2.1 Participants

26 dentists (13 female, years experience: $M = 10.46, SD = 11.26$) at the university hospital clinic participated in the current study. 46% of the participants see less than 10 patients per day and 54% see between 11 and 30 patients per day. Due to technical issues with the eye tracker, gaze data for two participants was not available, though their data for the drawing portions of the experiment was still recorded. Therefore, gaze data was available for 24 participants.

2.2 Eye Tracker

The eye tracker used was the SMI RED250 (Sensoric Motor Instruments, Germany) running at 250Hz. A 9-point calibration plus 4-point validation was performed prior to presentation. The experimental setup, including eye tracker and calibration, and design are similar to the one found in the study by Castner et al. [1], where subjects view OPT stimuli and are asked to mark where they detect anomalies. Our study employs the same structure, although we are measuring expert dentists working in a clinic and not dentistry students as in [1].

Fixations for the left eye were calculated using I-VT [8]: using a 40°/s velocity threshold and 50ms for minimum fixation duration. Where gaze points are considered one fixation if the point to point velocity is too slow (below the threshold) to be indicative of a rapid eye movement, or saccade, to another location.

Eye movement data for an image was removed if the tracking ratio was below 75%. This pruning was performed to control for any systematic offsets that could have potentially arose from head movements, and in turn would affect accuracy of the gaze points.

2.3 Data

2.3.1 Gaze and Drawing Protocol. The protocol consisted of one set of 15 OPTs with anomalies of varying difficulty and subtlety: Two images were negative controls with no anomalies. Similar to the protocol in [1], each OPT was viewed for an exploration phase,

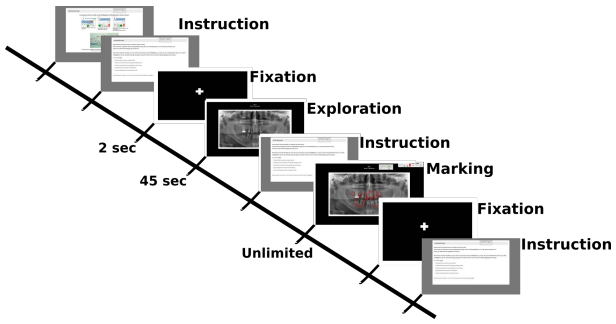

Figure 1: Experimental Session Protocol. The protocol comprised of a calibration, introduction, and instruction, then for the 15 OPTs, a fixation, exploration, and drawing. Image borrowed from [1].

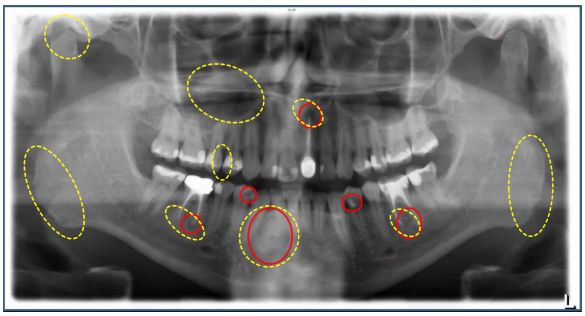

Figure 2: Drawing example. Drawings from a participant (Red) with predefined anomalies (Dotted Yellow), or targets, overlayed. In this example, the participant would have four hits and five misses and two false positives.

which was 45 seconds in duration, and again for a marking phase, which was unlimited in duration. Anomalies detected in the exploration phase[1] were then marked by drawing a red circle on-screen in a click-and-drag fashion. The instruction for the exploration phase was only to inspect the image for pathologies within the 45 seconds: Then, in the marking phase, only to mark the anomaly areas that were found in the exploration phase. Figure 1 illustrates the experimental protocol.

In addition to the gaze data, another interesting aspect is the participants' ability to detect anomalies. By employing an on-screen drawing phase, we were able to measure which areas participants determined as necessary for treatment.

Drawings obtained from the marking phase were compared to predefined anomalies determined for each image; Images had anywhere from four to fourteen anomalies. Participants' indication of an anomaly by marking it were hand-coded by trained evaluators in order to determine if the drawing matched that of the specific target anomaly. A correct mark on an anomaly was determined if the drawn circle overlapped or was within the predefined anomaly by the evaluators[2]. For simplicity, we will refer to the predefined

[1]e.g. Periodontal disease, cavities, insufficient fillings and abscesses, not including sufficient fillings, missing teeth needing no further treatment, or prosthetics.
[2]Inter-rater reliability: .94 and .934.

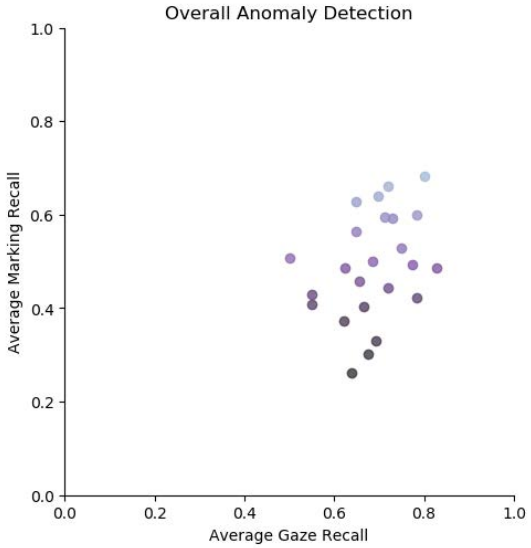

Figure 3: Relationship between overall gaze recall and marking recall. The lighter hues are indicative of higher marking recall.

anomalies as *targets* and the correct detection from a participant or participants as a *marked hit.*

Regarding targets and gaze, if the coordinates of a fixation were within or on the border of a target, it was considered a glance hit. Additionally, we measured how often glances were for per target.

2.3.2 Recall and Precision. In the following, we report the performance in terms of recall and precision. Recall (also known as sensitivity or true positive rate) is the number of true positives over the total of true positives and false negatives. Thus, if an image has a total of eight predefined anomalies and a participant finds six of the anomalies, meaning six true positives and two false negatives, the subject has a recall of 75%. The false negative rate, or miss rate, is the complement of the recall, being the number of false negatives over the total of false negatives and true positives. For the current example, the false negative rate would be 25%.

Precision is the true positives over the total of true positives and false positives. Though, the focus of this work is more on the recall, precision and recall affect the harmonic mean (F1 score). For the example shown in Figure 2, we have a recall of 50% and a precision of roughly 67% (four true positives and two false positives).

3 RESULTS

3.1 Recall

For the participants, marking recall averaged over all images ranged from 26% to 68%: $M = 49.99\%, SD = 11.12\%$ ($n = 26$). However, given that some of the images may have been more complex or harder to determine, this likely affected the overall recall rate per person. Considering each image separately, marking recall per person could be as high as 96% or even 0%. In addition, overall precision ranged from 53.85% to 96.43%; the mean F-Score was 60.89% (SD: 8.65%).

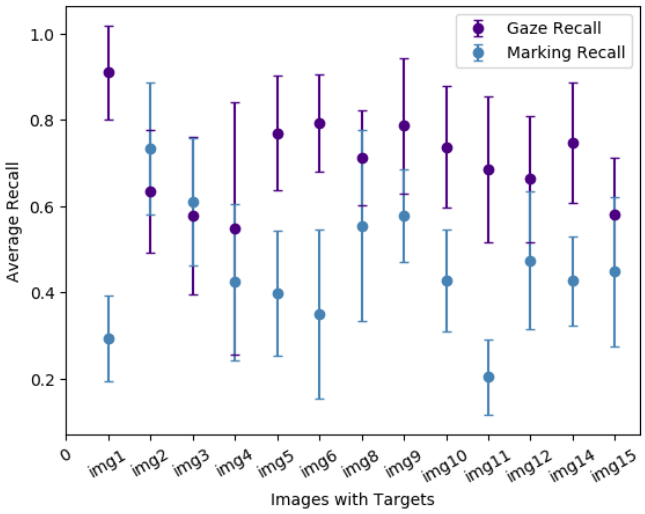

Figure 4: Frequency of Glances for marked and unmarked targets

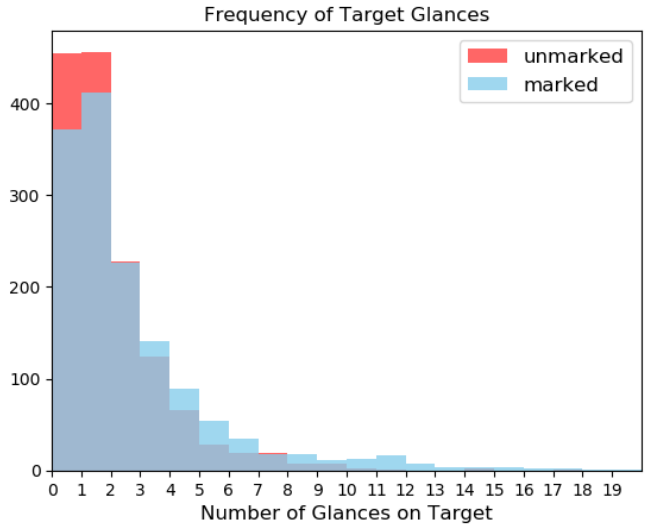

Figure 5: Frequency of glances per target for marked and unmarked targets for all images as depicted by the overlapping distributions for marked targets (blue bars) and unmarked targets (red bars). The frequencies when number of glances per target is 3 or more is overall higher for when the target was marked in contrast to when the target was not marked.

We measured the average gaze recall over all images for each participant; Where one or more glances on a target are considered a gaze hit and no glances on a target are considered a gaze miss. Gaze recall ranged from 50% to 83%: $M = 69.82\%, SD = 8.44\%$ ($n = 24$). Figure 5 shows the relationship between the gaze recall to the marking recall, where there is a slight positive correlation: $r = 0.33, p = 0.11$. Figure 4 shows the gaze and marking behavior on an image level. Once again, for image two and three there was a tendency toward extra searching within the marking phase as shown by the gaze recall being lower than the marking recall.

Table 1 shows the true positive and false negatives for all targets for all images for both gaze and marking data. Interestingly enough, there is a portion of instances where targets were marked even if no gaze was measured for those targets. This behavior could be attributed to extra searching in the marking phase of the experimental protocol, though participants were advised not to.

Table 1: Gaze and Marking Data: Absolute & (Percent) Values

Condition	Marked Target	Missed Target	Total
Gaze on Target	1067 (37.41%)	960 (33.66%)	2027 (71.07%)
No Gaze on Target	371 (13%)	454 (15.91%)	825 (28.93%)
Total	1438 (50.42%)	1414 (49.56%)	2852 (100%)

A chi-square test of independence was performed to examine the association between gaze recall and marking recall. The association between these variables was highly significant, $X^2\ (1, N = 2852) = 13.49, p < 0.01$.

More interesting, when we look at the gaze behavior per target the number of glances per target was significantly higher ($M = 2.34, SD = 3.25$) when the target was marked than when not marked ($M = 1.51, SD = 1.82$), $t\ (2850) = 8.35, p < 0.001$. Considering targets were not marked 49.56% of the time, zero glances on a target could be indicative of ineffective searching of the image. Whereas, when there are glances per target for the case target missed, this behavior could be indicative of an analysis error: Where a low number of glances on a target could indicate an error in recognition, and a high number of glances could indicate an error in interpretation.

3.2 Glance Frequency

The frequency of glances per target as seen in Figure 5 shows that for unmarked targets, there is a higher frequency for zero glances or one glance on a target. For marked targets, there is also a trend to glance once on a target. However, when there are three or more glances per target, there is a switch in the marking behavior, where the frequency is higher for targets marked compared to targets unmarked.

Due to the variability of the targets in the images, marking recall per image varied greatly, as seen in Figure 4. In particular, for image nine (see Figure 6), the average gaze recall is 80%. The number of glances per target for this image shows a distinction between glance behavior for target marked or target unmarked. Here, there are higher frequencies for glancing at a target three or more times when the target was marked, while when the target was not marked, there are higher frequencies for glancing at a target one or zero times. Overall, there was a higher true positive rate for target marking, which is also apparent in the gaze behavior.

Another example of different behavior respective of image is shown in Figure 7. Here, the gaze indicates a relatively high number of targets detected as false negatives were glanced at 2 or more times. Especially, when the number of glances increases to five, the frequencies are higher for targets unmarked in contrast to targets

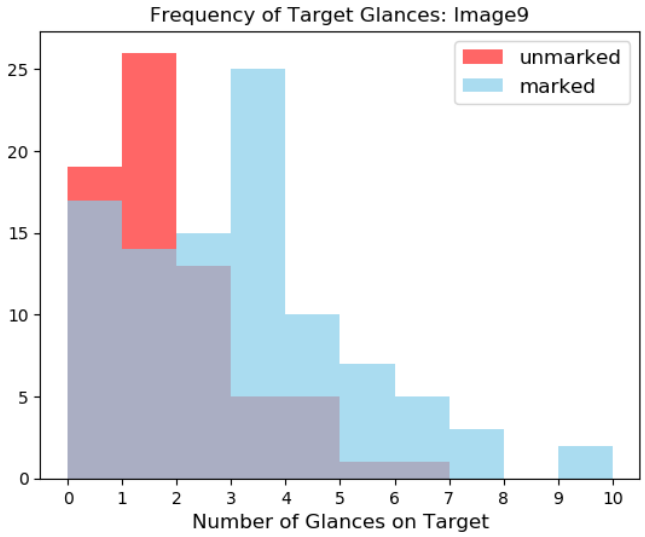

Figure 6: Histogram of number of glances per target for when the target was marked or missed. For this image in particular, the number of glances on a target was higher when the target was marked in contrast to when the target was not marked.

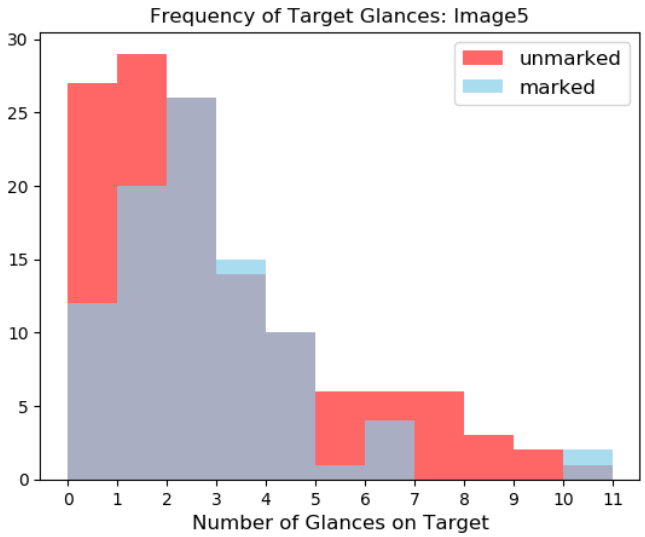

Figure 7: Example of an image where there is a high amount of false negatives in marking although there is a high frequency of higher glances per target.

marked. Thus, targets looked at often were possibly interpreted as not being an anomaly. This glance behavior could be indicative of interpretation errors.

4 DISCUSSION

For detecting anomalies, the sample of expert dentists we tested found roughly 50% of the target anomalies, though their performance varied over the images. The recall rates we found are roughly similar to those in the study by Diniz et al. [2], where the mean recall from the expert dentists was between 20 and 40%, depending on the nature of the anomaly. They attributed the experts' detection behavior to 'overlooking' anomalies where the cost (i.e. treatment cost) of detecting the anomaly as such would outweigh any long term benefit. One possible explanation for the recall of the experts in our study could be the nature of the experiment. They were instructed to mark only the anomalies they detected in the exploration phase and not mark anomalies detected additionally during the marking phase. Although, we could not control for additional searching, if the subjects adhered to this instruction, naturally recall would be lower than real world conditions where they may have unlimited time to inspect an OPT.

However, the allusion of 'overlooking' is apparent. We found there was a slight relationship ($r = 0.33$) between gaze on target anomalies and the detection of target anomalies. Although, more interesting was that gaze recall, or the rate of whether an anomaly target was glanced at, was overall higher than the recall of marking the anomalies. High sensitivity to looking at anomaly areas can be indicative of effective searching of the image and all possible areas where pathologies reside. Thus, experts often looked at an anomaly area, although they marked it roughly at chance level (50.42%).

It is known that experts often have more effective search strategies, where they fixate more often on relevant areas compared to their novice counterparts, and that experts are also better at detecting anomalies [9, 10]. However, when an expert does not mark an anomaly when he or she has seen it, which mechanisms determine that cognitive decision? Kundel et al. [6] proposed three types of decision errors. Based on the fixation duration, a false negative could be classified as either a search error (no fixation on target), a recognition error (short fixation duration on target), or a decision error (long fixation duration on target).

Fixation duration can be applied to distinguish different errors. However, we successfully applied the number of glances for determining the cognitive mechanisms behind false negatives. For experts, we found very few occurrences that could be similarly classified as a search error. Roughly 30% of targets missed were due to no gaze on the target, meaning an anomaly was not detected because it was not looked at. Similarly, a recognition error could be distinguished as glancing once or twice on the anomaly, where an expert may look over an anomaly and determine it is not worth further scrutiny. Whereas, a decision error may be characterized by more glances to the area. This high number of glances could indicate, that more cognitive processing may be involved for determining the nature of the anomaly.

Overall, when an anomaly was not detected as such, there were higher frequencies of one or two glances on the anomaly. Therefore, it is possible these were recognition errors. Decision errors were overall less frequent, where generally if an anomaly was looked at three or more times, it was more likely to be explicitly determined as such. However, this was not the case when we looked at each image separately. There were some images where unmarked

anomalies had high frequencies for three or more glances on an anomaly. The exact nature of how obvious or subtle anomalies were per image was out of the scope of this paper. However, future work could employ expert glance behavior as a predictor of how easy or hard an anomaly is to accurately detect. Furthermore, the scanpath, or order that the anomalies were fixated on, can offer insight into patterns indicative expert search behavior and is of great interest to our future research. In our future work we will therefore employ advanced algorithms for scanpath analysis (e.g., Subsmatch [5]) to relate expertise with performance. This understanding of the cognitive processes involved in effective medical image interpretation as illustrated by the gaze behavior can offer expert insight toward teaching effective decision making in novices.

REFERENCES

[1] Nora Castner, Enkelejda Kasneci, Thomas Kübler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, and Constanze Keutel. 2018. Scanpath comparison in medical image reading skills of dental students: distinguishing stages of expertise development. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ACM, 39.

[2] Michele B Diniz, Jonas A Rodrigues, Klaus Neuhaus, Rita C.L. Cordeiro, and Adrian Lussi. 2008. Influence of Examiners' Clinical Experience on the Reproducibility and Validity of Radiographic Examination in Detecting Occlusal Caries. *Caries research* 42 (2008), 227.

[3] Trafton Drew, Melissa Le-Hoa Vo, Alex Olwal, Francine Jacobson, Steven E Seltzer, and Jeremy M Wolfe. 2013. Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of vision* 13, 10 (2013), 3–3.

[4] Andreas Gegenfurtner, Erno Lehtinen, and Roger Säljö. 2011. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review* 23, 4 (2011), 523–552.

[5] Thomas C Kübler, Colleen Rothe, Ulrich Schiefer, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017. SubsMatch 2.0: Scanpath comparison and classification based on subsequence frequencies. *Behavior research methods* 49, 3 (2017), 1048–1064.

[6] Harold L Kundel, Calvin F Nodine, and Dennis Carmody. 1978. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative radiology* 13, 3 (1978), 175–181.

[7] Susan Mallett, Peter Phillips, Thomas R Fanshawe, Emma Helbren, Darren Boone, Alastair Gale, Stuart A Taylor, David Manning, Douglas G Altman, and Steve Halligan. 2014. Tracking eye gaze during interpretation of endoluminal three-dimensional CT colonography: visual perception of experienced and inexperienced readers. *Radiology* 273, 3 (2014), 783–792.

[8] Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. ACM, 71–78.

[9] Daniel P Turgeon and Ernest WN Lam. 2016. Influence of experience and training on dental students' examination performance regarding panoramic images. *Journal of dental education* 80, 2 (2016), 156–164.

[10] A Van der Gijp, CJ Ravesloot, H Jarodzka, MF van der Schaaf, IC van der Schaaf, Jan PJ van Schaik, and Th J Ten Cate. 2017. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Advances in Health Sciences Education* 22, 3 (2017), 765–787.