Scanpath comparison in medical image reading skills of dental students

Distinguishing stages of expertise development

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

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

Thomas Kübler*
Perception Engineering, University of Tübingen
Tübingen, Germany
thomas.kuebler@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

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

ABSTRACT

A popular topic in eye tracking is the difference between novices and experts and their domain-specific eye movement behaviors. However, very little is researched regarding how expertise develops, and more specifically, the developmental stages of eye movement behaviors. Our work compares the scanpaths of five semesters of dental students viewing orthopantomograms (OPTs) with classifiers to distinguish sixth semester through tenth semester students. We used the analysis algorithm SubsMatch 2.0 and the Needleman-Wunsch algorithm. Overall, both classifiers were able distinguish the stages of expertise in medical image reading above chance level. Specifically, it was able to accurately determine sixth semester students with no prior training as well as sixth semester students after training. Ultimately, using scanpath models to recognize gaze patterns characteristic of learning stages, we can provide more adaptive, gaze-based training for students.

CCS CONCEPTS

• **Applied computing** → **Psychology**; Interactive learning environments; • **Computing methodologies** → *Classification and regression trees*;

KEYWORDS

Remote Eye Tracking, Scanpath analysis, Medical image interpretation, Learning

ACM Reference Format:
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 *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3204493.3204550

1 INTRODUCTION

Experts ranging from Olympic athletes and chess players to surgeons, doctors, and teachers are often characterized by their proficient abilities. Their skills are built over time, through practice and developing the knowledge that accompanies their expertise. Not only does expertise relate to performance, but also eye movement behavior [Gegenfurtner et al. 2011]. Here, it has been consistently found that differences between experts' and novices' task related eye movements are indeed apparent and can be reflective of performance [Eivazi et al. 2017; Gegenfurtner et al. 2011; Kübler et al. 2015; Moran et al. 2002; Reingold et al. 2001; Van der Gijp et al. 2017]. Conventionally, most of the expertise literature focuses on this stark group contrast and, to an extent, the novice - intermediate - expert differences. In this work, we aim to determine whether eye

*Work of the authors is supported by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63)
†Department of Prosthodontics
‡Department of Radiology, Center of Dentistry, Oral Medicine and Maxillofacial Surgery

ETRA '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-1-4503-5706-7/18/06…$15.00
https://doi.org/10.1145/3204493.3204550

movement differences within the novice category become apparent and, if so, at what level of task-knowledge they appear.

1.1 Expert and Novice Differences

There are tenable theories for eye movement behavior differences in experts and novices. Task-relevant information gathered more rapidly [Haider and Frensch 1999], more rapid processing and retrieval of information stored in memory [Ericsson and Kintsch 1995], and more thorough global image analysis [Kundel et al. 2007] are considered by Gegenfurtner and colleagues [Gegenfurtner et al. 2011] to be the most supported by the literature.

In the medical domain, expertise is relevant to image interpretation; for instance, accurate detecting of anomalies in radiographs [Kundel et al. 2007; Van der Gijp et al. 2017, 2014]. Here, it has been found that experts employ fewer fixations than novices [Gegenfurtner et al. 2011; Nodine et al. 1996; Van der Gijp et al. 2017], as well as longer saccade lengths [Gegenfurtner et al. 2011; Van der Gijp et al. 2014] and they are overall faster and more accurate at detecting anomalies [Gegenfurtner et al. 2017, 2011; Kok et al. 2016; Kundel et al. 2007]. Efficient detection lies in the search strategy experts employ. For instance, a *global - to - focal* search strategy [Nodine et al. 1996; Van der Gijp et al. 2017], where the whole image is quickly scanned for overall assessment, then more subtle issues are focused in on. In contrast, novices show more initial centralized search that systematically covers an image and more attention to salient structures[Van der Gijp et al. 2017]. Van der Gijp and colleagues also looked at search patterns related to expertise and found that, within tasks (e.g. looking at chest x-rays or mammography), expert's visual patterns (e.g. diffusive, left-right comparison) are consistent [Van der Gijp et al. 2017].

To the best of our knowledge, only one study has looked at expert-novice gaze differences in the context of radiograph images specifically for dentistry (orthopantomogram, short: OPT). Turgeon and Lamm [Turgeon and Lam 2016] found that the complexity of the image affected search time regardless of expertise. Also, experts had fewer fixations on OPTs where the anomalies were more obvious compared to novices, though for images with no anomalies, scanning behavior for both groups was not significantly different [Turgeon and Lam 2016]. These findings could imply that visual search behavior in OPTs may have similar gaze behaviors to other types of radiographs, but the OPT visual search strategy patterns may differ.

1.2 Developing Expert Behavior

Although literature on gaze behavior in the particular context of OPTs is sparse, the majority of radiographs are taken in dental medicine[1]. In contrast to other medical fields, OPTs are major part of the routine diagnosis. However, given how critical OPTs are to dental medicine, like radiographs, they are susceptible to under-detections and missed information (dental OPTs:[Baghdady et al. 2014, 2009], non-dental radiographs:[Kok et al. 2016; Krupinski et al. 2006; Kundel et al. 1978, 2008]).

The rate of correct detection can be increased in both the dental and general medical fields. In dentistry for instance, patients benefit greatly from early detection of calcifications of the cervical vessels or pathologies of an inflammatory or neoplastic nature in the jawbones or maxillary sinuses. Thus, there is large potential for addressing methodologies in the teaching of radiologic feature identification and interpretation [Van der Gijp et al. 2017]. In addition, previous work provides evidence that eye-tracking can be successfully deployed to design training techniques [Van der Gijp et al. 2017]. Therefore, augmenting the learning material to promote how to read radiographs is a promising approach for novice training.

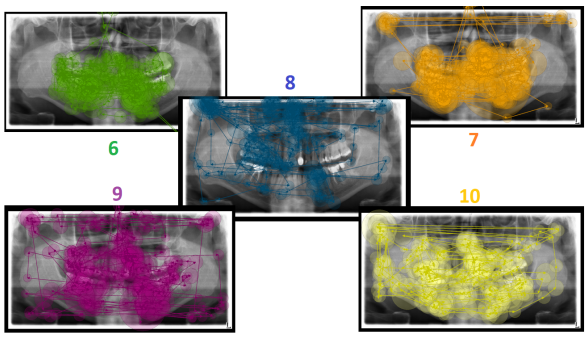


Figure 1: Visualization of fixations from a student in each semester evaluated in the current study as indicated by the colored numbers respectively. In this condition, the sixth semester student's data is prior to training.

The expert-novice discussion is important because it may have implications for the question of how to teach. Given what is known of an expert's eye movements, how can learning interventions impart expert eye movement patterns to a student? Jarodzka and colleagues [Jarodzka et al. 2010b] found that novices were more likely to focus on irrelevant information because they lacked the conceptual knowledge to filter out the extraneous details. As a training intervention, they found that displaying an expert gaze behavior model improved visual attention to the relevant information in visual stimuli [Jarodzka et al. 2010b]. Furthermore, Jarodzka and colleagues [Jarodzka et al. 2012] found that by combining verbal instruction and expert gaze overlay, these eye movement modeling examples (EMMEs) improved visual search behavior for medical students in a clinical reasoning task. Despite these encouraging results, it is yet an open question whether using a model that is only slightly ahead of the student and modeling of gaze behavior in a progressive fashion could be even more effective. For that question to be answered, one first needs to better understand the developmental stages of students.

The purpose of this work intends to address the visual search behavior related to the developmental stages of students. With their differences in mind, we can use these progressive models in learning interventions. Therefore, the future goal will be to detect when and where a student's visual search of an OPT deviates from a more advanced visual search model and, in real time, redirect him or her towards the gaze behavior most optimal for the best performance.

[1] According to the statistics of the Federal Agency for Radiation Protection, 39% of all x-rays in Germany were taken within dental medicine in 2012 (www.bfs.de).

1.3 Gaze Behavior

Gaze behavior differences between novice and experts have been reliably measured in multiple studies [Gegenfurtner et al. 2011]. However, it is interesting to see whether differences appear within one dimension: e.g. novices. Differences between students based on their conceptual knowledge may be apparent at the semester level. Figure 1 shows the scanpath of a student from each semester, six through ten, taken from the current study. Here, the sixth semester student's scanpath visualized is prior to the OPT analysis course; he or she has some basic anatomy knowledge, but not in the context of OPTs. His or her scanpath shows fixations only on the teeth and no peripheral area exploration. A change in exploratory behavior is seen from the sixth semester to the seventh semester, where scanning behavior that compares similar areas of the jaw on the left and the right is present. Then, eighth, ninth, and tenth semester students show more coverage of the OPT; specifically, less fixations on the teeth and longer saccades spanning the upper and lower jaw areas.

Differences in exploratory behavior, as characterized in the scanpaths of experts and novices, is often under-explored in the literature. Even more, scanpath differences relating to the developmental stages has yet to be measured: Such as scanpaths reflecting acquired knowledge in each semester. Understanding gaze behavior in an effort to find patterns determinant of a students' developmental level can ultimately build an adequate model representation of eye movements for the complete learning process. Therefore, we aim to distinguish exploratory behavior differences at the semester level.

1.4 Scanpath analysis

One of the most accepted methods for scanpath analysis is relabeling fixations to characters. Then, patterns of fixations are expressed as a string of characters. String representations are often constructed to provide information on how a subject views a stimulus relative to areas of interest (AOIs). Then, we can measure the similarity of one subject's scanpath to another's: For instance, via a distance score [Goldberg and Helfman 2010; Jarodzka et al. 2010a; Kübler et al. 2014]. The scores relate to how the sequences can be aligned. Thus, these metrics are known as sequence alignment techniques.

According to Jarodzka and Colleagues [Jarodzka et al. 2010a], AOIs can either be *semantic*, where they are manually defined, or *gridded*. The gridded-AOI approach divides the stimulus into blocks. This approach saves time compared to the former approach and maintains the sequential order, shape, and the length of the scanpaths [Jarodzka et al. 2010a]. An example of two scanpaths represented as strings, as well as their alignment, is depicted in Figure 2. In general, string alignment techniques are dependent on the AOIs, meaning they are susceptible to noise [Cristino et al. 2010; Holmqvist et al. 2011; Jarodzka et al. 2010a]. Aside from the sequence alignment approaches to scanpath comparison, there are other methods such as implementations of Hidden Markov Models [Ellis and Stark 1986; Goldberg and Helfman 2010; Hacisalihzade et al. 1992; Josephson and Holmes 2002] as well as vector-based approaches [Dewhurst et al. 2012; Jarodzka et al. 2010a]; though they are more complex and may be less sensitive to sequence order. This paper deals largely with sequence alignment.

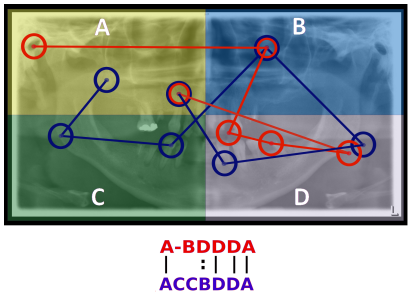


Figure 2: Scanpath comparison example with two scanpaths for same stimuli and AOI grid. Below the image is the global string alignment calculated with the Needleman-Wunsch algorithm. Matches, mismatches, and gaps are [| , : , -] respectively.

Global String Alignment Approach. As previously mentioned, string alignment methods score a scanpath against another based on their similarity. These methods can either align locally, where subsequence alignment takes precedence, or globally. One global alignment approach is the Needleman-Wunsch algorithm. For two sequences, a matrix is created, and each element is filled with either corresponding penalties for gaps or substitutions or rewards for matches. Compared to other sequence alignment techniques, the scoring system can offer more flexibility, such as limiting the penalties for either gaps or mismatches [Baichoo and Ouzounis 2017; Day 2010].

Originally used in bioinformatics, the Needleman-Wunsch algorithm was developed for genetic sequence alignments [Needleman and Wunsch 1970]. It has also become a staple of scanpath analysis. Since string alignment methods' first appearance in the eye-tracking world in the nineties [Brandt and Stark 1997; Hacisalihzade et al. 1992], the Needleman-Wunsch algorithm has been used for numerous studies. For instance, [Day 2010] used it to classify differing visual search behavior strategies during a decision making task. Pan and colleagues [Pan et al. 2004] determined that scanpath differences on web pages were affected by the complexity of the web page design. Additionally, an implementation of the Needleman-Wunsch algorithm supported that expert and novice programmers showed scanpath differences while reading lines of Java code [Busjahn et al. 2015]. In both [Busjahn et al. 2015; Pan et al. 2004], group and behavioral differences were measured by grouping similarity scores. Day and colleagues [Day 2010] validated it as a classifier rather than post hoc similarity grouping. They found that it was capable of distinguishing six decision making strategies at from 88% accuracy [Day 2010].

An issue with the Needlman-Wunsch and other sequence alignment algorithms is that they can be time costly [Goldberg and Helfman 2010]. Pairwise comparisons have $O(mn)$ complexity for both time and space for very large sequences m and n [Baichoo and Ouzounis 2017]. Furthermore, it does not account for fixation duration, though other implementations of the Needleman-Wunsch algorithm, as well as other string alignment approaches, have compensated for temporal information loss. [Cristino et al. 2010].

String Kernel Approach. SubsMatch [Kübler et al. 2014] combines string representation with transition frequency analysis. Contrary to transition matrices or Markov chains, transitions between multiple subsequent fixations can be handled, which can correspond to behavioral patterns. Initially, a scanpath string is constructed by assigning letters to fixations in a way that the final scanpath string contains roughly the same number of occurrences of each letter. Therefore, horizontal bins of different sizes are constructed so that each bin contains the same number of fixations [Kübler et al. 2014]. The number of such bins, and thereby of letters to use, is one of the parameters of the algorithm. Then, all possible subsequences of a given size (so-called n-grams, where n stands for the length of the sequence and is the second parameter in the algorithm) and their occurrence frequencies are calculated. A similarity metric between scanpaths can be calculated as the sum of differences between all subsequence frequencies.

Relatively new to scanpath analysis metrics, SubsMatch has demonstrated its versatility across task based eye movements [Braunagel et al. 2017a,b; Kübler et al. 2015, 2014, 2017]. Originally, it was developed and evaluated on dynamic driving scenarios to determine safe versus unsafe drivers [Kübler et al. 2014]. Moreover, Subsmatch was able to determine expert and novice microneurosurgeon viewing behavior for multiple images with significant between group differences compared to other metrics such as Scanmatch, Multimatch, and Eyenalysis [Kübler et al. 2015][2].

SubsMatch was further improved in the version SubsMatch 2.0 [Kübler et al. 2017] by replacing the similarity metric with a SVM classification. The frequencies of n-grams are then features used for a support vector machine (SVM) with a linear kernel. Feature weights are determined by their importance for distinguishing between two conditions during the training phase. Fundamentally, SubsMatch 2.0 sets out to determine the best-fit subsequence length in conjunction with the best-fit string representation in order to perform SVM classification based on subsequence occurrences. SubsMatch 2.0 was evaluated on four different data sets (see [Kübler et al. 2017]). It was capable of accurately distinguishing group based scanpath patterns in varying laboratory and real-world experiments [Kübler et al. 2017]. Reported accuracies ranged from approximately 20% to 90% for all experimental data evaluated. Where the highest classification accuracies were for experts and novices in MarioKart video game driving scenario and the lowest were for image prediction for both a conjunction search task and the Yarbus task. It should be noted, even the low accuracies were significantly above chance level [Kübler et al. 2017].

In general, sequence alignment algorithms can offer insight into the exploratory eye movement behavior of individuals and groups. The Needleman-Wunsch algorithm has shown great flexibility across fields in eye tracking and is regularly applied to determine scanpath similarity. We aim to distinguish exploratory behavior differences at the semester level; therefore, such an algorithm is applicable to our cause. Another interesting aspect is the subsequence patterns that may develop based on a student's level of understanding, i.e., a representation of the associations between different stimulus areas. The SubsMatch algorithm is able to analyze patterns of this nature. They can be substantially different from those found by global sequence alignment, and are an interesting addition. SubsMatch is less commonly used than the Needleman-Wunsch algorithm, but its versatility in classifying scanpaths in laboratory and real-world scenarios has been demonstrated and it can be interpreted as a generalization of the more commonly found transition matrices. From this analysis, we can further work towards developing a representative model of the stages of learning development.

[2]False Discovery Rate adjusted p-values of a permutation test were provided showing differences in gaze behavior detected for [Kübler et al. 2015],[Kübler et al. 2014]

2 METHODOLOGY

2.1 Participants

Dentistry students in the sixth, seventh, eighth, ninth, and tenth semesters from the University Hospital Center for Dentistry, Oral Medicine, and Maxillofacial Surgery were invited to participate in an assessment of their OPT analysis training. This assessment was held in a classroom equipped with 30 remote SMI RED250 eye trackers, each attached to a laptop[3]. Data from a total of 103 students were collected: Sixth semester (n = 17), seventh semester (n = 18), eighth semester (n = 26), ninth semester (n = 28), and tenth semester (n = 14). Students in the seventh through tenth semesters were invited to participate once during the semester, whereas the sixth semester students were assessed three times: At the beginning of the semester (n = 17), then again in the middle of the semester (n = 17), and lastly, at the end of the semester (n = 15). These students were measured on multiple occasions because the sixth semester is the first and only semester in the dentistry program where they receive explicit instruction and start massed practice OPT interpretation.

2.2 Eye Tracker

The SMI RED250 remote eye tracker is a commercial eye tracker with 250Hz sampling frequency. The experiment was created and controlled using the SMI software *ExperimentCenter 3.7.60*. Stimuli were web-based[4], with a 13-point[5] calibration prior to presentation. Analysis of the data was performed with the software *BeGaze*.

2.3 Data Collection

All students were presented with two sets of ten OPTs with varying anomalies, some more difficult than others. Each OPT was viewed twice: Once to explore, then again to draw and indicate any anomalies found (e.g. Periodontal disease, cavities, insufficient fillings and abscesses, not including sufficient fillings, missing teeth needing no further treatment, or prosthetics). Students fixated on a fixation cross for two seconds. Then, for the exploration phase, they had 1:30 minutes to look at the OPT. Here, they were instructed to search the OPT for anomalies[6]. For the marking phase, they were instructed to mark anomaly areas with a red circle[7]. A web-based tool bar was used with a paint-palette symbol in order to draw red circles on the OPT image presented on the screen. For this phase, they had

[3]Display: 1920 ×1080 pixel resolution.
[4] Mozilla Firefox version 45.9.0
[5]However, a 9-point calibration was used for pre-training sixth semester students.
[6]Exploration: "Das Panoramaröntgenbild lediglich betrachten und nach Auffälligkeiten mit Krankheitswert suchen."
[7]Marking: "...Nun sollen Sie Auffälligkeiten markieren."

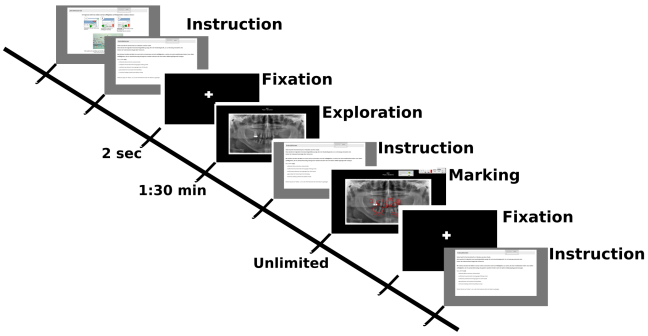


Figure 3: Outline of Experimental Session. After a calibration, there is an introduction to the task and a tutorial on marking the anomalies. After, a verbal instruction was presented with information on what kind of anomalies he should focus on. The subject is primed with a fixation cross. Then in the exploration phase, he has 1:30 minutes to search the image in a clinical context. After, there is another instruction slide for drawing anomalies. Then, in the drawing phase, he marks the issues using an on-screen drawing tool. Here, he has unlimited time and clicks a button on the top right corner to advance. There are 10 OPTs presented in a set, each in an visual exploration and marking phase.

as much time as they needed and could click the continue button to advance. In all, one set comprised of a calibration, introduction, and instruction, then for the ten images, a fixation, exploration, and drawing. Figure 3 illustrates the experimental protocol. In one testing session, two OPT sets were presented with a ten minute break in between.

2.4 Data Analysis

In the current study, eye movement data during the visual exploration phase of OPTs in the first set were evaluated. Fixations and saccades for the left eye, including tracking ratios per image, were calculated using the *BeGaze* software. Fixations were calculated using the standard SMI high-speed settings for the I-VT [Salvucci and Goldberg 2000]: 50ms for minimum duration and 40°/*s* peak velocity threshold and peak velocity start at 20% of the saccade length and peak velocity end at 80% of saccade length. Eye movement data was removed for images where the tracking ratio was below 80%. Furthermore, participants were removed if they had missing data for more than two of the ten images. Ultimately, for the scanpath comparison, eye movement data from 88 participants were used.

Scanpaths were evaluated in three conditions. First, six semester students prior to their first OPT analysis training course were compared to seventh, eighth, ninth, and tenth semester students (pre-training). Second, sixth semester students during the training course were compared to each of the higher semesters (mid-training). Third, sixth semester students at the semester end were compared to each of the higher semesters (post-training). By evaluating the pre-training condition, we can determine how distinguishable their gaze behavior is due to their lack of OPT exposure. For the post-training condition, we can determine how similar the gaze behavior of sixth semester students is to other semesters, e.g. seventh semester students. Since the time-course of each semester is a few months, with roughly two month difference between consecutive semesters, we also expect similarities in gaze behaviors in consecutive semesters, e.g. ninth and tenth semester students.

3 RESULTS

We aim to determine whether there are differences in OPT exploratory behavior of dentistry students at incremental levels of their training. We evaluated the SubsMatch algorithm and the Needleman-Wunsch algorithm on three conditions. Since the classifiers are trained on five semesters (and trials are almost balanced), guess chance level is roughly 20 percent. The accuracy of the classifier is measured as the total number of correctly predicted labels over the total data set.

Since both classifiers employ supervised learning, data is divided and used for either training or validation. For training, pre-, mid-, and post- conditions each had 73, 68, and 68 participants respectively. These values were the total students from each of the five semesters, with data differing only for the sixth semester students: since they were evaluated over three occasions. For the validation data, a total of 15 participants – three per each semester– were set aside. Each participant viewing up to ten OPTs would result in a maximum of 150 data sets, though after removal of data with low tracking ratios, 139 data sets were included. As per the training data, the validation data for all semesters was the same for each condition, with the sixth semester students' data differing.

3.1 SubsMatch 2.0 Algorithm Classification

Table 1: Model Classification Accuracy for Data

Condition	Subsmatch 2.0		Needleman-Wunsch	
	Test	Validation	Test	Validation
Pre-Training	37.20%	28.06%	37.20%	30.90%
Mid-Training	34.49%	20.14%	36.30%	20.14%
Post-Training	34.48%	25.18%	33.73%	23.74%

For training the SVM, both the percentile binning (from [Kübler et al. 2017]) and the gridded bins (from [Cristino et al. 2010]) were evaluated. We chose the latter approach for our data because it provided higher accuracies. However, it should be noted that the overall difference in classification accuracy for gridded and percentile binning was minimal and either approach could be employed.

After a leave one out cross validation on the training data, as described in [Kübler et al. 2017], the SVM model suggested the respective n-gram and alphabet size parameters for all conditions: 2 and 3 for the pre-training condition, 3 and 7 for the mid-training condition, and 2 and 7 for the post-training condition.

Table 1 details the overall accuracies for the models for both the test data and the validation data. The classifier is capable of distinguishing semesters above chance level for pre- and post-conditions. Above all, the classifier shows the highest accuracy for the pre-training condition, where the sixth semester students before their OPT analysis training.

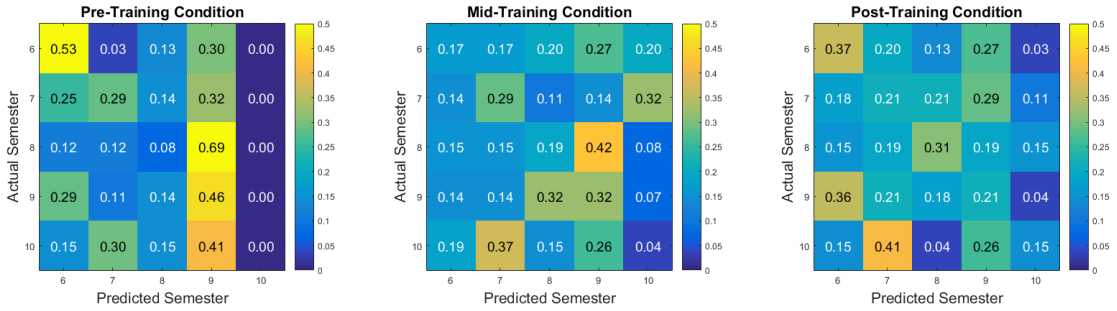


Figure 4: SubsMatch semester classification on the validation data. From left to right, confusion matrices for conditions pre-training, mid-training, and post-training are presented. With true positive rate for each semester along the diagonals. Note that the colorbar for all conditions is scaled at .5.

More important than overall performance is how the semesters were distinguished. Figure 4 shows the confusion matrices for each condition. From the first matrix in figure 4. The model accurately predicts pre-training sixth semester students (53.33%) and ninth semester students. However, it often predicts eighth semester students as ninth semester students (69.23%). Additionally, tenth semester were falsely classified as ninth semester or seventh semester students.

Concerning the mid-training condition, overall performance was at chance level. The middle confusion matrix in figure 4 also shows that misclassification was more often high for all semesters.

Similar to the pre- training condition, post-training sixth semester students were accurately classified (36.67%). Interestingly, the ninth and tenth semester were more likely to be misclassified as lower semesters (See last matrix in figure 4).

This error in classifying the tenth semester students was also apparent in all three conditions, where they are often misclassified as either seventh or ninth semester students. Moreover, eighth were more likely to be accurately classified, or misclassified as ninth semesters in all conditions. Sixth semester students were able to be accurately classified in both the pre- and post-training condition.

3.2 Needleman-Wunsch Algorithm 1-Nearest Neighbor Classification

We ran the Needleman-Wunsch algorithm for each scanpath in the training set against all others to create a matrix of similarity scores for each pair. For scoring, 2, −2, and −1 for matches, mismatches, and gaps respectively.

For the grid-overlay size, we divided the stimulus evenly into blocks: For example, a 10×8 size grid means ten blocks wide and eight blocks high. We ran a multiple-pairwise NW alignment on the training data for grid sizes from 5×5 to 10×10. The most optimal grid size was 6×5 width and height respectively[8]. Then, with the multiple-pairs similarity matrix, a one-nearest neighbor classifier determined the best matched similarity score for each scanpath. The idea is that the scanpaths in the same class will have the highest similarity score and will be classified accurately.

[8]For our stimuli: 320×216 pixels for each block size

Table 1 reports the overall accuracies for the Needleman-Wunsch classifier for both training and validation data. Figure 5 shows the confusion matrix for semester classification for each condition.

In the pre-training condition (first matrix of figure 5), sixth semester students are classified accurately 80% of the time; however, the model also tends to over-classify other semesters as sixth semester, such as the eighth semester and the tenth semester students. Otherwise, ninth semester students are accurately classified. Similar to SubsMatch, seventh semester students were also more likely to be classified as ninth semester.

In the mid-training condition (middle matrix of figure 5), again, performed overall at chance level and similar to SubsMatch. For example, the ninth semester students are accurately detected. Also, sixth semester students were more likely to be misclassified as ninth semester students. Finally, tenth semester students were highly likely to be classified as seventh semester students (51.85%).

Lastly, in the post-training condition (last matrix of figure 5), tenth semester students are again misclassified as seventh semester students (48.15%) which is similar to SubsMatch. More interesting, is the slight shift in the sixth and seventh semester students, where they were misclassified more often as higher semester students.

Moreover, there were no significant differences between semesters sixth through tenth regarding the overall fixation time on expert defined anomalies (p = .826). Moreover, differences in fixation time within the 6th semester (pre, mid, post-training) were not significant as well (p = .881). Thus, the classifiers were able to extract pattern information related to learning where the eye movement data alone could not. Both algorithms were highly capable of distinguishing sixth semester students in the pre-training condition, and if they falsely classified students in a semester, they were likely classified as either the preceding or successive semester.

4 DISCUSSION

Both SubsMatch and Needleman-Wunsch algorithms are similarly capable of distinguishing semesters from the scanpath data. Both are highly accurate at classifying sixth semester students with no prior training in OPT analysis as well as distinguishing sixth semester students at the end of the semester. These results indicate that

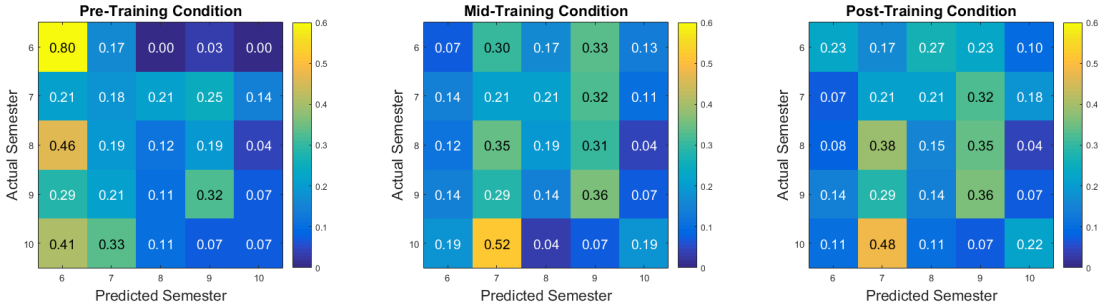


Figure 5: Needleman-Wunsch semester classification on the validation data. From left to right, confusion matrices for conditions pre-training, mid-training, and post-training are presented. With true positive rate for each semester along the diagonals. Note that the colorbar for all conditions is scaled at .6.

learning in the first semester (pre-training vs post-training condition) is very relevant. As previously mentioned, the sixth semester is where they are first exposed to OPT analysis and interpretation. This lack of previous exposure in the pre-training is clearly observable in the classifiers. The 1-nearest neighbor Needleman-Wunsch classifier is very sensitive to the pre-training sixth semester and, therefore, more likely to classify any trial as such. As apparent in the confusion matrix (first matrix in figure 5), where eighth and tenth semesters are frequently misclassified as sixth. With this consideration, SubsMatch performs better separation between pre-training sixth semester students and all others.

Regarding the mid-training condition, both classifiers performed similarly and barely above chance level. This behavior from the classifier could be an effect of heterogeneity in learning speed and success. In the framework proposed by [Van der Gijp et al. 2014], the initial stage of expertise development is multi-faceted. Not only is it a foundation of anatomy and pathology knowledge, but also spatial abilities and ability to mentally manipulate images. Possibly, some students advance in one of these areas, but not in another (i.e. high anatomy recall, but not yet in a clinical context), hence the overall behavior is not consistent enough to be easily distinguishable.

Sixth semester students at the end of the semester, the post-training condition, are distinguishable from higher semesters, but at a much lesser extent than they were prior to training. A possible effect seen in this condition could be the imminent final exams motivating students to study. Hence, these students were likely to be misclassified, as higher semesters as seen in the Needleman-Wunsch classifier and, to a lesser extent, in the SubsMatch classifier.

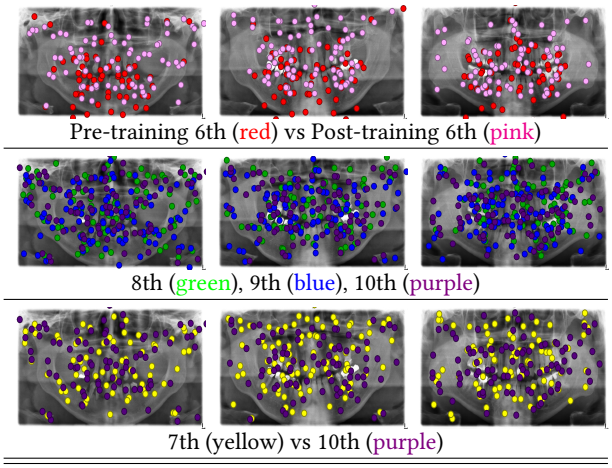
Al-Moteri and colleagues [Al-Moteri et al. 2017] comprised literature regarding eye movements and medical decision making and found that clinical experience was related to gaze behavior that was more *goal-driven* and less *stimulus-driven* [Al-Moteri et al. 2017; Krupinski et al. 2006]. This finding supports the research that experts are less drawn to salient features with no diagnostic relevance. However, differences in gaze behavior before and after massed training (i.e. within the novice level) could also be explained by their findings. For instance, less experienced students may still be more drawn to salient areas, such as the teeth, and may neglect more important areas that have more subtle cues in comparison to a more experience student in the same semester.

Overall, it is apparent that OPT exploratory behavior shows considerable initial change. However, these patterns become more homogeneous over the course of the higher semesters. This behavior can be inferred by the classifiers consistently misclassifying eighth, ninth, and tenth semester students. The gaze behavior differences between eighth through tenth semester may not be as large or clear as between other semesters. Thus, there seems to be a gaze *behavioral plateau* once students reach the later semesters, where visual search behavior of OPT does not appear to change drastically. For example, table 2 shows fixation clusters of the validation data for three of the ten OPTs. Even without the sequential information, we can see that image coverage differences are the most visible when comparing the sixth semester students with no prior OPT analysis training against the sixth semester students after OPT analysis training. More complicated to decipher are the clusters of the eighth, ninth, and tenth semester students; in the second row of table 2 we see minimal difference in image coverage between the semesters.

Due to the classifier's behavior, we decided to look at the data in another context: The content of the curriculum for each semester. The sixth semester students receive the OPT analysis and interpretation course alongside lectures on radiology protection and methods and clinical based lectures on dental, oral and maxillofacial diseases. In the seventh semester, the curriculum includes another radiology lecture as well as other courses dental care and orthodontology. After the seventh semester, the curriculum has no courses addressing OPT analysis, rather other concepts related to orthodontics, prosthetics, or diseases and treatment. Students in higher semesters also take practical training courses as well as supervised treatment of patients, though there is no requirement to review OPTs, nor is there further training targeted at OPT analysis.

Interesting enough, the tenth semester students are classified as seventh semesters relatively often (see third row of table 2). This finding could be due to lack of OPT exposure in the curriculum of the higher semesters. Whether their gaze behavior is similar to that of seventh semester students due to outstanding effects has yet to be determined. One possibility could be the expertise reversal

Table 2: Validation data fixation clusters per semester on three separate images



effect [Kalyuga et al. 2003], where at some point in their studies they have may have increased cognitive load (a prime example being their final medical school examinations). Another possibility could be that the tenth semester students start to slowly develop and test their own gaze shortcuts. Tenth semester students could be transitioning towards intermediate level, and their visual search strategies start becoming more personalized. Cooper and colleagues [Cooper et al. 2010] found that radiologist trainees, though more accurate than novices at identifying anomalies in magnetic resonance images, spend the same amount of time searching the image. The authors liken this behavior to constructing their own visual pattern; where more advanced trainees shows similar gaze patterns to experts[Cooper et al. 2010]. Future research could further compare students in their last semester at university against first year interning in order to determine if there are any changes in performance as well as visual search strategy.

In the present study, data was collected from only 14 participants in the tenth semester. Since each participant had scanpath data for ten different images, this sample size was determined to be adequate. There is a chance that the nearest neighbor classifier was affected by the group sizes, but the SVM classifier used in SubsMatch balanced class weights. However, more participants in this semester could improve the classifiers prediction accuracy for these students.

Although the fixation data did not show significant differences between students, both the SubsMatch and Needleman-Wunsch classifiers were able to detect patterns in the visual search behavior at the semester level. These patterns were more reflective of learning that occurs in the initial training course in the sixth semester in the curriculum. Even with only a few months between these semesters, subtle differences were still apparent.

The overall accuracy was relatively low when comparing to the previous work for both the Needleman-Wunsch and Subsmatch 2.0. In [Busjahn et al. 2015], the Needleman-Wunsch achieved distinguishable differences between of experts and novices. Where novices were 14 introduction to computer science students and experts were 6 experienced software engineers [Busjahn et al. 2015]. Based on much of the literature reviewed in [Gegenfurtner et al. 2011], we can also conclude that students compared to engineers or even, in our case, students compared to experienced radiologists would have highly contrasting behavior that would affect higher classification accuracy. [Day 2010] achieves high accuracy (88%) for classifying 6 decision strategies, but the authors specify that participants were trained in each strategy for two hours prior to evaluation.

Similarly, Subsmatch 2.0 was evaluated on varying data from the Yarbus task (66%) to MarioKart (92%), and consistently achieved high classification [Kübler et al. 2017]. More important, Kübler and colleagues note that the algorithm performs better when classifying stimuli differences or performed task, but performance differences (i.e. passing or failing a driving test) can be challenging [Kübler et al. 2017]. Given that our task used semester level as a measure of learning differences, classification in this context is very difficult. Moreover, eye movements, such as number of fixations, between semesters do not differ as dramatically as between novices and experts. Hence, our work was less intent on such high level abstraction and more on the complex pattern distinction. Considering the curriculum for dentistry students offers the OPT analysis course only in the sixth semester and that higher semester dentistry students have no mandatory OPT exposure, we were able to see the learning from this course as represented in the scanpaths.

5 CONCLUSION

With scanpath comparision, we were able to distinguish OPT exploratory gaze behavior at a semester level. Both models evaluated indicated that there was an initial effect in the sixth semester students, which is in line with the sixth semester curriculum. Additionally, higher semesters become less distinguishable in their gaze behavior, which could also be an effect of minimal OPT training in the curriculum of these semesters. Whether continuous routine OPT image interpretation in higher semesters would lead to more effective visual search strategies and ultimately performance poses further interesting future research questions.

Performance data of each semester, such as detection rate and number of false positives, were out of the scope of this paper since the main focus was scanpath analysis. However, this information would serve as an ideal baseline for comparing classifier behavior. Future research could measure performance of the semesters and how scanpath differences are intertwined. From previous literature, employing learning interventions to promote expert visual search strategies in students often neglects improving the performance [Gegenfurtner et al. 2017; Jarodzka et al. 2012, 2010b; Kok et al. 2016; Van der Gijp et al. 2017]. This discord is attributed to semantic knowledge or reasoning that novices have yet to develop. In order to coalesce both search strategy and performance of students, future research can concentrate more on the progressive behavior modeling rather than expert behavior modeling. Gaze-based learning interventions that model each stage of expertise development rather than the absolute end may provide promising outcomes regarding the performance. Consequently, adapting the model behavior to the level of the student may be more effective for dependable diagnoses later on in the dental and even medical fields.

REFERENCES

Modi Owied Al-Moteri, Mark Symmons, Virginia Plummer, and Simon Cooper. 2017. Eye tracking to investigate cue processing in medical decision-making: A scoping review. *Computers in Human Behavior* 66 (2017), 52–66.

Mariam T Baghdady, Heather Carnahan, Ernest WN Lam, and Nicole N Woods. 2014. Dental and dental hygiene students' diagnostic accuracy in oral radiology: effect of diagnostic strategy and instructional method. *Journal of dental education* 78, 9 (2014), 1279–1285.

Mariam T Baghdady, Michael J Pharoah, Glenn Regehr, Ernest WN Lam, and Nicole N Woods. 2009. The role of basic sciences in diagnostic oral radiology. *Journal of dental education* 73, 10 (2009), 1187–1193.

Shakuntala Baichoo and Christos A Ouzounis. 2017. Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. *Biosystems* (2017).

Stephan A Brandt and Lawrence W Stark. 1997. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of cognitive neuroscience* 9, 1 (1997), 27–38.

Christian Braunagel, David Geisler, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017a. Online Recognition of Driver-Activity Based on Visual Scanpath Classification. *IEEE Intelligent Transportation Systems Magazine* 9, 4 (2017), 23–36.

Christian Braunagel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017b. Ready for Take-Over? A New Driver Assistance System for an Automated Classification of Driver Take-Over Readiness. *IEEE Intelligent Transportation Systems Magazine* 9, 4 (2017), 10–22.

Teresa Busjahn, Roman Bednarik, Andrew Begel, Martha Crosby, James H Paterson, Carsten Schulte, Bonita Sharif, and Sascha Tamm. 2015. Eye movements in code reading: Relaxing the linear order. In *Program Comprehension (ICPC), 2015 IEEE 23rd International Conference on*. IEEE, 255–265.

Lindsey Cooper, Alastair G Gale, Janak Saada, Swamy Gedela, Hazel J Scott, and Andoni Toms. 2010. The assessment of stroke multidimensional CT and MR imaging using eye movement analysis: does modality preference enhance observer performance? (2010).

Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. 2010. ScanMatch: A novel method for comparing fixation sequences. *Behavior research methods* 42, 3 (2010), 692–700.

Rong-Fuh Day. 2010. Examining the validity of the Needleman–Wunsch algorithm in identifying decision strategy with eye-movement data. *Decision Support Systems* 49, 4 (2010), 396–403.

Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. 2012. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior research methods* 44, 4 (2012), 1079–1100.

Shahram Eivazi, Ahmad Hafez, Wolfgang Fuhl, Hoorieh Afkari, Enkelejda Kasneci, Martin Lehecka, and Roman Bednarik. 2017. Optimal eye movement strategies: a comparison of neurosurgeons gaze patterns when using a surgical microscope. *Acta Neurochirurgica* 159, 6 (2017), 959–966.

Stephen R Ellis and Lawrence Stark. 1986. Statistical dependency in visual scanning. *Human factors* 28, 4 (1986), 421–438.

K Anders Ericsson and Walter Kintsch. 1995. Long-term working memory. *Psychological review* 102, 2 (1995), 211.

Andreas Gegenfurtner, Erno Lehtinen, Halszka Jarodzka, and Roger Säljö. 2017. Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis. *Computers & Education* (2017).

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.

Joseph H Goldberg and Jonathan I Helfman. 2010. Scanpath clustering and aggregation. In *Proceedings of the 2010 symposium on eye-tracking research & applications*. ACM, 227–234.

Selim S Hacisalihzade, Lawrence W Stark, and John S Allen. 1992. Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Transactions on systems, man, and cybernetics* 22, 3 (1992), 474–481.

Hilde Haider and Peter A Frensch. 1999. Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25, 1 (1999), 172.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Halszka Jarodzka, Thomas Balslev, Kenneth Holmqvist, Marcus Nyström, Katharina Scheiter, Peter Gerjets, and Berit Eika. 2012. Conveying clinical reasoning based on visual observation via eye-movement modelling examples. *Instructional Science* 40, 5 (2012), 813–827.

Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. 2010a. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications*. ACM, 211–218.

Halszka Jarodzka, Katharina Scheiter, Peter Gerjets, Tamara van Gog, and Michael Dorr. 2010b. How to convey perceptual skills by displaying experts' gaze data. In *Proceedings of the 31st annual conference of the cognitive science society*. 2920–2925.

Sheree Josephson and Michael E Holmes. 2002. Visual attention to repeated internet images: testing the scanpath theory on the world wide web. In *Proceedings of the 2002 symposium on Eye tracking research & applications*. ACM, 43–49.

Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. 2003. The expertise reversal effect. *Educational psychologist* 38, 1 (2003), 23–31.

Ellen M Kok, Halszka Jarodzka, Anique BH de Bruin, Hussain AN BinAmir, Simon GF Robben, and Jeroen JG van Merriënboer. 2016. Systematic viewing in radiology: seeing more, missing less? *Advances in Health Sciences Education* 21, 1 (2016), 189–205.

Elizabeth A Krupinski, Allison A Tillack, Lynne Richter, Jeffrey T Henderson, Achyut K Bhattacharyya, Katherine M Scott, Anna R Graham, Michael R Descour, John R Davis, and Ronald S Weinstein. 2006. Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human pathology* 37, 12 (2006), 1543–1556.

Thomas Kübler, Shahram Eivazi, and Enkelejda Kasneci. 2015. Automated visual scanpath analysis reveals the expertise level of micro-neurosurgeons. In *MICCAI Workshop on Interventional Microscopy*.

Thomas C Kübler, Enkelejda Kasneci, and Wolfgang Rosenstiel. 2014. Subsmatch: Scanpath similarity in dynamic scenes based on subsequence frequencies. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 319–322.

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.

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.

Harold L Kundel, Calvin F Nodine, Emily F Conant, and Susan P Weinstein. 2007. Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology* 242, 2 (2007), 396–402.

Harold L Kundel, Calvin F Nodine, Elizabeth A Krupinski, and Claudia Mello-Thoms. 2008. Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. *Academic radiology* 15, 7 (2008), 881–886.

Aidan Moran, Alison Byrne, and Nicola McGlade. 2002. The effects of anxiety and strategic planning on visual search behaviour. *Journal of sports sciences* 20, 3 (2002), 225–236.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48, 3 (1970), 443–453.

Calvin F Nodine, Harold L Kundel, Sherri C Lauver, and Lawrence C Toto. 1996. Nature of expertise in searching mammograms for breast masses. *Academic radiology* 3, 12 (1996), 1000–1006.

Bing Pan, Helene A Hembrooke, Geri K Gay, Laura A Granka, Matthew K Feusner, and Jill K Newman. 2004. The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 symposium on Eye tracking research & applications*. ACM, 147–154.

Eyal M Reingold, Neil Charness, Marc Pomplun, and Dave M Stampe. 2001. Visual span in expert chess players: Evidence from eye movements. *Psychological Science* 12, 1 (2001), 48–55.

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.

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.

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.

A Van der Gijp, MF Van der Schaaf, IC Van der Schaaf, JCBM Huige, CJ Ravesloot, JPJ van Schaik, and Th J ten Cate. 2014. Interpretation of radiological images: towards a framework of knowledge and skills. *Advances in Health Sciences Education* 19, 4 (2014), 565–580.