Gaze and visual scanpath features for data-driven expertise recognition in medical image inspection.

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Nora Jane Castner
aus Flemington, New Jersey

Tübingen
2020

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 23.10.2020

	Stellvertretender Dekan:	Prof. Dr. József Fortágh
1.	Berichterstatter:	Prof. Dr. Enkelejda Kasneci
2.	Berichterstatter:	Jun.-Prof. Dr. Michael Krone

Acknowledgments

Throughout the past three years of my doctoral research, the knowledge gained from the challenges and unique oppourtunities would not be possible without the support from others.

I wish to express my sincere appreciation to my supervisor, Prof. Dr. Enkelejda Kasneci, whose guidance and ambition has been an invaluable assest in this work as well as in life. I am also thankful to the late Prof. Dr. Rosenstiel for his consistent support during this project and his work with the EAES Doctoral program. I would also like to thank Jun.-Prof. Dr. Krone for his continued support throughout this dissertation.

I would like to acknowledge the significant interdisciplinary contributions to this project, Priv.-Doz. Dr. Fabian Hüttig, Dr. Dr. Constanze Keutel, Prof. Dr. Katharina Scheiter, Dr. Juliane Richter and Thérése Eder. They have offered not only crucial knowledge from their own fields of expertise, but also creative ideas, critical feedback, and dedication. I would also like to thank Prof. Dr. Andrew Duchowski for his collaboration and counsel on an aspect of my research that was fairly new to me.

I thank my fellow colleagues past and present – Dr. Thomas Kübler, Dr. Wolfgang Fuhl, Dr. Thiago Santini, and Dr. David Geisler – I am indebted to you guys for taking me under you wing. You gave me great confidence in knowing that there were no stupid questions and you always gave me the time to help me understand complex topics. I am also grateful for the rest of the Human-Computer Interaction team – Benedikt Hosp, Tobias Appel, Dr. Shahram Eivazi, and others. Coffee, stimulating conversations, and even Thanksgiving are one level of camaraderie, but those collaborative late nights for deadlines is really what binds us.

Moreover, I would like to thank my parents John and Tara and my brother Jeff, and Erin and Curtis and Aunt Jane. You all have supported me in every aspect of my life and contributed to who I am. To all my extended family, you have also always been there for me and your vacations to Tübingen have always provided a wonderful break from work and the normal routine. Finally, I am deeply grateful to my proof-readers and my mental support through this research, Kai, Greer, and many more.

Summary

Expert medical professionals must visually examine medical images (MRI and CT scans, radiographs, ultrasounds etc.) with the utmost concern for a patient's health. Developing the perceptual abilities to distinguish an atypical shadow from an anatomical structure involves considerable training and time. Although students view a multitude of these images in their studies, often, they must receive further supervision upon entering their residencies or even early on in their careers. This current approach can exhaust expert resources allocated for supervision and leaves room for error.

This thesis sets out to investigate the gaze behavior as an effective tool for expert and novice anomaly recognition, specifically in the context of dental image inspection (Technical term: orthopantomograms, or OPTs). Our ability to go deeper into the predictive aspect of scanpath analysis makes our research truly innovative. Much of the current literature regarding experts and novices has found that domain specific tasks evoke different eye movements. However, research has yet to predict these behaviors and guide students towards expert behavior strategies. More important, advanced pattern recognition and analysis algorithms have not yet been employed to identify and quantify differences in the visual search strategy between advanced learners, residents, and expert practitioners.

The potential to integrate expertise model development from scanpath features into intelligent tutoring systems is the ultimate inspiration for our research. This novel approach to training dentistry students with gaze-based learning environments can offer insight into the training of students in other medical domains. Currently, the training of OPT interpretation in dental students exhibits a deficit of systematic learning approaches and can vary between universities. Moreover, there are no known user-aware intervention techniques that address the improvement of image reading performance in students or advanced learners.

By employing machine learning-based scanpath classification, we found features in the gaze indicative of expertise and expert cognitive processes. We were also able to distinguish gaze behavior related to a student's level of understanding. The culmination of these findings provide support for a robust classification algorithm we developed to extract semantic features of the gaze and cluster experts and novices based on feature similarities in the scanpath with high accuracy.

Zusammenfassung

Diagnoseprozesse in der Medizin basieren zunehmend auf bildgebenden Verfahren, wie z.B. Magnetresonanztomographie, Computertomographie, Röntgenaufnahmen oder Ultraschall. Effiziente Entwicklung von Interpretationskompetenz, um beispielweise einen atypischen Schatten von einer anatomischen Struktur zu unterscheiden, erfordert umfangreiches Training und viel Zeit. Obwohl Studierende während ihrer Ausbilung eine Vielzahl solcher Bilder betrachten und zu interpretieren lernen, müssen sie beim Eintritt in die Berufspraxis von Experten weiterbetreut werden. Dieser derzeit gängige Ansatz kann jedoch Expertenressourcen für die Betreuung erschöpfen und lässt Raum für Diagnosefehler.

Aktuelle Literatur über die Manifestation von Expertise bei Studien mit medizinischen Experten und Studierenden berichtet von unterschiedlichen und domänenspezifischen Augenbewegungen. Obwohl die zugrundeliegenden Verhaltensweisen noch nicht abschließend geklärt sind, könnten durch datengetriebene Analyseverfahren und Mustererkennungsalgorithmen erfolgreiche Suchstrategien identifiziert und quantifiziert werden.

In dieser Arbeit untersucht daher zunächst das Blickverhalten als wirksames Instrument Feststellung medizinischer Expertise, insbesondere im Zusammenhang mit der zahnärztlichen Bildinspektion (Fachbegriff: Orthopantomogramme oder OPGs, engl.: OPTs). Gegenwärtig variieren Ausbildungsverfahren in den Zahnmedizin Studiengängen hinsichtlich der Interpretation von OPTs von Universität zu Universität. Es mangelt an systematischen Lernansätzen. Wichtiger noch, es sind keine auf Nutzerverhalten basierenden Interventionstechniken bekannt, die sich mit der Verbesserung der Leistung von Studierenden oder gar Fortgeschrittenen befassen.

Durch den Einsatz der auf maschinellem Lernen basierenden Klassifikation visueller Suchpfade (sog. Scanpaths) konnten im Rahmen dieser Arbeit Merkmale im Blickverhalten, die auf Fachwissen und Expertise hinweisen, identifiziert werden. Basierend auf diesen Ergebnissen wurde ein robuster Klassifikationsalgorithmus entwickelt, um anhand von semantischen Blickbewegungsmerkmalen Experten und Novizen mit hoher Genauigkeit zu kennen.

Diese Forschung stellt eine wichtige Basis für die Entwicklung von Expertenmodellen aus Scanpath-Features dar und dient letztlich der Integration in intelligente Tutoring-Systeme. Dieser Ansatz bietet potential für Ausbildung mithilfe blickbasierter Lernumgebungen kann nicht nur über den zahnmedizinischen Anwendungsbereich hinaus Erkenntnisse für die personalisierte Ausbildung von Studierenden in anderen medizinischen Bereichen geben, sondern auch Hinweise zur Optimierung von automatisierten Entscheidungsunterstützungssystemen liefern.

Contents

List of Figures

List of Tables

List of Abbreviations

ACM	Association for Computer Machinery
AI	Artifical Intelligence
AOI	Area of Interest
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
CT	Computed Tomography
ECG	Electrocardiogram
ETRA	Eye Tracking and Research Applications
GAN	Generative Adversarial Network
GBVS	Graph-Based Visual Saliency
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HMM	Hidden Markov Model
I-VT	Velocity-Threshold Fixation Identification
LSTM	Long Short-Term Memory
MRI	Magnetic Resonance Imaging
OPT	Orthopantomographs
PNG	Portable Network Graphic
PPV	Positive Predictive Value
RNN	Recurrent Neural Network
ReLu	Rectified Linear Unit
RQA	Recurrence Quantification Analysis
SGD	Subtle Gaze Direction
SMI	SensoMotoric Instruments
SS & WS	Summer & Winter Semester
SVM	Support Vector Machine
TPR	True Postive Rate

1 List of Publications

Publications relevant to thesis

[1] *Scanpath comparison in medical image reading skills of dental students.* **Castner, N.**, Kasneci, E., Kübler, T., Scheiter, K., Richter, J., Eder, T., Hüttig, F., & Keutel, C. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (2018)*.

[2] *Development and Evaluation of a Gaze Feedback System Integrated into EyeTrace.* Otto, K., **Castner, N.**, Geisler, D., & Kasneci, E. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (2018)*.

[3] *Overlooking: The nature of gaze behavior and anomaly detection in expert dentists.* **Castner, N.**, Klepper, S., Kopnarski, K., Hüttig, F., Keutel, C., Scheiter, K., Richter, J., Eder, T., & Kasneci, E. *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data (2018)*.

[4] *Pupil diameter differentiates expertise in dental radiography visual search.* **Castner, N.**, Appel, T., Eder, T., Richter, J., Scheiter, K., Keutel, C., Hüttig, F., Duchowski, A., & Kasneci, E. *Plos One (2020)*.

[5] *Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing.* **Castner, N.**, Kübler, T., Scheiter, K., Richter, J., Eder, T., Hüttig, F., Keutel, C. & Kasneci, E. *Proceedings of the 2020 ACM Symposium on Eye Tracking Research & Applications (2020)*.

Accepted Articles

[6] *Towards expert gaze modeling and recognition of a user's attention in realtime.* **Castner, N.**, Gessler, L., Geisler, D., Hüttig, F., & Kasneci, E. *24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (2020)*.

Publications relevant to fundamentals

[7] *How to support dental students in reading radiographs: effects of a gaze-based compare-and-contrast intervention.* Eder, T., Richter, J., Scheiter, K., Keutel, C., **Castner, N.**, Kasneci, E., & Hüttig, F. *Advances in Health Sciences Education: Theory and Practice (2020)*.

[8] *Histogram of Oriented Velocities for Eye Movement Detection.* Fuhl, W., **Castner, N.**, & Kasneci, E. *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data (2018)*.

[9] *Rule-Based Learning for Eye Movement Type Detection.* Fuhl, W., **Castner, N.**, & Kasneci, E. *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data (2018)*.

[10] *Ferns for Area of Interest Free Scanpath Classification.* Fuhl, W., **Castner, N.**, Kübler, T., Lotz, A., Rosenstiel, W., & Kasneci, E. *Proceedings of the 2019 ACM Symposium on Eye Tracking Research & Applications (2019)*.

[11] *Encodji: Encoding Gaze Data into Emoji Space for an Amusing Scanpath Classification Approach ;).* Fuhl, W., Bozkir, E., Hosp, B., **Castner, N.**, Geisler, D., Santini, T.C., & Kasneci, E. *Proceedings of the 2019 ACM Symposium on Eye Tracking Research & Applications (2019)*.

[12] *A MinHash Approach for Fast Scanpath Classification.* Geisler, D., **Castner, N.**, Kasneci, G., & Kasneci, E. *Proceedings of the 2020 ACM Symposium on Eye Tracking Research & Applications (2020)*.

[13] *Exploiting the GBVS for Saliency Aware Gaze Heatmaps.* Geisler, D., Weber, D., **Castner, N.**, & Kasneci, E. *Proceedings of the 2020 ACM Symposium on Eye Tracking Research & Applications (2020)*.

Scientific Contribution

This work promotes the use of gaze as a distinguishing feature in expert and novice recognition in the context of medical image inspection. Six scientific publications from 2018 to 2020 support gaze behavior related to expert and novice decision making as well as scanpath classification, and are each detailed in chapter 5. In addition to these publications, the thesis author has 7 additional scientific publications that contribute to fundamental aspects in eye tracking data analysis. Three of these publications are method-based contributions to state of the art scanpath analysis and are detailed in chapter 3.

2 Introduction

> We are all experts in our own little niches
>
> –*Alex Trebek*

The focus of this work is eye movements as a feature for expertise distinction. Specifically, how to use gaze to measure expertise and its development from a data driven perspective. This thesis will start with a brief overview of expertise research in section 2.1. Although for a more comprehensive investigation of expertise from the psychological and educational perspective, the author recommends *The Cambridge Handbook of Expertise and Expert Performance* [14]. The following sections in this chapter will also provide an overview of eye movement behavior (section 2.2), specifically how these movements relate to expertise and expert visual search (section 2.3), and how these gaze features can be a guiding mechanism in the educational context (section 2.4). Chapter 3 provides an overview of the state of the art in scanpath analysis, with a deeper investigation into medical expert recognition (section 3.5). Following these overviews, chapter 4 lays out the current objectives of this thesis, while the major findings are discussed in detail in chapter 5. Chapter 6 concludes with the outlook and further implications of this research.

2.1 A brief overview of expertise, its development, and the visual aspects

Experts are renowned for their abilities. Many novices work towards becoming experts. While the actual development of expertise is highly examined, the simple question often remains: How does one *really* become an expert? Ten years[1] or 10,000 hours[2] of practicing a trade are the common colloquial suggestions. Indeed, developing the knowledge and the skill set to be an expert takes time, but many factors are involved to make one truly stand out as an expert in their field.

Someone is classified as an expert if they " ... are recognized within their profession as having the necessary skills and abilities to perform at the highest level" according to Shanteau [17, p. 255], i.e. judged on exceptional performance [18]–[20]. It is evident that experts perform faster and more accurately than their novice

[1] First suggested by Bryan and Harter in 1899 [15].
[2] First suggested by Chase and Simon in 1973 [16].

counterparts in domain-specific tasks [19], [21]–[25]. Compared to novices, expert chess players can determine a winning strategy more quickly [16], [26], [27], expert radiologists can more rapidly recognize anomalies [28]–[31], and expert athletes score more points in a competition [32]–[35]. Peers, researchers, and laymen recognize this superior performance, but the underlying mechanics are not as easily perceived. The actions – cognitive and procedural – that dictate success need to be understood as well.

Other definitions of expertise attempt to explain superior performance by focusing on the inner cognitive features, where intuition, effortlessness, and optimal use of minimal resources are essential [21], [36]–[38]. Naturally, a cognitive model is harder to evaluate than a tangible metric such as performance. Some techniques that make the cognitive process apparent are self-reports, think aloud protocols [39]–[41], and testing memory and recall [16], [42], [43]. For example, due to effective recall, experts can view a radiograph for just a few milliseconds and tell if there was an anomaly present with extremely high accuracy [44], [45].

The differences between experts and novices are attributed to more structured cognitive approaches employed by experts, whereas novices lack the knowledge and practice [24], [29], [46]. Then, as expertise develops, so does the cognitive ability, which affects fluency in operations, resistance to distraction, and dual-task ability [47]. For example, expert microsurgeons have more stable hand movements during surgical microscopy procedures compared to novices [48], [49]. Moreover, expert cognitive ability facilitates focus on relevant features for a cohesive decision (see example in figure 2.1). The extent of an expert's focus is manifested in their visual behavior, and many expert domains rely on effective visual processing [24].

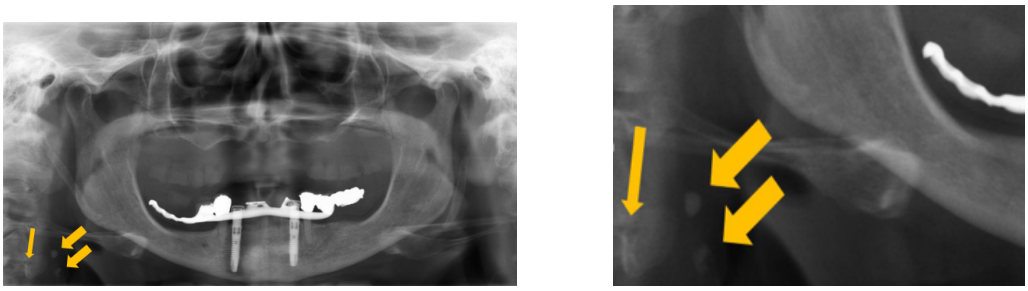


Figure 2.1: Examples of findings that possibly indicate calcification of the carotid arteria (diagonal arrows). Novices may focus on non-relevant features, i.e. the dental status and the setting of the teeth within the bone. An expert observer might be interested in the situation of the canalis mandibularis for anaesthesia or implantology reasons (down pointing orange arrow). However, findings such as this calcification are often overlooked.

Experts look at their work differently than novices or laymen. They also perceive differently than novices [50]–[52], e.g. expert radiologists are more sensitive to

low contrast features in radiographs. Perceptual expertise is an inherent quality in many medical professionals. Their extensive training results in fast recognition and decision making with high sensitivity and specificity. Often, medical experts tackle complicated situations – e.g. cases, operations, diagnoses – where recognizing the best solution is a challenge. One area prone to error is radiological image interpretation [53]–[56]. For instance, statistics on diagnostic error show 10-30 % of overlooked breast cancer recognition in mammograms [57] and 45-50 % of errors in lesion and nodule detection in chest radiographs [58], [59] (See [55], [60] for further statistics on diagnostic error).

Radiographs are highly prevalent in dental medicine [61]–[66]. Orthopantomographs – or panoramic radiographs (OPTs) – are complex projections of the maxillomandibular region on a single film. OPTs include the entire dentition, alveolar bone, temporomandibular joints as well as adjacent structures such as maxillary sinus, hyoid, styloid processes, vertebral bodies, and even soft tissues of the head and neck (esp. arteries and adenoids) [67] (see left figure 2.1 for example). Detection of certain anomalies (e.g. caries on the molars) is often harder in OPTs compared to other dental radiographs that capture only a portion of the teeth [68]–[71]. Under detection and misdiagnosis for other anomalies (e.g. periapical lesions) can be as high as 20 % [72], [73]. OPTs are also more prone to technological errors (e.g. positional errors, exposure/contrast issues) [64], [66], [74]. However, given the complicated nature of OPTs and radiographs in general, experts still outperform their novice counterparts [53], [75], [76].

An expert's visual search strategy offers insight into his or her thought processes, which provides a new means for teaching and improving expertise. The end goal of understanding expert cognition has always been to measure progress [46]. Expert gaze models can augment the well-known mantra of *practice makes perfect* by bringing the focus to relevant components in a task. Assessing progress, however, is twofold: Providing an expert model and recognizing a novice's understanding of it. Therefore, gaze-based intelligent tutoring systems can bring together traditional targeted practice and user-awareness. In order to work successfully, these systems need to both evaluate gaze behavior automatically and detect strategic patterns from that behavior. Thus, expertise recognition through scanpath analysis is a crucial step toward gaze-based training. Robust recognition of a student's level of understanding through gaze can provide the appropriate level of content. This solution has the ability to smooth the transition between residency and professional environments for students by minimizing the knowledge gap. More important, human expert resources, often allocated to time-consuming training efforts, are freed up for actual work.

2.2 Fundamentals of eye movements

Types of eye movements Eye movements offer insight into how we perceive the world. The eyes move to accommodate the vast wealth of information, since

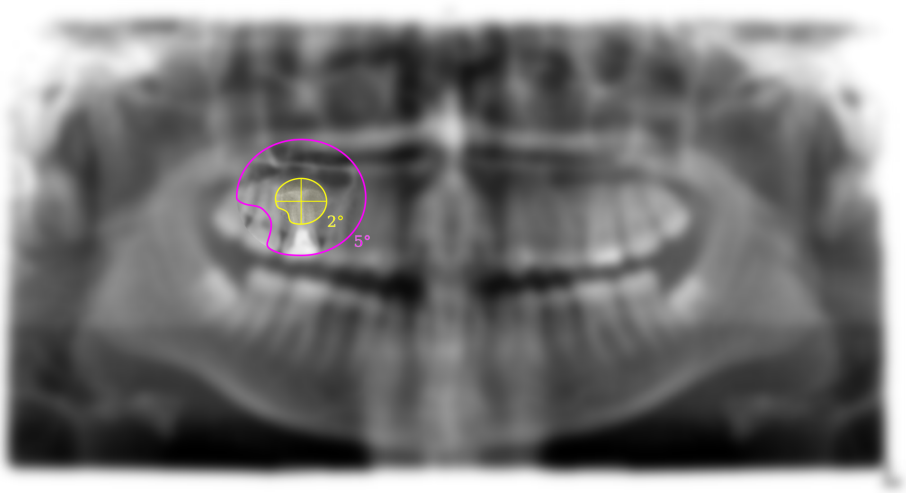


Figure 2.2: An example of the visual field, with foveal (our sharpest) vision less than 2° and para-foveal vision around 2-5°. At a viewing distance of 60cm to a 17″1920 × 1080 pixel monitor, 100 pixels is approximately 1.82° of our visual field.

the sharpest vision – from the fovea on our retina – is less than 2° (illustrated in figure 2.2 in yellow) of the visual field [77]–[80]. With increasing eccentricity from foveal vision, visual acuity decreases [81]. At roughly 2 - 5° eccentricity is para-foveal vision (illustrated in figure 2.2 in pink), where simple features (e.g. orientation, shape, texture) can be recognized without high acuity [82], [83]. The remaining visual field is the periphery, with the lowest visual acuity and with the least color sensitivity [78], [80], [81], [84]. Additionally, we have a blind spot (approx. 10 - 15°) where the optic nerve originates at the retina, which corresponds to the nasal part in our field of view [78]. If we take the example in figure 2.2 of a dentist inspecting an OPT sitting approximately 60cm away from a 17″monitor with 1920 × 1080 pixel dimension, foveal vision (shown in yellow) would be less than 110 pixels. This area does not include a whole tooth. The rest of the visual input would appear blurred, and in para-foveal vision (approximately within 275 pixels), only the tooth outline and a neighboring tooth is distinguishable. Considering a radiograph's grayscale format and potential for technical errors, the idea of accurately detecting an anomaly based on a small window for sharp feature recognition becomes hard to fathom for a naive viewer.

Since nearly half the visual information sent to our brain comes from the fovea, the eyes move to encompass 170° vertical and 200° horizontal of the visual field [77]–[79]. Fixations occur when the eye remains relatively stable to perceive available visual information. More information can influence the duration of the fixa-

(a) Original image. (b) The salient features. (c) Given the task: "Find the cat".

Figure 2.3: Visual processing styles of the *The overturned Bouquet* painting by Abraham Mignon (1660-1679), oil on canvas. public domain `https://commons.wikimedia.org/wiki/`.

tion [80], [85], [86].[3] Saccades are the rapid eye movements to a new area, occurring between fixations. They can occur at velocities of 30 to 500°/s and our brains do not process the visual intake during this movement [80]. Thresholds for fixation/saccade detection can be temporal and dispersion- or velocity-based [85]. Recently, event detection approaches that are Bayesian [87]–[90] or machine learning-based, such as the approaches by Fuhl, **Castner** et al. [8], [9] can recognize task dependent events. Other eye movements such as microsaccades, smooth pursuits, glissades, vestibular-ocular reflex, etc. are not the focus to this work, though for additional details the reader should see Holmqvist et al. [80].

Eye movements linked to attention and processing Eye movements allude to key aspects in perception, with fixations providing the most basic unit for understanding visual attention [79], [91], [92]. Attention is organized by two intertwined processes. *Bottom-up* processing is linked to salient features that exogenously catch our initial attention. For example, the painting in figure 2.3a has bright flowers that pop out against a shadowed background. Saliency-models, based on this processing style, dictate that color and contrast features (e.g. figure 2.3b) would be the first to capture our attention [93], [94]. This model of attention processing is biologically based, but lacks higher cognitive processes that are context dependent [94], [95]. It is argued that in visual search, early fixations may be more bottom-up inclined, but through the course of the search task demands influence the gaze pattern to a greater extent [95], [96].

When medical professionals inspect a radiograph, their expertise helps them know what to look for. This context-dependent gaze behavior is the main feature

[3] In general, there are other ways to define these eye movements that account for the type of task (i.e. fixations respective to areas or moving objects in the ego-perspective), but the current work follows these simple definitions.

of visual search, i.e. “Can we rule out cancer?” or, in the case of figure 2.3c, “find the cat”. This type of attentional processing is *top-down*, which can be knowledge or task dependent [95]. The pioneering works of Buswell [97] and Yarbus [98] found that gaze behavior adapts to differing context or task. This corresponding gaze to task behavior is known as the *scanpath*. Scanpaths during visual search show a relationship between saccade and fixation behavior, and adapt to changes in search strategy [86], [95], [96], [99], [100].

Pupillary response in visual processing Not only are eye movements indicative of how we process visual information, but the pupillary response can indicate cognitive processing. A greater dilation change from a baseline measurement can alludes to the requirement of more mental resources – or cognitive load – during the task [101]–[112]. Task difficulty as well as uncertainty affect cognitive load, where pupil diameter increases in visual search tasks when targets are hard to find [23], [113]–[116].

Due to these aspects, cognitive load has become a staple in evaluating learning environments for students [117]. Pupil diameter has shown to decrease with learning [107], [118]. Students may not be exposed to difficult tasks, but accumulate more experiences overtime, which can reduce cognitive load. Furthermore, pupillary response serves as a way to further distinguish the cognitive processes of experts and novices. For instance, experts show smaller pupillary response to domain tasks than novices [110], [119], [120]. However, even experts exhibit cognitive load when tasks become more difficult [108], [121]. One contribution of the current work (detailed in section 5.2) furthers this research by coupling expert pupillary response and fixations on relevant anomalies, whereas previous research has handled these components separately.

2.3 Expertise and eye movements: State of the art

Tracking a subject’s eye movements has become increasingly pervasive over the past twenty years [122]–[125]. Before, eye tracking was highly constrained to a lab – and even invasive at times. Now eye trackers can unobtrusively measure behavior in naturalistic settings. Research in the 1950s was first starting to recognize the expert gaze in sports [126] and aviation [127]. Since then it has spread to juggling [128], sailing [129], forgery detection [130], and organic chemistry [131], to name only a few. This overview is confined to medical expertise, with a stronger focus on medical image inspection. However, for a more comprehensive overview, the author refers the reader to Gegenfurtner et al. [24] and Brahms et al. [132].

Expertise differences in eye movement behavior can be explained against the backdrop of three theories. (1) *Long-Term Working memory* [37] justifies how experts have consolidated memory structures that allow for fast extraction of areas of importance and rapid analysis of these areas. (2) *Information-reduction hypothesis* [133] similarly indicates that experts are more attuned to relevant areas for the

problem at hand, and they effectively ignore irrelevant areas. (3) *Holistic image processing* [134] states that experts first rapidly obtain a global impression of the problem, by quick scanning of the whole, then they hone into areas, that require deeper investigation.

2.3.1 Expert fixation and saccades

The aforementioned theories all regard the rapid and accurate ability of experts to process task specific visual information [24]. The fact that average fixation duration during a task is shorter for experts than it is for novices further supports this assertion [24], [28], [56], [135]. For chest X-ray inspection, however, there is no difference in the overall fixation duration [28], [136], [137], though experts can still detect anomalies faster. This behavior could imply that other factors in medical image interpretation play a role.

Although the fixation count between experts and novices, in general, is similar, expert radiograph inspection is characterized by fewer fixations. Naturally, if an expert is faster at inspecting an image, less fixations occur as a result of less time spent on the task.

More interesting to expertise understanding is how they extract relevant information. Experts appear to form a global representation of the whole image at a glance [138]. They generally have a shorter time to first fixation on areas that are relevant to the task than novices, i.e. an anomaly in a radiograph [28], [30], [132], [134], [139]–[142]. This suggests that their experience and knowledge provides shortcuts that are more sophisticated than novice comprehension. Experts also have more fixations and of longer durations on relevant areas and less fixations and of shorter durations on irrelevant areas than novices [24], [132]. This can be attributed to reducing extraneous processing demands, i.e. cognitive load [133], [143], [144]. For example, in laparoscopic surgery, experts have more fixations on target locations than on the surgical tools being used, whereas novices shift their gaze more often between tools and target locations [145]–[148].

Specifically for medical image interpretation, image content has a significant impact on expert eye movements [24], [28], [137], [149]. Obvious and easy to spot anomalies do not require as many fixations for experts than harder to detect anomalies [28], [150], [151]. In mammograms, dental CTs, and OPTs, experts have less fixations for more obvious anomalies compared to novices, but have more fixations than novices for more subtle anomalies [149], [152], [153]. Conversely, in a periapical radiograph inspection study with a combined subject pool of experts and novices, first fixations and area revisits were highly affected by the image content, e.g. cavities or restorations [154]. This finding suggests that more obvious anomalies also direct initial attention and need further investigation, though it is unclear the extent to which expertise affects this behavior.

Generally, the literature has provided a clear picture of an expert novice distinction in overall fixation behavior. However, this distinction in the saccade behavior

is less apparent. Experts exhibit overall longer saccades [24], [28], [135], [136], [155].[4] Although one study found saccade velocities were lower for experts compared to novices and decreased as a result of increased years experience [156]. Concerning the number of saccades, experts need less saccades than novices to inspect chest x-rays as well as angiograms [56], [157]. Similar to fixation behavior, the image content has an effect on the saccades. For instance, experts performed larger saccades than novices to detect lung nodules or chest lesions in chest x-rays [137], [155], [158] and in slide microscopy [156], but shorter saccades to detect visceral abnormalities and enlarged lymph nodes in CT scans [136]. In contrast, experts had fewer saccades than novices when inspecting pathological OPTs regardless of difficultly, but this difference was not found when inspecting non-pathological OPTs [149].

Although research appears to have an extensive understanding of the differences between expert and novice eye movements, these summaries are still quite limited. One considerable limitation is the sample sizes investigated. The review by Gegenfurtner et al. [24] evaluated 73 sources and points out that the majority of these studies evaluate five, maybe ten experts. In general, experts are hard to acquire; they have busy schedules, and must devote resources to their professional domains. Additionally, there is always the concern that the investigated task [24], [159] or even the expert's own motivation will affect validity. Thus, one expert who does not take the task seriously can heavily affect the outcome of the research. Another limitation is the under-representation of intermediate eye movements [24], [28]. This information regarding the in-between stages can also be crucial in understanding student proficiency and developing appropriate learning interventions. In the present work these limitations are addressed by (1) investigating an extensive subject pool of students ranging from introductory to advanced in addition to multiple dental professionals and (2) distinguishing scanpath differences between the levels of students using state of the art classifications algorithms. The respective details can be found in sections 4.1 and 5.4.

2.3.2 State of the art in visual search strategy

Experts have shorter search times when inspecting medical images compared to novices [28], [65], [132], [137], [140], [142], [155], [158]. Additionally, they employ similar search strategies, such as a global-to-focal order, supporting the holistic processing theory [28], [56], [134], [137], [160]. Meaning, experts scan the outer periphery and central areas, then hone in on areas needing more scrutiny. This behavior has also been linked to effective visual search in general [96].[5] It is initialized with a period of long saccades and short fixations *to get the gist* [86]. Then, fixations become longer whereas saccadic amplitudes decrease when certain

[4] Either reported as saccadic amplitude in degrees, or length respective to image dimensions in pixels.

[5] Known in other literature areas as *ambient scene processing* [161].

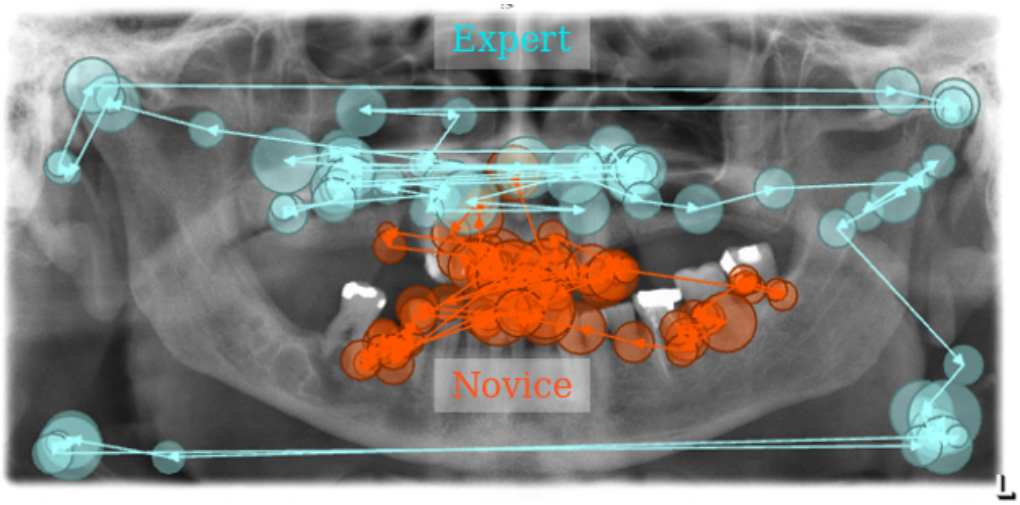


Figure 2.4: Differences in gaze behavior of the first 18 seconds of visual inspection of a panoramic radiograph. The gaze pattern in blue is from one of the expert dentists involved in the project and the gaze pattern in orange is from an incoming dental student (sixth semester).

features are further investigated [96].

Novices, on the other hand, tend to exhibit primarily focal search behavior [162]. They attend to more central and salient regions as exemplified by shorter saccade lengths and longer and more frequent fixations [28], [56], [137], [163]. This contrast in expert and novice search strategies is illustrated in figure 2.4. The expert scanpath – in blue – shows long saccadic sweeps of the peripheral structures before inspecting a certain region with multiple fixations, whereas the novice scanpath – in orange – appears to jump from tooth to tooth.

These search differences can be attributed to experts being more goal-driven, and novices being more stimulus-driven due to inexperience [156], [164], [165]. However, there are some cases where experts can also be saliency-driven, e.g. brain CTs, which creates a focal-then-global search strategy [166]. Thus, the spatial pattern of an expert search can reflect the specialty. Experts prefer circular patterns for mammograms [134], [139], though spiraling out for hand x-rays [167], or drilling downwards in 3D chest CTs [168], [169](See [56], [162], [165] for further descriptions of the search patterns during medical image inspection).

Specifically for dental radiographs, tooth-by-tooth and circular search strategies were preferred depending on the nature of anomalies present in periapical projections [154]. However for OPTs, it was found that more experienced clinicians employed more systematic scanning over less image areas for OPTs. Their less experienced counterparts covered more areas, but overall showed less of a clear scanning strategy. The main strategies they employed were spiraling inward (periphery areas first, then dental areas) and circular (going back and forth between

central and periphery) techniques [65]. These studies are only starting to suggest that gaze behavior in dental medical image interpretation is worth further investigation, branching away from generalizing gaze behavior over all medical images.

Not only can visual search patterns indicate expertise, it can also allude to the decision making process. Kundel et al. [58] recognized three types of diagnostic errors that were evident in the gaze. For instance, a search error was apparent when no fixations were on an anomaly. This was generally exhibited more so by novices [29], [56], [170]. Recognition errors are evident when there are a few fixations on an anomaly though it is not properly detected as such. Then, decision errors are evident when there are multiple fixations on an anomaly and, ultimately, it is still detected as a false negative. This type of fixation behavior has also been linked to uncertainty in medical diagnoses, e.g. the expert is unsure which pathologies to rule out [171]. To further understand experts' fixation patterns in relation to diagnostic decision making, the current work investigated glance frequencies for recognized and overlooked anomalies. The details for this work can be found in section 5.1.

2.4 Expert models for learning

Research in the field of expertise sets out to understand how factors contribute to expertise in order to improve teaching novices. The current training strategy is based on students analyzing and interpreting large quantities of cases that represent variations of normality and abnormality [29]. Evidence supports that this "massed practice" improves perceptual sensitivity [50], [172]. However, more refined training procedures are still scarce. Even though it has been available for decades, eye tracking has yet to deliver the promises for adaptive training.

One approach towards gaze-based training that has become more prominent is presenting a gaze model. In the educational literature, simply instructing systematic search effectively improves features of students' gaze behavior (e.g. image coverage, strategy), but not performance [158], [173], [174]. Building off this concept, showing an expert model (i.e. a scanpath illustration or dynamic representation it) improves performance, but only when coupled with expert instruction [174]–[176]. Similar findings were found concerning OPT inspection in Eder,...,**Castner** et al. [7]. They presented students with heatmaps modeling the gaze behavior of other students as well as a student's own gaze. Students were asked to compare their own gaze to that of the peer model. This instruction was found to improve image coverage and draw more attention to low-prevalent anomalies, which are generally harder for novices to recognize. Although anomaly detection did not improve, this approach promotes investigation into guiding attention towards features for training perceptual sensitivity. The current work follows up on this concept by developing an attention-aware adaptive gaze feedback, which is detailed in section 5.6.

Current gaze-based training of students can lack expert intention when directing a student's attention to certain regions rich in semantics. One effort towards smarter interaction with training systems is automated scanpath analysis, which aims at revealing intentions of the task. This information relies on expert intention recognition and effective extraction of semantics: Specifically, visual search and interpretation of radiographs for this work. However, in order to apply the scanpath information to subject prediction that is not constrained to one image, conventional approaches (described in chapter 3) are not feasible. Herein lies the necessity of the current work's contribution to scanpath classification using deep learning to extract semantic features (see section 5.5).

3 Scanpath Analysis

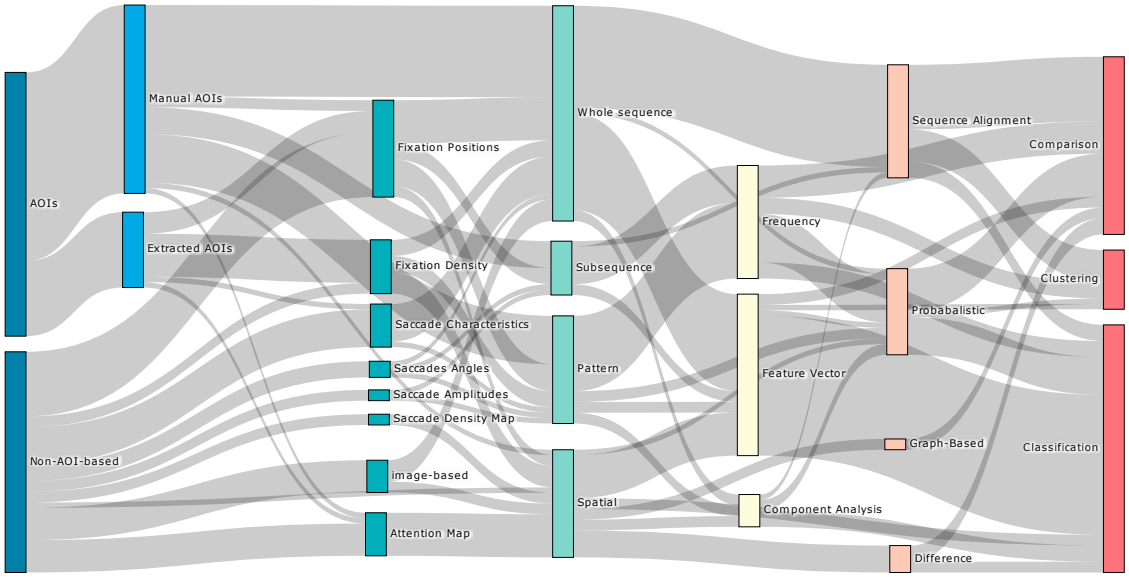
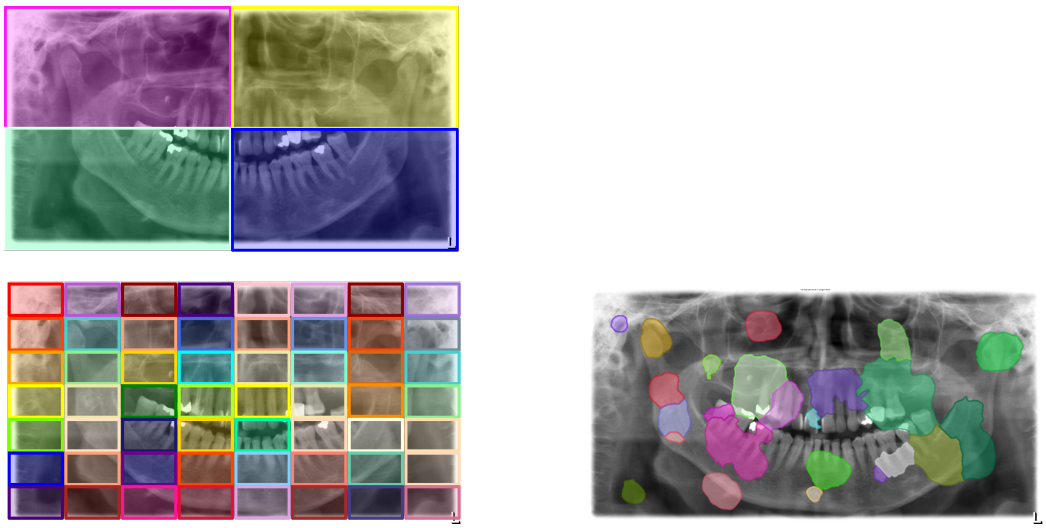


Figure 3.1: Broad overview of scanpath taxonomy over the last 33 years. 72 methods proposed and employed that deal with the scanpath as a temporal sequence of information are the input for this taxonomy and referenced in this chapter.

The literature of expert eye movement behavior involves high level abstraction: i.e. assuming complex cognitive processes from simple metrics such as fixation duration and saccade length. However, the scanpath differences can offer even further insight into the complex strategies that expert performance exhibits. At its core, scanpath comparison determines how similar one scanpath is to another. Already, figure 2.4 illustrates that there are two distinct scanpaths just from the spatial representation alone. Comparison can then extend to encompass more features to create a temporal, procedural, and even semantic understanding. Comparing multiple scanpaths can lead to groupings based on sequence similarity, patterns, etc. Scanpath classification then predicts which group a scanpath belongs to based on the learned patterns attributed to specific groups. The benefit of classification



(a) Gridded AOIs of varying granularity (b) AOIs segmented based on attention

Figure 3.2: AOI Types that are representative of spatial (left) of attentional (right).

is that models can be developed for online recognition [88], [89], [177], [178], i.e. novice level recognition to present the appropriate learning intervention.

Figure 3.1 illustrates the common approaches to scanpath analysis as comprised of the literature reviewed in this chapter, which goes back over 30 years. This review will only highlight traditional approaches to scanpath comparison, providing a stronger focus on the fundamentals for the results and discussion (Chapter 5) of this work. It will provide an overview of how spatial representation is handled with AOIs (section 3.1), which serves as the backbone for string alignment approaches (section 3.2). Then, how scanpath comparison extracts the subsequences to gain insight on the transitional behavior is detailed (section 3.3). This gives way to pattern recognition, which machine learning approaches attempt to automatically extrapolate (section 3.4). Recent approaches with deep learning have been able to successfully include attentional awareness and image semantics and are overviewed. Finally, the state of the art for medical expert scanpath classification is described (section 3.5). The contribution to the state of the art from the current work adds to this collection, though is described in detail in section 5.5. For more comprehensive overviews of the traditional approaches, the reader is referred to Anderson et al. [179], Dewurst et al. [180], Kübler et al. [181], Coutrot et al. [182], and Fahimi and Bruce [183].

3.1 Areas of interest

One of the more straightforward approaches to scanpath comparison is to delineate fixations relative to areas. These areas of interest (AOIs) can range from purely spatial to areas of attentional or semantic interest. The most low-level grid

construction of AOIs is depicted in figure 3.2a. These grid-based approaches are easy to implement and are stimuli independent, providing only fixation position relative to an area, i.e. top-right, center, left-middle, etc. However, the resolution of the grids can have an effect on scanpath comparison [184], [185]. For instance, figure 3.2a shows two different grid resolutions that would output completely different scanpath strings for the same scanpath.

AOIs can also be constructed from attentional information in many ways [186], [187]: Averaging scanpaths [188], [189], fixation density maps [190], or heatmap-gradient segmenting (as depicted in figure 3.2b) [191], [192]. Purely salience-based AOIs can also be produced from renowned models [186], [193], e.g. Itti-Koch [93] and GBVS [194]. They discern the bottom up-attentional properties, and offer a way to predict percievability [13], [51], [195]. For example, Geisler, Weber, **Castner** et al. [13] incorporated gaze behavior for the attention map calculation step in the GBVS model. This novel approach adds scene semantics to previously bottom-up models. More important, this approach and other attention approaches are data-driven, resolving the limitations that grid-based AOIs present in scanpath analysis [186], [196].

To provide high-level task information, AOIs can be segmented based on the image semantics. In the two images in figure 3.3, we see two levels of task semantics: AOIs based on the teeth and jaw structures visible in an OPT (top) and AOIs indicating specific pathologies – and their level of difficulty to detect (bottom). The latter AOIs were defined by the two expert dentists involved in this project. Herein lies the limitation of researcher subjectivity. AOIs respective to pathological errors require an expert or experts to point out, and given the statistics for underdections in radiology, this can affect your semantic segmentation. Overall, they are more tedious because they are manually defined and may not transfer to all stimuli [10], [184].

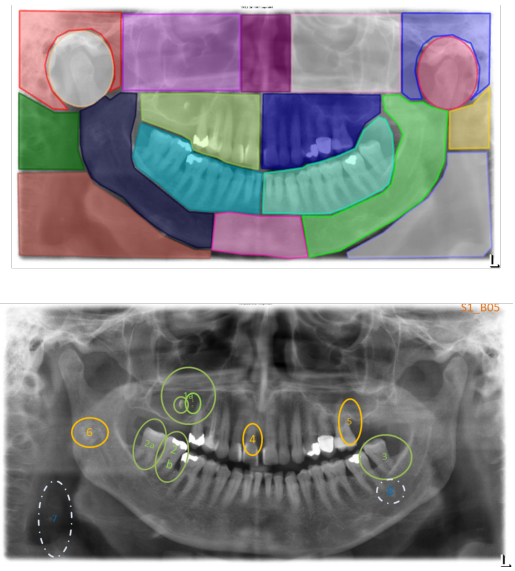


Figure 3.3: Semantic AOIs

3.2 String alignment

One approach to scanpath comparison is to characterize the scanpaths as an AOI-representative string sequence as illustrated in figure 3.4. This approach is one of the more prominent, even though it was originally developed for gene sequence alignment in bioinformatics [179], [197], [198].

The most basic metric is the Levenshtein Distance [199], where a cost is calculated based on the minimum number of transformations needed to make the strings similar [179], [197], [200], [201]. The Levenshtein distance has been used for scanpath comparison in web page inspection [202], face perception in autism [203], problem solving strategies [204], map reading [205], and systematic mammogram inspection [158]. However, such a simple character comparison is limited because it scores on the basis of match/mismatch and can overlook similarities in the scanpaths that happen on different timescales.

Building on similar concepts, global string alignment uses dynamic programming to determine the most optimal alignment for the entirety of two sequences. Here, one of the most dependable algorithms is the Needleman-Wunsch [198]. It consists of three steps: ① Matrix initialization,② similarity scoring, and ③ traceback. In ①, a scoring matrix, $M(m+1, n+1)$ is created, where m is the length of one scanpath (A) and n is the length of the other scanpath (B). The first row and column of M are then filled with the gap penalty. This penalty is the cost (gap) of adding a space in one or the other string for proper alignment. Then in ②, scoring the character i in scanpath A against character j in scanpath B proceeds iteratively using equation 3.1 sans stop criteria:

$$M_{ij} = \max \begin{cases} M_{i-1,j-1} + \ S(n_i, m_j), & \text{Match} \\ M_{i,j-1} + gap, & \text{Gap in A} \\ M_{i-1,j} + gap, & \text{Gap in B} \\ 0 & \text{Stop criteria.} \end{cases} \tag{3.1}$$

If the characters match, this is generally positively scored via S as in the first case of this equation, but if the characters do not match, a mismatch penalty takes the place of S. However, single character comparisons are not being assessed, rather their relative similarity is considered [206]–[208]. Due to this notion, the gap costs need to be considered, i.e. if the optimal alignment benefits from a gap in A or B. Hence, equation 3.1 takes the maximum of the cases at each comparison. Once the scoring matrix is filled, step ③ outputs the similarity score ($M_{m+1,n+1}$) and proceeds backwards to achieve the most optimal alignment given the direction of the scores. This scoring system offers more flexibility, such as limiting the penalties for either gaps or mismatches [206], [207], [209]. Taking the scanpaths in figure 3.4, and using 1, -1, and -2 for match, mismatch, and gap respectively, we can see the global alignment (in blue), with two mismatches and one gap to determine the best alignment.

The Needleman-Wunsch has been used for scanpath comparison in decision-making strategies [207], [210] and programming expertise [185], [211]. An implementation of the Needleman-Wunsch, *ScanMatch* [200] has been successfully used for physics problem solving strategies [212], surveillance [213], and webpage inspection [202]. The benefit of this approach is the overall impression of how similar two sequences really are, which compared over multiple scanpaths, offers information to aid in, i.e., clustering. However, the Needleman-Wunsch strug-

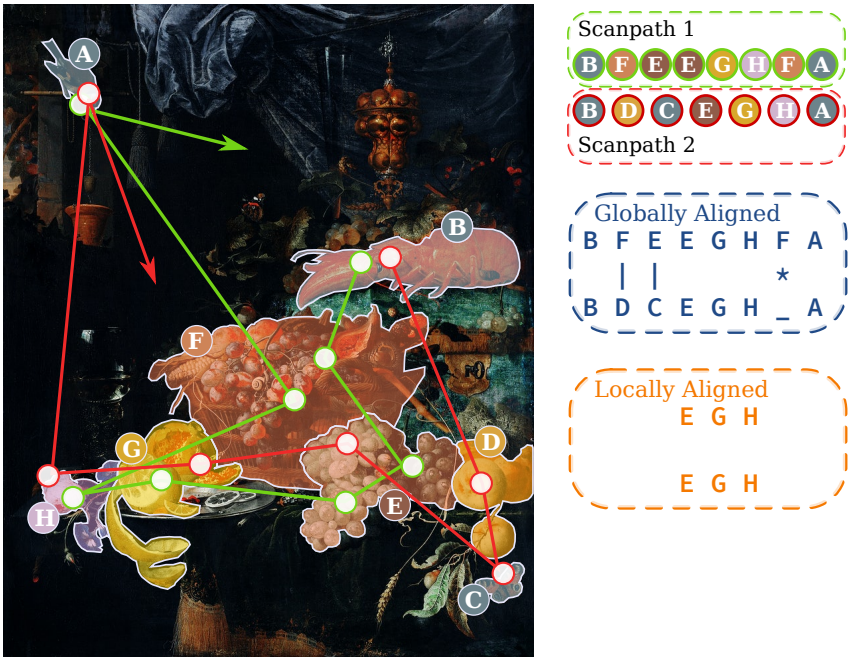


Figure 3.4: Two scanpaths represented by a sequence of characters. Comparison of these two scanpaths can be performed by string alignment approaches: Global alignment (blue) or local alignment (orange). *Still life with fruit, a lobster and a goldfinch* painting by Abraham Mignon (1660-1679), oil on canvas. public domain `https://commons.wikimedia.org/wiki/`

gles with normalizing scores when comparing different length scanpaths, which is often done by dividing the score by the length of the longer sequence.

Rather than deal with the entirety of two sequences, local alignment determines the most optimal aligned subsequence as illustrated in orange in figure 3.4. The most common local alignment is the Smith-Waterman algorithm [214]. Using the same equation 3.1, this algorithm adds an extra case, 0, to terminate the backtrace for subsequence extraction. Local alignment has also proved its versatility in scanpath comparison. For example, comparison of medical undergrads clinical reasoning performance [215] and aptitude in reading interactive map displays [216]. The benefit of this approach is that it compensates for sequences of differing lengths and temporal shifts, e.g. a similar pattern at the beginning of one scanpath and at the end of another can be coupled.

These alignment techniques, though commonly used, are simple to implement, but slow: Having $\mathcal{O}(nm)$ time and space complexity.[1] Given that most scanpath research involves comparing multiple scanpaths, processing increases quadratically. Additionally, they are restricted to the AOIs defined, which can be subjective [184],

[1] $\mathcal{O}(nm)$ overall, but also for filling the scoring matrix alone for both the Needleman-Wunsch and the Smith-Waterman Algorithms [206].

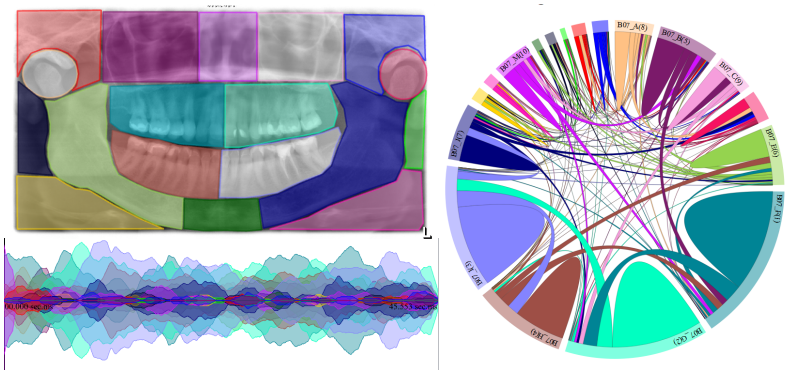


Figure 3.5: Expert AOI transitional behavior for control (no anomalies) OPT with transition behavior on the right and transition overtime behavior on the bottom left.

[185], [200], and the scoring and gap weights need to be set initially. Although these approaches generally do not compensate for fixation duration information, some approaches have incorporated temporal binning. *ScanMatch* [200], for example, solves the issue of fixation duration by adding the AOI label a number of times respective to time bins.

Given their limitations, string alignment algorithms are often the benchmark for many researchers. They do not require many parameters and, in the optimal case, extract scene or search semantics from AOIs. Ultimately, they work great for a large variety of tasks given they generate the best output from all possible combinations.

3.3 Transitional behavior

The transitional information can reveal the most common transitions and what catches the first fixation. Figure 3.5 shows how the transitional behavior for the AOIs (top left) can be visualized: An interpretation of a transition matrix[2] (right), and AOI transitions over time (bottom). Moreover, scanpath comparison can also be expressed in terms of AOI transitions overtime. For instance, Holmqvist et al. [218] used AOI transitional information with a sliding window to get sequence information for common and unique transition frequencies.

The transitional behavior of scanpaths can be best highlighted with pattern analysis, e.g. T-Pattern [219], [220] or Recurrence Quantification Analysis (RQA) [221]. These approaches point out areas of recurring fixations or fixation patterns and handle patterns that can be temporally shifted. Moreover, these approaches offer a new take on scanpath comparison that is not necessarily confined to AOIs. Rather, they have shown that transitions can also be extracted from spatial information [222] or Fixation Density [182], [223]. This transitional behavior can

[2]A transition matrix denotes the number of transitions from each AOI to all other AOIs [217].

further generate probabilistic states for e.g. Markov Models [182], [220], [224], component analysis [222], [225], and entropy- or graph-based approaches [226]–[229]. The elegance of these methods is that they recapture the long-term dynamics. Markov Model-based scanpath comparisons have been used for face processing [203], [230], task prediction [223], and image recognition [231].

Building off similar ideologies, prior assumptions from the transitional behavior moves scanpath comparison towards classification. Bayesian-based classification works with notion of the *likelihood that this scanpath belongs to a certain group given its features* [87], [232]. It has produced models for neurodegenerative disease recognition [233] and for detecting cognitive style [234] among other classification models [196], [235], [236].

3.4 Clustering and classification

There are machine learning models w.r.t. summary statistics (mean fixation count or saccade length, etc.) for feature input, but they lack abstraction regarding, i.e. variations during the search process (global vs. local) or detecting the task context [237]. Hence, this review will focus on scanpath classification methods that do not compromise the temporal integrity. There are two approaches: supervised, where the data labels are known, and unsupervised, where groups are created with no label information. Supervised learning can predict e.g. task type (e.g. the Yarbus task [98], [223], [237]–[239]), where unsupervised learning can cluster scanpaths based on dividing the data in a way that meaningful representations of groups become evident [240].

3.4.1 Sequence and transition features

String-alignment approaches can also be an informative feature for model learning. For instance, the Levenshtein distance has been used to classify problem solving strategy [204] and face processing in mentally disabled individuals [203], and for clustering based on general gaze similarity [205], [216]. The Needleman-Wunsch has also been used for problem-solving strategy classification [207], [210] and similarity clustering [216]. With clustering, common subsequence patterns and their relations can be easily visualized with a dendrogram from hierarchical clustering [241], [242] or be used to distinguish groups, i.e. air traffic control strategies of experts and novices [243]. One recent approach by Koch et al. [244] combined string-alignment scoring (tested the Levenshtein distance and Needleman-Wunsch) with image information (as slit-scans) for agglomerative hierarchical clustering [245], [246] and found accurate clustering of a memory recall task based on sequence similarity scores.

Recent approaches have introduced data mining techniques to scanpath analysis [12], [196], [241], [247]. For instance, *MinHash* from Geisler, **Castner**, et al. [12] tackles the issue of string alignment's time complexity by matching AOI

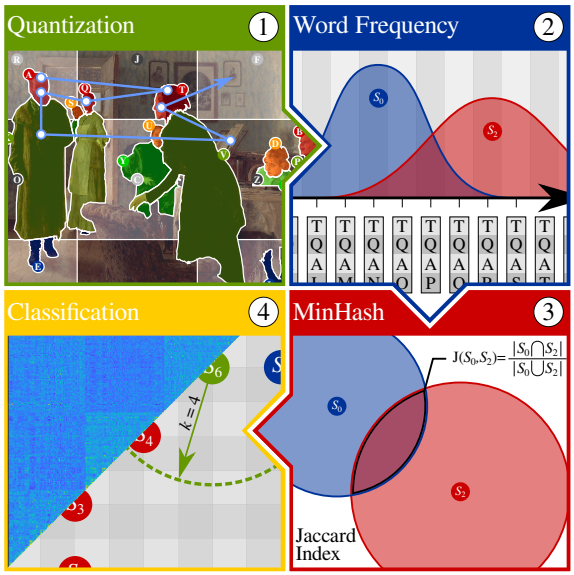


Figure 3.6: Outline of the *MinHash* algorithm for scanpath comparison from Geisler, Castner, Kasneci, and Kasneci [12]. *Best Paper award at ETRA 2020.*

subsequences using the same-named popular data mining approach [248]. The workflow of the algorithm – as shown in Figure 3.6 – uses string subsequencing (step ①) , to extract word frequencies (step ②) to go into a set of distinct hash functions. Here, runtime is dramatically decreased by merely assessing the word's frequency at matches in the strings' minimal hash response (step ③) [12]. Then, an approximation of the Jaccard Index is provided. The kNN classifies (step ④) an unseen scanpath's pairwise Jaccard index to a training/model set. This method of classification showed competitive results to a standard Needleman-Wunsch similarity score as input, yet with significantly less runtime [12].

However, AOIs being tedious and subjective should be avoided. There are some efforts that maintain string-based scanpath encoding but with a data-driven approach. For example, methods [250], [251] that use string-similarity clustering with attentional AOI extraction from [187]. One renowned metric is *SubsMatch 2.0* [181]. It couples sequence patterns and Support Vector Machines (SVMs).[3] It works with the features from its predecessor, *SubsMatch* [249], [254]. As illustrated in figure 3.7, it combines string representation with transition frequency analysis to handle multiple subsequent fixations, which can correspond to behavioral patterns [181]. Initially, a scanpath string is constructed by assigning letters (bins) to fixations in a way that the final scanpath string contains roughly the same number of occurrences of each letter (step 1.). Then, all possible subsequences of a given size (so-called n-grams, where n stands for the length of the sequence) and their occurrence frequencies are calculated (step 2.). A similarity metric between scanpaths can be calculated as the sum of differences between all subsequence frequencies (step 3.) [249]. Classification was realized in *SubsMatch*

[3] SVMs use a hyperplane with the intent of maximal feature separation and are known for being highly robust – even for small datasets; though they can rely on proper kernel selection [252], [253].

2.0, where the similarity metric was replaced by an SVM with a linear kernel that takes frequencies of n-grams as input. Fundamentally, it sets out to determine the best-fit subsequence length in conjunction with the best-fit string representation in order to perform classification based on subsequence occurrences. *SubsMatch* and *SubsMatch 2.0* show high generalizabitily over a range of tasks (i.e. Yarbus, SuperMarioKart, target detection [181]) and have also been used for classification of expert and novice microneurosurgeons [254], driving scenarios [249] and detection of distractions while driving [177].

3.4.2 Approaches that avoid AOIs

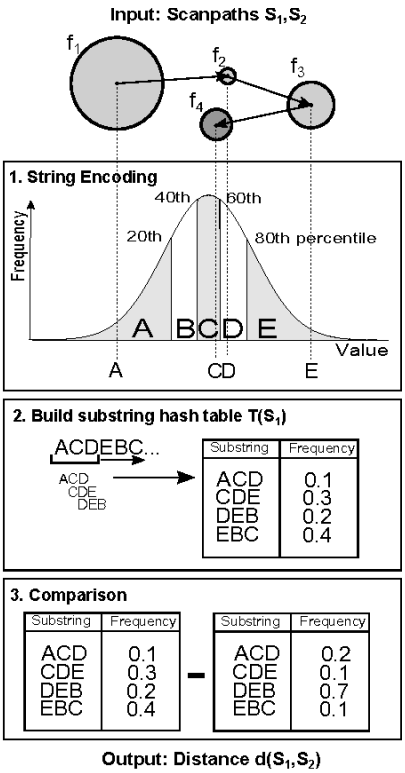


Figure 3.7: Workflow of *SubsMatch* [249]. Permission to use this image granted by the authors [181], [249].

Geometric and vector based Other scanpath classification methods have successfully employed SVMs without relying on AOIs, instead they use more geometric-based features. For instance, saccade vectors have been employed for detecting neurological disorders [255], reading behavior [256], [257], fatigue detection [258], and many others. *MultiMatch* [180], [184], [259] is a well-known scanpath comparison metric that is based on the scanpaths' geometry and saccade vector patterns. Therefore, it does not require AOIs and can handle spatial offsets that can arise from positional errors. Showing its versatility for comparison [260]–[262], it has also recently extended problem-solving classification with both an SVM and neural network [204].

Saccade sequence features have also proved to be a robust feature for clustering [263] and other classifiers. For instance, random forests have been used for scanpath-based preference classification [264] using saccades and spatial features at time segments.[4] Similarly, one approach by Fuhl, **Castner**, et al. [10] employed random ferns for scanpath classification, though for saccade sequence features (see work flow in figure 3.8). A fern represents features as a binary encoding at an index. Then, a conditional distribution is created, i.e. each class receives a probability based on the occurrences of features therein [267]–[269]. It

[4]Random Forests are an ensemble of decision trees, which can determine the best model [265], [266].

learns features from patterns in successive saccades. The features are then scored to obtain the best features representative of a class. The benefit of this approach is the use of saccade successions to avoid AOI dependencies and compensation of positional and rotational shifts.

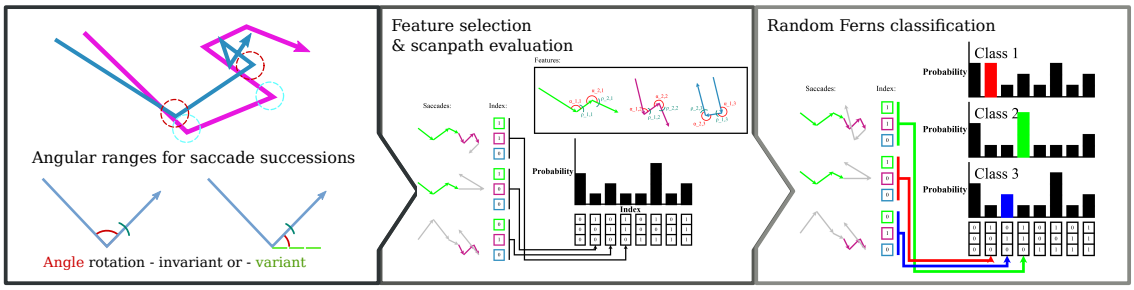


Figure 3.8: Workflow of scanpath classification with random ferns using saccade angle features from Fuhl, **Castner** et al. [10].

Spatial and attentional representation Another way to avoid AOIs yet still achieve a similar understanding of the fixation sequence is to go straight to the positional information. Such approaches regard the spatial information, such as positional features at temporal intervals for scanpath clustering [270] or classification [196], [222], [255], [264]. Therefore spatial representation over time builds off the early comparison methods that used the fixation distributions – or attention maps. Extracting the differences in these attention maps for scanpath analysis is a common approach (e.g. Attention maps [271], [272], iComp [251], iMap [189]), though out of the scope of this current review because they lack the temporal information of the scanpath.[5] However as previously mentioned, these maps hint at bottom-up scene semantics and perception. Using attention or saliency maps linked to a time period has also been for SVM classification in mental workload [258] and neurological disorders [255]. The saliency maps benefit scanpath classification by automating the feature selection. Moreover, this interplay of scene semantics and attentional effects provides input for deep learning that works towards human perception [274].

3.4.3 Deep learning approaches

More recently, the use of Neural Networks for research in object classification [275], image segmentation [276], [277], speech and gesture recognition [278], [279], and now scanpath classification has expanded over the last few years due to recent advances in hardware and data availability [280]. These networks are advantageous in that they can handle high dimensional data and are able to easily recognize patterns [281]. However, they are often a black box. Understanding how

[5] See [183], [273] for a more detailed overview.

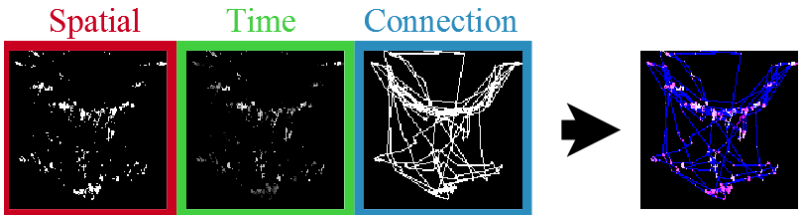
the model learned e.g. to predict the animal as a cat from the features cannot be deciphered. Neural Networks are layers of connected nodes, that take data in an input layer. Learning happens through the weighted connections throughout some hidden layers, and outputs the patterns recognized.

In scanpath classification, French et al. [204] classified adult and children gaze patterns using a feed-forward back propagation network with features from multi-dimensional scaling plots representing similarity. This approach was able to reduce error in the training weights (the back-propagation step) after forward traversal of the architecture. Another approach that can handle more sequential information in the input are recurrent neural networks (RNNs). One type of RNN that has been used quite recently in scanpath classification is Long-Short Term Memory (LSTM). These models have exhibited their aptitude for handling time series data and forecasting [282], [283]. Their architecture handles learning relevant information from long-term dependencies better [284]. For example, Sims and Conati [285] found that a simplified LSTM architecture was better at classifying confusion from scanpaths compared to a random forest classifier.

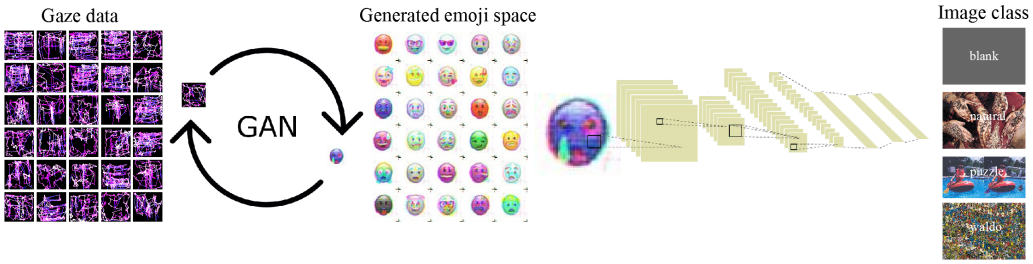
Convolutional Neural Networks (CNNs) [286] can provide information of image semantics that can be used for segmentation [287], [288] or classification [289] and saliency prediction [290], [291], and many other applications. In the field of eye-tracking research, they have also provided robust performance in eye movement behavior and scanpath generation [292], [293]. For instance, methods using probabilistic models and deep learning techniques coupled with ground truth gaze behavior have been shown to predict fixation behavior [294], [295] and regions of interest [296]. Mishra et al. [297] created images depicting scanpath information as input for a CNN sarcasm detector.

Both Tao and Shyu [298] and Chen and Zhao [299] employ CNN-LSTM networks that run on scanpath-based patches from a saliency-predicted map and classify autism spectrum disorder gaze behavior. In [298], square patches are defined based on fixation positions as they occur in the scanpath. Then, each patch is run through a shallow CNN, and the patch feature vector with the duration information provides an LSTM network input for classification from a dense layer from each patch. Additonally, Sodoké et al. [300] used eye movement sequences related to a number of tasks – defined by specific orders of AOI glances – as input for their CNN-LSTM for expert novice classification.

One limitation of CNNs can be their need for large amounts of annotated data for training. An alternative solution is to train CNNs to generate new training data out of existing images and easily generated data [301] – i.e., generative adversarial networks (GANs). Their training process requires an image generator and a discriminator that decides whether the generated image is real or simulated [301]. However, to reduce the effort required to determine the right learning parameters, a cyclic loss function [302] is included, which also makes unpaired training possible (i.e. unsupervised learning). Recently, GANs have been used for unsupervised feature learning from scanpaths for intention prediction [303], [304]. Fuhl et al. [11] combines this unsupervised approach to feature learning for further su-



(a) Raw gaze encoded as 3D image.



(b) Work flow from encodji [11].

Figure 3.9: Use of unsupervised learning to create a feature space (generated emojis in [11]) for task classification with a CNN

pervised task classification using a CNN. Initially, it encodes gaze data as a compact image (see figure 3.9a) with the spatial, temporal, and connectivity represented as pixel values in the red, green, and blue channels respectively – similar to [297]. However, the unconventional aspect to this approach is its use of a cycle GANs to generate emojis based on the scanpaths (see figure 3.9b). Afterwards, the generated Emojis are used for scanpath classification using a CNN to predict the task type.

Scanpath comparison and classification seeks to accurately determine task and group differences in the gaze behavior. Over the years, models have become more sophisticated to handle the a variety of tasks and subjects. The end goal is to employ these models for challenging real-world tasks, i.e. expertise performance for their domain specific tasks, which will offer immense potential for training tools.

3.5 State of the art in medical expertise scanpath analysis

Much of the expert eye movement research has relied heavily on summary statistics as indicated in the overview in section 2.3. The literature so far has yet to create a comprehensive understanding of an *effective* scanpath. Scanpath analysis can offer advanced pattern recognition to quantify differences in the visual search strategy between the many stages up to – and including – expertise. Research

on scanpath classification for medical expertise is becoming more realized as a viable approach for teaching interventions (see contribution in section 5.6). With the recent propensity towards artificial intelligence, expert gaze models can even support accurate model training [305]–[308]. Therefore, medical expert scanpath models offers insight for both human training and automated system development. Current approaches towards robust scanpath classification for medical expertise is established from many of the aforementioned methods.

Building off traditional string alignment, Kok et al. [158] looked at the Levenshtein distance of expert and novices inspecting chest x-rays. Students had higher scores (more dissimilar) compared to experts indicating less systematic scanpaths among students. Davies et al. [309] looked at expert ECG readers and compared the subjects with correct interpretation of the images to those with incorrect interpretation. Overall, there were highly varying search strategies as seen by the scanpaths. However, the average Levenshtein distance of all participants in a group showed higher variability for subjects in the incorrect group. Whereas, subjects who interpreted ECGs correctly had more similar scanpaths as indicated by lower Levenshtein distance scores [309]. Using *ScanMatch*, Crowe et al. [310] found experts had the highest level of scanpath similarity for brain tumor detection in MRIs, while intermediates had the least similarity.

Regarding the scanpath structure, Drew et al. [168] found 2 main scanpaths for volumetric images. The one *scanning* strategy was found to increase false negatives and the other *drilling* strategy promoted less errors in diagnoses and more areas covered. The scanpaths were recognized by tracking the depth of fixations over time relative to anatomical quadrants. To quantify the patterns as either zig-zagging or going deeper over time, they looked at fixation behavior relative to the quadrants. However, they were able to robustly classify scanning and drilling [311].

Concerning scanpath features related to expert search, Li et al. [312], [313] used autoregressive HMMs to cluster expert dermatologists eye movement into patterns indicative of specific diagnostic tasks. They looked at expert, intermediate, and novice dermatologists and how they classified skin problems. Experts and intermediates had longer fixation durations and had decreasing saccade amplitudes as an effect of viewing time. A dynamic hierarchical Hidden Markov Model (HMM) based on the level of expertise could infer the scanpath patterns for particular moments in the image interpretation such as the primary morphology and differential diagnosis; these patterns were also diagnosis specific [312]. Also using HMMs, Ahmidi et al. [314] could distinguish expertise for varying minimally invasive surgery tasks. Richstone et al. [315] looked at eye movement events (e.g. fixation and blink rate, ICA [316]) for one second intervals during a surgical task and found a shallow neural network with back propagation fed this input could predict expertise with over 90% accuracy. Their results indicated that expert surgeons remain more focused for longer periods of time compared to novices.

From scanpath similarity, patterns indicative of expert scanpaths have been investigated. Concerning scanpath similarity, Kübler et al. [254] found that *Subs-*

Match[249] outperformed the State of the Art (*ScanMatch*,HMMs, *iMap*, *FuncSim*, etc.) in classifying expert and novice microneurosurgeon scanpaths. They found that novices show more repetitive viewing to certain areas, where experts employ quicker, more repetitive fixation behavior in certain image regions before a more broader exploration of other areas [147], [254]. Regarding dermatologists, similar gaze behavior was also found based on RQA [317].

Specifically for radiograph examination, Gandomkar et al. [318], [319] classified experience in mammogram interpretation using SVM with the features from RQA, achieving 86% accuracy. They found experience corresponded with more unique scanpath dynamics compared to novices [162]. This finding regarding experts' dynamicity from RQA was also found in orthopedic radiographs of the hips [135]. Medical decision making coupled with attention-level in mammogram inspection was modeled with deep neural networks in [305], [320]. These approaches offer insight into how certain expert gaze behaviors can be linked to false negative likelihood.

Recently, the subsequences transitions from expert and novice radiologist were clustered for expert and novices using data mining techniques (contrast mining with temporal binning) for subsequence extraction [241]. Hierarchical clustering was then based off the similarity of patterns; experts had more similar patterns, and novices less so [241].

This work furthers the research of scanpath classification, contributing to expert strategies in OPT image inspection. Although, summary statistics have supported that expert OPT gaze behavior aligns with the findings of medical image inspection in general [149], research has started to unsurface scanpath characteristics distinct to OPT inspection [65], [154]. We support scanpath classification from previous approaches, but in the context of distinguishing novice dentists, which is described in section 5.4. Moreover, we introduce a new method for scanpath comparison that combines deep learning and traditional sequence alignment and evaluate it on expert dentist classification from the gaze behavior during OPT inspection. This novel approach achieves an accuracy of 94% in expert novice recognition. Details of the algorithm's workflow and its evaluation are in section 5.5.

4 Current Approach to Understanding Expertise in OPT Inspection

OPTs are highly prevalent in dental medicine and require high-level perceptual expertise in conjunction with proper clinical understanding. Analysis of gaze features can provide insight into the cognitive processes of experts and the development of expertise. In addition, the scanpath offers a rich input for learning and expert models as it goes further than gaze summary statistics by formulating sequences that characterize expertise levels. This work's contribution is twofold: 1) gaze features are linked to expert cognitive strategies and 2) expertise development is classified with advanced scanpath classification metrics. Moreover, a novel scanpath classification method is presented and further used to develop an expert model for an attention-aware gaze intervention. Chapter 5 details the outcome of this work.

To support the outcome, this chapter details the comprehensive investigation of student and expert OPT inspection. Section 4.1 reports on the spectrum of expertise sampled. Already, this work stands apart from previous literature because of its extensive sample size, which further supports the evaluated approaches. Section 4.2 explains the study protocol of how the images were presented to participants and how their pathology detection could be compared to ground truth information. The eye movements analyzed and pupillary response pre-processing are described in section 4.3. Finally, a brief overview of other factors that were controlled for pupillary response is given in section 4.4.

4.1 The data collected

The data collection took place over multiple semesters from 2017 to 2019. It was done in the context of a larger project in collaboration with the Universitätsklinik für Zahn-, Mund- und Kieferheilkunde (Collaborators: Dr. Fabian Hüttig and Dr. med. Dr. med. dent. Constanze Keutel) and the Leibniz-Institut für Wissensmedien (Collaborator: Prof. Dr. Katharina Scheiter) in Tübingen.

The Ethical Review Board of the Leibniz-Institut für Wissensmedien Tübingen approved the student cohort of the study with the project number LEK 2017/016. All participants were informed in writing and consented with a signature that their pseudonymous data can be analyzed and published. Due to a self constructed pseudonym, they had the option to revoke this consent until the date

Table 4.1: The Project Data. First column indicates in which semester the data was collected. Data from students sixth through tenth as well as experts was collected over four separate collections. The sixth semester students were measured on threes separate occasions (M1 - M3).

	Sixth			*Seventh*	*Eighth*	*Ninth*	*Tenth*	*Expert*
Measure	*M1*	*M2*	*M3*					
SS17	17	17	15	17	26	28	14	
WS1718	16	18	16	4	19	25	8	
SS18	24	24	19	2	12	5	28	26
SS19								3
Glasses*	18	6	12	-	5	8	11	9
Total	57	59	50	23	57	58	50	29

* indicates there was unknown data for some of the student data collections

of anonymization of the data after data collection was finished. The Independent Ethics Committee of the Medical Faculty and University Hospital Tübingen approved the expert cohort of the study with the project number 394/2017BO2. All participants were informed in writing and consented verbally that their anonymous data can be analyzed and published. Due to a self-constructed pseudonym, they had the option to revoke this consent at any time.

Novices Dentistry students in semesters six through tenth from the university were invited to participate and have gaze and performance data recorded during an OPT inspection task. For reference, sixth semester students are in the second half of their third year and the tenth semester is in the fifth year of their studies, which is last semester before they continue on to the equivalent of a residency. The sixth semester students are incoming dental students from their initial premed studies. They had no prior training in dental radiograph interpretation, but basic conceptual knowledge in general medical concepts. This group was evaluated three times (beginning, during, and end as M1, M2, and M3 respectively) in each period of data collection due their curriculum requirement of an OPT interpretation training course. This is the only semester that includes this obligatory course, which covers instruction and starts massed practice OPT interpretation.

At the time of the contribution to student scanpath classification [1] there was only one data collection from students from semesters six through ten. Additionally, only sixth semester students at the end of their OPT course (M3) were evaluated against experts regarding their pupillary response in the contribution [4]. This decision was due to potentially having higher cognitive load as well as conceptual knowledge. Finally, for the evaluation of a novel scanpath method [5], only incoming sixth semester students were evaluated against experts to avoid

over-representing students in the model.

Experts Experts from the university clinic volunteered their expertise for the current project. Experience was defined as professional years working as a dentist. All experts had the necessary qualifications to practice dentistry and or any other dental related specialty: e.g. Prosthodontics, Orthodontics, Endodontics, etc. In total, their years experience was an average of 10.16 and ranged from 1 to 43 years. 50 % of experts reported seeing between 11 and 30 patients on a typical work day and the remainder saw less than 10 patients a day. At the time of [3], there were only 26 dentists, whereas in later evaluations [4]–[6], the total data set of experts was available for evaluation.

4.2 Study protocol

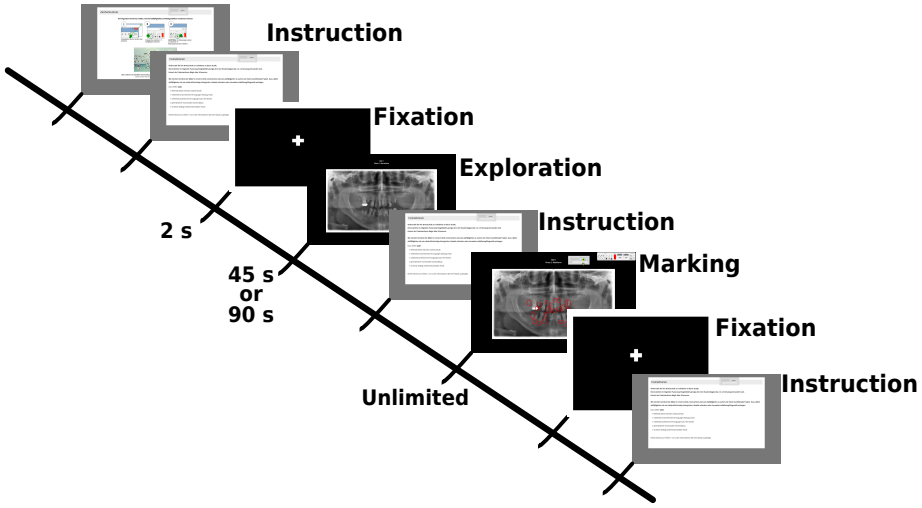


Figure 4.1: Outline of Experimental Session. For each image, there is a fixation cross for baseline data, then an exploration phase (45s duration for experts and 90s for students), and marking phase (unlimited time). Students received two sets of 10 OPTs with a break in between and experts received one set of 15 OPTs with a break after the first seven.

The experimental protocol for students and experts consisted of an initial calibration, task instruction, then two image phases: Exploration and Marking (see figure 4.1). Prior to the exploration phase, a two second fixation cross was presented. This served as baseline for our analysis concerning pupillary response in [4]. Then, in the exploration phase, particpants were instructed to search the OPT for anomalies that could be indicative of a pathology. Pathologies examples were periodontal disease, cavities, insufficient fillings and abscesses, not including sufficient fillings, missing teeth needing no further treatment, or prosthetics.

Students had 90 seconds to inspect each OPT, where experts had 45 seconds to inspect each OPT. This shortened duration was determined because, much of the previous literature has shown that experts are faster at scanning radiographs [24], [28], [29], [138], [149], [155], [156], [158].

Then, following the exploration phase, the same image was presented in the marking phase. Here, they were instructed to indicate any areas they found in the prior phase that could be indicative of pathologies. This phase was self-paced, and they marked the areas using an on-screen painting tool. The markings from this phase served as detection performance data for [3], [5], [6], which is detailed later in this section.

Students inspected two blocks of 10 OPTs in one experimental run and experts – due to their hard-pressed schedules – inspected 15 OPTs. Both students and experts were unrestrained during the experiment, although they were instructed to move their head as little as possible.

Environment Data collection for students took place in a digital classroom equipped with 30 remote eye trackers attached to laptops. This setup allowed for data collection of up to 30 participants simultaneously, minimizing the overall time needed for collection. For this study, verbal instructions were given en masse pertaining to a brief overview of the protocol and an explanation of eye tracking, then individual calibrations were performed with a supervised quality check; students could then run the experiment self-paced.

Data collection for the experts took place in the university hospital so the experts could conveniently participate during work hours. There, the room used for data collection was dedicated for radiograph reading. The same model remote eye tracker was used for expert data collection and was run with the same sampling frequency on a laptop.

Eye tracker and pre-processing Data collection was done using an SMI RED250 remote eye tracker with 250 Hz sampling frequency. We used the included software for both the experiment design (*Experiment Center*) and event analysis (*BeGaze*). A quality assessed validation of the calibration was performed for each participant prior to the experiment as well as after breaks between sets. An acceptable calibration was if is was less than one degree average deviation from a four point validation.

Eye movement data was removed for images where the tracking ratio – proportion of valid gaze points – was below 80 %. Furthermore, participants were removed if they had missing data for more than two of the ten images.

4.2.1 Stimuli

Images had anywhere from four to fourteen anomalies types, though an anomaly type (e.g. Caries) could be found in multiple areas in a single OPT. There were also

control OPTs with no anomalies present. The OPTs were chosen from the university clinic database by the two expert dentists involved in this research project, and were determined to have no artifacts and technological errors. Both dentists independently examined the OPTs and the patient work-ups and further consolidated together to determine ground truths for each image.

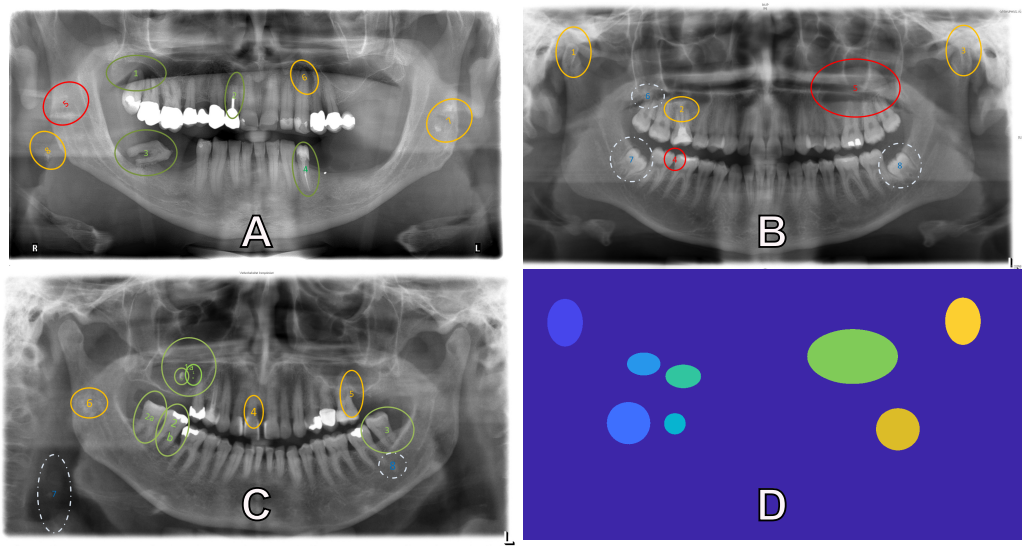


Figure 4.2: Example of the OPTs used in the experiment. Pre-determined ground truths are indicated by the ellipses and their colors indicate the level of difficulty each anomaly is: Green (least difficult), yellow (intermediary), red (most difficult) and white (nature of difficulty unclear). Image (D) is the ground truth map for image (B). Each anomaly is segmented and given a distinguishing integer.

Additionally, the level of difficulty for each anomaly was pre-determined. Fig. 4.2 shows three OPT examples. Anomalies are illustrated in green, yellow, and red, and represent easy, medium, and difficult, respectively. For example, the green anomalies in Fig. 4.2.A are a dental cyst (1) and insufficient root canal fillings. (2a,b) in Fig. 4.2.C are an example of elongated lower molars due to missing antagonists. The yellow anomalies in Fig. 4.2.B are irregular forms of the madibular condyle (1,3) and (2) is an apical translucency indicative of inflammation due to a contagious (bacterially colonized) root canal filling. The red anomalies in this image are approximal caries (4) and a maxillary sinus mass. Anomalies indicated by the white dashed circles were determined as ambiguous, e.g. the nature of their difficulty and or pathology is unclear. For example, in Fig. 4.2.B (7,8) are impacted wisdom teeth, though it is uncertain whether this will become a problem for the patient and therefore is regarded as potentially pathologic. (6) is an apical translucency at the mesial root apex and it is unclear whether it is indicative of an inflammation. Therefore, they were kept in this analysis even though the nature of their difficulty is uncertain.

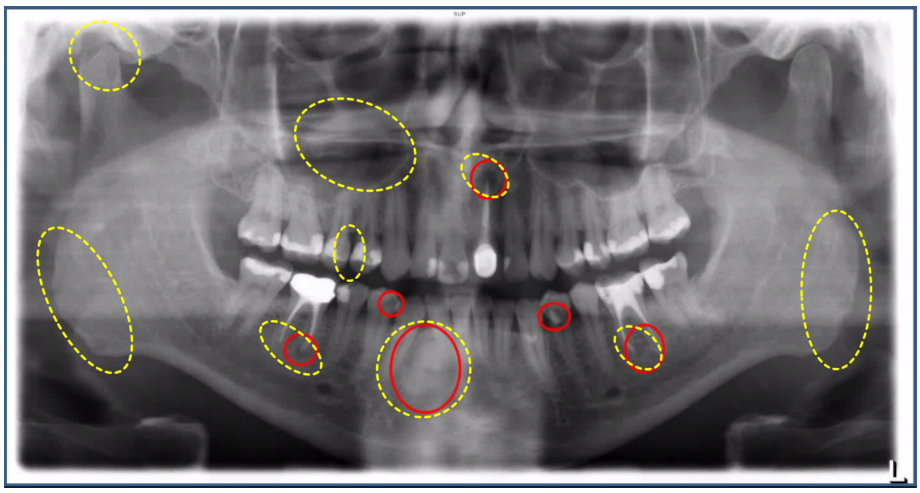


Figure 4.3: Drawings from a participant (Red) with predefined anomalies (Dotted Yellow), or targets, overlayed. In this example, the participant would have four hits and five misses and two false positives.

4.2.2 Ground truth maps

We created maps for the OPTs (See Fig. 4.2.C) using Matlab 2018. As input, OPTs were loaded as png files with their respective anomalies – all colored red. Thresholding for red values was performed to automatically get the pixel coordinates of the ellipse edges. Then, the ellipses were filled with the `poly2mask()` function. Anomalies automatically extracted from this process were double checked for overlapping and had their boundaries corrected. Similar anomalies inside of another, such as (2a,b) in Fig. 4.2.C, were grouped together as one anomaly. Other anomalies too close together and too different in pathology, such as (3,8) in Fig. 4.2.C, were excluded from the analysis, due to possible spatial accuracy errors in the gaze. Similarly, anomalies that were denoted by too small of an ellipse were padded to have a larger pixel area,e.g. (4) in Fig. 4.2.B, to account for the spatial accuracy errors in the gaze. Each segmented anomaly is given a distinguishing integer for its respective pixels.

Drawings obtained from the marking phase were compared to predefined anomalies determined for each image. 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.[1]

We report the performance in anomaly recognition with 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. For example, four anomalies were marked by the subject shown in figure 4.3 and five were not recognized

[1]Inter-rater reliability: 0.94 and 0.93.

(false negatives); the recall is then 44%. Precision is the true positives over the total of true positives and false positives. False positives are determining there is an anomaly when there is none. The same subject in figure 4.3 has two false positive, thus precision is 67%. Precision and recall affect the harmonic mean (F1 score).

4.3 Eye movement behavior

Since the eye tracker has a high sampling frequency, both stable (fixations) and rapid (saccadic) eye movement events for static stimuli can be measured. Only events during the exploration phase of OPTs were evaluated, because the marking phase data had too many spatial offset errors.[2] Fixations and saccades for the left eye, including tracking ratios per image, were calculated using the standard SMI high-speed settings for the I-VT [85]: 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.

To measure gaze behavior on the ground truth anomaly AOIs, x and y fixation coordinates with timestamps are plotted to the ground truth maps (D in figure 4.2). If the coordinates of a fixation were within or on the border of a target, it was considered a glance hit. This concept was also applied in [3] to investigate how often experts glance at anomalies.

To further investigate the mental processing for anomalies of varying difficulty in [4], analysis for the pupillary response from the raw gaze data was used. The raw data points also have pupil diameter output in millimeters [321].[3] For further signal processing, we removed gaze coordinates and pupil data for the raw data points labeled as saccades. Data points with a pupil diameter of zero or labeled as a blink were also removed. Additionally, data points 100 ms before and after blinks were removed, due to pupil size distortions from partial eye-lid occlusion. Lastly, the first and last 125 data points in the stimulus presentation were removed due to stimulus flickering [322]–[324]. The remaining data was smoothed with a third order low-pass Butterworth filter with a 2 Hz cutoff as illustrated by the purple data points in Fig. 4.4.

Similar to [3], the processed gaze points that land in each anomaly are a gaze hit, and that anomaly's level of difficulty is extracted. For all gaze hits on an anomaly for a participant, we calculated the median pupil diameter. The median pupil diameter for each anomaly was then subtracted from the respective baseline data for that image. We performed subtractive baseline correction because it has been found to be a more robust metric and have higher statistical power [325]. Therefore, the difference from baseline could indicate diameter increase (positive value) or diameter decrease (negative value) compared to baseline.

[2] Too many participants moved for this phase.
[3] Millimeters extrapolated from pupil height and width dimensions in pixels [321].

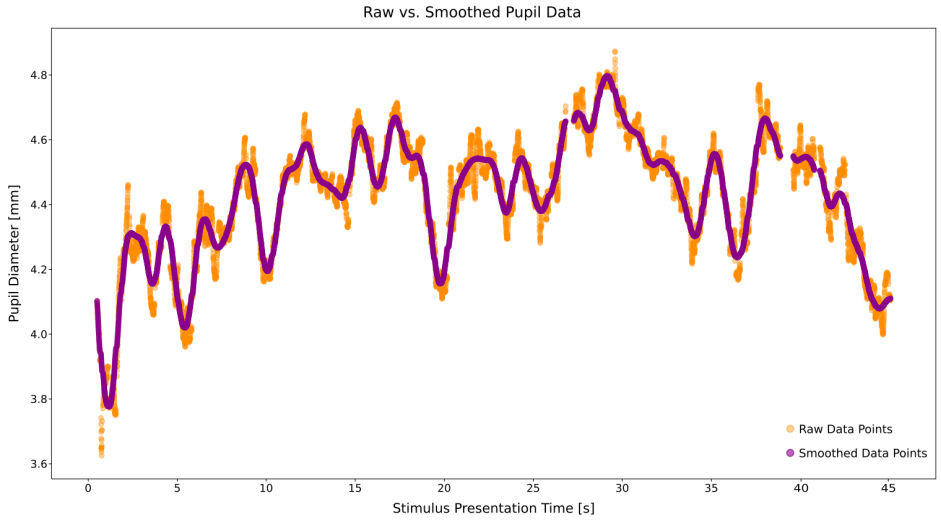


Figure 4.4: Raw signal from the left eye (orange) and the smoothed signal (purple) with a Butterworth filter with 2 Hz cutoff.

With the gaze hits on anomalies of varying difficulties, we can evaluate the pupillary response of both experts and students during anomaly fixations. The pupillary response, as measured by change from baseline, can then provide insight into the cognitive load both groups are undergoing while interpreting the anomalies.

4.4 Context relevant to pupillary response measurement

More important to the cognitive load study [4], both data collection environments had the room illumination levels controlled with no effects from sunlight or other outdoor light. The standard maintained illuminance for experimental sessions was between 10 to 50.[4] It is advised that environment illumination during radiograph reading should be ambient (25–50 lux) for the best viewing practices [326] and to optimize contrast perception in radiographs [327]–[329]. Therefore, with room illumination controlled, we can evaluate pupillary response independent of environmental illumination changes.

Regarding the screen display, radiograph reading is not affected by the luminance of the display [328]. However, both the laptop models used for the experimental sessions abide by the multiple medical and radiology commission stan-

[4]lux was measured with a lux sensor (Gossen Mavo-Max illuminance sensor, MC Technologies, Hannover, Germany).

dards [156], [326], [330].[5] The HP Z Book 15 (for students) has screen brightness averages approx. 300cd/m^2 [331]. The Dell Precision m4800 (for experts) averages approx. 380cd/m^2 [332]. While the screen luminance was also controlled and followed the standard protocols for viewing radiographs, the exact effect of the screen brightness on the pupillary response is out of the scope of [4].

[5]Pixel density affects comfortable viewing distances of 30 to 60 cm and a monitor luminance should be at least 200 cd/m^2 to 420 cd/m^2.

5 Major Results and Discussion

This work derives expert visual search strategies through highly data-driven approaches. The contributions that support expert and novice cognitive processing are detailed in sections 5.1 and 5.2. These contributions use expert-defined anomalies (detailed in section 4.2.2), which offer ground truth comparison for further investigation of decision making. However, the contributions on scanpath analysis (found in sections 5.4 and 5.5) move towards extracting behaviors without AOIs and image-independent semantic understanding. Ultimately, these contributions can be generalized to visual search behavior beyond the evaluated context and towards general medical image inspection. The present proposal also works towards gaze-based intelligent systems by providing a framework for attention-aware real-time guiding systems for medical image inspection (section 5.6).

5.1 Fixation behavior related to expert performance

It is known that experts have more effective search strategies (see section 2.3.2) and are better at detecting anomalies [28], [133], [149]. However, when an expert does not mark an anomaly after seeing it, which mechanisms determine that cognitive decision? Fixation duration can be applied to distinguish different errors [58]. For instance, a false negative can 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). Building on the findings that attention allocation can affect misclassification during decision making process, we looked at glance behavior of expert dentists.

5.1.1 Relation between fixation and recall

Experts detected around 50 % ($SD = 11.12$) of anomalies in the OPTs. Although, there was high variability between images: Recall ranged from 96 % to 0 and precision ranged from 96 % to 54 %. The average harmonic mean (F-score) was 60.89 % ($SD = 8.65$). Diniz et al. [76] found comparable recall rates (20 to 40 %) in dentists and attributed it to *overlooking* certain anomalies where the treatment cost could possibly outweigh any long term benefit.[1]

[1] Cost not only implying price, but also effort, pain evoked, rehabilitation, etc. [76].

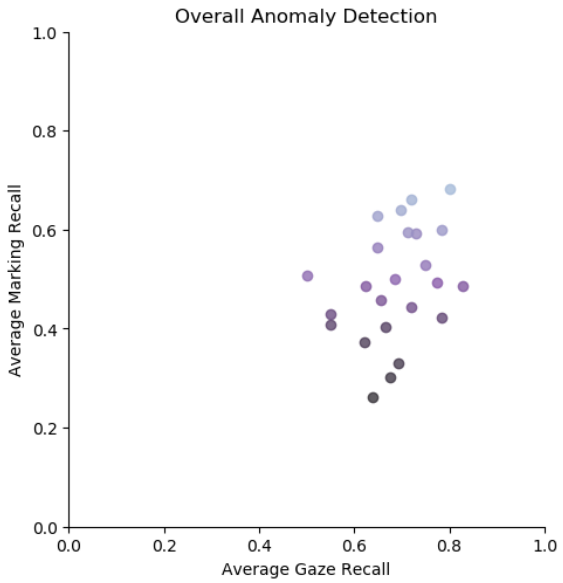


Figure 5.1: Relationship between overall gaze recall and marking recall. The lighter hues are indicative of higher marking recall.

This allusion of *overlooking* was further supported by the findings concerning gaze behavior. There was a slight, though insignificant, relationship between the gaze and detection of anomalies ($r = 0.33, p = 0.11$; see figure 5.1). More interesting, gaze on anomalies was overall higher ($M = 69.82\,\%, SD = 8.44\,\%$) than the recall behavior. High sensitivity to looking at anomaly areas can indicate effective searching of possible areas where pathologies reside. Thus, experts often looked at an anomaly area, although they marked it roughly at chance level.

Moreover, the amount of gaze on an anomaly affected its likelihood to be detected. The number of glances on a detected anomaly ($M = 2.34, SD = 3.25$) was significantly higher than an anomaly that was not detected ($M = 1.51, SD = 1.82$), $t\ (2850) = 8.35, p < 0.001$. The frequency of glances per anomaly as seen in figure 5.2 shows that for unmarked anomalies, there is a higher frequency for zero glances or one glance. For marked anomalies, there is also a trend to glance once. However, when there are three or more glances on an anomaly, there is a switch in the marking behavior, where the frequency is higher for anomalies marked compared to anomalies unmarked.

5.1.2 Implications for decision making

In line with the findings for fixation duration [28], [58], the number of glances can determine the cognitive mechanisms behind false negatives. For experts, we found very few occurrences that could be similarly classified as a search error, i.e. an anomaly was not detected because it was not looked at. A recognition error, on the other hand, occurred when an anomaly was not detected as such, but glanced

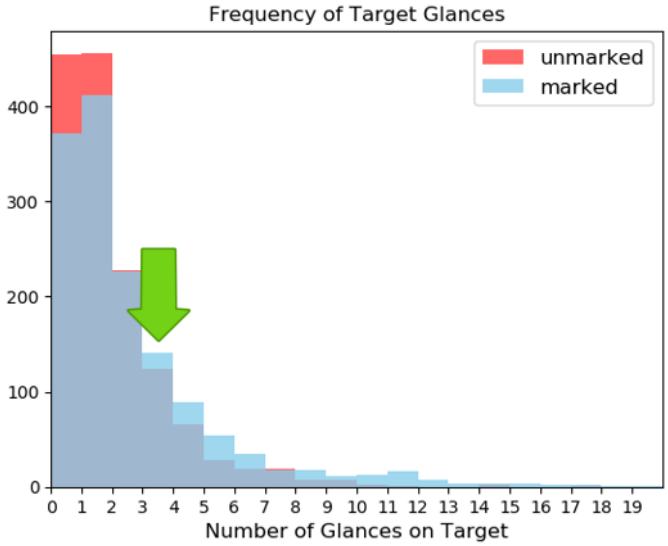


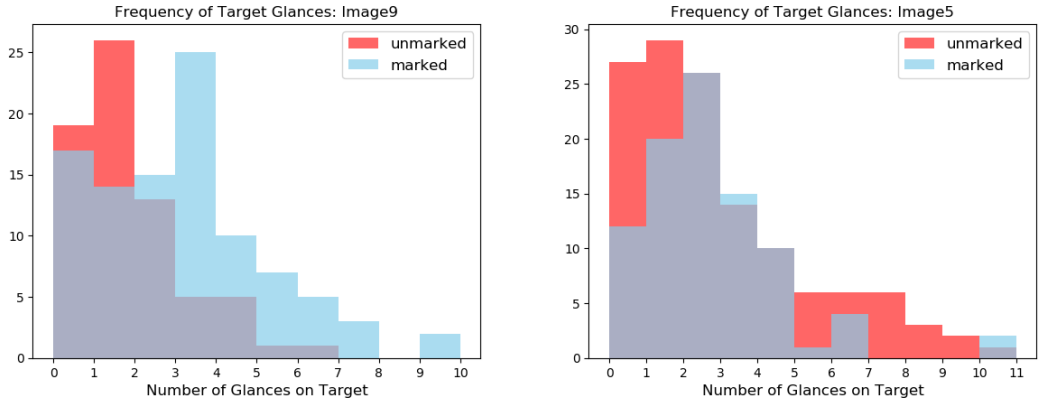
Figure 5.2: Frequency of glances for marked (detected: blue bars) and unmarked (not detected: red bars). The frequencies when number of glances per anomaly is 3 (green arrow) or more is overall higher for when an anomalies recognized in contrast to when not.

at once or twice. In this case, an expert may look over an anomaly and determine that it is not worth further scrutiny. Then, a decision error is characterized by frequent glances to the area. This high number of glances could indicate that more cognitive processing is involved for determining the nature of the anomaly (see section 5.2 for further support).

5.1.3 Insight towards effect of anomaly difficulty

Decision errors were generally less frequent in our findings, though presented an interesting direction for further research. As previously mentioned, we found high variability in the performance between images. This behavior was also apparent in the gaze. Figure 5.3 shows that the image content varies the glance frequency for detected anomalies (marked) and false negatives (unmarked). One image (Figure 5.3a) elicits higher frequencies for glancing at an anomaly (3+ glances) to properly detect it. This gaze behavior could indicate that the anomalies in this image are more difficult to detect and, thus, require more examination. Conversely, another image (Figure 5.3b) has an overall higher tendency for false negatives. Here, experts had 8 or more glances on an anomaly and still made a false decision.

Since it has been supported that the image content effects expert and novice gaze [28], [137], [149], [154], these findings prompted further investigation into the nature of anomaly difficulty on cognitive strategies. The idea behind these findings would be to employ expert glance behavior as a predictor of how easy or hard an anomaly is to accurately detect. Furthermore, the scanpath, or order



(a) An image where the number of glances on anomalies were higher when marked in contrast to when not marked.

(b) An image where there is a high amount of false negatives in marking although there is a high frequency of higher glances per anomaly.

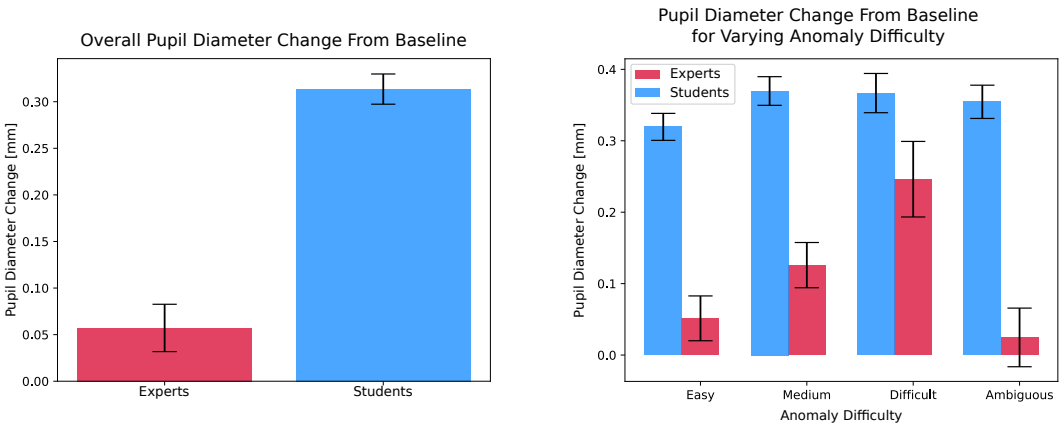
Figure 5.3: Two image examples variability in the glance behavior shown by histograms of the glance frequency.

that the anomalies were fixated on, can highlight the patterns that indicate expert search behavior and is of great interest for further research. Understanding the cognitive processes involved in effective medical image interpretation through the gaze can offer expert insight into teaching novices effective decision making skills.

5.2 Cognitive load indication in visual search of experts and novices

In general, as task difficulty increases, so does the workload [121] and correspondingly, the pupil dilation [103], [116], [316], [333], [334]. Uncertainty as well as perceived task difficulty have been found to affect the pupillary response, and acquired knowledge has been shown to reduce these apprehensions [23], [114]. Prior knowledge of a problem has also been shown to reduce cognitive load [104], [110], [335], [336].

The current work was built on the premise that gaze behavior in expert dentists changes with difficult images [3], [149]. Thus cognitive processing may adapt to handle these difficulties. This work goes one step further by examining the adaptability of cognitive processes during visual inspection of multiple features in decision making. We measured pupil diameter change from baseline when gazing on anomalies of varying difficulty during visual search of dental panoramic radiographs. We found changes within the visual search of an OPT in contrast to the overall response to image interpretation.



(a) Median Pupil Change From Baseline for Experts and Novices.

(b) Median Pupil Change From Baseline for Gaze on Anomalies.

Figure 5.4: Pupillary Response of Experts and Novices During Visual Inspection. The median pupil diameter change from baseline for students (blue bars) and experts (red bars) for the overall image behavior (5.4a) and when gazing on anomalies of varying difficulty (5.4b). Standard errors are indicated in black. Students had larger pupillary response from baseline compared to experts, but this effect was homogeneous for the differing anomalies. Whereas experts showed an increased pupillary response behavior as an effect of increasing difficulty.

5.2.1 Support for cognitive load in students

Initially, our findings were supported by the previous literature [108], [110], [119], [120], [170], [171], where students' pupillary response from baseline is higher that experts (see figure 5.4a). Independent of anomaly difficulty, students' pupillary response ($M = 0.314$, $SD = 0.315$) had a larger increase from baseline than experts ($M = 0.057$, $SD = 0.353$, $t(568) = -8.824$, $p < 0.001$). Cognitive load is often used to explain this behavioral response regarding learning [117], [335]–[337]. For instance, Tien et al. [119] found that novices reported higher memory load compared to experts performing the same task. This behavior can be likened to students' lack of conceptual knowledge and experience, leading them to "think harder" to interpret these images [338], [339].

One of the more interesting findings is the lack of influence of anomaly gradation on student cognitive processing. Students showed larger and more homogeneous pupil size change from baseline for all anomaly gradations compared to experts (see figure 5.4b). One would imagine that even the most pronounced of anomalies would make the recognition process easier. However, our findings from student pupillary response indicate that, regardless of how conspicuous, the level of mental workload remains constant. Figure 5.4b details the pupillary response of experts and novices on the varying anomaly difficulties. A $2(expertise) \times 4(anomaly)$ factor ANOVA found a main effect for expertise ($F(1, 1388) = 161.68$, $p < 0.001$),

though there were no significant effects of anomaly difficulty on student pupillary response.

5.2.2 Experts' adaptability to difficulty

The most interesting finding is that there were significant effects of anomaly difficulty on expert pupillary response. The ANOVA revealed a significant interaction between expertise and anomaly difficulty ($F(3, 1388) = 2.76$, $p = 0.041$). Post hoc analyses with Bonferroni correction for anomaly difficulty on the expert data revealed significant differences for the more difficult anomalies ($M = 0.246$, $SD = 0.370$) compared to least difficult ($M = 0.0514$, $SD = 0.396$, $t(207) = -3.0582$, $p = 0.003$) and ambiguous ($t(150) = 3.1796$, $p = 0.002$). There were no significant differences for medium ($M = 0.1259$, $SD = 0.3904$) compared to the difficult anomalies ($t(200) = 1.8989$, $p = 0.059$). Therefore, experts had the largest pupil size change from baseline for more difficult anomalies, especially compared to least difficult and ambiguous anomalies.

As the gradation of difficulty increases so did the pupillary response in experts. The red bars in figure 5.4b shows the least pupil size change from baseline for the least difficult anomalies and the largest change for the more difficult anomalies, with the medium difficult anomalies producing a response in between. This behavior, however, was not evident for the ambiguous anomalies, which showed the smallest response change from baseline. This effect may lie in the uncertainty of these anomalies as determined by the two experts involved in the project.[2] Therefore, it is uncertain how difficult, easy, or even existing these anomalies were (see description in section 4.2).

Our findings suggest that students may employ similar cognitive strategies that evoke higher load for all anomaly gradations. Comparatively, experts employ more efficient strategies [116]; however they are more sensitive to task features. They generally know where anomalies are prevalent and how they are illustrated in the image features. When inspecting these specific areas, pupil dilation fluctuation can be indicative of changes in their cognitive processes to accommodate more complex features. Additionally, the pupillary response can be indicative of their selective attention allocation [133], which promotes quick recognition of anomaly specific features. Depending on the gradation of the area in focus, proper interpretation can be instantaneous (low workload) or may need to evoke adaptations in the decision-making strategies.

Interpretation of medical images is not trivial and certain image or pathology features can hinder the true diagnosis. Experts are more robust at determining more difficult or subtle anomalies [76], [149], [156], [158], [340]. Although when anomalies become harder to interpret, experts evoke pupillary response indicative of increasing task-difficulty, leading to behavior that is likely of more thor-

[2]This category was a mixture of potential areas that may or may not have included an anomaly, or anomaly, but with no cause for alarm.

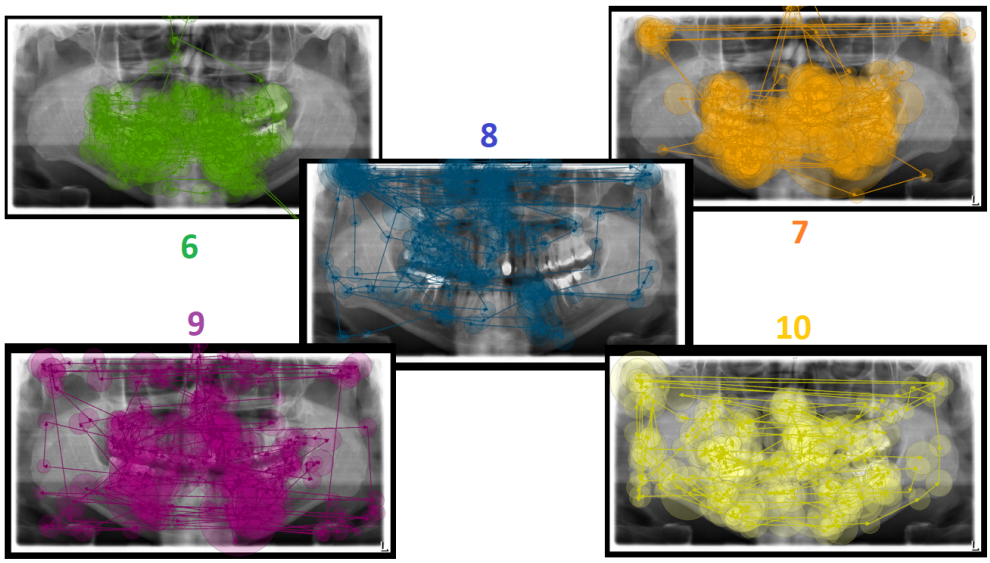


Figure 5.5: Visualization of fixations from a student in each semester evaluated in the current study as indicated by the colored numbers respectively. In this condition, the sixth semester student's data is prior to training.

ough inspection. With more insight into expert decision making during visual search, appropriate learning interventions can be developed. These interventions can incorporate not only the scanpath behavior, but also the cognitive load during appropriate detection of pathologies. From this combination, image semantics can be better conveyed to the learner. Training sessions that convey the appropriate information through adaptive gaze interventions by detecting the gaze and cognitive load offer a promising direction for medical education.

5.3 Cognitive processes: From gaze features towards the scanpath

Gaze features (i.e. eye movement events [24], [28], pupillary response [4], etc.) are capable of explaining differences in the cognitive-processes of experts and novices. Conventionally, expertise literature (see section 2.2) employs standard key metrics in quantitative analysis and tries to fit high-level pattern hypotheses by looking at such metrics, e.g. longer fixations indicate more focused viewing. Although this approach is well suited to the analysis of differences between experts and students, advanced pattern recognition and analysis algorithms are needed to identify and quantify differences in the visual search strategy between advanced learners, residents, and expert practitioners.

Differences in the scanpaths of experts and novices is starting to gain visibility in the literature (See overview in Chapter 3). Yet, scanpath differences relating to

the developmental stages is severely under-investigated. It is interesting to observe whether scanpath differences already appear between novices of different levels, since student curriculum each semester builds off of existing knowledge structures. Figure 5.5 shows the evolution of students' scanpaths taken from the current study. Here, the sixth semester student 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 with the seventh semester student, where scanning behavior that compares similar areas of the jaw on the left and the right is present. Then, the eighth, ninth, and tenth semester students exhibit more coverage of the OPT; specifically, less fixations on the teeth and longer saccades spanning the upper and lower jaw areas.

Understanding the scanpaths in an effort to find patterns determinant of a developmental level can ultimately build an adequate representation of eye movements for the complete learning process. Therefore, a model initially should recognize gaze patterns (i.e. subsequences) that are characteristic for a dentist at a respective expertise level. Then, building off accurate recognition, scanpath components can further be clustered based on patterns representative of key phases in effective visual search, i.e. systematic, comparative or explorative. Such patterns are likely to contain highly discriminative information, which is not bound to e.g. one specific OPT. Rather, the patterns can be linked to the specific semantics of a certain structure or anomaly.

The strength of the scanpath classification methods we implement is that they are completely data driven and do not rely on any expert labeling (e.g. AOIs). Therefore, we are also able to capture potential novice effects that an expert may not realize, such as unsystematic scanning of regions that experts would not find worthy of labeling as an AOI.

5.4 Distinguishing dental students through scanpath classification

In this work, we set out to determine whether eye movement differences among novices become apparent and, if so, at what level of task-knowledge. Using *SubsMatch2.0* [181] and a Nearest-Neighbor classifier with Needleman-Wunsch scores, we evaluated each of the sixth semester datasets (pre-, mid-, post- or M1, M2, M3 in table in chapter 4) against the data from semesters seven through ten. Since the classifiers are trained on five groups with datasets roughly balanced, guess chance level was around 20 %. After removal of data with low tracking ratios, 139 data sets were used: With 73, 68, and 68 participants for training the respective pre-, mid-, and post-models, and a total of 15 participants – three per each semester – were set aside for validation.

Table 5.1 details the overall accuracies for the models. Both classifiers are capable of distinguishing semesters above chance level for pre- and post-conditions.

Table 5.1: 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 %

Above all, the highest accuracy is for the pre-training condition, where the sixth semester students are measured before their OPT analysis training. The results show that the curriculum in this semester – specifically, the OPT training course – is very relevant to the gaze behavior.

More important than overall performance is how the semesters were distinguished. Figures 5.6 and 5.7 show both classifiers' confusion matrices for each condition. A breakdown of the classifiers and their performance is detailed below.

5.4.1 *SubsMatch 2.0* algorithm classification

For training the SVM, both the percentile binning (from [181]) and the gridded bins (from [200]) 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, so either approach could be employed.

After a leave one out cross validation on the training data, as described in [181], the SVM model suggested the respective n-gram and alphabet size parameters for all conditions: 2 and 3 for pre-training, 3 and 7 for mid-training, and 2 and 7 for post-training.

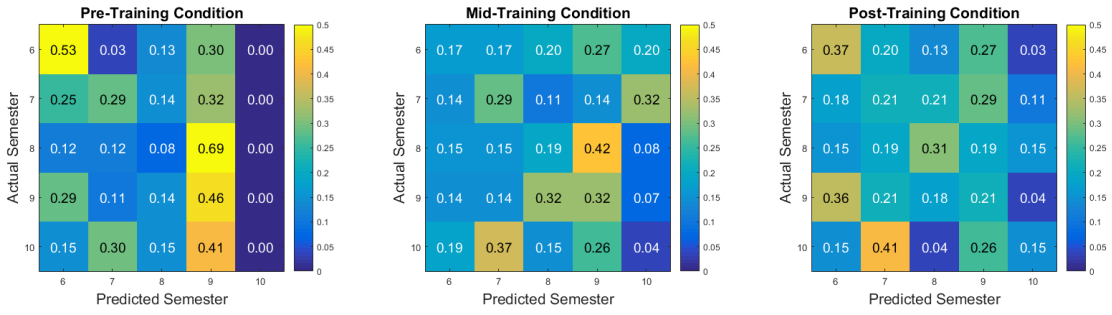


Figure 5.6: *SubsMatch 2.0* semester classification on the validation data. From left to right, confusion matrices for conditions pre-training, mid-training, and post-training are presented. With TPR for each semester along the diagonals. Note that the colorbar for all conditions is scaled at .5.

Figure 5.6 details the *Subsmatch 2.0* classifier's performance with the confusion matrices for the pre-, mid-, post-training conditions. *Subsmatch 2.0* accurately distinguishes pre-training sixth semester students (53.33 %, first matrix) and post-training sixth semester students (36.67 %, last matrix) from higher semester students. However, mid-training sixth semester students affect the model by producing high misclassification rates.

The models were not able to consistently discern students in higher semester. In the pre-training condition, ninth semester students were accurately classified; but eighth semester and tenth semester students were highly misclassified as the ninth semester (69.23 % and 41 % respectively). Even in the mid- and post-training conditions, tenth semester students were also misclassified as seventh or ninth semester. Thus, tenth semester students for all three models were always misclassified and the seventh through ninth semester students were more likely to be misclassified as either the previous semester or the following semester.

5.4.2 Needleman-Wunsch similarity classification

Using a 6×5 AOI grid[3], multiple pairwise scanpath similarity scores were calculated with the Needleman-Wunsch algorithm: 2, -2, and -1 for matches, mismatches, and gaps, respectively. Then, a one-nearest neighbor classifier determined the class label of the scanpath with the highest similarity to the current scanpath. This approach follows the assumption that scanpaths of students in the same class will have higher similarity to each other.

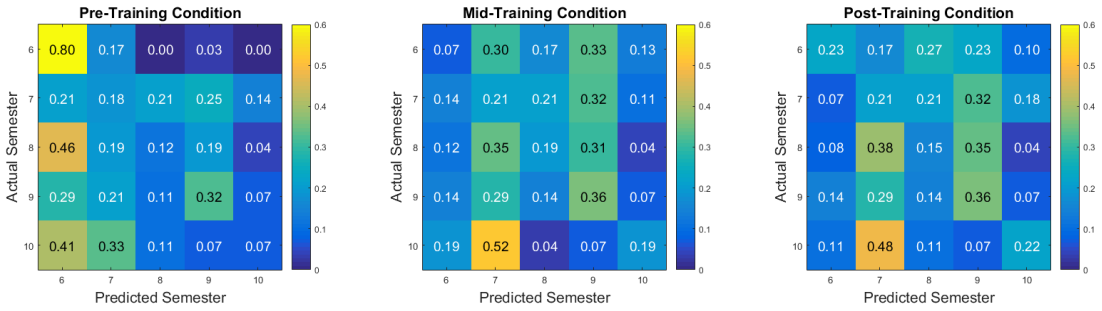


Figure 5.7: Needleman-Wunsch semester classification on the validation data. From left to right, confusion matrices for conditions pre-training, mid-training, and post-training are presented. With TPR for each semester along the diagonals. Note that the colorbar for all conditions is scaled at .6.

Figure 5.7 details the Needleman-Wunsch classifier's performance with the confusion matrices for the pre-, mid-, post-testing conditions. Similar to *Subsmatch*

[3] As determined to be the most optimal resolution from training. One AOI corresponds to 320×216 pixels.

2.0, sixth semester students in the pre-training model (first matrix) are accurately classified (80 %). However, the model highly misclassified students from other semesters as this semester, e.g. eighth and tenth semester students. Otherwise, the ninth semester students are accurately classified, and this also was the case for the mid- (middle matrix) and post-training (last matrix) models. The tenth semester students are again misclassified as seventh semester students, which is similar to the *SubsMatch* classifier. More interesting is the slight shift in the sixth (mid- and post-training models) and seventh semester (all models) students, where they were misclassified more often as higher semester students. In all models, seventh and eighth semester students were more likely to be classified as ninth semester.

5.4.3 Classification compared to statistical analysis

Both algorithms were highly capable of distinguishing sixth semester students in the pre-training condition and, if they falsely classified students, the students were likely classified as either the preceding or successive semester. More interesting, there were no significant differences between semesters six through ten regarding the overall fixation duration on expert defined anomalies ($p = 0.826$). Moreover, differences in fixation duration within the 6th semester (pre-, mid-, post-training) were not significant ($p = 0.881$). Thus, the classifiers were able to distinguish the effect of targeted training on the scanpath behavior where summary statistics fell short. With *SubsMatch*'s focus on high-frequency subsequences and the Needleman-Wunsch's global similarity, both classifiers could extract temporal patterns. Considering that there is only a few months difference between semesters, the models are sensitive to subtle differences in the sequences.

5.4.4 Scanpaths revealed pattern information related to learning

Both *SubsMatch 2.0* 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 after their OPT training course. This distinction is well in line with the curriculum received by sixth semester students. Thus, the scanpath behavior is highly illustrative of knowledge level prior to targeted OPT training. Accordingly, the understanding developed as a result of this course is very relevant to the gaze behavior. The one-nearest neighbor Needleman-Wunsch classifier is very sensitive to the pre-training sixth semester and, therefore, more likely to classify any dataset as such, e.g. eighth and tenth semester students. With this consideration, *SubsMatch 2.0* performs better separation between the pre-training sixth semester students and all others.

The classifiers were similarly at chance level for the mid-training model. This effect could stem from heterogeneity in learning speed and success. In the frame-

work proposed by [341], 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 the ability to mentally manipulate images. It is possible that 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. However, the classifiers appropriately distinguish the post-training sixth semester students from higher semesters students, though to a lesser extent than the pre-training. Imminent final exams motivating students to study could be a possible effect seen in this condition. Hence, these students were likely to be misclassified, as was the case for high semesters with the Needleman-Wunsch classifier and, to a lesser extent, with the *SubsMatch 2.0* classifier.

Al-Moteri and colleagues [164] found that clinical experience evoked gaze behavior that was more *goal-driven* and less *stimulus-driven* [156], [164]. Their research supports that experts are less drawn to salient features with no diagnostic relevance. However, the scanpath differences we found before and after targeted training could also be explained by this notion. For instance, less experienced students may be more drawn to salient areas, such as the teeth, and may neglect more important areas that have more subtle cues when compared to more advanced students 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, resulting in the classifiers consistently misclassifying eighth, ninth, and tenth semester students. The gaze behavior differences between these semesters are not as large or clear as when compared to sixth semester students. Thus, there seems to be a gaze *behavioral plateau* once students reach these later semesters, where visual search behavior of OPT does not appear to change drastically.

However, this behavioral plateau aligns with the curriculum for higher semesters. Only the sixth semester students receive this OPT targeted training alongside lectures with a focus on radiology and clinical knowledge. Seventh semester students receive one other radiology lecture, but then the curriculum focuses more on dental care and orthodontology. After the seventh semester, there are no courses addressing OPT analysis, rather other concepts related to orthodontics, prosthetics, or diseases and treatment. Students in the 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.

Moreover, the tenth semester students are misclassified as seventh semester students relatively often. This finding could be due to lack of OPT exposure in the curriculum of 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 effect [159], where, at some point in their studies, they 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 et al. [140] found that radiologist trainees, though more accurate than novices at identifying anomalies in MRIs, 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 [140]. Future research could further compare students in their last semester at university against first year interns in order to determine if there are any changes in performance as well as visual search strategy.

Furthermore, we were able to distinguish OPT exploratory gaze behavior at a semester level through methods of scanpath classification. 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 semester students become less distinguishable in their gaze behavior, which could also be caused by 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 better performance poses additional, interesting questions for future research.

.

5.5 A deep semantic gaze embedding approach to scanpath classification

Even though conventional scanpath classification approaches were able to distinguish the slightest behavior difference between students, these methods do not realize the image content's effect on the gaze. As mentioned in section 2.3, conspicuous and subtle anomalies can highly affect expert and novice scanpaths as they need to adapt their decision making strategies. Certain areas can also be more prone to anomalies, which affects the context of the search. Scanpath classification has yet to recognize gaze patterns linked to these features (e.g. anatomical structure, subtle anomaly, technical error, etc.) unconstrained to a specific image.

In this work, we propose a method that incorporates high-level, deep neural network-generated image patch representations into classical scanpath comparison measures. This novel approach allows similar features across images to be recognized (see figure 5.9). We apply our method DeepScan to decode expertise from eye movements during dental radiograph inspection.

It is worth noting that this metric is not confined to dental expertise recognition, but rather developed with the intention for various use cases. It offers the future potential to assess a student's learning progress in realtime and to adapt stimulus material based on current aptitude, while not being restricted to the stimulus material used during creation of the classifier.

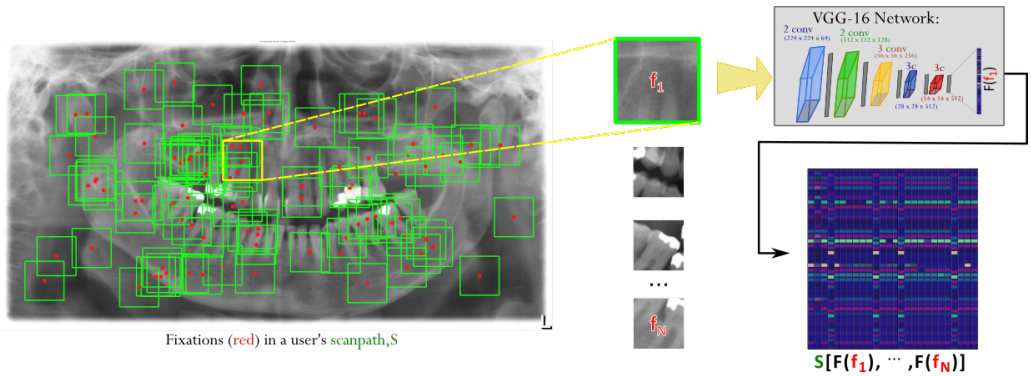


Figure 5.8: Proposed Model: DeepScan. For a scanpath, we extract the fixation locations and, using the VGG-16 CNN architecture, we create a feature corresponding to an image patch relative to the ith fixation F(f_i). The resulting vector illustrating the scanpath S can then be compared to another scanpath vector. The pre-trained VGG-16 network consist of 5 blocks of convolutions with ReLus with max-pooling between each layer.

5.5.1 Proposed approach: *DeepScan*

Scene semantics extracted from fixations Each individual fixation corresponds to the visual intake of a certain stimulus region. We then encode each fixation location on the specific stimulus image using a vector that describes the local image region. We generate such encodings via the output from the VGG-16 architecture [342]. Accordingly, for each fixation location on the stimulus image, we extract a patch of 100×100 pixels as input to the network. This step is relatively similar to [298], although we determined that using a fixed size bounding box is adequate for our stimuli. The fixation coordinates indicate the center of the image patch's bound box, unless a fixation is too close to the stimulus borders. Then, appropriate shifting of the box along the x or y axis is necessary.

The architecture we employed for patch processing originally takes 224×224 RGB input images. For the current evaluation, our stimuli were grayscale with pixel dimensions 1920×1080. In development, we determined that patch sizes of 224×224 for our stimuli were too large (e.g. four or more teeth would be in this sized area). Smaller patches were preferable so that enough information from an entity is extracted. Therefore, we rescaled the 100×100 image patches to the desired input size for the network, and replicated the one channel image information to get three channels that can utilize the weights pre-trained on ImageNet [343].

Image patch input size and channels can be adapted for other stimuli or any other preferred network for the fixation encodings. The takeaway from this image patch approach is that through only the gaze: 1) we map the image features of interest in temporal order, and 2) we can extract the semantics from these features.

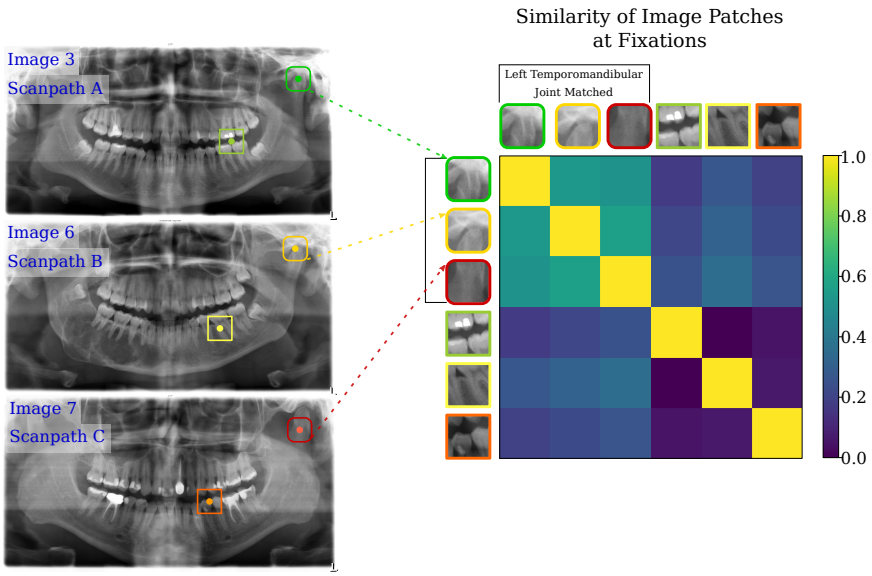


Figure 5.9: Matching image patch descriptors are recognized as similar across stimuli. When three different participants fixate on the left temporomandibular joint, the feature descriptors from DeepScan value them as similar. In contrast to when these participants fixate elsewhere, e.g. teeth, roots, etc.

CNN architecture For patch descriptor extraction, we employed a VGG-16 network [342] as implemented in keras and pre-trained on ImageNet.[4] Figure 5.8 shows the network: Consisting of five blocks of convolutions, with each block followed by rectified linear units (ReLUs) and max-pooling.

Since we are only interested in the features, we omit the fully-connected and prediction output layers of the model and use the output after max-pooling, which has $7 \times 7 \times 512$ dimensions, and flatten it to a $1 \times 1 \times 25088$ vector. This feature description from the final convolutional layer, $\mathbf{F}(f_i)$, represents the image patch at the ith fixation f_i.

The feature descriptors provide the semantic information for each fixation in a user's scanpath and are the equivalent of a symbol representation in the traditional string-sequence representation. Below, we discuss the changes required in the alignment algorithm in order to work with alignment scores generated by comparing these image features to one another. Figure 5.9 shows an example of how similar features accross different stimuli can be determined as similar.

We chose the VGG-16 in contrast to a network pre-trained on radiology images since it generalizes to a variety of tasks and stimuli. Additionally, it is publicly available and easily integrated for replication purposes. Pre-trained networks for medical images are often not publicly available due to data sensitivity and protec-

[4] Using Python 3.6 with GPU compatibility [344].

tion. Furthermore, any existing architectures for these images are not yet up to par with the generic image trained architectures. Choosing a network that is trained for a specific stimulus category, e.g., panoramic radiographs or other x-rays, might improve results. However, it introduces the risk of limiting data analysis to specific elements, which is comparable to manual AOI selection. Ultimately, though our approach is evaluated on medical image expertise, we developed it for generalizabitily in multiple applications.

Local alignment Once we have descriptors for each fixation, we assemble them as a scanpath. The resulting matrix of image features at each fixation creates a scanpath matrix. $S_A = (F_{f_1}, F_{f_2}, \cdots, F_{f_N})$. With this matrix representation, we can compare it to the matrix representing another scanpath.

For scanpath comparison, we perform local alignment using a variant of the Smith-Waterman Algorithm. We preferred local alignment scoring over global alignment due to its ability to find similar subsequences, even if the scanpaths may otherwise be vary greatly [215]. Moreover, we did not want to enforce strict global alignment due to different viewing times required by students and experts. In sequence alignment, the penalty system can have a major effect on values in the scoring matrix, which effects the similarity score [206]. Our scoring choice prioritizes finding long rather than short similar subsequences by accumulating scores. Equation 5.1 details the scoring system used for the current evaluation:

$$M_{ij} = \max \begin{cases} M_{i-1,j-1} + \ c - \sum_{i,j} |A_{:F_j} - B_{:F_i}|, & \text{Match} \\ M_{i,j-1} - gap, & \text{Gap in A} \\ M_{i-1,j} - gap, & \text{Gap in B} \\ 0 & \text{Stop Criteria.} \end{cases} \tag{5.1}$$

Where M is the scoring matrix of size $(n+1) \times (m+1)$ for two scanpaths A and B with n and m fixations respectively. Element $M_{i,j}$ takes the maximum value based on if there is a match between the values at index j of scanpath A and index i of scanpath B. The original algorithm scores matches as the score value added to the value at the previous indices: $M_{i-1,j-1} + score(a_j, b_i)$. Then, if there is no match, it determines whether the value of a gap penalty (gap) in either scanpath, or no similarity (0) are more optimal for the score.

The interesting part of our approach is contained in the calculation of the match value. Since it is highly unlikely that two features will be exactly the same, we cannot explicitly match or mismatch. Therefore, we calculate this value by taking the sum of absolute differences in feature descriptor j of scanpath A and descriptor i of scanpath B as shown in the first line of equation 5.1. This is simple to implement and cheap to compute, but other metrics such as cosine or Euclidean distance could also be used. This procedure leads to a dissimilarity value between the image patches. The more dissimilar the image patches, the larger the value.

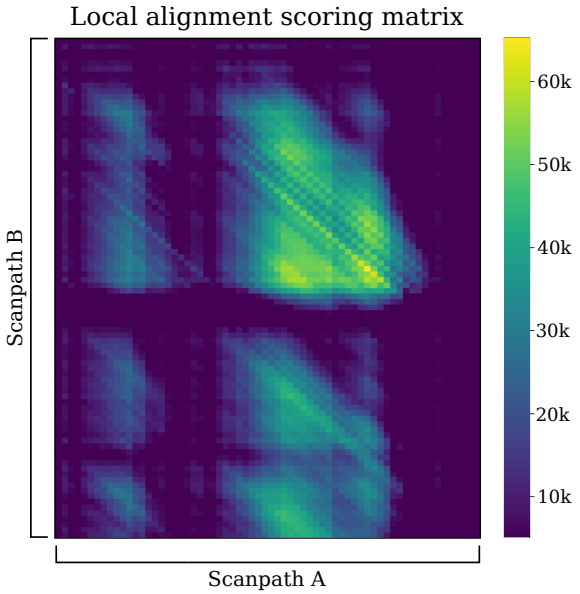


Figure 5.10: Scoring matrix of the local alignment. Backtracing from the index with the highest value (yellow) will give you the optimal local alignment of two scanpaths.

In order to evaluate the current image patch comparison to the stimuli context, we subtract the dissimilarity value from a constant c. We calculated c in equation 5.1 by averaging the sum of the differences for all features between all scanpaths of one random image in the dataset. Therefore, c was $21,049$ in the evaluation of our proposed approach. This constant affects highly similar image patches positively, but highly dissimilar image patches are penalized negatively with the same weight. Meaning it functions similar to a match/mismatch threshold. Additionally, we set our gap penalty (lines 2 and 3 in equation 5.1) to $42,000$ to highly penalize gaps, therefore almost double c.

This choice of c makes the algorithm consider about half of the image patches relatively dissimilar to each other. Furthermore, gaps are penalized quite strongly, resulting in compact alignments that are not drastically influenced by large differences in sequence lengths. Figure 5.10 shows an example of the similarity matrix created from the local alignment performed for two scanpaths. The maximum value in the matrix is the similarity score [214]. In figure 5.10, the highest yellow color indicates the final similarity score and backtracing from this index till 0 will give the optimal local alignment of both sequences.

The resulting similarity score for the two scanpaths is $\max(M)$. Then, we normalize this score based on the length of the shorter scanpath, thus:

$$similarity = \frac{\max(M)}{\min\left(|S_A|, |S_B|\right)}. \tag{5.2}$$

We compared the performance of our DeepScan method to the Smith-Waterman local alignment of hand-labeled semantic AOIs (see top example in figure 3.3), which serves as the benchmark for image context. These AOIs indicate specific anatomical structures and regions across dental radiographs, producing paramount semantic information that can be represented in a scanpath. For scoring the semantic scanpath comparisons, we used a simple, standard scoring system: 1 for matches, -1 for mismatches, and -2 for gaps.

For compatibility, we chose to evaluate gaze data from the first 45 seconds of each student participant, in line with the experts' total viewing time. Additionally, our model is only evaluated on gaze data for the 10 OPTs that both groups viewed. Gaze data was lost for two expert participants due to software failure. Five participants were also excluded due to having high data loss, leaving 25 experts and 54 students for the final analysis. The resulting total for all participants for all images was 733 scanpaths.

5.5.2 Results

We performed local alignment of the scanpath vectors with patch features for each participant for all images. In order to get the scanpath behavior representative of each participant, we averaged a participants' similarity output for all images. Figure 5.11 shows the similarity scores from DeepScan of each participants' scanpath behavior over the images viewed in pairwise comparison to other participants. The diagonal of the matrix indicates the highest similarity value, which is a participants' gaze behavior compared to his or herself.

From the similarity matrix, a trend is apparent where experts (labeled green in figure 5.11) show higher similarity scores among each other, as visible by the more yellow values. Conversely, students' gaze behavior shows less similarity, especially when compared to experts.

Hierarchical clustering We clustered the similarity scores of all participants using agglomerative hierarchical clustering [242], [245], [246]. As the similarity matrix can easily be inverted to a distance matrix, the unsupervised clustering approach was straightforward; however, one could introduce additional weighting factors or more complex classification methods on top as well. This approach evaluates the distance between data points and links those clusters closer in distance until one cluster remains [245]. Partitioning the clusters is then determined by the linkage distance. We used Ward's [245] method for proximity definition, which minimizes the sum of the squared distances of points from the cluster centroid.

Average gaze behavior of each subject For the scores of each student and expert summed over all images, the resulting dendrogram (2-dimensional tree view of the nested clusters) is shown on the y-axis in figure 5.11.

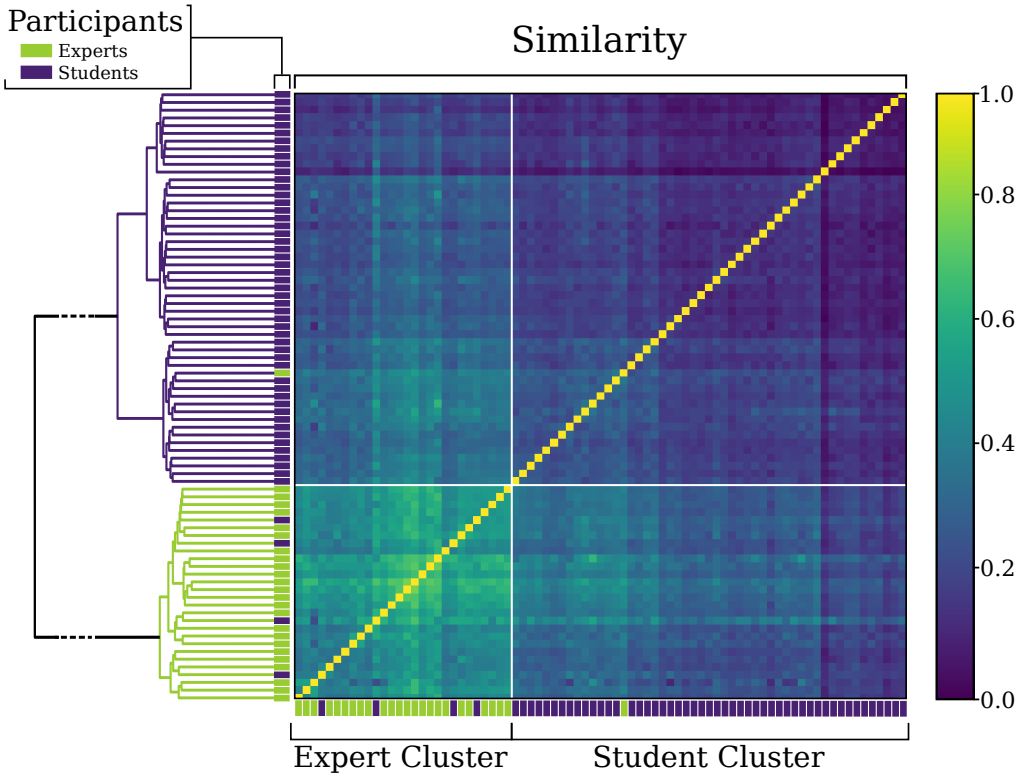


Figure 5.11: Similarity matrix of subjects' average scanpath behavior. Purple labels indicate students' gaze behavior. Green labels indicate experts' gaze behavior. Values closer to yellow indicate higher similarity, where the diagonal is a participant compared against themselves. Values shown on the diagonal are rescaled relative to values off-diagonal solely for perceivability. On the y-axis is the resulting clustering of the dendrogram, which recognized 2 clusters. One cluster (purple) with mainly students and the other cluster (green) with mainly experts.

Table 5.2: Performance of linkage clustering for our approach (*Feature*) and Semantic AOIs as measured by the True Positive Rate (TPR). Two main clusters were found based upon the gaze behavior for both approaches.

	Student		*Expert*		***Accuracy***	
	Feature	*Semantic*	*Feature*	*Semantic*	*Feature*	*Semantic*
Student	50	44	1	1		
Expert	4	10	24	24		
TPR	92.5 %	81.5 %	96.0 %	96.0 %	**93.7 %**	86.06 %

The clustering seen in figure 5.11 recognizes two main clusters evident in the gaze behavior with the majority of students in one cluster (purple cluster, purple labels) and the majority of experts (green cluster, green labels) in the other. Table 5.2 calculates the true positive rate (TPR) when utilizing the clustering as a classification for both students and experts as well as the overall accuracy. We achieved 93.7 % accuracy. We also found two clusters evident in the traditional local alignment with manual AOIs; however, more students were misplaced in the expert cluster (as seen in table 5.2), resulting in an overall accuracy of 86 %.

Gaze behavior on the image level We then ran the hierarchical clustering for participants' gaze at the image level (over all 733 datasets and not the average similarities for each participant as above). The dendrogram also recognized two clusters, therefore, we calculated the number of experts in one cluster and the number of students in the other. The achieved accuracy for our approach was 68.62 %: Experts had 85.65 % TPR and students had 61.18 % TPR. The achieved accuracy for the traditional, semantic approach was 64.39 %: Experts had 51.76 % TPR and students had 93.27 % TPR. This slight dip in performance could be attributed to pathological differences in the stimuli. Previous literature has also found that gaze behavior of expert and novice dentists can be highly stimulus dependent, where dental radiographs considered easy to interpret evoke similar gaze behavior in experts and novices [65], [149].

Cross-image classification To further see whether we could predict classification performance on an image level, we performed a leave one subject and one image out cross-validation using the similarity scores from DeepScan. We performed classification to 1) see whether we could predict a participant's expertise from their scanpath on a new image, i.e. not contained in the compared set. 2) to confirm that certain stimuli may affect the similarities more than others. For each subject, we ran a 3-Nearest Neighbor classifier, trained on the remaining subjects and images. Table 5.3 shows the performance for each image. Here, it is clear that for some images, distinguishing expert and student scanpaths becomes more difficult. For instance, image 1 shows a heavy tendency to classify all participants'

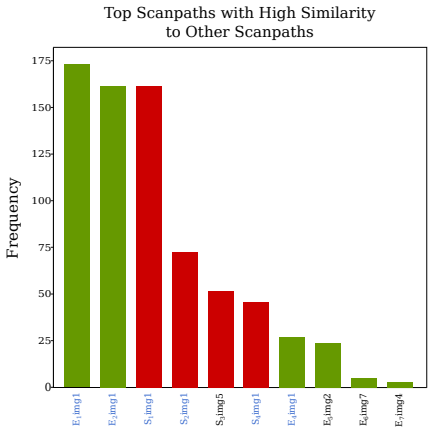
Table 5.3: Performance of kNN classifier when one image is left out and each participants' expertise for that image is predicted. Note that chance level is not 50 %, therefore we provide Cohen's Kappa (κ) as an indicator of performance, with bold text indicating fair performance.

	Expert TPR		*Student TPR*		***Accuracy***			
	Feature	*Semantic*	*Feature*	*Semantic*	*Feature*		*Semantic*	
Chance:	32 %		68 %		*Overall*	κ	*Overall*	κ
Image 1	100.0 %	75.0 %	20.4 %	76.6 %	44.9 %	0.14	**78.2 %**	0.52
Image 2	59.1 %	68.2 %	83.3 %	85.4 %	**75.7 %**	0.43	**80.0 %**	0.54
Image 3	28.6 %	66.7 %	93.5 %	80.4 %	73.1 %	0.26	**76.1 %**	0.46
Image 4	52.4 %	57.1 %	89.8 %	83.7 %	**78.6 %**	0.45	**75.7 %**	0.41
Image 5	76.2 %	53.4 %	68.6 %	88.2 %	70.8 %	0.39	**77.8 %**	0.43
Image 6	66.7 %	75.0 %	67.9 %	81.1 %	65.5 %	0.31	**79.2 %**	0.54
Image 7	60.9 %	30.4 %	86.5 %	90.4 %	**78.7 %**	0.49	72.0 %	0.24
Image 8	73.9 %	91.3 %	88.2 %	68.6 %	**83.8 %**	0.62	**75.7 %**	0.51
Image 9	45.8 %	58.3 %	92.6 %	96.3 %	**78.2 %**	0.43	**84.6 %**	0.60
Image 10	30.0 %	80.0 %	96.2 %	65.4 %	77.8 %	0.32	69.4 %	0.37
Overall	60.1 %	65.5 %	78.2 %	82 %	72.7 %	0.37	**76.9 %**	0.46

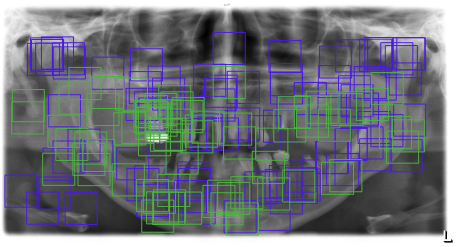
scanpaths for that image as experts, and image 3 shows a tendency to over-classify as students. Nevertheless, five images allowed us to determine expertise of a new subject on a new stimulus that was not contained in the data we used for the classification. In particular, image 8 shows the highest accuracy in classifying level of expertise, meaning this OPT and its semantics can possibly trigger experts to inspect the image in a distinctive way.

The cross-validation for the traditional local alignment scoring for the scanpaths with manual AOIs, showed better performance on the image level than DeepScan, and slightly better overall (77 % versus 73 % respectively). Therefore, it is possible that we cannot yet utilize the full potential of semantic encoding using the feature approach. However, given that DeepScan is purely data driven, its results are comparable and lessens the tedious process of manual AOI labeling. Retraining the network on OPT data might help the encoding come closer to manually-defined semantic labels.

Additionally, we sorted the similarity scores of all scanpaths from DeepScan to isolate those that expose especially high similarity values to many other scanpaths. We hoped to extract archetype-scanpaths this way. The histogram in figure 5.12a shows that two expert scanpaths had the highest similarity scores to most other scanpaths. Interestingly enough, both these scanpaths and a number of the other high similarity scanpaths are for image 1. Thus, from the local alignment similarity, certain scanpaths from image 1 (as seen in figure 5.12a) offer highly similar subsequences to other scanpaths regardless of image. Image 1 was one of the stimuli



(a) The top scanpaths who have the highest frequencies of similarities to other scanpaths.



(b) The two experts' scanpaths on image 1 with highly similar scanpaths to other subjects.

Figure 5.12: The two experts' scanpaths (illustrated by their image patches in blue and green) with the most highest similarities to each other and many other subjects' scanpaths based on the data in 5.12a. In 5.12a, expert scanpaths are in green and students' are in red. The majority of theses scanpaths are for image 1, as indicated by the blue text

that made a distinction between expertise levels difficult to discern. It could represent a standard scanpath for checking OPTs that abstracts over special attributes of individual stimuli.

5.5.3 Expert dentists exhibit highly similar attention to features

We were able to successfully extract similarities in scanpath behaviors between experts as well as their difference from student gaze behavior while interpreting panoramic dental radiographs. Our developed scanpath comparison approach uses temporal scanpath information to extract image features at the fixation level. The resulting similarity comparison of scanpaths, therefore, incorporates this image information into the traditional approach of sequence alignment to detect patterns between the behaviors.

From traditional local alignment techniques using image features, we found that experts showed highly similar behavior to each other and, as a result were more likely to be clustered together. Moreover, students' similarity scores indicated that their scanpaths were not highly similar to those of experts. In addition, there was no distinct homogeneity among students' similarity scores (see figure 5.13). One possible reason for their low similarity to one another could be that they

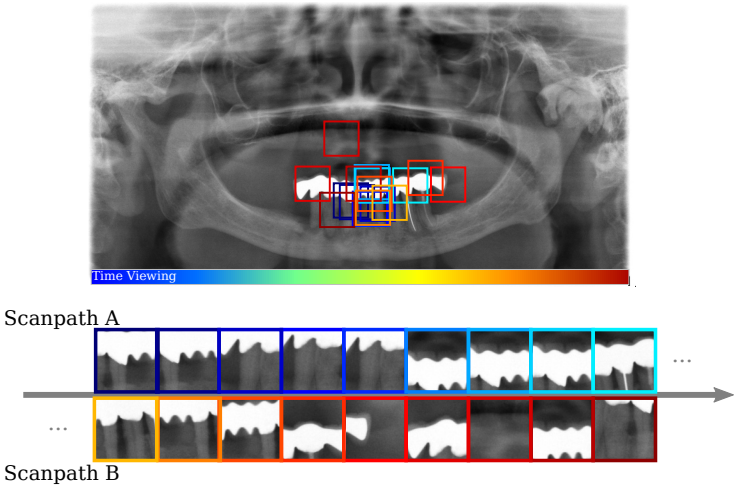


Figure 5.13: Two relatively dissimilar scanpaths from students. The local alignment finds the optimal matching subsequence starting in scanpath A at the twentieth fixation (far left top) and in scanpath B at the fiftieth fixation (far left bottom).

are incoming students with some conceptual background; however, they had no training on radiograph interpretation. Previous research has found that students evoke more systematic search strategies after training, resulting in more similar gaze behaviors [28], [158]. Additionally, the heterogeneity of background and training can affect scanpath similarity [345]. Students potentially have varying levels of conceptual knowledge or familiarity with radiographs before entering their first year of dental studies.

Our algorithm was able to accurately classify unseen scanpaths given scanpaths from other participants and other images. Although, we found that, depending on the image, it could be easier or more difficult to differentiate levels of expertise from the scanpath similarities. This finding is, however, in alignment with previous studies specifically on dentists and dental radiograph examination. For instance, [149] found that radiographs defined as easy to interpret offered no differences in the gaze behavior of experts and novices. We [3] also found that even among experts some images evoked highly differing gaze behavior to achieve accurate anomaly detection.

With the system at hand, we could classify expertise of dental students in an adaptive feedback setting from viewing a single stimulus (with decent accuracy), even if the stimulus itself is an arbitrary OPT that is unknown to the classifier. This could be used to guide students through the learning process and to adapt the difficulty of stimulus material to their current knowledge level. When viewing multiple stimuli (which students do in the current mass practice approach), classification accuracy can be increased.

Furthermore, we observed that some stimuli allowed for a classification of expertise, while others did not. We could utilize this information as a hint to which stimuli are likely to induce a training effect and to differentiate from stimuli that are too easy (for the current student).

We designed DeepScan to handle image variability. One image feature descriptor of a patch in one image can match to similar patches in other images (see figure 5.9); This way, scanpaths can be more effectively compared cross-stimuli, but this process also replaces a manual AOI-annotation. By the assumption that similar semantic meaning in a visual task corresponds to similar looking features in the stimulus, we have introduced a notion of stimulus semantics into the automated scanpath interpretation. A similar workflow could be used to compare data where the annotation of dynamic AOIs is typically unfeasible, e.g., recordings of mobile eye-tracking devices to each other. Furthermore, we do not restrict the algorithm to individual annotated AOIs, but represent each fixation by its feature descriptor, no matter whether a data analyst would deem it relevant for the analysis at hand or not.

5.6 Toward developing gaze-based interventions

In this work, we focused on realtime gaze-contingent feedback that visualizes already observed regions and incorporates more information from the periphery. Additionally, we introduced a novel software system for performing eye tracking experiments, which allows for realtime feedback to the subject. We successfully validated the implementation in a visual search task study. The current system was integrated into an already existing eye-tracking analysis platform – *EyeTrace* [346].

5.6.1 Gaze-contingent software

Software The *Experimenter* plug-in for EyeTrace [346] was developed for creating and performing remote eye-tracking experiments. It offers the following capabilities that are controllable in the designer widget as illustrated in Figure 5.14:

- Create and modify the experiment design where each index block is highly customizable ① & ②.
- Import and export experiment designs as CSV file ②.
- Record subject data together with name, group and dominant eye ③.
- Select the eye tracker to be used ④.
- Select an interruption key ⑤.
- Start/cancel the experiment run ⑤.

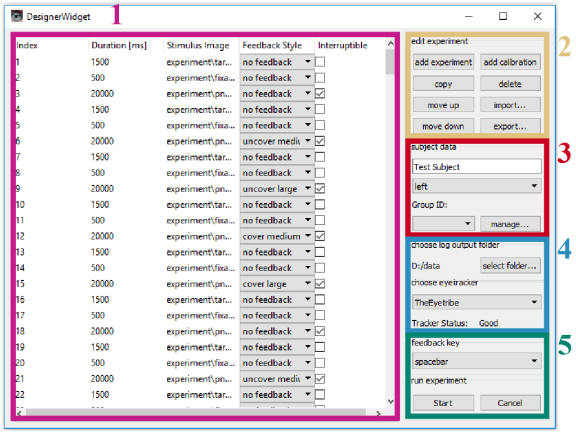


Figure 5.14: Designer Widget GUI. Here, experiments can be designed, and managed. The workflow of the experiment is organized (1) and can be modified (2) and each participant's data is defined (3). For each experiment, an eye tracker is selected (4) as well as a key for interruption (5).

Additionally, a researcher can manually organize an experiment design, offering customization of stimuli, time of presentation, gazefeedback, and keypress interruptions. These experimental designs can be exported and saved as a CSV file, for additional data collection. The ability to import experiment designs in CSV file format allows for the option of auto-generating randomized experiment designs with a simple script in any programming language with CSV parsing libraries or text editor.

During an experiment, the stimuli and gaze-contingent feedback are visualized in the presenter widget. This widget handles any pre-defined presentation timeouts, our key-event triggers to present a new stimuli.

Gaze-contingent feedback For the realtime, gaze-contingent feedback, the user's gaze is visualized on the screen as he or she is performing a task. In order to achieve low and relatively constant response times from the feedback system, intermediate results are stored in a cache. Then, the system only has to process new gaze data when it is repeatedly called. Triggered by a timer, every 7 ms (approximately 144 Hz), the screen drawing method of the presenter widget gets called to update the screen content. This trigger calls the currently active realtime feedback implementation to draw over the stimulus.

Two feedback algorithms were implemented, plus the default 'no feedback' condition. First, the 'cover' feedback occludes the user's gaze coordinates on the stimulus with opaque circles as illustrated in figure 5.15a. Second, the 'uncover' feedback unoccludes a semitransparent cover in a similar manner to the former condition as illustrated in figure 5.15b. Essentially, this feedback is the complement of the former applied to the mask overlay. For both conditions, there is no decay of

Figure 5.15: Screenshot of an experiment trial, showcasing the 'cover' (a) and 'uncover' (b) feedback condition. Stimulus: Ilya Repin, "Unexpected Visitors", 1884-1888. Oil on canvas. public domain `https://commons.wikimedia.org/wiki/`.

feedback for older gaze points.

Both feedback conditions use a white mask-like image overlayed over the original stimuli, and the feedback effects the masks' alpha channel, meaning its opacity is changed. For each event where the feedback class is called, the list of new gaze points is run through and circles are drawn on the mask overlay for each new gaze point coordinate.

In both feedback conditions, the mask is either transparent (for covered) or semi-transparent (for uncovered). The alpha channel on this map is then changed based on the gaze coordinates. The compositing method adds or subtracts the circle's alpha values to the existing mask, corresponding to gaze coordinates. This creates the effect of decreasing or increasing transparency the longer the subject looks at a certain spot. However, a lower bound threshold is given to the circle's alpha value to prevent any part of the stimulus from becoming invisible. Each circle consists of a radial gradient, projecting outwards from the circle's center to its edge. This effect makes the feedback appear smoother, removing distracting, sharp edges (see figure 5.15). The compositing method updates the mask with new gaze points each time the trigger timer event takes place, then, the mask gets drawn over the stimulus.

Gaze-contingency in a visual search paradigm The visual search task was performed with images consisting of either 80 distractors (target absent) or 79 distractors plus the target item (target present). In total, 100 images were generated.[5] Order of stimuli presentation was randomized for each participant. A total

[5]The images had equal distribution of color and shape of the target item, and its absence/presence.

of 18 participants (17 university students; five wore glasses) took part. They were positioned roughly 60 cm away from the screen and gaze position was collected using an EyeTribe commercial eye tracker. A 9-point calibration was performed using the EyeTribe's calibration software. Following the experiment, participants filled out a self-report regarding perceived performance and experience.

Generally, it took subjects longer to correctly decide if a target was absent, than it took them to decide if a target was present. The difference was highly significant; this experimental result reproduces a well documented effect of target presence or absence on reaction time [347], [348].

Concerning how the intervention influenced subject behavior, we can see that even with our rather simple feedback methods we were able to induce a change in subjects. There is more periphery incorporated, which is evident by longer saccades for both the covering and uncovering interventions. However, only the cover large condition, where a semi-transparent circle with a 200 pixel diameter was overlayed on the gaze coordinates, increased reaction times when the target was accurately determined as absent. There was also a trend for less fixations when determining the target was absent for both cover and uncover (where the circle uncovers a semi-transparent overlay) large feedback conditions, though significantly less fixations were found only in the cover small (100 pixel diameter) feedback condition. Therefore, the realtime, gaze-based feedback algorithms produced an effect on the gaze behavior in the visual search task.

Interestingly enough, self-reports from the participants did not indicate that the feedback helped or improved their performance. In contrast, their reaction times, as well as their eye movement differences, showed that gaze feedback indeed affected their behavior compared to no gaze feedback. Participants also reported that none of the feedback conditions were distracting in any way. Therefore, the gaze feedback system we developed appears to be unobtrusive, yet effective.

A large cover feedback could be a more effective gaze contingency model as it decreased reaction time for when a target was correctly determined as absent. However, overall correct detection was extremely high at 96 %. Future work into effective gaze modeling could look into more complex visual search tasks to see whether gaze modeling improves performance.

5.6.2 Towards attention awareness: Gaze-aware subtle feedback intervention

Building off the aforementioned framework, we combined domains that have previously run in parallel: Expert gaze modeling for learning and user-attention awareness. We designed a framework for gaze guiding based on expert viewing behavior of dental radiographs while recognizing a user's realtime gaze. Our interests are two-fold: 1) whether we can effectively guide a user's gaze to relevant regions of an image without occluding any information and 2) whether expert gaze guiding can improve perceptibility of anomaly features for non-experts. We present an ex-

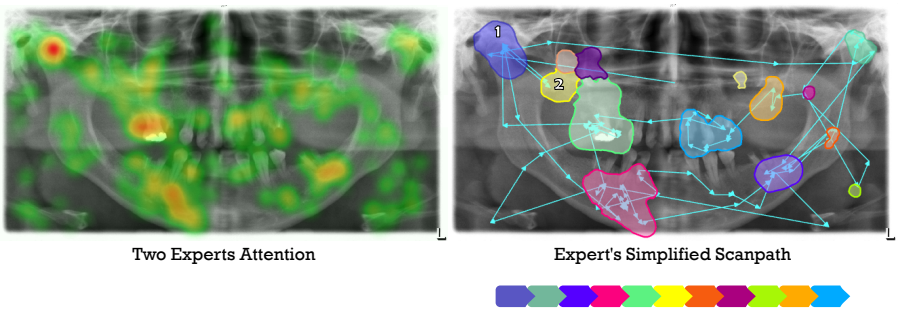


Figure 5.16: Gaze guiding through experts' attention. AOIs are calculated from the heatmap. Then, the simplified transitional behavior becomes the order of presentation.

ploratory evaluation of the intervention design with naive participants, accessing its efficacy in guiding the gaze unobtrusively as well as in providing usability feedback. Additionally, we look at detected anomaly features; however, we are aware that diagnostic performance would be more appropriately evaluated with students and advanced trainees, who have a more appropriate skill set for pathology interpretation.

The expert gaze model We employ subtle gaze direction [349] to present expert attention while examining panoramic dental radiographs. Our method does not occlude relevant areas in the foveal vision, as it recognizes when attention is directed towards the area. We could successfully guide the gaze to relevant image features and promoted further inspection. Our findings with naive participants showed that the gaze feedback could not develop successful dental radiograph diagnosis, but elicited gaze transitions similar to the expert model. Participants also felt more confident and stated that the framework helped them to properly inspect radiographs. This aspect suggests further research to promote SGD as a suitable way to illustrate expert gaze behavior in learning interventions with students or advanced trainees.

To create the AOIs, we chose gaze data from two experts from a previous data collection with expert OPT inspection. Experts from this data collection had an average of 10 years of experience. Through similarity clustering, two experts were found to have scanpaths highly similar to all other experts' scanpath (see [5] for further details); their data was chosen to develop the expert model. From their heatmap, areas with higher concentration of gaze are segmented as illustrated in the right image in figure 5.16. We chose the scanpath of the more accurate (higher detected anomalies) of the two experts to provide transitional behavior. We preferred a simplified version of the transition, denoting the first glance into an AOI and not including revisits, since it was determined that revisits would be too hard to follow. An example of a simplified scanpath is also found in figure 5.16:

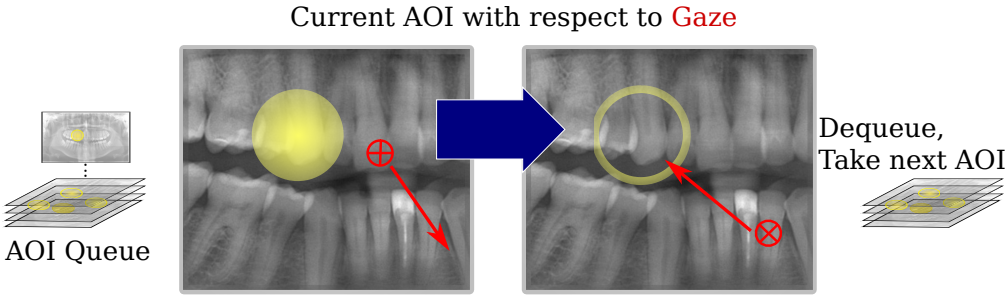


Figure 5.17: Illustration of feedback animation. When the gaze attention (red cross-hair) is not directed towards the AOI, it pops up as a semi-translucent yellow circle (left image). When the gaze attention goes towards or is in the AOI, it presents the feedback as a translucent yellow ring (right image).

The first blue AOI is looked at (1) then transitions to four other AOIs were made before going back to the first AOI, we omit the revisit and set the next transition to the yellow AOI (2). Without revisits, scanpaths ranged from 9 to 23 transitions, and with revisits, they ranged from 88 to 175 transitions.

Software We incorporated the AOIs and the ability to recognize attention towards them. Our method is based off of the subtle gaze direction (SGD) method from [349]. We added a short delay of 5 seconds, before the first AOI pops up, so participants could scan the image briefly.

AOIs for a certain feedback are placed into a queue. Upon an animation timer timeout, the current AOI is dequed and painted over the stimulus. For this work, we set the timer to timeout every 3.8 seconds so participants would not feel rushed, as they were non-experts. The AOI is initially illustrated as yellow ($RGB: 252, 252, 103$) with a translucent radial gradient (left image in figure 5.17). We chose this color as we felt it would be salient against our grayscale stimuli.

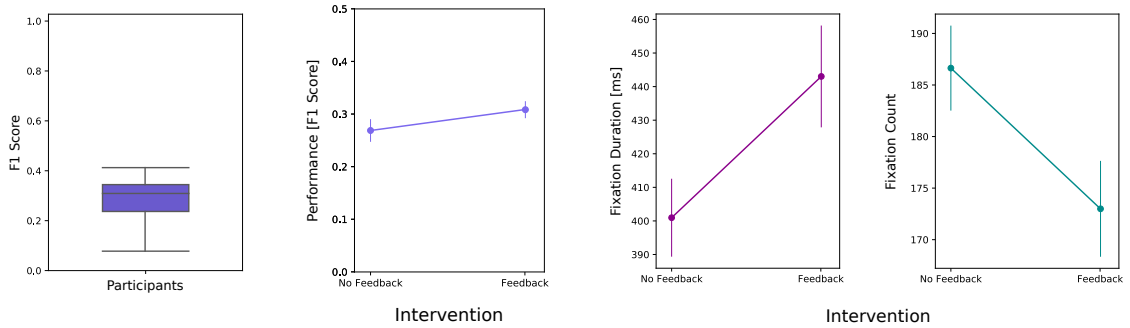
In order to avoid occlusion of important image features, we repaint the AOI area with a translucent yellow ring (right image in figure 5.17), when our SGD implementation detects the gaze angle as going towards the AOI. Where the angle, α, is calculated as follows:

$$\alpha = \cos^{-1}\left(\frac{\overrightarrow{v} \cdot \overrightarrow{t}}{|\overrightarrow{v}| \cdot |\overrightarrow{t}|}\right), \tag{5.3}$$

where $\overrightarrow{v}$ indicates the vector from the previous gaze point to the current gaze point and $\overrightarrow{t}$ indicates the vector from the previous gaze point to the target AOI. We calculate α five times using equation 5.3: with one $\overrightarrow{t}$ to center coordinates of the AOI and then $\overrightarrow{t}$ for each of the corner coordinates of its bounding box. We

calculate the previous gaze as the average of the last two gaze coordinates stored in a buffer. We take the minimum of the five angles and subtract it from $360°$ if it is larger than $180°$.

Then, if α is between 0 and $10°$, the AOI updates from the circle to the ring. This threshold was used in [349], and was determined stable when testing our implementation. For gaze input, we used the SMI RED250 remote eye tracker running at 60Hz



(a) Non-Expert Performance

(b) Performance with Intervention

(c) Gaze Behavior with Intervention

Figure 5.18: Performance as measured by the F1 Score (a) overall images (b) comparing the intervention of expert gaze feedback against feedback, and (c) the gaze behavior with respect to feedback or no feedback. Means (circles) and standard errors (tails) are plotted for all figures.

Performance, gaze, and attention We calculated the sensitivity and precision of the participants over all images, then calculated the harmonic mean (F1 score) between the metrics. As was expected with non experts, performance in OPT anomaly detection was relatively low: The average F1 score overall was $M = 28.42\,\%$, $SD = 8.45$. The distribution is shown in figure 5.18a.

To see if there were any effects of the expert gaze feedback intervention, we ran a repeated measures t-test on both the performance and the gaze behavior for "feedback" versus "no feedback" conditions. No major effect was found for the intervention on performance ($t(26) = -2.021$, $p = 0.054$), though performance with the feedback was slightly better ($M = 30.80\,\%$, $SD = 8.23$) than without it ($M = 26.85\,\%$, $SD = 10.97$). Figure 5.18b shows the performance with respect to the intervention. The low sample size may explain the high variance in the gaze behavior for both the intervention and no intervention condition. Further research with an appropriate sample size to observe a significant difference is necessary.

However, the intervention had a stronger effect on the gaze behavior. Average fixation durations were higher for the feedback condition ($M = 443.03$, $SD = 78.76$) compared to the no feedback condition ($M = 400.96ms$, $SD = 60.29$, $t(26) = -4.704$, $p < 0.0001$). Additionally, the average fixation count for the

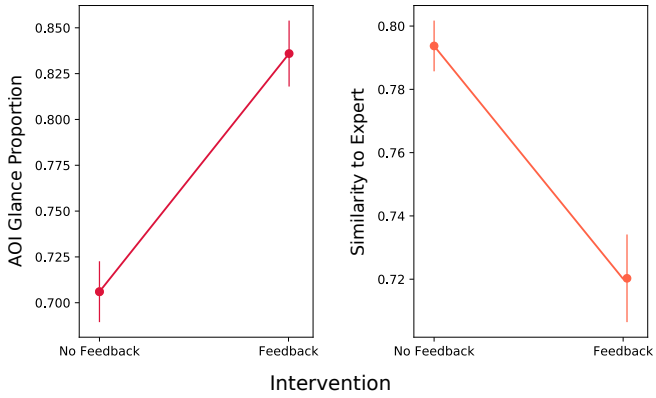


Figure 5.19: Attention to AOIs as measured by the AOI glance proportion (left) and the scanpath similarity (Levenshtein distance) to the expert model (right) with respect to feedback or no feedback.

feedback condition was lower ($M = 173.0$, $SD = 24.15$) than the no feedback condition ($M = 186.64$, $SD = 21.41$, $t(26) = 4.502$, $p = 0.00012$). Therefore, when presented with the expert gaze model, participants exhibited fewer fixations, but longer fixation durations. This behavior could be indicative of more information processing and, as a result, associated with novices [24], [133], [137], [158].

To assess whether the intervention successfully guided the gaze behavior, we looked at subjects' gaze behavior in relation to the AOIs as shown in figure 5.19. We ran repeated measures t-test for AOI glances and transition similarity. The gaze model elicited a higher proportion of AOI glances ($M = 0.8359$, $SD = 0.0935$) than without the model ($M = 0.7060$, $SD = 0.0863$, $t(26) = -8.165$, $p < 0.0001$). Consequently, there was more attention to relevant areas of the image.

We looked at the effect of the intervention on the similarity of subjects' AOI transitions to the expert's transition. Similarity was calculated with the Levenshtein distance [208] for subjects' scanpaths compared to the expert's scanpath and normalized to the length of the longest scanpath. We found that with the feedback, subjects had significantly more similarity to the expert (M $= 0.7203$, SD $= 0.072$) than without the feedback (M $= 0.7937$, SD $= 0.0416$, $t(26) = 4.791, p < 0.0001$). Figure 5.20 shows subjects' transitional information for one image with (middle) and without (right) the intervention compared to the expert's gaze transitions relative to the AOIs (Left). Here, the similarity of the subjects who received the gaze feedback was closer to the expert's gaze behavior than the subjects who received no feedback compared to the expert: Note the transitions to (lines originating) and from (lines landing) AOI 5 (burgundy).

User response Regarding usability, we asked subjects to fill out a short questionnaire about the task and the gaze feedback. Average responses for the questions are plotted in figure 5.21. Overall, the subjects found the task difficult and were not

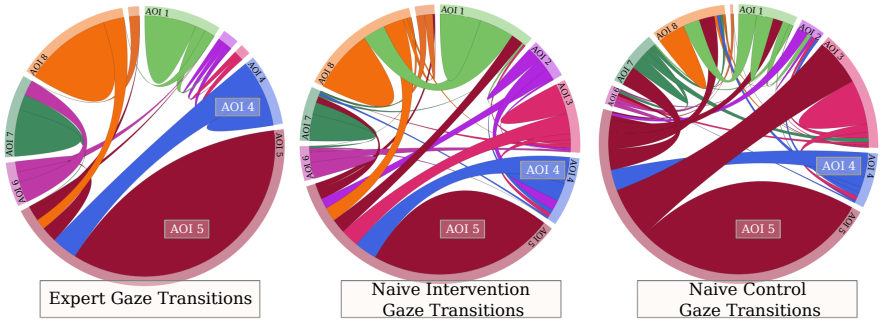


Figure 5.20: Example of AOI transitions for one image. Where the left most diagram is the expert's transitional information and the middle is the transitional information of subjects who received the gaze intervention and the right most is the transitional information of subjects who received no gaze intervention .

confident in their performance. This could be expected as the nature of anomalies in these images are likely to be very subtle to the untrained eye. Moreover, they shared overall positive feedback regarding the intervention, finding it beneficial and depending on it to complete the task. Some participants made informal comments to the researchers that, after a few images with interventions, they started to recognize features (e.g. dark shadows in the gums), which they felt could be indicative of something abnormal (peridontitis). They did, however, find the task a bit too long and slightly rushed. These responses will be helpful for future testing and system development.

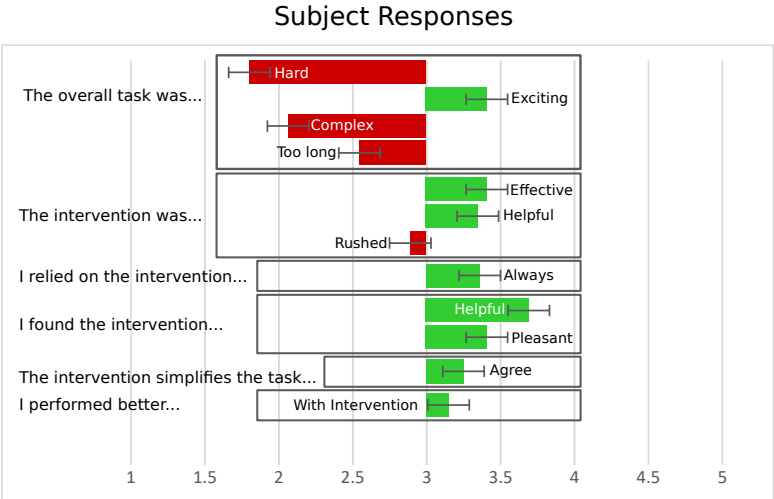


Figure 5.21: Average responses for questionnaire regarding the task and the gaze intervention. The task was reported as difficult for the non-experts. They did report that the feedback was helpful and they used it.

Participants reported feeling more confident with the gaze intervention and relied on it to complete the task. They successfully followed the expert gaze model and their AOI transitional behavior was more similar to the expert's. Although they lacked the conceptual knowledge that facilitates the proper interpretation of relevant features. Previous research has also indicated that search pattern training draws attention to relevant areas, but does not affect performance [158], [174], [350]. Waite et al. [51] highlights the reciprocity of perception and cognition in diagnostic performance. For instance, initial feature localization, then conceptual knowledge, facilitates the decision that this feature needs further inspection (e.g. difference in contrast, and area prone to anomalies, etc.) and whether it is recognized as a specific pathology or can be ruled out.

Implications for novice training

Dental radiographs, as with all medical images, are highly complex in nature and require some form of conceptual knowledge to interpret reliably. Presenting only ten OPTs may not have been enough for a significant training effect. Considering the low number of OPTs, naive participants seemed to recognize features the intervention highlighted in later images, as reported. To achieve improved performance in naive observers, [351] used around 800 images to improve hip fracture detection. Further research is needed that will address the optimal number of images required to improve interpretation without inducing fatigue and still providing ample time to interact with the gaze model. In our study, we were limited to investigating short term effects of training naive participants. A longitudinal study regarding the gaze-aware feedback system on naive subjects' or novices' learning over time would be an interesting aspect for further research.

Moreover, we show a potentially effective learning intervention for either novices or more advanced dentists. Students undergo intense studying and exposure to get to the level of professional expertise that leads to success later in their careers. More effective learning interventions can smooth a student's transition to residency and professional environments by minimizing the knowledge gap between each stage. With better preparation, less professional resources need to be expended on supervising incoming residents and early professionals. Even then, *expert* is never a final state, experts should always be open to further learning and improvement. Generally, it has been found that experts and more advanced trainees greatly benefit from gaze interventions [28], [174]. Our implementation of the SGD with expert AOIs has the potential to be catered to advanced learners, in hopes of further fine-tuning established skills.

6 Outlook

> Talent is a pursued interest . . . Anything that you're willing to practice, you can do
>
> –*Bob Ross*

Research will not find the key to make someone an expert overnight. Expert skills cannot simply be just handed to anyone, rather they need to be developed through "rich instrumental experiences" [46]. However, research on what makes experts excel in their domain specific tasks offers insight for learning interventions. In medical education, research in expert-novice differences has already produced *problem-based learning*, which supplements conventional coursework [352]–[354]. Even the study of expert eye movements and the creation of eye movement models have started to reveal the potential for expert gaze research to support conventional teaching [174], [176], [350].

The tracking of expert gaze behavior creates an awareness of their cognitive processing and decision making. This thesis has contributed towards this awareness by investigating expert glance behavior and the recognition of anomalies [3], which supports previous work on false negative and attention [58]. Furthering expert decision making, fresh insight is given into the versatility of expert cognitive processing when inspecting varying anomalies [4]. Moving towards advanced scanpath analysis from a data-driven perspective, subtle differences in student scanpaths due to targeted training are distinguished with conventional scanpath classification approaches. Moreover, an innovative classification method to recognize attention to semantic features rather than an image shows robust distinction between experts and novices in [5]. This work paves the way for classifying expertise in an adaptive feedback setting, even for unseen stimuli. Working towards intelligent tutoring systems that use gaze as indicator of user perception, a framework was developed in [2] and then later evaluated with an expert gaze model [6]. This system shows high usability and effective attention guiding, without occlusion of relevant features.

Understanding the scanpaths in an effort to find patterns determinant of a developmental level can ultimately build an adequate representation of eye movements for the complete learning process. Therefore, a model initially should recognize gaze patterns (i.e. subsequences) that are characteristic for a dentist at a respective expertise level. Then, building off accurate recognition, scanpath components can further be clustered based on patterns representative to key phases in effective vi-

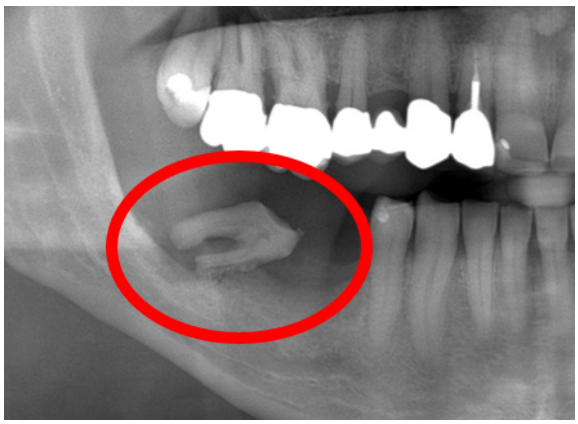


Figure 6.1: Example of an obvious and highly salient anomaly (circled in red) that experts and novices alike recognize. Even a layman could determine this tooth is not properly positioned. Regardless of expertise level, this pathology was often the first fixated.

sual search, i.e. systematic, comparative or explorative. Such patterns are likely to contain highly discriminative information, which are not bound to e.g. one specific OPT, rather that can be linked to the specific semantics of a certain structure or anomaly. This thesis provides a substantial contribution towards expertise cognition through the gaze and robust scanpath classification towards gaze augmented intelligent tutoring.

Furthermore, this thesis and the observations reported herein shows the usefulness of gaze and robust scanpath classification and its application for gaze-augmented tutoring in training of expertise cognition. It approaches online recognition of expert or novice, and even more fine-grained sub-divisions within a group. Not only can the findings of this work be targeted toward students, but also how advanced students, upcoming professionals and even experts can be affected. For instance, interventions could be designed to train more successful strategies and experts. One interesting question whether a manifested and observed strategy in an expert is pliable and open for change, to what degree can we improve expertise by building on that strategy.

Research involving data-driven approaches for bias recognition can highly benefit from the gaze. Already, understanding bias from data has major contributions social and economic factors can contribute to decision making in AI [355]. Therefore, another further implication of this research is revealing bias in expert decision making through the gaze. The cognitive processes during expert decision making (pupil change adaptability and glance frequency during anomaly investigation) coupled with specific scanpath strategies can hint at expert bias. Gaze literature has shown that although experts generally attend less to salient features, they still recognize these obvious, highly salient anomalies such as the one in figure 6.1. Psychologically-based research has already started to uncover expert bias in their investigative processes. One area recognizes a phenomenon known as satisfaction of search, where they successfully recognize an anomaly thus terminate any further investigation [356]–[359]. For instance, one interesting study found experts radiologist asked to detect lung-nodules did not recognize or even fixate on a gorilla illustrated onto the radiograph [360]. Conversely, hindsight bias in

expert radiologists can lead them to more easily recognize previously overlooked anomalies [361]. Expert specialty has also shown to reflect bias in diagnostic hypotheses [362]. A cognitive model was developed in [363] for bias in diagnostic accuracy based on expert and novice pathologists, where a speed-accuracy tradeoff as well as prior expectation were evident in the model. Experts expected to receive slides already viewed by a technician or resident, thus a pathology would be present [363]. One benefit of using expert gaze for bias recognition could be misdetections. For instance, the exact interplay between highly salient anomalies and specialization can affect the global search of a radiograph, leading to overlooking a more subtle anomaly not generally of interest to said specialist though still problematic. Therefore, it is worth further investigation into how the gaze can represent expert bias, which can offer better assessment of the quality of the diagnosis.

References

[1] N. Castner, E. Kasneci, T. Kübler, K. Scheiter, J. Richter, T. Eder, F. Hüttig, and C. Keutel, "Scanpath comparison in medical image reading skills of dental students: Distinguishing stages of expertise development," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–9.

[2] K. Otto, N. Castner, D. Geisler, and E. Kasneci, "Development and evaluation of a gaze feedback system integrated into eyetrace," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–5.

[3] N. Castner, S. Klepper, L. Kopnarski, F. Hüttig, C. Keutel, K. Scheiter, J. Richter, T. Eder, and E. Kasneci, "Overlooking: The nature of gaze behavior and anomaly detection in expert dentists," in *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, ser. MCPMD '18, Boulder, Colorado: Association for Computing Machinery, 2018, ISBN: 9781450360722. DOI: 10.1145/3279810.3279845. [Online]. Available: https://doi.org/10.1145/3279810.3279845.

[4] N. Castner, T. Appel, T. Eder, J. Richter, K. Scheiter, C. Keutel, F. Hüttig, A. Duchowski, and E. Kasneci, "Pupil diameter differentiates expertise in dental radiography visual search," *Plos One*, vol. 15, no. 5, e0223941, 2020.

[5] N. Castner, T. C. Kuebler, K. Scheiter, J. Richter, T. Eder, F. Huettig, C. Keutel, and E. Kasneci, "Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Full Papers, Stuttgart, Germany: Association for Computing Machinery, 2020, ISBN: 9781450371339. DOI: 10.1145/3379155.3391320. [Online]. Available: https://doi.org/10.1145/3379155.3391320.

[6] N. Castner, L. Geßler, D. Geisler, F. Hüttig, and E. Kasneci, "Towards expert gaze modeling and recognition of a user's attention in realtime," in *24th International Conference on Knowledge-based and Intelligent Information & Engineering Systems*.

[7] T. F. Eder, J. Richter, K. Scheiter, C. Keutel, N. Castner, E. Kasneci, and F. Huettig, "How to support dental students in reading radiographs: Effects of a gaze-based compare-and-contrast intervention," *Advances in Health Sciences Education: Theory and Practice*, 2020.

[8] W. Fuhl, N. Castner, and E. Kasneci, "Histogram of oriented velocities for eye movement detection," in *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, ser. MCPMD '18, Boulder, Colorado: Association for Computing Machinery, 2018, ISBN: 9781450360722. DOI: 10.1145/3279810.3279843. [Online]. Available: https://doi.org/10.1145/3279810.3279843.

[9] W. Fuhl, N. Castner, and E. Kasneci, "Rule-based learning for eye movement type detection," in *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, ser. MCPMD '18, Boulder, Colorado: Association for Computing Machinery, 2018, ISBN: 9781450360722. DOI: 10.1145/3279810.3279844. [Online]. Available: https://doi.org/10.1145/3279810.3279844.

[10] W. Fuhl, N. Castner, T. Kübler, A. Lotz, W. Rosenstiel, and E. Kasneci, "Ferns for area of interest free scanpath classification," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '19, Denver, Colorado: Association for Computing Machinery, 2019, ISBN: 9781450367097. DOI: 10.1145/3314111.3319826. [Online]. Available: https://doi.org/10.1145/3314111.3319826.

[11] W. Fuhl, E. Bozkir, B. Hosp, N. Castner, D. Geisler, T. C. Santini, and E. Kasneci, "Encodji: Encoding gaze data into emoji space for an amusing scanpath classification approach ;)" in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '19, Denver, Colorado: Association for Computing Machinery, 2019, ISBN: 9781450367097. DOI: 10.1145/3314111.3323074. [Online]. Available: https://doi.org/10.1145/3314111.3323074.

[12] D. Geisler, N. Castner, G. Kasneci, and E. Kasneci, "A minhash approach for fast scanpath classification," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Full Papers, Stuttgart, Germany: Association for Computing Machinery, 2020, ISBN: 9781450371339. DOI: 10.1145/3379155.3391325. [Online]. Available: https://doi.org/10.1145/3379155.3391325.

[13] D. Geisler, D. Weber, N. Castner, and E. Kasneci, "Exploiting the gbvs for saliency aware gaze heatmaps," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Short Papers, Stuttgart, Germany: Association for Computing Machinery, 2020, ISBN: 9781450371346. DOI: 10.1145/3379156.3391367. [Online]. Available: https://doi.org/10.1145/3379156.3391367.

[14] K. A. Ericsson, R. R. Hoffman, and A. Kozbelt, Eds., *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, 2018.

[15] L. B. William and N. Harter, "Studies on the telegraphic language: The acquisition of a hierarchy of habits.," *Psychological Review*, vol. 6, no. 4, p. 345, 1899.

[16] W. G. Chase and H. A. Simon, "Perception in chess," *Cognitive Psychology*, vol. 4, no. 1, pp. 55–81, 1973.

[17] J. Shanteau, "Competence in experts: The role of task characteristics," *Organizational Behavior and Human Decision Processes*, vol. 53, no. 2, pp. 252–266, 1992.

[18] K. A. Ericsson, R. S. Perez, D Eccles, L. Lang, E Baker, J Bransford, K. VanLehn, and P. Ward, "The measurement and development of professional performance: An introduction to the topic and a background to the design and origin of this book," *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments*, pp. 1–26, 2009.

[19] K. A. Ericsson and A. C. Lehmann, "Expert and exceptional performance: Evidence of maximal adaptation to task constraints," *Annual Review of Psychology*, vol. 47, no. 1, pp. 273–305, 1996.

[20] K. A. Ericsson and J. Smith, *Toward a general theory of expertise: Prospects and limits*. Cambridge University Press, 1991.

[21] M. T. Chi, "Two approaches to the study of expert' characteristics," in, K. A. Ericsson, R. R. Hoffman, and A. Kozbelt, Eds., Cambridge University Press, 2006, pp. 21–30.

[22] B. Adelson, "When novices surpass experts: The difficulty of a task may increase with expertise.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 10, no. 3, p. 483, 1984.

[23] C. E. Lance, J. W. Hedge, and W. E. Alley, "Joint relationships of task proficiency with aptitude, experience, and task difficulty: A cross-level, interactional study," *Human Performance*, vol. 2, no. 4, pp. 249–272, 1989.

[24] A. Gegenfurtner, E. Lehtinen, and R. Säljö, "Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains," *Educational Psychology Review*, vol. 23, no. 4, pp. 523–552, 2011.

[25] M. T. Chi, R. Glaser, and M. J. Farr, Eds., *The nature of expertise*. Psychology Press, 2014.

[26] D De Groot Adriann, *Thought and Choice in Chess*. Ishi Press International, 1946.

[27] A. D. De Groot, *Thought and choice in chess*. Walter de Gruyter GmbH & Co KG, 2014, vol. 4.

[28] A Van der Gijp, C. Ravesloot, H Jarodzka, M. van der Schaaf, I. van der Schaaf, J. P. van Schaik, and T. J. Ten Cate, "How visual search relates to visual diagnostic performance: A narrative systematic review of eye-tracking research in radiology," *Advances in Health Sciences Education*, vol. 22, no. 3, pp. 765–787, 2017.

[29] C. F. Nodine, H. L. Kundel, C. Mello-Thoms, S. P. Weinstein, S. G. Orel, D. C. Sullivan, and E. F. Conant, "How experience and training influence mammography expertise," *Academic Radiology*, vol. 6, no. 10, pp. 575–585, 1999.

[30] C. F. Nodine, H. L. Kundel, S. C. Lauver, and L. C. Toto, "Nature of expertise in searching mammograms for breast masses," *Academic Radiology*, vol. 3, no. 12, pp. 1000–1006, 1996.

[31] R. Nakashima, C. Watanabe, E. Maeda, T. Yoshikawa, I. Matsuda, S. Miki, and K. Yokosawa, "The effect of expert knowledge on medical search: Medical experts have specialized abilities for detecting serious lesions," *Psychological Research*, vol. 79, no. 5, pp. 729–738, 2015.

[32] A. M. Williams and P. R. Ford, "Expertise and expert performance in sport," *International Review of Sport and Exercise Psychology*, vol. 1, no. 1, pp. 4–18, 2008.

[33] G. Tenenbaum, N. Levy-Kolker, S. Sade, D. G. Liebermann, and R. Lidor, "Anticipation and confidence of decisions related to skilled performance.," *International Journal of Sport Psychology*, 1996.

[34] J. Sampaio, T. McGarry, J. Calleja-González, S. Jiménez Sáiz, X. Schelling i del Alcázar, and M. Balciunas, "Exploring game performance in the national basketball association using player tracking data," *Plos One*, vol. 10, no. 7, e0132894, 2015.

[35] A. Ali, "Measuring soccer skill performance: A review," *Scandinavian Journal of Medicine & Science in Sports*, vol. 21, no. 2, pp. 170–183, 2011.

[36] M. Polanyi, "Personal knowledge: Towards a post-critical," *Philosophy*, 1962.

[37] K. A. Ericsson and W. Kintsch, "Long-term working memory.," *Psychological Review*, vol. 102, no. 2, p. 211, 1995.

[38] H. L. Dreyfus and S. E. Dreyfus, "The power of human intuition and expertise in the era of the computer," *Mind Over Machine. Nueva York: the Free Press*, 1986.

[39] K. Duncker and L. S. Lees, "On problem-solving.," *Psychological Monographs*, vol. 58, no. 5, p. i, 1945.

[40] K. A. Ericsson, "Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks," in, K. A. Ericsson, R. R. Hoffman, and A. Kozbelt, Eds., Cambridge University Press, 2006, pp. 223–241.

[41] S. A. Vitak, J. E. Ingram, A. T. Duchowski, S. Ellis, and A. K. Gramopadhye, "Gaze-augmented think-aloud as an aid to learning," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12, Montréal, Québec, Canada: ACM, 2012, pp. 1253–1262, ISBN: 1-59593-372-7. DOI: http://doi.acm.org/10.1145/1124772.1124961.

[42] A. de Groot, “Thought and choice in chess.,” 1978.

[43] H. A. Simon and F. Gobet, “Expertise effects in memory recall: Comment on vicente and wang (1998).,” 2000.

[44] H. L. Kundel and C. F. Nodine, “Interpreting chest radiographs without visual search,” *Radiology*, vol. 116, no. 3, pp. 527–532, 1975.

[45] K. K. Evans, D. Georgian-Smith, R. Tambouret, R. L. Birdwell, and J. M. Wolfe, “The gist of the abnormal: Above-chance medical decision making in the blink of an eye,” *Psychonomic Bulletin & Review*, vol. 20, no. 6, pp. 1170–1175, 2013.

[46] P. J. Feltovich, M. J. Prietula, and K. A. Ericsson, “Studies of expertise from psychological perspectives.,” in *The Cambridge Handbook of Expertise and Expert Performance*, K. A. Ericsson, R. R. Hoffman, and A. Kozbelt, Eds., Cambridge University Press, 2006, pp. 41–67.

[47] M. I. Posner, C. R. Snyder, and R Solso, “Attention and cognitive control,” *Cognitive Psychology: Key Readings*, vol. 205, 2004.

[48] M. Uemura, M. Tomikawa, R. Kumashiro, T. Miao, R. Souzaki, S. Ieiri, K. Ohuchida, A. T. Lefor, and M. Hashizume, “Analysis of hand motion differentiates expert and novice surgeons,” *Journal of Surgical Research*, vol. 188, no. 1, pp. 8–13, 2014.

[49] A. Ghasemloonia, Y. Maddahi, K. Zareinia, S. Lama, J. C. Dort, and G. R. Sutherland, “Surgical skill assessment using motion quality and smoothness,” *Journal of Surgical Education*, vol. 74, no. 2, pp. 295–305, 2017.

[50] P. T. Sowden, I. R. Davies, and P. Roling, “Perceptual learning of the detection of features in x-ray images: A functional role for improvements in adults’ visual sensitivity?” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, no. 1, p. 379, 2000.

[51] S. A. Waite, A. Grigorian, R. G. Alexander, S. L. Macknik, M. Carrasco, D. Heeger, and S. Martinez-Conde, “Analysis of perceptual expertise in radiology–current knowledge and a new perspective,” *Frontiers in Human Neuroscience*, vol. 13, p. 213, 2019.

[52] A. Lesgold, H. Rubinson, P. Feltovich, R. Glaser, D. Klopfer, and Y. Wang, “Expertise in a complex skill: Diagnosing x-ray pictures.,” in *The Nature of Expertise*, M. T. Chi, R. Glaser, and M. J. Farr, Eds., Lawrence Erlbaum Associates, Inc, 1988, pp. 311–342.

[53] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, “Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction,” *Radiographics*, vol. 35, no. 6, pp. 1668–1676, 2015.

[54] Y. W. Kim and L. T. Mansfield, "Fool me twice: Delayed diagnoses in radiology with emphasis on perpetuated errors," *American Journal of Roentgenology*, vol. 202, no. 3, pp. 465–470, 2014.

[55] A. Pinto and L. Brunese, "Spectrum of diagnostic errors in radiology," *World Journal of Radiology*, vol. 2, no. 10, p. 377, 2010.

[56] A. Ganesan, M. Alakhras, P. C. Brennan, and C. Mello-Thoms, "A review of factors influencing radiologists' visual search behaviour," *Journal of Medical Imaging and Radiation Oncology*, vol. 62, no. 6, pp. 747–757, 2018.

[57] A. S. Majid, E. S. de Paredes, R. D. Doherty, N. R. Sharma, and X. Salvador, "Missed breast carcinoma: Pitfalls and pearls," *Radiographics*, vol. 23, no. 4, pp. 881–895, 2003.

[58] H. L. Kundel, C. F. Nodine, and D. Carmody, "Visual scanning, pattern recognition and decision-making in pulmonary nodule detection.," *Investigative Radiology*, vol. 13, no. 3, pp. 175–181, 1978.

[59] B. Brogdon, C. Kelsey, and R. Moseley Jr, "Factors affecting perception of pulmonary lesions.," *Radiologic Clinics of North America*, vol. 21, no. 4, pp. 633–654, 1983.

[60] E. S. Berner and M. L. Graber, "Overconfidence as a cause of diagnostic error in medicine," *The American Journal of Medicine*, vol. 121, no. 5, S2–S23, 2008.

[61] F. Huettig and D. Axmann, "Reporting of dental status from full-arch radiographs: Descriptive analysis and methodological aspects," *World Journal of Clinical Cases: WJCC*, vol. 2, no. 10, p. 552, 2014.

[62] M. T. Baghdady, M. J. Pharoah, G. Regehr, E. W. Lam, and N. N. Woods, "The role of basic sciences in diagnostic oral radiology," *Journal of Dental Education*, vol. 73, no. 10, pp. 1187–1193, 2009.

[63] M. T. Baghdady, H. Carnahan, E. W. Lam, and N. N. Woods, "Dental and dental hygiene students' diagnostic accuracy in oral radiology: Effect of diagnostic strategy and instructional method," *Journal of Dental Education*, vol. 78, no. 9, pp. 1279–1285, 2014.

[64] M. Dhillon, S. M. Raju, S. Verma, D. Tomar, R. S. Mohan, M. Lakhanpal, and B. Krishnamoorthy, "Positioning errors and quality assessment in panoramic radiography," *Imaging Science in Dentistry*, vol. 42, no. 4, pp. 207–212, 2012.

[65] T. Grünheid, D. A. Hollevoet, J. R. Miller, and B. E. Larson, "Visual scan behavior of new and experienced clinicians assessing panoramic radiographs," *Journal of the World Federation of Orthodontists*, vol. 2, no. 1, e3–e7, 2013.

[66] F. Gijbels, A.-M. De Meyer, C. B. Serhal, C Van den Bossche, J Declerck, M Persoons, and R. Jacobs, "The subjective image quality of direct digital and conventional panoramic radiography," *Clinical Oral Investigations*, vol. 4, no. 3, pp. 162–167, 2000.

[67] S Perschbacher, "Interpretation of panoramic radiographs," *Australian Dental Journal*, vol. 57, pp. 40–45, 2012.

[68] C. W. Douglass, R. W. Valachovic, A. Wijesinha, H. H. Chauncey, K. K. Kapur, and B. J. McNeil, "Clinical efficacy of dental radiography in the detection of dental caries and periodontal diseases," *Oral Surgery, Oral Medicine, Oral Pathology*, vol. 62, no. 3, pp. 330–339, 1986.

[69] Z. Akarslan, M Akdevelioglu, K Gungor, and H Erten, "A comparison of the diagnostic accuracy of bitewing, periapical, unfiltered and filtered digital panoramic images for approximal caries detection in posterior teeth," *Dentomaxillofacial Radiology*, vol. 37, no. 8, pp. 458–463, 2008.

[70] M. Abdinian, S. M. Razavi, R. Faghihian, A. A. Samety, and E. Faghihian, "Accuracy of digital bitewing radiography versus different views of digital panoramic radiography for detection of proximal caries," *Journal of Dentistry (Tehran, Iran)*, vol. 12, no. 4, pp. 290–297, 2015.

[71] D. Y. Yeler and M. Koraltan, "Diagnostic accuracy of five different techniques for detection of approximal caries.," *Acta Odontologica Turcica*, vol. 35, no. 1, 2018.

[72] B. Molander, M. Ahlqwist, and H.-G. Gröndahl, "Panoramic and restrictive intraoral radiography in comprehensive oral radiographic diagnosis," *European Journal of Oral Sciences*, vol. 103, no. 4, pp. 191–198, 1995.

[73] M. Hedesiu, A. Serbanescu, and C. Ciolan, "Interobserver variability of the diagnosis of apical periodontitis on panoramic radiography assessment," *Mædica A Journal of Clinical Medicine*, vol. 2, no. 4, pp. 289–293, 2007.

[74] R. H. N. Rondon, Y. C. L. Pereira, and G. C. do Nascimento, "Common positioning errors in panoramic radiography: A review," *Imaging Science in Dentistry*, vol. 44, no. 1, pp. 1–6, 2014.

[75] J. J. Donald and S. A. Barnard, "Common patterns in 558 diagnostic radiology errors," *Journal of Medical Imaging and Radiation Oncology*, vol. 56, no. 2, pp. 173–178, 2012.

[76] M. B. Diniz, J. A. Rodrigues, K. Neuhaus, R. C. Cordeiro, and A. Lussi, "Influence of examiners' clinical experience on the reproducibility and validity of radiographic examination in detecting occlusal caries," *Clinical Oral Investigations*, vol. 14, no. 5, pp. 515–523, 2010. DOI: 10.1007/s00784-009-0323-z. [Online]. Available: https://doi.org/10.1007/s00784-009-0323-z.

[77] J. M. Provis, A. M. Dubis, T. Maddess, and J. Carroll, "Adaptation of the central retina for high acuity vision: Cones, the fovea and the avascular zone," *Progress in Retinal and Eye Research*, vol. 35, pp. 63–81, 2013.

[78] F. Schaeffel, "Processing of information in the human visual system," *Handbook of Machine Vision*, pp. 1–33, 2007.

[79] J. R. Bergstrom and A. Schall, Eds., *Eye Tracking in User Experience Design*. Elsevier, 2014.

[80] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye Tracking: A Comprehensive Guide to Methods and Measures*. OUP Oxford, 2011.

[81] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson, "Human photoreceptor topography," *Journal of Comparative Neurology*, vol. 292, no. 4, pp. 497–523, 1990.

[82] U. Neisser, *Cognitive Psychology*. New York: Appleton-century-crofts, 1967.

[83] B. Wang and K. J. Ciuffreda, "Blur discrimination of the human eye in the near retinal periphery," *Optometry and Vision Science*, vol. 82, no. 1, pp. 52–58, 2005.

[84] H. Strasburger, I. Rentschler, and M. Jüttner, "Peripheral vision and pattern recognition: A review," *Journal of Vision*, vol. 11, no. 5, pp. 13–13, 2011.

[85] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, ACM, 2000, pp. 71–78.

[86] S. Pannasch, J. R. Helmert, K. Roth, A.-K. Herbold, and H. Walter, "Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions," *Journal of Eye Movement Research*, vol. 2, no. 2, pp. 1–19, 2008.

[87] E. Tafaj, G. Kasneci, W. Rosenstiel, and M. Bogdan, "Bayesian online clustering of eye movement data," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2012, pp. 285–288.

[88] E. Kasneci, G. Kasneci, T. C. Kübler, and W. Rosenstiel, "The applicability of probabilistic methods to the online recognition of fixations and saccades in dynamic scenes," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2014, pp. 323–326.

[89] E. Kasneci, G. Kasneci, T. C. Kübler, and W. Rosenstiel, "Online recognition of fixations, saccades, and smooth pursuits for automated analysis of traffic hazard perception," in *Artificial Neural Networks*, Springer, 2015, pp. 411–434.

[90] T. Santini, W. Fuhl, T. Kübler, and E. Kasneci, "Bayesian identification of fixations, saccades, and smooth pursuits," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 163–170.

[91] S. Martinez-Conde, S. L. Macknik, and D. H. Hubel, "The role of fixational eye movements in visual perception," *Nature Reviews Neuroscience*, vol. 5, pp. 229–240, 2004. DOI: 10.1038/nrn1348. [Online]. Available: http://dx.doi.org/10.1038/nrn1348.

[92] M. A. Just and P. A. Carpenter, "Using eye fixations to study reading comprehension," *New Methods in Reading Comprehension Research*, pp. 151–182, 1984.

[93] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[94] J. M. Wolfe and T. S. Horowitz, "Five factors that guide attention in visual search," *Nature Human Behaviour*, vol. 1, no. 3, pp. 1–8, 2017.

[95] W. Einhäuser, U. Rutishauser, C. Koch, *et al.*, "Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli," *Journal of Vision*, vol. 8, no. 2, pp. 1–19, 2008.

[96] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[97] G. T. Buswell, *How people look at pictures: a study of the psychology and perception in art.* Univ. Chicago Press, 1935.

[98] A. Yarbus, *Eye Movements and Vision (B. Haigh, Trans.)* 1967.

[99] D. Noton and L. Stark, "Scanpaths in saccadic eye movements while viewing and recognizing patterns," *Vision Research*, vol. 11, no. 9, 929–IN8, 1971.

[100] J. R. Antes, "The time course of picture viewing.," *Journal of Experimental Psychology*, vol. 103, no. 1, p. 62, 1974.

[101] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966.

[102] J. E. Birren, R. C. Casperson, and J. Botwinick, "Age changes in pupil size," *Journal of Gerontology*, vol. 5, no. 3, pp. 216–221, 1950.

[103] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources.," *Psychological Bulletin*, vol. 91, no. 2, p. 276, 1982.

[104] E. Granholm, R. F. Asarnow, A. J. Sarkin, and K. L. Dykes, "Pupillary responses index cognitive resource limitations," *Psychophysiology*, vol. 33, no. 4, pp. 457–461, 1996.

[105] J. Hyönä, J. Tommola, and A.-M. Alaja, "Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks," *The Quarterly Journal of Experimental Psychology*, vol. 48, no. 3, pp. 598–612, 1995.

[106] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational Psychologist*, vol. 38, no. 1, pp. 63–71, 2003.

[107] C. Sibley, J. Coyne, and C. Baldwin, "Pupil dilation as an index of learning," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 55, 2011, pp. 237–241.

[108] T. T. Brunyé, M. D. Eddy, E. Mercan, K. H. Allison, D. L. Weaver, and J. G. Elmore, "Pupil diameter changes reflect difficulty and diagnostic accuracy during medical image interpretation," *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, p. 77, 2016.

[109] T. Appel, C. Scharinger, P. Gerjets, and E. Kasneci, "Cross-subject workload classification using pupil-related measures," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, ACM, 2018, pp. 1–8.

[110] A. Szulewski, N. Roth, and D. Howes, "The use of task-evoked pupillary response as an objective measure of cognitive load in novices and trained physicians: A new tool for the assessment of expertise," *Academic Medicine*, vol. 90, no. 7, pp. 981–987, 2015.

[111] P. W. Van Gerven, F. Paas, J. J. Van Merriënboer, and H. G. Schmidt, "Memory load and the cognitive pupillary response in aging," *Psychophysiology*, vol. 41, no. 2, pp. 167–174, 2004.

[112] B. Winn, D. Whitaker, D. B. Elliott, and N. J. Phillips, "Factors affecting light-adapted pupil size in normal human subjects.," *Investigative Ophthalmology & Visual Science*, vol. 35, no. 3, pp. 1132–1137, 1994.

[113] G. Porter, T. Troscianko, and I. D. Gilchrist, "Effort during visual search and counting: Insights from pupillometry," *The Quarterly Journal of Experimental Psychology*, vol. 60, no. 2, pp. 211–229, 2007.

[114] J. J. Geng, Z. Blumenfeld, T. L. Tyson, and M. J. Minzenberg, "Pupil diameter reflects uncertainty in attentional selection during visual search," *Frontiers in Human Neuroscience*, vol. 9, p. 435, 2015.

[115] R. W. Backs and L. C. Walrath, "Eye movement and pupillary response indices of mental workload during visual search of symbolic displays," *Applied Ergonomics*, vol. 23, no. 4, pp. 243–254, 1992.

[116] J. Veltman and A. Gaillard, "Physiological workload reactions to increasing levels of task difficulty," *Ergonomics*, vol. 41, no. 5, pp. 656–669, 1998.

[117] P. Gerjets, K. Scheiter, and R. Catrambone, "Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures," *Instructional Science*, vol. 32, no. 1-2, pp. 33–58, 2004.

[118] T. Takeuchi, T. Puntous, A. Tuladhar, S. Yoshimoto, and A. Shirama, "Estimation of mental effort in learning visual search by measuring pupil response," *Plos One*, vol. 6, no. 7, e21973, 2011.

[119] T. Tien, P. H. Pucher, M. H. Sodergren, K. Sriskandarajah, G.-Z. Yang, and A. Darzi, "Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair," *Surgical Endoscopy*, vol. 29, no. 2, pp. 405–413, 2015.

[120] A. Szulewski, D. Kelton, and D. Howes, "Pupillometry as a tool to study expertise in medicine.," *Frontline Learning Research*, vol. 5, no. 3, pp. 53–63, 2017.

[121] B. Zheng, X. Jiang, and M. S. Atkins, "Detection of changes in surgical difficulty: Evidence from pupil responses," *Surgical Innovation*, vol. 22, no. 6, pp. 629–635, 2015.

[122] T. Santini, H. Brinkmann, L. Reitstätter, H. Leder, R. Rosenberg, W. Rosenstiel, and E. Kasneci, "The art of pervasive eye tracking: Unconstrained eye tracking in the austrian gallery belvedere," in *Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*, 2018, pp. 1–8.

[123] E. Kasneci, "Towards pervasive eye tracking," *IT-Information Technology*, vol. 59, no. 5, pp. 253–257, 2017.

[124] A. Bulling, "Pervasive attentive user interfaces," *Computer*, no. 1, pp. 94–98, 2016.

[125] D. W. Hansen and A. E. Pece, "Eye tracking in the wild," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 155–181, 2005.

[126] A. Slater-Hammel, "Vision and success in motor performance," *Physical Educator*, vol. 6, no. 1, p. 8, 1949.

[127] P. M. Fitts, R. E. Jones, and J. L. Milton, "Eye movements of aircraft pilots during instrument-landing approaches.," *Aeronautical Engineering Review*, 1950.

[128] P. J. Beek, *Juggling dynamics*. Free University Press, 1989.

[129] A. Manzanares, R. Menayo, and F. Segado, "Visual search strategy during regatta starts in a sailing simulation," *Motor Control*, vol. 21, no. 4, pp. 413–424, 2017.

[130] A. G. Dyer, B. Found, and D. Rogers, "An insight into forensic document examiner expertise for discriminating between forged and disguised signatures," *Journal of Forensic Sciences*, vol. 53, no. 5, pp. 1154–1159, 2008.

[131] J. J. Topczewski, A. M. Topczewski, H. Tang, L. K. Kendhammer, and N. J. Pienta, "Nmr spectra through the eyes of a student: Eye tracking applied to nmr items," *Journal of Chemical Education*, vol. 94, no. 1, pp. 29–37, 2017.

[132] S. Brams, G. Ziv, O. Levin, J. Spitz, J. Wagemans, A. M. Williams, and W. F. Helsen, "The relationship between gaze behavior, expertise, and performance: A systematic review.," *Psychological Bulletin*, vol. 145, no. 10, p. 980, 2019.

[133] H. Haider and P. A. Frensch, "Eye movement during skill acquisition: More evidence for the information-reduction hypothesis.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 25, no. 1, p. 172, 1999.

[134] H. L. Kundel, C. F. Nodine, E. F. Conant, and S. P. Weinstein, "Holistic component of image perception in mammogram interpretation: Gaze-tracking study," *Radiology*, vol. 242, no. 2, pp. 396–402, 2007.

[135] D. Assaf, E. Amar, N. Marwan, Y. Neuman, M. Salai, and E. Rath, "Dynamic patterns of expertise: The case of orthopedic medical diagnosis," *Plos One*, vol. 11, no. 7, e0158820, 2016.

[136] R. Bertram, L. Helle, J. K. Kaakinen, and E. Svedström, "The effect of expertise on eye movement behaviour in medical image perception," *Plos One*, vol. 8, no. 6, e66169, 2013.

[137] E. M. Kok, A. B. De Bruin, S. G. Robben, and J. J. Van Merrienboer, "Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology," *Applied Cognitive Psychology*, vol. 26, no. 6, pp. 854–862, 2012.

[138] C. F. Nodine and C. Mello-Thoms, "The nature of expertise in radiology," *Handbook of Medical Imaging. SPIE*, pp. 859–895, 2000.

[139] E. A. Krupinski, "Visual scanning patterns of radiologists searching mammograms," *Academic Radiology*, vol. 3, no. 2, pp. 137–144, 1996.

[140] L. Cooper, A. G. Gale, J. Saada, S. Gedela, H. J. Scott, and A. Toms, "The assessment of stroke multidimensional ct and mr imaging using eye movement analysis: Does modality preference enhance observer performance?," vol. 7627, 76270B, 2010.

[141] S. Mallett, P. Phillips, T. R. Fanshawe, E. Helbren, D. Boone, A. Gale, S. A. Taylor, D. Manning, D. G. Altman, and S. Halligan, "Tracking eye gaze during interpretation of endoluminal three-dimensional ct colonography: Visual perception of experienced and inexperienced readers," *Radiology*, vol. 273, no. 3, pp. 783–792, 2014.

[142] A. L. Warren, T. L. Donnon, C. R. Wagg, H. Priest, and N. J. Fernandez, "Quantifying novice and expert differences in visual diagnostic reasoning in veterinary pathology using eye-tracking technology," *Journal of Veterinary Medical Education*, vol. 45, no. 3, pp. 295–306, 2018.

[143] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive Psychology*, vol. 8, no. 4, pp. 441–480, 1976.

[144] S. P. Marshall, "Identifying cognitive state from eye metrics," *Aviation, Space, and Environmental Medicine*, vol. 78, no. 5, Seciton II, B165–B175(11), 2007, Supplement 1.

[145] M. Wilson, J. McGrath, S. Vine, J. Brewer, D. Defriend, and R. Masters, "Psychomotor control in a virtual laparoscopic surgery training environment: Gaze control parameters differentiate novices from experts," *Surgical Endoscopy*, vol. 24, no. 10, pp. 2458–2464, 2010.

[146] S. Eivazi, A. Hafez, W. Fuhl, H. Afkari, E. Kasneci, M. Lehecka, and R. Bednarik, "Optimal eye movement strategies: A comparison of neurosurgeons gaze patterns when using a surgical microscope," *Acta Neurochirurgica*, vol. 159, no. 6, pp. 959–966, 2017.

[147] S. Eivazi, R. Bednarik, M. Tukiainen, M. von und zu Fraunberg, V. Leinonen, and J. E. Jääskeläinen, "Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2012, pp. 377–380.

[148] B. Law, M. S. Atkins, A. E. Kirkpatrick, and A. J. Lomax, "Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2004, pp. 41–48.

[149] D. P. Turgeon and E. W. Lam, "Influence of experience and training on dental students' examination performance regarding panoramic images," *Journal of Dental Education*, vol. 80, no. 2, pp. 156–164, 2016.

[150] G. Wood, K. M. Knapp, B. Rock, C. Cousens, C. Roobottom, and M. R. Wilson, "Visual expertise in detecting and diagnosing skeletal fractures," *Skeletal Radiology*, vol. 42, no. 2, pp. 165–172, 2013.

[151] T. Donovan and D. Litchfield, "Looking for cancer: Expertise related differences in searching and decision making," *Applied Cognitive Psychology*, vol. 27, no. 1, pp. 43–49, 2013.

[152] K. Suwa, A. Furukawa, T. Matsumoto, and T. Yosue, "Analyzing the eye movement of dentists during their reading of ct images," *Odontology*, vol. 89, no. 1, pp. 0054–0061, 2001.

[153] F. Alamudun, H.-J. Yoon, K. B. Hudson, G. Morin-Ducote, T. Hammond, and G. D. Tourassi, "Fractal analysis of visual search activity for mass detection during mammographic screening," *Medical Physics*, vol. 44, no. 3, pp. 832–846, 2017.

[154] B. P. Hermanson, G. C. Burgdorf, J. F. Hatton, D. M. Speegle, and K. F. Woodmansey, "Visual fixation and scan patterns of dentists viewing dental periapical radiographs: An eye tracking pilot study," *Journal of Endodontics*, vol. 44, no. 5, pp. 722–727, 2018.

[155] D. Manning, S. Ethell, T. Donovan, and T. Crawford, "How do radiologists do it? the influence of experience and training on searching for chest nodules," *Radiography*, vol. 12, no. 2, pp. 134–142, 2006.

[156] E. A. Krupinski, A. A. Tillack, L. Richter, J. T. Henderson, A. K. Bhattacharyya, K. M. Scott, A. R. Graham, M. R. Descour, J. R. Davis, and R. S. Weinstein, "Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience," *Human Pathology*, vol. 37, no. 12, pp. 1543–1556, 2006.

[157] T. T. Brunyé, B. K. Nallamothu, and J. G. Elmore, "Eye-tracking for assessing medical image interpretation: A pilot feasibility study comparing novice vs expert cardiologists," *Perspectives on Medical Education*, vol. 8, no. 2, pp. 65–73, 2019.

[158] E. M. Kok, H. Jarodzka, A. B. de Bruin, H. A. BinAmir, S. G. Robben, and J. J. van Merriënboer, "Systematic viewing in radiology: Seeing more, missing less?" *Advances in Health Sciences Education*, vol. 21, no. 1, pp. 189–205, 2016.

[159] S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller, "The expertise reversal effect," *Educational Psychologist*, vol. 38, no. 1, pp. 23–31, 2003.

[160] S. Brams, G. Ziv, I. T. Hooge, O. Levin, T. De Brouwere, J. Verschakelen, S. Dauwe, A. M. Williams, J. Wagemans, and W. F. Helsen, "Focal lung pathology detection in radiology: Is there an effect of experience on visual search behavior?" *Attention, Perception, & Psychophysics*, pp. 1–14, 2020.

[161] P. J. A. Unema, S. Pannasch, M. Joos, and B. M. Velichkovsky, "Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration," *Visual Cognition*, vol. 12, no. 3, pp. 473–494, 2005. DOI: 10.1080/13506280444000409. [Online]. Available: http://dx.doi.org/10.1080/13506280444000409.

[162] Z. Gandomkar and C. Mello-Thoms, "Visual search in breast imaging," *the British Journal of Radiology*, vol. 92, no. 1102, p. 20 190 057, 2019.

[163] N. Koide, T. Kubo, S. Nishida, T. Shibata, and K. Ikeda, "Art expertise reduces influence of visual salience on fixation in viewing abstract-paintings," *Plos One*, vol. 10, no. 2, e0117696, 2015.

[164] M. O. Al-Moteri, M. Symmons, V. Plummer, and S. Cooper, "Eye tracking to investigate cue processing in medical decision-making: A scoping review," *Computers in Human Behavior*, vol. 66, pp. 52–66, 2017.

[165] S. E. Fox and B. E. Faulkner-Jones, "Eye-tracking in the study of visual expertise: Methodology and approaches in medicine.," *Frontline Learning Research*, vol. 5, no. 3, pp. 29–40, 2017.

[166] H. Matsumoto, Y. Terao, A. Yugeta, H. Fukuda, M. Emoto, T. Furubayashi, T. Okano, R. Hanajima, and Y. Ugawa, "Where do neurologists look when viewing brain ct images? an eye-tracking study involving stroke cases," *Plos One*, vol. 6, no. 12, e28928, 2011.

[167] C. H. Hu, H. L. Kundel, C. F. Nodine, E. A. Krupinski, and L. C. Toto, "Searching for bone fractures: A comparison with pulmonary nodule search," *Academic Radiology*, vol. 1, no. 1, pp. 25–32, 1994.

[168] T. Drew, M. L.-H. Vo, A. Olwal, F. Jacobson, S. E. Seltzer, and J. M. Wolfe, "Scanners and drillers: Characterizing expert visual search through volumetric images," *Journal of Vision*, vol. 13, no. 10, pp. 3–3, 2013.

[169] E. Mercan, L. G. Shapiro, T. T. Brunyé, D. L. Weaver, and J. G. Elmore, "Characterizing diagnostic search patterns in digital breast pathology: Scanners and drillers," *Journal of Digital Imaging*, vol. 31, no. 1, pp. 32–41, 2018.

[170] T. T. Brunyé, T. Drew, D. L. Weaver, and J. G. Elmore, "A review of eye tracking for understanding and improving diagnostic interpretation," *Cognitive Research: Principles and Implications*, vol. 4, no. 1, p. 7, 2019.

[171] T. T. Brunyé and A. L. Gardony, "Eye tracking measures of uncertainty during perceptual decision making," *International Journal of Psychophysiology*, vol. 120, pp. 60–68, 2017.

[172] A. Rozenshtein, G. D. Pearson, S. X. Yan, A. Z. Liu, and D. Toy, "Effect of massed versus interleaved teaching method on performance of students in radiology," *Journal of the American College of Radiology*, vol. 13, no. 8, pp. 979–984, 2016.

[173] S. J. Vine, R. S. Masters, J. S. McGrath, E. Bright, and M. R. Wilson, "Cheating experience: Guiding novices to adopt the gaze strategies of experts expedites the learning of technical laparoscopic skills," *Surgery*, vol. 152, no. 1, pp. 32–40, 2012.

[174] A. Gegenfurtner, E. Lehtinen, H. Jarodzka, and R. Säljö, "Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis," *Computers & Education*, 2017.

[175] H. Jarodzka, T. Balslev, K. Holmqvist, M. Nyström, K. Scheiter, P. Gerjets, and B. Eika, "Conveying clinical reasoning based on visual observation via eye-movement modelling examples," *Instructional Science*, vol. 40, no. 5, pp. 813–827, 2012.

[176] H. Jarodzka, K. Scheiter, P. Gerjets, T. van Gog, and M. Dorr, "How to convey perceptual skills by displaying experts' gaze data," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2010, pp. 2920–2925.

[177] C. Braunagel, D. Geisler, W. Rosenstiel, and E. Kasneci, "Online recognition of driver-activity based on visual scanpath classification," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 23–36, 2017.

[178] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel, "Driver-activity recognition in the context of conditionally autonomous driving," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, IEEE, 2015, pp. 1652–1657.

[179] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof, "A comparison of scanpath comparison methods," *Behavior Research Methods*, vol. 47, no. 4, pp. 1377–1392, 2015.

[180] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist, "It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach," *Behavior Research Methods*, vol. 44, no. 4, pp. 1079–1100, 2012.

[181] T. C. Kübler, C. Rothe, U. Schiefer, W. Rosenstiel, and E. Kasneci, "Subsmatch 2.0: Scanpath comparison and classification based on subsequence frequencies," *Behavior Research Methods*, vol. 49, no. 3, pp. 1048–1064, 2017.

[182] A. Coutrot, J. H. Hsiao, and A. B. Chan, "Scanpath modeling and classification with hidden markov models," *Behavior Research Methods*, vol. 50, no. 1, pp. 362–379, 2018.

[183] R. Fahimi and N. D. Bruce, "On metrics for measuring scanpath similarity," *Behavior Research Methods*, pp. 1–20, 2020.

[184] H. Jarodzka, K. Holmqvist, and M. Nyström, "A vector-based, multidimensional scanpath similarity measure," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, ACM, 2010, pp. 211–218.

[185] F. Deitelhoff, A. Harrer, and A. Kienle, "The influence of different aoi models in source code comprehension analysis," in *2019 IEEE/ACM 6th international Workshop on Eye Movements in Programming (EMIP)*, IEEE, 2019, pp. 10–17.

[186] W. Fuhl, T. Kuebler, T. Santini, and E. Kasneci, "Automatic generation of saliency-based areas of interest for the visualization and analysis of eye-tracking data," in *Proceedings of the Conference on Vision, Modeling, and Visualization*, 2018, pp. 47–54.

[187] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 970–982, 2000.

[188] H. Hembrooke, M. Feusner, and G. Gay, "Averaging scan patterns and what they can tell us," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2006, pp. 41–41.

[189] R. Caldara and S. Miellet, "Imap: A novel method for statistical fixation mapping of eye movement data," *Behavior Research Methods*, vol. 43, no. 3, pp. 864–878, 2011.

[190] D. S. Wooding, "Eye movements of large populations: Ii. deriving regions of interest, coverage, and similarity using fixation maps," *Behavior Research Methods, Instruments, & Computers*, vol. 34, no. 4, pp. 518–528, 2002.

[191] W. Fuhl, T. Kübler, K. Sippel, W. Rosenstiel, and E. Kasneci, "Arbitrarily shaped areas of interest based on gaze density gradient," in *European Conference on Eye Movements*, vol. 1, 2015, p. 5.

[192] N. Castner, S. Eivazi, K. Scheiter, and E. Kasneci, "Using eye tracking to evaluate and develop innovative teaching strategies for fostering image reading skills of novices in medical training," in *Workshop on Eye Tracking Enhanced Learning (ETEL '17)*, 2017.

[193] U. Rajashekar, I. Van Der Linde, A. C. Bovik, and L. K. Cormack, "Gaffe: A gaze-attentive fixation finding engine," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 564–573, 2008.

[194] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 545–552.

[195] M. Cerf, J. Harel, A. Huth, W. Einhäuser, and C. Koch, "Decoding what people see from where they look: Predicting visual stimuli from scanpaths," in *International Workshop on Attention in Cognitive Systems*, Springer, 2008, pp. 15–26.

[196] C. Mills, R. Bixler, X. Wang, and S. K. D'Mello, "Automatic gaze-based detection of mind wandering during narrative film comprehension.," *Proceedings of the International Educational Data Mining Society*, pp. 30–37, 2016.

[197] S. A. Brandt and L. W. Stark, "Spontaneous eye movements during visual imagery reflect the content of the visual scene," *Journal of Cognitive Neuroscience*, vol. 9, no. 1, pp. 27–38, 1997.

[198] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[199] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet Physics Doklady*, vol. 10, 1966, pp. 707–710.

[200] F. Cristino, S. Mathôt, J. Theeuwes, and I. D. Gilchrist, "Scanmatch: A novel method for comparing fixation sequences," *Behavior Research Methods*, vol. 42, no. 3, pp. 692–700, 2010.

[201] D. Noton and L. Stark, "Eye movements and visual perception," *Scientific American*, vol. 224, no. 6, pp. 34–43, 1971.

[202] S. Eraslan, Y. Yesilada, and S. Harper, "Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison," *Journal of Eye Movement Research*, vol. 9, no. 1, 2015. DOI: 10.16910/jemr.9.1.2. [Online]. Available: https://bop.unibe.ch/JEMR/article/view/2430.

[203] F. Galgani, Y. Sun, P. L. Lanzi, and J. Leigh, "Automatic analysis of eye tracking data for medical diagnosis," in *2009 IEEE Symposium on Computational Intelligence and Data Mining*, IEEE, 2009, pp. 195–202.

[204] R. M. French, Y. Glady, and J.-P. Thibaut, "An evaluation of scanpath-comparison and machine-learning classification algorithms used to study the dynamics of analogy making," *Behavior Research Methods*, vol. 49, no. 4, pp. 1291–1302, 2017.

[205] J. Dolezalova and S. Popelka, "Scangraph: A novel scanpath comparison method using visualisation of graph cliques," *Journal of Eye Movement Research*, vol. 9, no. 4, 2016. DOI: 10.16910/jemr.9.4.5. [Online]. Available: https://bop.unibe.ch/JEMR/article/view/2522.

[206] S. Baichoo and C. A. Ouzounis, "Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment," *Biosystems*, 2017.

[207] R.-F. Day, "Examining the validity of the needleman–wunsch algorithm in identifying decision strategy with eye-movement data," *Decision Support Systems*, vol. 49, no. 4, pp. 396–403, 2010.

[208] J. H. Goldberg and J. I. Helfman, "Scanpath clustering and aggregation," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, ACM, 2010, pp. 227–234.

[209] J. M. Henderson, J. R. Brockmole, M. S. Castelhano, and M. Mack, "Visual saliency does not account for eye movements during visual search in real-world scenes," in *Eye Movements*, Elsevier, 2007, pp. 537–562.

[210] L. Zhou, Y.-Y. Zhang, Z.-J. Wang, L.-L. Rao, W. Wang, S. Li, X. Li, and Z.-Y. Liang, "A scanpath analysis of the risky decision-making process," *Journal of Behavioral Decision Making*, vol. 29, no. 2-3, pp. 169–182, 2016.

[211] T. Busjahn, R. Bednarik, A. Begel, M. Crosby, J. H. Paterson, C. Schulte, B. Sharif, and S. Tamm, "Eye movements in code reading: Relaxing the linear order," in *2015 IEEE 23rd International Conference on Program Comprehension (ICPC)*, IEEE, 2015, pp. 255–265.

[212] A. Madsen, A. Larson, L. Loschky, and N. S. Rebello, "Using scanmatch scores to understand differences in eye movements between correct and incorrect solvers on physics problems," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2012, pp. 193–196.

[213] M. E. Frame, R. Warren, and A. M. Maresca, "Scanpath comparisons for complex visual search in a naturalistic environment," *Behavior Research Methods*, vol. 51, no. 3, pp. 1454–1470, 2019.

[214] T. F. Smith, M. S. Waterman, *et al.*, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[215] A. B. Khedher, I. Jraidi, and C. Frasson, "Local sequence alignment for scan path similarity assessment," *International Journal of Information and Education Technology*, vol. 8, no. 7, pp. 482–490, 2018.

[216] A. Çöltekin, S. I. Fabrikant, and M. Lacayo, "Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings," *International Journal of Geographical Information Science*, vol. 24, no. 10, pp. 1559–1575, 2010.

[217] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "State-of-the-art of visualization for eye tracking data.," in *Eurographics Conference on Visualization (EuroVis)*, 2014, pp. 63–82.

[218] K. Holmqvist, C. Andrà, P. Lindström, F. Arzarello, F. Ferrara, O. Robutti, and C. Sabena, "A method for quantifying focused versus overview behavior in aoi sequences," *Behavior Research Methods*, vol. 43, no. 4, pp. 987–998, 2011.

[219] M. S. Magnusson, "Discovering hidden time patterns in behavior: T-patterns and their detection," *Behavior Research Methods, Instruments, & Computers*, vol. 32, no. 1, pp. 93–110, 2000.

[220] M. Burmester and M. Mast, "Repeated web page visits and the scanpath theory: A recurrent pattern detection approach," *Journal of Eye Movement Research*, vol. 3, no. 4, 2010. DOI: 10.16910/jemr.3.4.5. [Online]. Available: https://bop.unibe.ch/JEMR/article/view/2305.

[221] N. C. Anderson, W. F. Bischof, K. E. Laidlaw, E. F. Risko, and A. Kingstone, "Recurrence quantification analysis of eye movements," *Behavior Research Methods*, vol. 45, no. 3, pp. 842–856, 2013.

[222] C. Kanan, N. A. Ray, D. N. Bseiso, J. H. Hsiao, and G. W. Cottrell, "Predicting an observer's task using multi-fixation pattern analysis," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2014, pp. 287–290.

[223] A. Haji-Abolhassani and J. J. Clark, "An inverse yarbus process: Predicting observers' task from eye movement patterns," *Vision Research*, vol. 103, pp. 127–142, 2014.

[224] S. R. Ellis and L. Stark, "Statistical dependency in visual scanning," *Human Factors*, vol. 28, no. 4, pp. 421–438, 1986.

[225] T. R. Hayes, A. A. Petrov, and P. B. Sederberg, "A novel method for analyzing sequential eye movements reveals strategic influence on raven's advanced progressive matrices," *Journal of Vision*, vol. 11, no. 10, pp. 10–10, 2011.

[226] I. T. Hooge and G. Camps, "Scan path entropy and arrow plots: Capturing scanning behavior of multiple observers," *Frontiers in Psychology*, vol. 4, p. 996, 2013.

[227] K. Krejtz, T. Szmidt, A. T. Duchowski, and I. Krejtz, "Entropy-based statistical analysis of eye movement transitions," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2014, pp. 159–166.

[228] P. Yazdan-Shahmorad, N. Sammaknejad, and F. Bakouie, "Graph-based analysis of visual scanning patterns: A developmental study on green and normal images," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[229] V. Cantoni, C. Galdi, M. Nappi, M. Porta, and D. Riccio, "Gant: Gaze analysis technique for human identification," *Pattern Recognition*, vol. 48, no. 4, pp. 1027–1038, 2015.

[230] T. Chuk, A. B. Chan, and J. H. Hsiao, "Understanding eye movements in face recognition using hidden markov models," *Journal of Vision*, vol. 14, no. 11, pp. 8–8, 2014.

[231] I. A. Ebeid, N. Bhattacharya, J. Gwizdka, and A. Sarkar, "Analyzing gaze transition behavior using bayesian mixed effects markov models," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–5.

[232] K. P. Murphy *et al.*, "Naive bayes classifiers," *University of British Columbia*, vol. 18, p. 60, 2006.

[233] D. P. Crabb, N. D. Smith, and H. Zhu, "What's on tv? detecting age-related neurodegenerative eye disease using eye movement scanpaths," *Frontiers in Aging Neuroscience*, vol. 6, p. 312, 2014.

[234] G. E. Raptis, C. Katsini, M. Belk, C. Fidas, G. Samaras, and N. Avouris, "Using eye gaze data and visual activities to infer human cognitive styles: Method and feasibility studies," in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 2017, pp. 164–173.

[235] A. Borji, A. Lennartz, and M. Pomplun, "What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations," *Neurocomputing*, vol. 149, pp. 788–799, 2015.

[236] F. Lethaus, M. R. Baumann, F. Köster, and K. Lemmer, "A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data," *Neurocomputing*, vol. 121, pp. 108–130, 2013.

[237] M. R. Greene, T. Liu, and J. M. Wolfe, "Reconsidering yarbus: A failure to predict observers' task from eye movement patterns," *Vision Research*, vol. 62, pp. 1–8, 2012.

[238] M. Lipps and J. B. Pelz, "Yarbus revisited: Task-dependent oculomotor behavior," *Journal of Vision*, vol. 4, no. 8, p. 115, 2004.

[239] A. Borji and L. Itti, "Defending yarbus: Eye movements reveal observers' task," *Journal of Vision*, vol. 14, no. 3, pp. 1–22, 2014.

[240] P. Dayan, M. Sahani, and G. Deback, "Unsupervised learning," *The MIT Encyclopedia of the Cognitive Sciences*, pp. 857–859, 1999.

[241] Y. Li, C. Allen, and C.-R. Shyu, "Quantifying and understanding the differences in visual activities with contrast subsequences," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–5.

[242] J. M. West, A. R. Haake, E. P. Rozanski, and K. S. Karn, "Eyepatterns: Software for identifying patterns and similarities across fixation sequences," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, ACM, 2006, pp. 149–154.

[243] Z. Kang and S. J. Landry, "An eye movement analysis algorithm for a multielement target tracking task: Maximum transition-based agglomerative hierarchical clustering," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 1, pp. 13–24, 2014.

[244] M. Koch, K. Kurzhals, and D. Weiskopf, "Image-based scanpath comparison with slit-scan visualization," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–5.

[245] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[246] F. Corpet, "Multiple sequence alignment with hierarchical clustering," *Nucleic Acids Research*, vol. 16, no. 22, pp. 10 881–10 890, 1988.

[247] A. Li, Y. Zhang, and Z. Chen, "Scanpath mining of eye movement trajectories for visual attention analysis," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2017, pp. 535–540.

[248] M. E. Manaa and G. Abdulameer, "Web documents similarity using k-shingle tokens and minhash technique," *Journal of Engineering and Applied Sciences*, vol. 13, no. 6, pp. 1499–1505, 2018.

[249] T. C. Kübler, E. Kasneci, and W. Rosenstiel, "Subsmatch: Scanpath similarity in dynamic scenes based on subsequence frequencies," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, ACM, 2014, pp. 319–322.

[250] A. T. Duchowski, J. Driver, S. Jolaoso, W. Tan, B. N. Ramey, and A. Robbins, "Scanpath comparison revisited," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2010, pp. 219–226.

[251] J. Heminghous and A. T. Duchowski, "Icomp: A tool for scanpath visualization and comparison," in *Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization*, 2006, pp. 152–152.

[252] B. Schlkopf, A. J. Smola, and F. Bach, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018.

[253] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[254] T. C. Kübler and E. Kasneci, "Automated comparison of scanpaths in dynamic scenes," in *SAGA-International Workshop on Solutions for Automatic Gaze Data Analysis: Proceedings*, 2015, pp. 1–3.

[255] P.-H. Tseng, I. G. Cameron, G. Pari, J. N. Reynolds, D. P. Munoz, and L. Itti, "High-throughput classification of clinical populations from natural viewing eye movements," *Journal of Neurology*, vol. 260, no. 1, pp. 275–284, 2013.

[256] C. Kelton, Z. Wei, S. Ahn, A. Balasubramanian, S. R. Das, D. Samaras, and G. Zelinsky, "Reading detection in real-time," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–5.

[257] K. Kunze, Y. Utsumi, Y. Shiga, K. Kise, and A. Bulling, "I know what you are reading: Recognition of document types using mobile eye tracking," in *Proceedings of the 2013 International Symposium on Wearable Computers*, 2013, pp. 113–116.

[258] Y. Yamada and M. Kobayashi, "Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults," *Artificial Intelligence in Medicine*, vol. 91, pp. 39–48, 2018.

[259] T. Foulsham, R. Dewhurst, M. Nyström, H. Jarodzka, R. Johansson, G. Underwood, and K. Holmqvist, "Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach," *Journal of Eye Movement Research*, vol. 5, no. 4, pp. 1–14, 2012.

[260] K. A. Wilson, P. L. Heinselman, and Z. Kang, "Comparing forecaster eye movements during the warning decision process," *Weather and Forecasting*, vol. 33, no. 2, pp. 501–521, 2018.

[261] S. Stranc and K. Muldner, "Scanpath analysis of student attention during problem solving with worked examples," in *International Conference on Artificial Intelligence in Education*, Springer, 2020, pp. 306–311.

[262] R. Dewhurst, T. Foulsham, H. Jarodzka, R. Johansson, K. Holmqvist, and M. Nyström, "How task demands influence scanpath similarity in a sequential number-search task," *Vision Research*, vol. 149, pp. 9–23, 2018.

[263] J. Chakraborty and M. P. McGuire, "Directional scan path characterization of eye tracking sequences: A multi-scale approach," in *2016 Future Technologies Conference (FTC)*, IEEE, 2016, pp. 51–61.

[264] Y. Sugano, H. Kasai, K. Ogaki, and Y. Sato, "Image preference estimation from eye movements with a data-driven approach," *Proceedings of the Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*, 2013.

[265] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests.," *Psychological Methods*, vol. 14, no. 4, pp. 323–348, 2009.

[266] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[267] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *2007 IEEE 11th International Conference on Computer Vision*, IEEE, 2007, pp. 1–8.

[268] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, "Fast keypoint recognition using random ferns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 448–461, 2009.

[269] Y. Dong, Y. Zhang, J. Yue, and Z. Hu, "Comparison of random forest, random ferns and support vector machine for eye state classification," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11 763–11 783, 2016.

[270] M. J. Haass, L. E. Matzen, K. M. Butler, and M. Armenta, "A new method for categorizing scanpaths from eye tracking data," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 35–38.

[271] N. Ouerhani, H. Hügli, R. Müri, and R. Von Wartburg, "Empirical validation of the saliency-based model of visual attention," in *Electronic Letters on Computer Vision and Image Analysis*, 2003, pp. 13–23.

[272] J. F. Boisvert and N. D. Bruce, "Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features," *Neurocomputing*, vol. 207, pp. 653–668, 2016.

[273] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: Strengths and weaknesses," *Behavior Research Methods*, vol. 45, no. 1, pp. 251–266, 2013.

[274] K. Simonyan, A. Vedaldi, and A. Zisserman, *Deep inside convolutional networks: Visualising image classification models and saliency maps*, 2013. arXiv: 1312.6034 [cs.CV].

[275] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Advances in Neural Information Processing Systems*, 2013, pp. 2553–2561.

[276] G. Dong and M. Xie, "Color clustering and learning for image segmentation based on neural networks," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 925–936, 2005.

[277] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, IEEE, 2016, pp. 565–571.

[278] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[279] K. Murakami and H. Taguchi, "Gesture recognition using recurrent neural networks," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1991, pp. 237–242.

[280] M. A. Nielsen, *Neural networks and deep learning*. Determination press San Francisco, CA, 2015.

[281] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[282] Y. Wang, M. Long, J. Wang, Z. Gao, and S. Y. Philip, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," in *Advances in Neural Information Processing Systems*, 2017, pp. 879–888.

[283] H. Shao and B.-H. Soong, "Traffic flow prediction with long short-term memory networks (lstms)," in *2016 IEEE Region 10 Conference (TENCON)*, IEEE, 2016, pp. 2986–2989.

[284] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[285] S. Sims and C. Conati, *A neural architecture for detecting confusion in eye-tracking data*, 2020. arXiv: `2003.06434 [cs.CV]`.

[286] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[287] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[288] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[289] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[290] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 262–270.

[291] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *International Conference on Machine Learning*, 2015, pp. 597–606.

[292] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Saltinet: Scan-path prediction on 360 degree images using saliency volumes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2331–2338.

[293] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 362–370.

[294] M. Kümmerer, T. S. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 16 054–16 059, 2015.

[295] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[296] Z. Chen and W. Sun, "Scanpath prediction for visual attention using ior-roi lstm.," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 642–648.

[297] A. Mishra and P. Bhattacharyya, "Automatic extraction of cognitive features from gaze data," in *Cognitively Inspired Natural Language Processing*, Springer, 2018, pp. 153–169.

[298] Y. Tao and M. Shyu, "Sp-asdnet: Cnn-lstm based asd classification model using observer scanpaths," in *2019 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2019, pp. 641–646. DOI: 10.1109/ICMEW.2019.00124.

[299] S. Chen and Q. Zhao, "Attention-based autism spectrum disorder screening with privileged modality," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1181–1190.

[300] K. Sodoké, R. Nkambou, A. Dufresne, and I. Tanoubi, "Toward a deep convolutional lstm for eye gaze spatiotemporal data sequence classification," in *Proceedings of the 13th International Conference on Educational Data Mining*, 2020, pp. 672–676.

[301] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, MIT Press, 2014, pp. 2672–2680.

[302] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision*, IEEE, 2017, pp. 2223–2232.

[303] Z. Yang, L. Huang, Y. Chen, Z. Wei, S. Ahn, G. Zelinsky, D. Samaras, and M. Hoai, "Predicting goal-directed human attention using inverse reinforcement learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[304] G. J. Zelinsky, Y. Chen, S. Ahn, H. Adeli, Z. Yang, L. Huang, D. Samaras, and M. Hoai, *Predicting goal-directed attention control using inverse-reinforcement learning*, 2020. arXiv: 2001.11921 [cs.CV].

[305] S. Mall, P. C. Brennan, and C. Mello-Thoms, "Can a machine learn from radiologists' visual search behaviour and their interpretation of mammograms—a deep-learning study," *Journal of Digital Imaging*, vol. 32, no. 5, pp. 746–760, 2019.

[306] C. Garcia-Vidal, G. Sanjuan, P. Puerta-Alcalde, E. Moreno-García, and A. Soriano, "Artificial intelligence to support clinical decision-making processes," *Ebiomedicine*, vol. 46, pp. 27–29, 2019.

[307] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, 2018.

[308] S. Waite, Z. Farooq, A. Grigorian, C. Sistrom, S. Kolla, A. Mancuso, S. Martinez-Conde, R. G. Alexander, A. Kantor, and S. L. Macknik, "A review of perceptual expertise in radiology-how it develops, how we can test it, and why humans still matter in the era of artificial intelligence," *Academic Radiology*, vol. 27, no. 1, pp. 26–38, 2020.

[309] A. Davies, M. Vigo, S. Harper, and C. Jay, "Using simultaneous scanpath visualization to investigate the influence of visual behaviour on medical image interpretation," *Journal of Eye Movement Research*, vol. 10, no. 5, 2018. DOI: 10.16910/jemr.10.5.11. [Online]. Available: https://bop.unibe.ch/JEMR/article/view/3723.

[310] E. M. Crowe, I. D. Gilchrist, and C. Kent, "New approaches to the analysis of eye movement behaviour across expertise while viewing brain MRIs," *Cognitive Research: Principles and Implications*, vol. 3, no. 1, p. 12, 2018.

[311] G. Wen, A. Aizenman, T. Drew, J. M. Wolfe, T. M. Haygood, and M. K. Markey, "Computational assessment of visual search strategies in volumetric medical images," *Journal of Medical Imaging*, vol. 3, no. 1, p. 015 501, 2016.

[312] R. Li, P. Shi, J. Pelz, C. O. Alm, and A. R. Haake, "Modeling eye movement patterns to characterize perceptual skill in image-based diagnostic reasoning processes," *Computer Vision and Image Understanding*, vol. 151, pp. 138–152, 2016.

[313] R. Li, J. Pelz, P. Shi, C. O. Alm, and A. R. Haake, "Learning eye movement patterns for characterization of perceptual expertise," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2012, pp. 393–396.

[314] N. Ahmidi, G. D. Hager, L. Ishii, G. Fichtinger, G. L. Gallia, and M. Ishii, "Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, 2010, pp. 295–302.

[315] L. Richstone, M. J. Schwartz, C. Seideman, J. Cadeddu, S. Marshall, and L. R. Kavoussi, "Eye metrics as an objective assessment of surgical skill," *Annals of Surgery*, vol. 252, no. 1, pp. 177–182, 2010.

[316] A. T. Duchowski, K. Krejtz, I. Krejtz, C. Biele, A. Niedzielska, P. Kiefer, M. Raubal, and I. Giannopoulos, "The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18, Montréal QC, Canada: ACM, 2018, 282:1–282:13, ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173856. [Online]. Available: http://doi.acm.org/10.1145/3173574.3173856.

[317] P. Vaidyanathan, J. Pelz, C. Alm, P. Shi, and A. Haake, "Recurrence quantification analysis reveals eye-movement behavior differences between experts and novices," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2014, pp. 303–306.

[318] Z. Gandomkar, K. Tay, P. C. Brennan, and C. Mello-Thoms, "A model based on temporal dynamics of fixations for distinguishing expert radiologists' scanpaths," in *Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment*, International Society for Optics and Photonics, vol. 10136, 2017, p. 1 013 606.

[319] Z. Gandomkar, K. Tay, P. C. Brennan, and C. Mello-Thoms, "Recurrence quantification analysis of radiologists' scanpaths when interpreting mammograms," *Medical Physics*, vol. 45, no. 7, pp. 3052–3062, 2018.

[320] S. Mall, E. Krupinski, and C. Mello-Thoms, "Missed cancer and visual search of mammograms: What feature-based machine-learning can tell us that deep-convolution learning cannot," in *Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment*, International Society for Optics and Photonics, vol. 10952, 2019, p. 1 095 216.

[321] *BeGaze Manual*, Version 3.7, SensoMotoric Instruments, 2017.

[322] B. Hoeks and W. J. Levelt, "Pupillary dilation as a measure of attention: A quantitative system analysis," *Behavior Research Methods, Instruments, & Computers*, vol. 25, no. 1, pp. 16–26, 1993.

[323] J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the task-evoked pupillary response with a remote eye tracker," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, ACM, 2008, pp. 69–72.

[324] P. Kiefer, I. Giannopoulos, A. Duchowski, and M. Raubal, "Measuring cognitive load for map tasks through pupil diameter," in *the Annual International Conference on Geographic Information Science*, Springer, 2016, pp. 323–337.

[325] S. Mathôt, J. Fabius, E. Van Heusden, and S. Van der Stigchel, "Safe and sensible preprocessing and baseline correction of pupil-size data," *Behavior Research Methods*, vol. 50, no. 1, pp. 94–106, 2018.

[326] A. C. of Radiology *et al.*, *Acr technical standard for digital image data management*, Reston (VA): American College of Radiology, 2002.

[327] P. C. Brennan, M. McEntee, M. Evanoff, P. Phillips, W. T. O'Connor, and D. J. Manning, "Ambient lighting: Effect of illumination on soft-copy viewing of radiographs of the wrist," *American Journal of Roentgenology*, vol. 188, no. 2, W177–W180, 2007.

[328] J. M. Goo, J.-Y. Choi, J.-G. Im, H. J. Lee, M. J. Chung, D. Han, S. H. Park, J. H. Kim, and S.-H. Nam, "Effect of monitor luminance and ambient light on observer performance in soft-copy reading of digital chest radiographs," *Radiology*, vol. 232, no. 3, pp. 762–766, 2004.

[329] B. J. Pollard, E. Samei, A. S. Chawla, C. Beam, L. E. Heyneman, L. M. H. Koweek, S. Martinez-Jimenez, L. Washington, N. Hashimoto, and H. P. McAdams, "The effects of ambient lighting in chest radiology reading rooms," *Journal of Digital Imaging*, vol. 25, no. 4, pp. 520–526, 2012.

[330] G. C. Kagadis, A. Walz-Flannigan, E. A. Krupinski, P. G. Nagy, K. Katsanos, A. Diamantopoulos, and S. G. Langer, "Medical imaging displays and their use in image interpretation," *Radiographics*, vol. 33, no. 1, pp. 275–290, 2013.

[331] T. Winkler, *NotebookCheck review dell precision m4800 notebook*, 2013. [Online]. Available: `https://www.notebookcheck.net/Review-Dell-Precision-M4800-Notebook.104416.0.html` (visited on 08/01/2019).

[332] A. Ngo, *NotebookCheck review hp zbook 15 workstation*, 2014. [Online]. Available: https://www.notebookcheck.net/Review-HP-ZBook-15-Workstation.108229.0.html (visited on 08/01/2019).

[333] S. Chen and J. Epps, "Using task-induced pupil diameter and blink rate to infer cognitive load," *Human–Computer Interaction*, vol. 29, no. 4, pp. 390–413, 2014. DOI: 10.1080/07370024.2014.892428. [Online]. Available: http://dx.doi.org/10.1080/07370024.2014.892428.

[334] A. Duchowski, K. Krejtz, J. Żurawska, and D. House, "Using microsaccades to estimate task difficulty during visual search of layered surfaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 213, 2019. DOI: 10.1109/TVCG.2019.2901881.

[335] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988.

[336] F. Paas and P. Ayres, "Cognitive load theory: A broader view on the role of memory in learning and education," *Educational Psychology Review*, vol. 26, no. 2, pp. 191–195, 2014.

[337] M. A. Just and P. A. Carpenter, "A capacity theory of comprehension: Individual differences in working memory.," *Psychological Review*, vol. 99, no. 1, p. 122, 1992.

[338] S. Ahern and J. Beatty, "Pupillary responses during information processing vary with scholastic aptitude test scores," *Science*, vol. 205, no. 4412, pp. 1289–1292, 1979.

[339] S. P. Verney, E. Granholm, and S. P. Marshall, "Pupillary responses on the visual backward masking task reflect general cognitive ability," *International Journal of Psychophysiology*, vol. 52, no. 1, pp. 23–36, 2004.

[340] H. L. Kundel, C. F. Nodine, E. A. Krupinski, and C. Mello-Thoms, "Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms," *Academic Radiology*, vol. 15, no. 7, pp. 881–886, 2008.

[341] A Van der Gijp, M. Van der Schaaf, I. Van der Schaaf, J. Huige, C. Ravesloot, J. van Schaik, and T. J. ten Cate, "Interpretation of radiological images: Towards a framework of knowledge and skills," *Advances in Health Sciences Education*, vol. 19, no. 4, pp. 565–580, 2014.

[342] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2014. arXiv: 1409.1556 [cs.CV].

[343] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2009, pp. 248–255.

[344] F. Chollet, *Keras*, https://github.com/fchollet/keras, 2015.

[345] A. Davies, G. Brown, M. Vigo, S. Harper, L. Horseman, B. Splendiani, E. Hill, and C. Jay, "Exploring the relationship between eye movements and electrocardiogram interpretation accuracy," *Scientific Reports*, vol. 6, p. 38 227, 2016.

[346] T. C. Kübler, K. Sippel, W. Fuhl, G. Schievelbein, J. Aufreiter, R. Rosenberg, W. Rosenstiel, and E. Kasneci, "Analysis of eye movements with eyetrace," in *International Joint Conference on Biomedical Engineering Systems and Technologies*, Springer, 2015, pp. 458–471.

[347] M. M. Chun and J. M. Wolfe, "Just say no: How are visual searches terminated when there is no target present?" *Cognitive Psychology*, vol. 30, no. 1, pp. 39–78, 1996.

[348] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 3, p. 419, 1989.

[349] R. Bailey, A. McNamara, N. Sudarsanam, and C. Grimm, "Subtle gaze direction," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 4, pp. 1–14, 2009.

[350] A. van der Gijp, K. L. Vincken, C. Boscardin, E. M. Webb, O. T. J. Ten Cate, and D. M. Naeger, "The effect of teaching search strategies on perceptual performance," *Academic Radiology*, vol. 24, no. 6, pp. 762–767, 2017.

[351] W. Chen, D. HolcDorf, M. W. McCusker, F. Gaillard, and P. D. Howe, "Perceptual training to improve hip fracture identification in conventional radiographs," *Plos One*, vol. 12, no. 12, 2017.

[352] H. S. Barrows, R. M. Tamblyn, *et al.*, *Problem-based learning: An approach to medical education*. Springer Publishing Company, 1980, vol. 1.

[353] J. E. Froyd, "Problem-based learning and adaptive expertise," in *2011 Frontiers in Education Conference (FIE)*, IEEE, 2011, S3B–1–S3B–5.

[354] C. C. Farnsworth, "Measuring the effects of problem-based learning on the development of veterinary students' clinical expertise.," *Academic Medicine*, vol. 72, no. 6, pp. 552–554, 1997.

[355] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, *et al.*, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, e1356, 2020.

[356] S. Samuel, H. L. Kundel, C. F. Nodine, and L. C. Toto, "Mechanism of satisfaction of search: Eye position recordings in the reading of chest radiographs.," *Radiology*, vol. 194, no. 3, pp. 895–902, 1995.

[357] K. S. Berbaum, E. Franken Jr, D. D. Dorfman, R. T. Caldwell, and E. A. Krupinski, "Role of faulty decision making in the satisfaction of search effect in chest radiography," *Academic Radiology*, vol. 7, no. 12, pp. 1098–1106, 2000.

[358] K. S. Berbaum, K. M. Schartz, R. T. Caldwell, M. T. Madsen, B. H. Thompson, B. F. Mullan, A. N. Ellingson, and E. A. Franken Jr, "Satisfaction of search from detection of pulmonary nodules in computed tomography of the chest," *Academic Radiology*, vol. 20, no. 2, pp. 194–201, 2013.

[359] E. A. Krupinski, K. S. Berbaum, K. M. Schartz, R. T. Caldwell, and M. T. Madsen, "The impact of fatigue on satisfaction of search in chest radiography," *Academic Radiology*, vol. 24, no. 9, pp. 1058–1063, 2017.

[360] T. Drew, M. L. H. Võ, and J. M. Wolfe, "The invisible gorilla strikes again: Sustained inattentional blindness in expert observers," *Psychological Science*, vol. 24, no. 9, pp. 1848–1853, 2013.

[361] J. Chen, S. Littlefair, R. Bourne, and W. M. Reed, "The effect of visual hindsight bias on radiologist perception," *Academic Radiology*, vol. 27, no. 7, pp. 977–984, 2020.

[362] A. Hashem, M. T. Chi, and C. P. Friedman, "Medical errors as a result of specialization," *Journal of Biomedical Informatics*, vol. 36, no. 1-2, pp. 61–69, 2003.

[363] J. S. Trueblood, W. R. Holmes, A. C. Seegmiller, J. Douds, M. Compton, E. Szentirmai, M. Woodruff, W. Huang, C. Stratton, and Q. Eichbaum, "The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making," *Cognitive Research: Principles and Implications*, vol. 3, no. 1, pp. 1–14, 2018.

Appendix

The following publications are the contributions detailed in chapter 5 and appended in chronologial order from earliest to most recent. The publications with supplementary material have this material included after the respective last page. All publication articles appended are in their final publicized form except for the last article, *Towards expert gaze modeling and recognition of a user's attention in realtime*. At the time of this thesis, this article was accepted at a workshop as part of the *KES* conference, and had not yet been offically published. All published articles appendend, partial or complete, have permission to be included in this dissertation.

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.

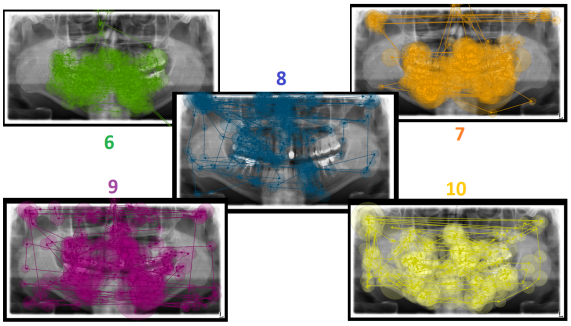


Figure 1: Visualization of fixations from a student in each semester evaluated in the current study as indicated by the colored numbers respectively. In this condition, the sixth semester student's data is prior to training.

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.

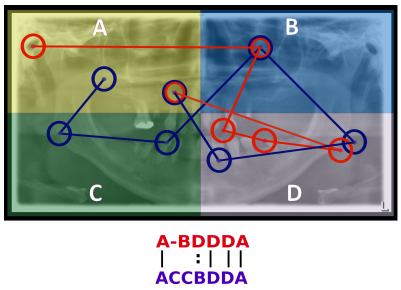


Figure 2: Scanpath comparison example with two scanpaths for same stimuli and AOI grid. Below the image is the global string alignment calculated with the Needleman-Wunsch algorithm. Matches, mismatches, and gaps are [| , : , -] respectively.

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."

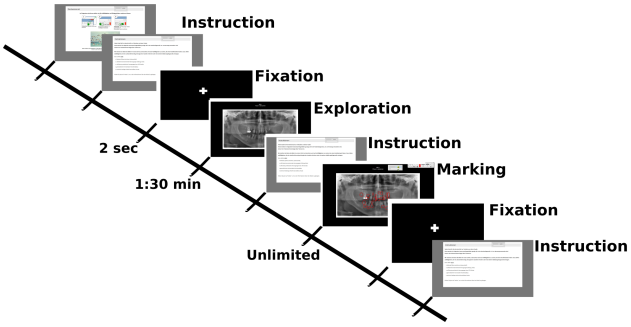


Figure 3: Outline of Experimental Session. After a calibration, there is an introduction to the task and a tutorial on marking the anomalies. After, a verbal instruction was presented with information on what kind of anomalies he should focus on. The subject is primed with a fixation cross. Then in the 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 $40^{\circ}/s$ 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.

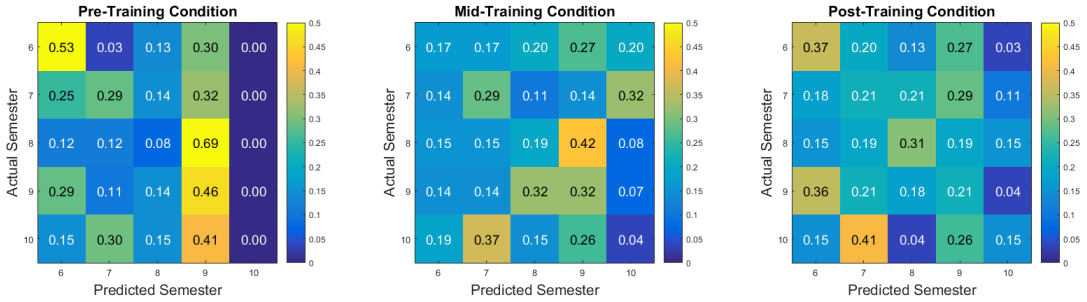


Figure 4: SubsMatch semester classification on the validation data. From left to right, confusion matrices for conditions pre-training, mid-training, and post-training are presented. With true positive rate for each semester along the diagonals. Note that the colorbar for all conditions is scaled at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

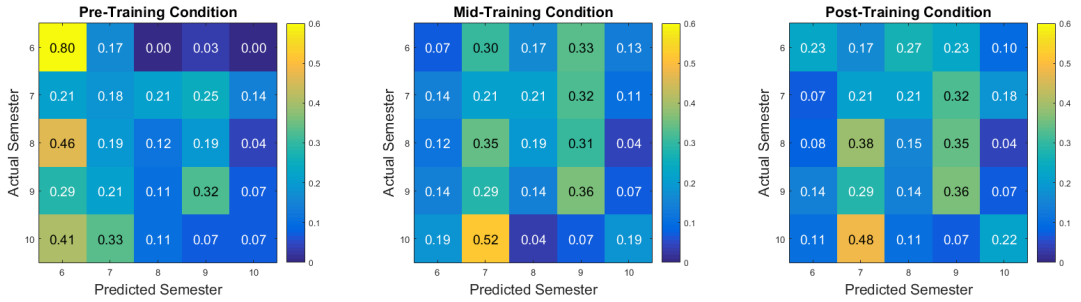


Figure 5: Needleman-Wunsch semester classification on the validation data. From left to right, confusion matrices for conditions pre-training, mid-training, and post-training are presented. With true positive rate for each semester along the diagonals. Note that the colorbar for all conditions is scaled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

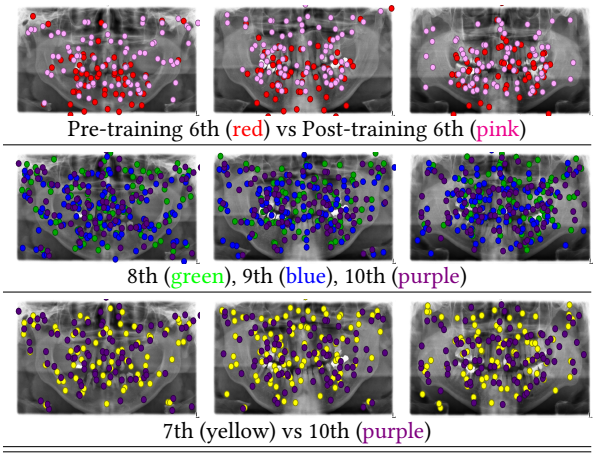
Al-Moteri and colleagues [Al-Moteri et al. 2017] comprised literature regarding eye movements and medical decision making and found that clinical experience was related to gaze behavior that was more *goal-driven* and less *stimulus-driven* [Al-Moteri et al. 2017; Krupinski et al. 2006]. This finding 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.

Supplemental Material

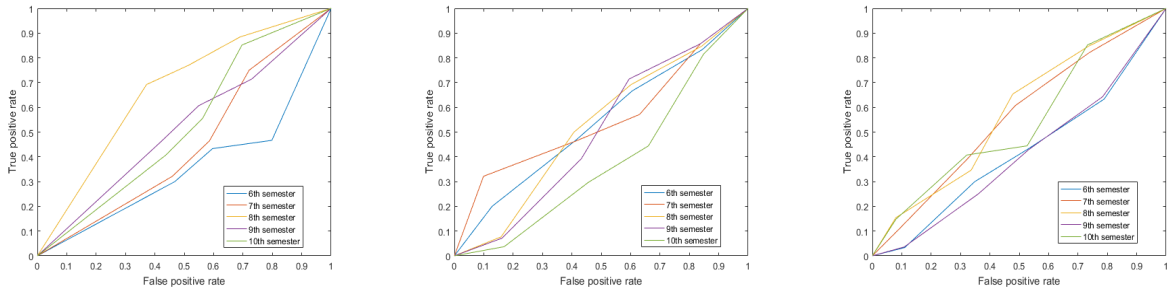
Table 1: Validity Of Multiple Class Classifiers for Sixth Semester Versus the Rest.

Condition	Pre-	Mid-	Post-	Pre-	Mid-	Post-
Classifier	Subsmatch 2.0			Needleman-Wunsch		
Sensitivity[a]	0.5333	0.1667	0.3667	0.8000	0.0667	0.2333
Specificity[b]	0.7982	0.8440	0.7890	0.6606	0.8532	0.8991
Precision[c]	0.4211	0.2273	0.3225	0.3934	0.1111	0.3889
F-Score[d]	0.4706	0.1923	0.3438	0.5275	0.0833	0.2917

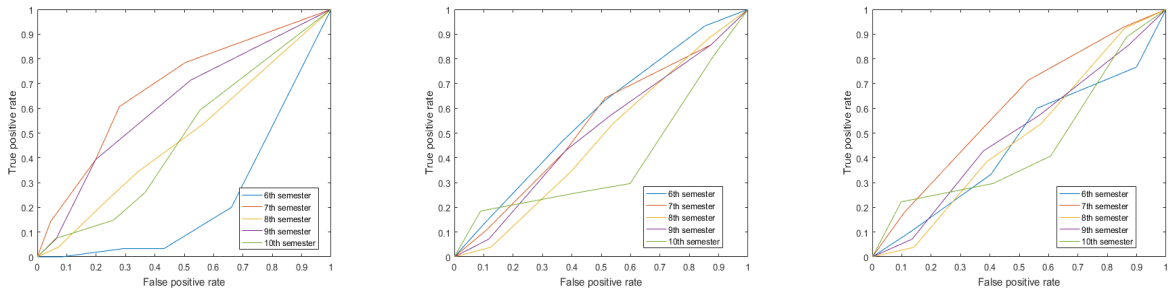
[a]Recall, True Positive Rate, or Hit Rate
[b]True Negative Rate
[c]Positive Predictive Value
[d]Harmonic Mean of precision and recall

Table 2: ROC Curves for Five Semesters

Pre- Mid- Post-



SubsMatch 2.0



Needleman-Wunsch

Development and Evaluation of a Gaze Feedback System Integrated into EyeTrace

Kai Otto
Perception Engineering, University of Tübingen
Tübingen, Germany
kai.otto@student.uni-tuebingen.de

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

David Geisler
Perception Engineering, University of Tübingen
Tübingen, Germany
david.geisler@uni-tuebingen.de

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

ABSTRACT

A growing field of studies in eye-tracking is the use of gaze data for realtime feedback to the subject. In this work, we present a software system for such experiments and validate it with a visual search task experiment. This system was integrated into an eye tracking analysis tool. Our aim was to improve subject performance in this task by employing saliency features for gaze guidance. This realtime feedback system can be applicable within many realms, such as learning interventions, computer entertainment, or virtual reality.

CCS CONCEPTS

• **Human-centered computing** → **Interaction paradigms**; • **Applied computing** → **E-learning**;

KEYWORDS

E-Learning, eyetracking, gaze-based feedback

ACM Reference Format:
Kai Otto, Nora Castner, David Geisler, and Enkelejda Kasneci. 2018. Development and Evaluation of a Gaze Feedback System Integrated into EyeTrace. In *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3204493.3204561

1 INTRODUCTION

In many real-world situations, we are confronted with visual displays where information needs to be extracted and interpreted. Often, it is almost unavoidable that important details are missed, as the human eye and brain can only process parts of the screen at a time [Jarodzka et al. 2012]. We are limited due to the fovea restricting our field of view: where we see sharply within two degrees [Holmqvist et al. 2011]. Therefore, visual search merges our fixations in order to perceive all present information. To stimulate perception, we can learn how to spread our attention more effectively. Effective gaze guiding can be implemented as graphical user interfaces or other visual feedback forms and has shown promising outcomes in a range of professions. For instance, in air traffic control [Mackworth 1948], piloting a vehicle [Wetzel et al. 1998], and reading medical imagery [Jarodzka et al. 2012]. Here, task detection and interpretation under certain circumstances is not only time consuming to learn, but can also be safety critical.

Gaze guidance or supportive highlighting of on-screen information can help in a number of scenarios. For instance, teaching systematic search of medical x-ray images [Kok et al. 2016; Kundel and La Follette Jr 1972; Van der Gijp et al. 2017]. Additionally, in air traffic control simulation, where stimuli is dynamic, gaze guidance highlights the relevant information as it appears [Mackworth 1948].

These attentional guiding systems not only highlight relevant information areas, but also needs to account already perceived information [Jarodzka et al. 2013]. From the literature, it is known that eye tracking offers insight into a user's perception through their gaze behavior [Holmqvist et al. 2011]. Thus in this work, we focus on effective visualization of online gaze behavior. Specifically, realtime gaze feedback that visualizes already viewed regions and incorporates more information from the periphery.

2 RELATED WORK

Employing eye movement data in the educational context has offered insight into how to model gaze. Most notable are the eye movement modeling examples (EMMEs); Where visual guidance to directly influence gaze behavior was employed by Jarodzka et al., in order to increase subjects' interpretation performance of medical records [Jarodzka et al. 2012] and a biological classification task [Jarodzka et al. 2009, 2013]. For [Jarodzka et al. 2012], eye movement data of experts were visualized by blurring areas they did not look at: i.e. non-relevant information. For [Jarodzka et al. 2009], experts' gaze was visualized as yellow circles on a stimulus image. For both studies, the model example incorporated gaze data post hoc.

Qvardfort and colleagues [Qvarfordt et al. 2010] found that applying white circular occlusions to fixations from a previous free-viewing over the stimuli was able to reduce the workload during a visual search task while increasing the true positive rate of targets

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

ACM ISBN 978-1-4503-5706-7/18/06…$15.00
https://doi.org/10.1145/3204493.3204561

found [Qvarfordt et al. 2010]. Thus, participants were able to notice areas that they did not search in the free-viewing phase because the already search areas were occluded.

ScreenMasker [Orlov and Bednarik 2016] is a open source software by Orlov and Bednarik intent on developing a customizable system that visualizes gaze behavior. Their gaze contingent system creates a pattern mask over the on screen stimuli. Then, it uses gaze coordinates from the eye tracker to subtract the pattern, or unmask, where the subject is gazing in real time. For this system, an NVIDIA graphics card with CUDA framework was used and was shown to perform with very low latencies [Orlov and Bednarik 2016]. Thus, offering low to none temporal offset that could disturb a user.

We propose a platform integrated in a publicly available eye tracking analysis tool. The multiple plugins integrated offer an experimental center and a realtime gaze feedback option. Our system was tested and capable of running an a standard computer.

3 SOFTWARE DEVELOPMENT

Eyetrace [Kübler et al. 2015] is a software providing state-of-the-art algorithms for eye tracking data visualization, statistical analysis, event detection, AOI generation, saccade clustering, and scanpath analysis and supports a variety of eye trackers. All algorithms are parameterizable and the parameters together with the visualization and statistics can be exported. Therefore, we decided to extend this existing software, which is publicly available at http://www.ti.uni-tuebingen.de/Eyetrace.eyetrace.0.html

For our experiment (detailed in section 4), we used the EyeTribe eye tracker [Ooms et al. 2015] since it was already supported by Eyetrace. We extended this plug-in to support online usage whereas previously, only recording and importing the eye tracking data was available. The developed application interface also allows for extending the plugin to other eye trackers and online calibration.

3.1 The Experimenter

The *Experimenter* plug-in for EyeTrace was developed for creating and performing remote eye tracking experiments. The central part is the Designer widget, shown in figure 1, which has the following capabilities:

- Create and modify the experiment design where each index block is highly customizable (Figure 1 area 1 and 2).
- Import and export experiment designs as CSV file (Figure 1 area 2).
- Record subject data together with name, group and dominant eye (Figure 1 area 3).
- Select the Eyetracker to be used (Figure 1 area 4).
- Select an interruption key (Figure 1 area 5).
- Start/cancel the experiment run (Figure 1 area 5).

In the *Experimenter*, a researcher can manually organize an experiment design offering customization of stimuli, time of presentation, gazefeedback, and keypress interruptions. These experimental designs can be exported and saved as a CSV file, for additional data collection. Additionally, the ability to import experiment designs in CSV file format allows for the option of autogenerating randomized experiment designs with a simple script in any programming language with CSV parsing libraries or text editor.

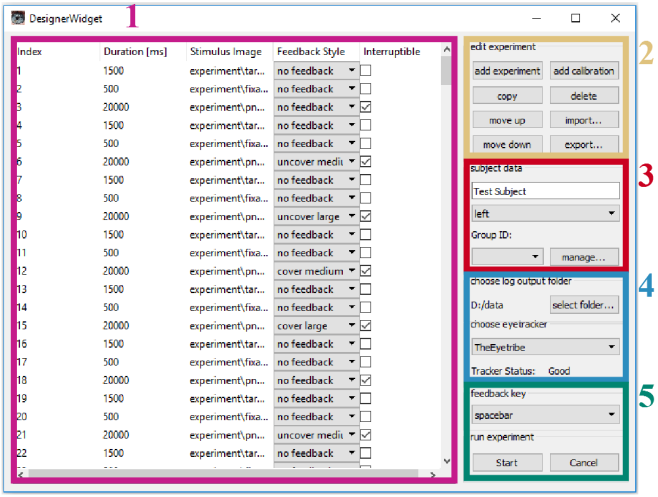


Figure 1: Designer Widget GUI. Here, experiments can be designed, and managed. The workflow of the experiment is organized (1) and can be modified (2) and each participant's data is defined (3). For each experiment, an eye tracker is selected (4) as well as a key for interruption (5).

Each step of the experiment design can either be a calibration, or a stimulus presentation/recording. In calibration, the eyetracker's calibration from the API is employed. In stimulus presentation, the durations (in milliseconds), the filepaths (if none is chosen, a white screen will be displayed), and whether the step is interruptible through keypress are customizable. The interruptible option is optimal for experimental designs where reaction-time or decision-making tasks are evaluated. Additionally, the researcher can also present the online gaze feedback for any number of stimuli, as described further in section 3.2.

The 'Start' button in the Designer widget initiates the experiment: Data logging starts here as well. Stimuli are shown in the Presenter widget, a second full screen widget that gets called. Ideally, if the researcher has two monitors, the main window of Eyetrace and the Designer widget can be displayed to the researcher and the participant only sees the Presenter widget. The researcher can always cancel the experiment with the Designer widget's 'Cancel' button. Otherwise, the Experimenter runs through the designed stimuli list and terminates at the end, closing the Presenter widget.

On the data handling side, timestamps, gaze coordinates, and keypresses are recorded in a log file. Internally, they are stored in a data structure for the experimental session for later calculations and analyses.

3.2 Realtime Feedback

In the realtime feedback, the user's gaze data is visualized on the screen as he or she is performing a task. In order to achieve low and relatively constant response times of the feedback system, intermediate results are stored in a cache. Then, the system only has to process new gaze data when it is repeatedly called.

Triggered by a timer, every 7 ms (approximately 144 Hz) the screen drawing method of the Presenter widget gets called to update

(a) (b)

Figure 2: Screenshot of an experiment trial, showcasing the 'cover' (a) and 'uncover' (b) feedback condition. Stimulus: Ilya Repin, "Unexpected Visitors", 1884-1888. Oil on canvas. public domain https://commons.wikimedia.org/wiki/.

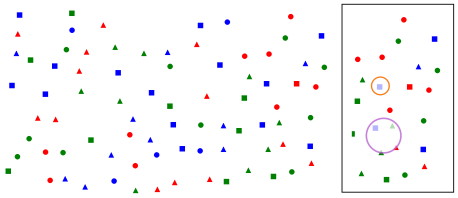


Figure 3: An example stimuli, as used in the experiment. In this case, subjects had to decide whether there is a red square visible or not. Stimuli was presented at full resolution. On the right, are the respective sizes for cover small (orange) and cover large (purple) conditions.

the screen content. This trigger calls the currently active realtime feedback implementation to draw over the stimulus.

Presently, two feedback algorithms are implemented, plus the default 'no feedback' condition. First, the 'cover' feedback occludes the user's gaze coordinates on the stimulus with opaque circles as illustrated in figure 2a. Second, the 'uncover' feedback unoccludes a semitransparent cover in a similar manner to the former condition as illustrated in figure 2b. Essentially, this feedback is the complement of the former applied to the mask overlay. For both conditions, there is no decay of feedback for older gaze points.

Both feedback conditions use a white mask-like image overlayed over the original stimuli, and the feedback effects the masks' alpha channel, meaning its opacity is changed. Each event where the Realtime Feedback class is called, the list of new gaze points is run through and circles are drawn on the mask overlay for each new gaze point coordinate. For a video recording illustrating both the 'cover' and the 'uncover' feedback methods, please refer to the supplementary materials.

In both feedback conditions, the mask is either transparent (for covered) or semi-transparent (for uncovered). the alpha channel on this map is then changed based on the gaze coordinates. The compositing method adds or subtracts the circle's alpha values to the existing mask corresponding to gaze coordinates, giving the effect of decreasing or increasing transparency the longer the subject looks at a certain spot. However, a lower bound threshold is given to the circle's alpha value to prevent any part of the stimulus becoming invisible. Each circle consists of a radial gradient, projecting outwards from the circle's center to its edge. This effect makes the feedback appear smoother, removing distracting, sharp edges (see figure 2). The compositing method updates the mask with the new gaze points each time the trigger timer event takes place. Then, the mask gets drawn over the stimulus.

3.3 Gaze Behavior with Feedback

In order to evaluate our online gaze feedback system, we measured performance in a visual search task with the feedback as an independent variable. We propose that both the cover and uncover feedback conditions will affect gaze behavior compared to no feedback at all.

The cover feedback method could have two effects. One, subjects may be less likely to look a second time at areas of the stimulus they already looked at, as the saliency gets decreased after looking the first time. Meaning their time to scan the image is shorter resulting in shorter reaction times. Two, subjects will have longer saccades, as the feedback includes distractors in the periphery of what is currently being fixated, which otherwise might have been the area of the next fixation.

Additionally, the uncover feedback could result in subjects' scanning behavior becoming more systematic because the saliency is being reduced, resulting in fewer inconsistent saccades across the stimulus. This systematic search behavior effect was also found in [Jarodzka et al. 2013], the spotlight condition (where non relevant information is blurred) resulted in faster, more efficient detection of relevant information.

4 METHODS

For our experiment, we used a Windows 10 computer with a 27 inch monitor (resolution 1920×1080 pixel) as display device, and the Eyetribe eyetracker. Evaluation of the Eyetribe with regards to scientific usability can be found in the literature [Dalmaijer 2014; Ooms et al. 2015].

The visual search task was performed with images consisting of either 80 distractors (target absent) or 79 distractors plus the target item (target present). An example image is shown in figure 3. In total, 100 images were generated [1]. Order of stimuli presentation was randomized for each participant.

For each stimuli, the target item was centered on screen for 1.5 seconds. Following target presentation, a fixation cross was visualized for 0.5 seconds, and then the stimulus presentation. To signal their decision, participants pressed a button on the keyboard: Either keypress `m` (right index finger) for target present, or keypress `y` (left index finger) for target absent.

The feedback methods, as introduced in section 3.2, were parameterized the following way. For each of the two feedback types (uncovering, covering), we chose two diameters: 100 or 200 pixel diameter. The control condition was no gaze feedback. Conditions, in conjunction to order of stimuli, were randomized to allow for a within-subject evaluation.

A total of 18 participants (17 university students; five wore glasses) took part. They were positioned roughly 60 cm away from the screen. A 9-Point calibration was performed using the EyeTribe's calibration software. Following the experiment, participants

[1] The images had equal distribution of color and shape of the target item, and its absence/presence.

Table 1: Reaction times for correct answers when target was absent. t-test is between intervention and control condition. '*' indicates a significant result.

condition	μ [ms]	σ [ms]	t-value	p
Target absent				
control	5256.49	2715.04	n.a.	
uncover small	5144.21	2531.49	0.39	0.70
uncover large	5213.20	2430.17	0.16	0.88
cover small	5198.08	2370.27	0.21	0.83
cover large	4693.89	1897.47	2.28	0.02*

filled out a self-report regarding perceived performance and experience.

Eye movement data was evaluated in Eyetrace. Fixations were calculated with the I-DT algorithm [Salvucci and Goldberg 2000] with the following parameters: minimum duration of 50 ms, maximum radius of 20 pixel, maximum outliers of 0. Saccades were then calculated as the spatial representation between two fixations. In addition to eye movement data, response error rate and reaction time for were calculated. Reaction time was defined as time between onset of stimulus to keypress.

5 RESULTS

From the questionnaire responses, it was found that self reports of effectiveness and helpfulness for both feedback conditions did not significantly differ compared to the control condition of no feedback.

Participant behavior for each experimental trial where they responded correctly was evaluated. Only 4.3% of the total trials were excluded because they were incorrect responses. A low correlation ($r = 0.31$) ruled out any effect of target distance from center fixation cross on reaction time.

5.1 Performance

The reaction times for target absent trials ($\mu_{\text{absent}} = 5095.12$ ms, $\text{SD}_{\text{absent}} = 2405.44$ ms) were significantly longer than for the target present trials ($\mu_{\text{present}} = 2254,12$ ms, $\text{SD}_{\text{present}} = 1346.08$ ms: $t = -30.16$, $p < 0.001$).

Regarding feedback intervention and reaction time, it was found that when the target was absent, the cover large (200 pixel diameter) condition had significant differences in reaction time. The Welch's unequal variances t-test[2] as shown in Table 1 found that this condition had significantly shorter reaction times ($t = 2.28, p = 0.023$).

When the target was present, reaction times were overall shorter, though there was no significant differences between feedback conditions here.

5.2 Gaze Behavior

Similar to reaction time, the effect of target absent or present on numbers of saccades was highly significant. Where there were more saccades when the target was present ($\mu_{\text{present}} = 16.75$, $\text{SD}_{\text{present}} = 9.32$) than when the target was absent ($\mu_{\text{absent}} = 7.44$, $\text{SD}_{\text{absent}} = 3.47$: Welch's unequal variances test, $t = 8.85$, $p < 0.001$). However, feedback conditions showed no significant effect on number of saccades.

[2]Welch's unequal variances t-test pools together all values for each condition, meaning sample sizes are larger, which increases the statistical power.

Table 2: Mean and standard deviation for saccade length. t-test is between intervention and control condition. '*' indicates a significant result with $p < 0.05$, '' for significance level $p < 0.005$.**

condition	μ_{sac} [px]	σ_{sac} [px]	t-value	p
control	313.84	148.48	n.a.	
uncover small	329.26	157.14	-2.16	0.045*
uncover large	333.11	152.79	-2.52	0.022*
cover small	344.18	164.12	-3.11	0.006*
cover large	333.19	150.23	-3.31	0.004**

Table 3: Mean and standard deviation for number of fixations needed to complete the task for each of the five conditions, split for target is absent. t-test between control and intervention condition, '*' indicates a significant result.

condition	μ	σ	t-value	p
Target absent				
control	17.36	9.53	n.a.	
uncover small	17.19	10.99	0.22	0.83
uncover large	16.86	8.99	0.78	0.45
cover small	16.30	9.67	2.31	0.03*
cover large	16.08	8.07	1.61	0.13

More interesting, saccade length was affected by the feedback. Here, the Welch's unequal variances t-test (values in table 2) also reported significant differences for feedback conditions compared to control, where the feedback conditions had longer saccade lengths.

Concerning fixations, fixation duration was not significantly different between control ($\mu_{\text{control}} = 118.47$ms, $\text{SD}_{\text{control}} = 49.45$ms) and all feedback conditions (For example, $\mu_{\text{coverLarge}} = 116.67$ms, $\text{SD}_{\text{coverLarge}} = 40.34$ms: $t = 0.99, p = 0.34$).

Similar to saccades and reaction times, the number of fixations was higher for target absent. Where for both controls: target present ($\mu_{\text{present,control}} = 7.22$, $\text{SD}_{\text{present,control}} = 3.07$) and target absent ($\mu_{\text{absent,control}} = 17.36$, $\text{SD}_{\text{absent,control}} = 9.53$).

There were no significant differences between feedback and control condition when the target was present. However, when the target was absent (see table 3), an effect for the cover small feedback (diameter 100px) condition was found. Although there were no significant differences, a trend can also be seen for less fixations in the cover and uncover large conditions compared to the control.

6 DISCUSSION

Generally, it took subjects longer to correctly decide if a target is absent, than it took them to decide if a target is present. The difference was highly significant; this experimental result reproduces a

well documented effect of target presence or absence on reaction time [Chun and Wolfe 1996; Wolfe et al. 1989].

Concerning how the intervention influenced subject behavior, we can see that even with our rather simple feedback methods, we were able to induce a change in subjects. An increase in periphery employed is apparent from the longer saccades for both covering and uncovering interventions. However, only the cover large condition, where a semi-transparent circle with a 200 pixel diameter overlayed on the gaze coordinates, increased reaction times when the target was accurately determined as absent. There was also a trend for less fixations when determining the target was absent for both cover and uncover (where the circle uncovers a semi-transparent overlay) large feedback conditions, though significantly less fixations were only found in the cover small (100 pixel diameter) feedback condition. Therefore, the realtime gaze based feedback algorithms developed for the system produced and effect on gaze behavior in the visual search task.

Interestingly enough, the self-reports from the participants did not indicate that the feedback helped or improved their performance. In contrast, their reaction times as well as their eye movement differences showed that gaze feedback indeed affected their behavior compared to no gaze feedback. Participants also reported that none of the feedback conditions were distracting in any way. Therefore, the gaze feedback system we developed appears to be unobtrusive, yet effective.

Regarding a more effective gaze model, the current experiment found that the large cover affected reaction time for target absent being correctly determined. However, overall correct detection was extremely high at 96%. Future work into effective gaze modeling could look into more complex visual search tasks to see whether gaze modeling improves performance.

Jarodzka and colleagues [Jarodzka et al. 2013] found that using either the spotlight condition (where non relevant information is blurred) or the dot condition for EMMEs were both effective in modeling gaze behavior. However, each condition affected a certain aspect of learning and performance, where the spotlight condition affected visual search and the dot condition affected interpretation [Jarodzka et al. 2013]. In our system, the uncover feedback algorithm is relatively similar to their spotlight condition, where both present a clear unaffected gaze area, and occlude the other areas ([Jarodzka et al. 2013]: blurring, ours: opaque mask). Additionally, [Jarodzka et al. 2013]'s dot condition is similar to our cover condition; however, ours covers the gaze area with a semi-transparent mask that does not hide the stimulus information underneath. Whether our online gaze feedback would be beneficial for learning environments is of great interest in future research.

7 CONCLUSION

In this work, we introduced a novel software system for eyetracking experimentation, which allows realtime feedback to the subject. We successfully validated the implementation in a visual search task study. The current system was integrated into our eye tracking analysis tool EyeTrace. Now, this tool provides experimental design and testing in addition to the analysis and visualization of eye tracking data.

REFERENCES

Marvin M Chun and Jeremy M Wolfe. 1996. Just say no: How are visual searches terminated when there is no target present? *Cognitive psychology* 30, 1 (1996), 39–78.

Edwin Dalmaijer. 2014. *Is the low-cost EyeTribe eye tracker any good for research?* Technical Report. PeerJ PrePrints.

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, Katharina Scheiter, Peter Gerjets, Tamara van Gog, and Michael Dorr. 2009. How to convey perceptual skills by displaying experts' gaze data. In *Proceedings of the 31st annual conference of the cognitive science society*. 2920–2925.

Halszka Jarodzka, Tamara van Gog, Michael Dorr, Katharina Scheiter, and Peter Gerjets. 2013. Learning to see: Guiding students' attention via a Model's eye movements fosters learning. *Learning and Instruction* 25 (2013), 62–70.

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.

Thomas C Kübler, Katrin Sippel, Wolfgang Fuhl, Guilherme Schievelbein, Johanna Aufreiter, Raphael Rosenberg, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2015. Analysis of eye movements with Eyetrace. In *International Joint Conference on Biomedical Engineering Systems and Technologies*. Springer, 458–471.

Harold L Kundel and Paul S La Follette Jr. 1972. Visual search patterns and experience with radiological images. *Radiology* 103, 3 (1972), 523–528.

Norman H Mackworth. 1948. The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology* 1, 1 (1948), 6–21.

Kristien Ooms, Lien Dupont, Lieselot Lapon, and Stanislav Popelka. 2015. Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental setups. *Journal of eye movement research* 8, 1 (2015).

Pavel A Orlov and Roman Bednarik. 2016. ScreenMasker: An open-source gaze-contingent screen masking environment. *Behavior research methods* 48, 3 (2016), 1145–1153.

Pernilla Qvarfordt, Jacob T Biehl, Gene Golovchinsky, and Tony Dunningan. 2010. Understanding the benefits of gaze enhanced visual search. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 283–290.

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.

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.

Paul A Wetzel, Gretchen M Anderson, and Barbara A Barelka. 1998. Instructor use of eye position based feedback for pilot training. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 42. SAGE Publications Sage CA: Los Angeles, CA, 1388–1392.

Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. 1989. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance* 15, 3 (1989), 419.

Supplemental Material

Table 1: Reaction times of participants in trials with correct answer, for each of the feedback conditions where target is present and absent. Welch's unequal variances t-test is between intervention and control condition. '*' indicates a significant result.

condition	μ [ms]	σ [ms]	t-value	p
Target present				
control	2169.33	1282.56	n.a.	
uncover small	2366.83	1451.12	-1.37	0.17
uncover large	2284.86	1306.01	-0.8	0.42
cover small	2290.51	1443.52	-0.82	0.41
cover large	2130.13	1176.14	0.29	0.78
Target absent				
control	5256.49	2715.04	n.a.	
uncover small	5144.21	2531.49	0.39	0.70
uncover large	5213.20	2430.17	0.16	0.88
cover small	5198.08	2370.27	0.21	0.83
cover large	4693.89	1897.47	2.28	0.02*

Table 2: Mean and standard deviation for number of fixations needed to complete the task for each of the five conditions, split for 'target is present' and 'target is absent'. Welch's unequal variances t-test between control and intervention condition, '*' indicates a significant result.

condition	μ	σ	t-value	p
Target present				
control	7.22	3.07	n.a.	
uncover small	7.93	3.88	-1.65	0.12
uncover large	7.55	3.30	-0.72	0.48
cover small	7.42	3.59	-0.46	0.65
cover large	7.05	3.40	0.29	0.78
Target absent				
control	17.36	9.53	n.a.	
uncover small	17.19	10.99	0.22	0.83
uncover large	16.86	8.99	0.78	0.45
cover small	16.30	9.67	2.31	0.03*
cover large	16.08	8.07	1.61	0.13

Table 3: Mean fixation duration and standard deviation for Welch's unequal variances t-test between intervention and control condition. Values for target present and target absent combined

condition	μ_{fix} [ms]	σ_{fix} [ms]	t-value	p
control	118.47	49.45	n.a.	
uncover small	119.27	41.64	-0.44	0.67
uncover large	120.10	45.97	-0.75	0.46
cover small	121.11	48.70	-1.05	0.30
cover large	116.47	40.34	0.99	0.34

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,

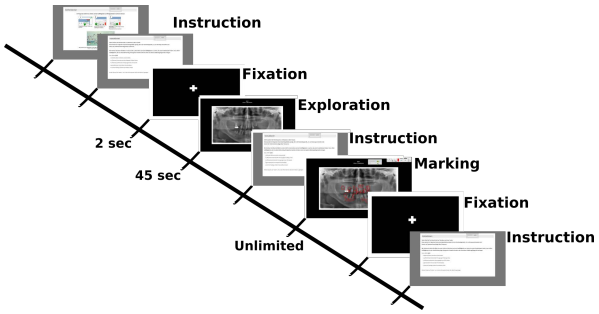


Figure 1: Experimental Session Protocol. The protocol comprised of a calibration, introduction, and instruction, then for the 15 OPTs, a fixation, exploration, and drawing. Image borrowed from [1].

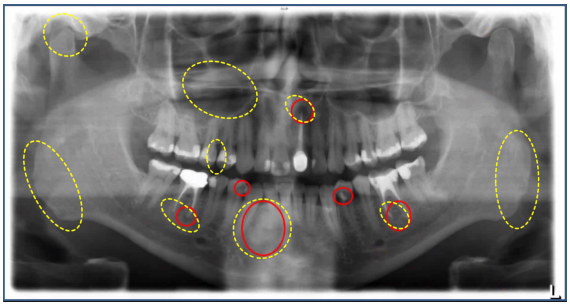


Figure 2: Drawing example. Drawings from a participant (Red) with predefined anomalies (Dotted Yellow), or targets, overlayed. In this example, the participant would have four hits and five misses and two false positives.

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.

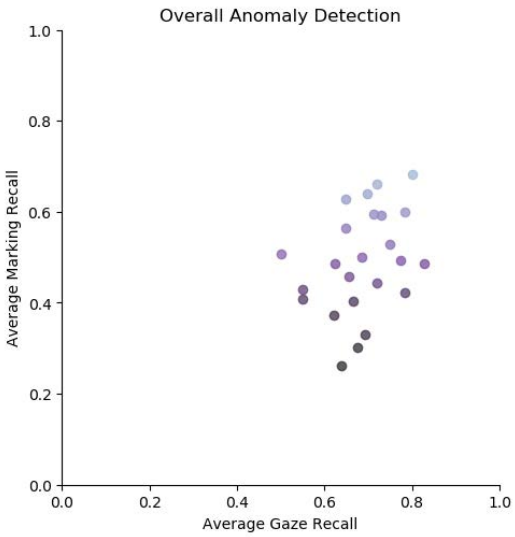


Figure 3: Relationship between overall gaze recall and marking recall. The lighter hues are indicative of higher marking recall.

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%).

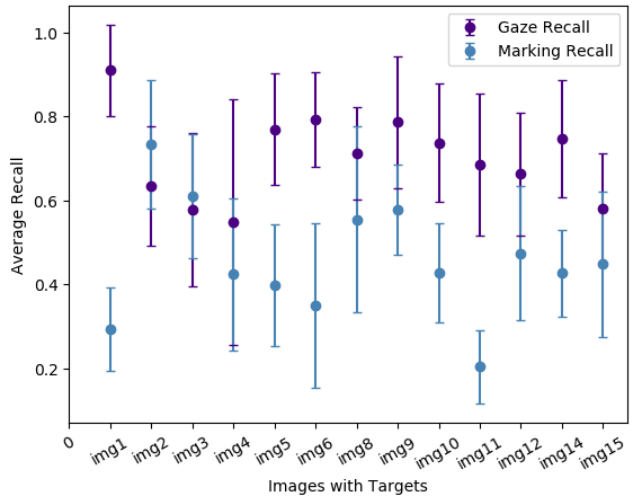


Figure 4: Frequency of Glances for marked and unmarked targets

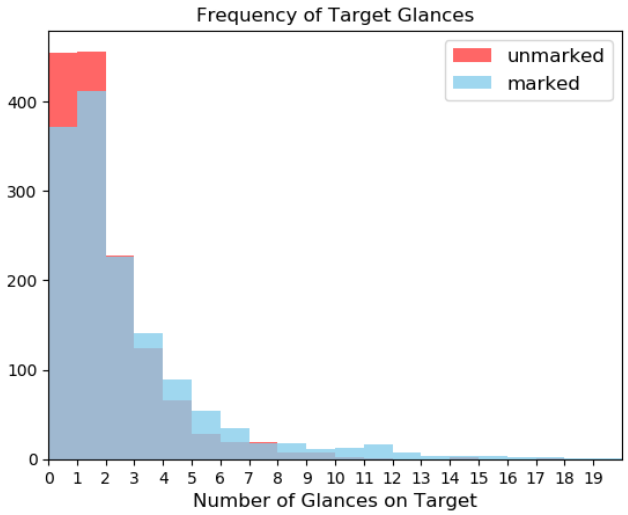


Figure 5: Frequency of glances per target for marked and unmarked targets for all images as depicted by the overlapping distributions for marked targets (blue bars) and unmarked targets (red bars). The frequencies when number of glances per target is 3 or more is overall higher for when the target was 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

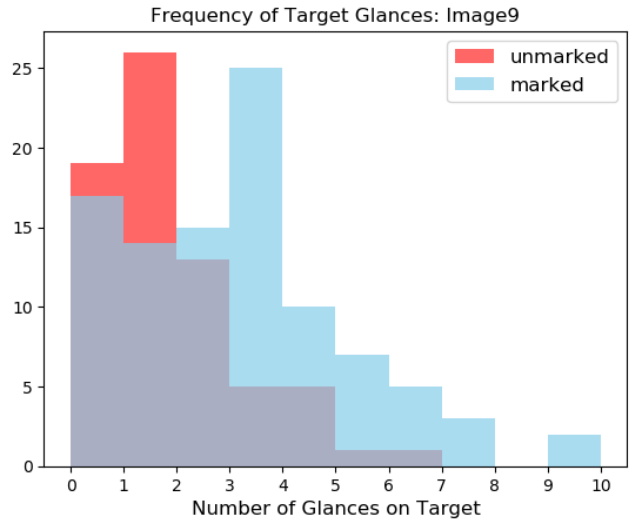


Figure 6: Histogram of number of glances per target for when the target was marked or missed. For this image in particular, the number of glances on a target was higher when the target was marked in contrast to when the target was not marked.

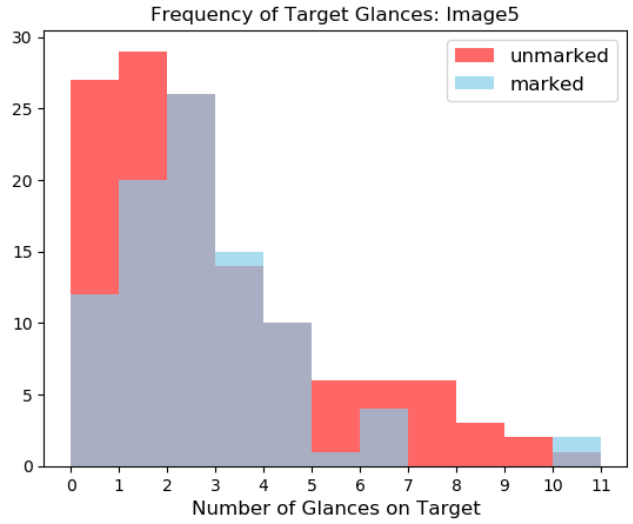


Figure 7: Example of an image where there is a high amount of false negatives in marking although there is a high frequency of higher glances per target.

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.

PLOS ONE

RESEARCH ARTICLE

Pupil diameter differentiates expertise in dental radiography visual search

Nora Castner[1]*, **Tobias Appel**[1☯], **Thérése Eder**[2‡], **Juliane Richter**[2‡], **Katharina Scheiter**[2,3‡], **Constanze Keutel**[4‡], **Fabian Hüttig**[5‡], **Andrew Duchowski**[6☯], **Enkelejda Kasneci**[1☯]

1 Human-Computer Interaction, Institute of Computer Science, University Tübingen, Tübingen, Germany, **2** Multiple Representations Lab, Leibniz-Institut für Wissensmedien, Tübingen, Germany, **3** University Tübingen, Tübingen, Germany, **4** Department of Oral- and Maxillofacial Radiology, University Clinic for Dentistry, Oral Medicine, and Maxillofacial Surgery, University of Tübingen, Tübingen, Germany, **5** Department of Prosthodontics, University Clinic for Dentistry, Oral Medicine, and Maxillofacial Surgery, University of Tübingen, Tübingen, Germany, **6** Visual Computing, Clemson University, Clemson, South Carolina, United States of America

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* nora.castner@uni-tuebingen.de

Citation: Castner N, Appel T, Eder T, Richter J, Scheiter K, Keutel C, et al. (2020) Pupil diameter differentiates expertise in dental radiography visual search. PLoS ONE 15(5): e0223941. https://doi.org/10.1371/journal.pone.0223941

Editor: Susana Martinez-Conde, State University of New York Downstate Medical Center, UNITED STATES

Received: September 30, 2019

Accepted: May 13, 2020

Published: May 29, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0223941

Data Availability Statement: The data and analysis scripts from the presented study are publicly available at: ftp://peg-public:peg-public@messor.

Abstract

Expert behavior is characterized by rapid information processing abilities, dependent on more structured schemata in long-term memory designated for their domain-specific tasks. From this understanding, expertise can effectively reduce cognitive load on a domain-specific task. However, certain tasks could still evoke different gradations of load even for an expert, e.g., when having to detect subtle anomalies in dental radiographs. Our aim was to measure pupil diameter response to anomalies of varying levels of difficulty in expert and student dentists' visual examination of panoramic radiographs. We found that students' pupil diameter dilated significantly from baseline compared to experts, but anomaly difficulty had no effect on pupillary response. In contrast, experts' pupil diameter responded to varying levels of anomaly difficulty, where more difficult anomalies evoked greater pupil dilation from baseline. Experts thus showed proportional pupillary response indicative of increasing cognitive load with increasingly difficult anomalies, whereas students showed pupillary response indicative of higher cognitive load for all anomalies when compared to experts.

Introduction

Visual inspection is a commonly performed task in many contemporary professions, e.g. radiologists and other medical personnel frequently examine medical radiographs to diagnose and treat patients, airport security scan X-rays of luggage for prohibited items, etc. [1, 2]. In such tasks, expert visual inspection is derived from domain knowledge and is optimized for a short period of search. Thus, understanding the search process and measuring mental workload are fundamental in expert research towards developing computer-based metrics. Generally, visual performance, e.g. during search, has been characterized by metrics derived from the

informatik.uni-tuebingen.de/peg-public/norac/VisualExpertiseRadiology.zip.

Funding: The student study is funded by the WissenschaftsCampus "Cognitive Interfaces" Tübingen (Principle Investigators: KS, CK and FH). The expert study with specialists and part of the data evaluation runs on budget of the University Hospital Tübingen / Department of Prosthodontics (Eberhard Karls University). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

discrimination of fixations and saccades. Fixations are the period when eye movements are relatively still, indicating focus of attention, usually on areas prone to a specific goal [3, 4]. Saccades, the rapid eye movements, are usually made when scanning over irrelevant areas to a specific goal [5].

Of particular interest is estimation of cognitive load during visual search used in demanding real-world tasks. Images with complex features can affect performance, especially in visual search, and so selection of measurement techniques to assess human performance is paramount [6]. One especially important factor in performance is workload, where feature complexity has a measurable effect. This research focuses on the objective, non-invasive, physiological measure of cognitive load [7] via eye tracking. Consequently, we expect that cognitive load measures will manifest significant responses during the decision-making aspect of the visual search task.

We examined the differences between expert and novice inspectors of dental panoramic radiographs. Orthopantomograms (OPTs), which are information-dense 2D superimpositions of the maxillomandibular region and used frequently in all aspects of dental medicine [8]. Due to their heavy reliance on OPTs, dentists undergo professional training and licensing; however, they are still highly susceptible to under-detections and missed information [9–13]. Coupled with concern for patients' health, accurate interpretation in spite of complex imagery is crucial. Specifically, OPTs have been shown to be less sensitive imagery for certain anomaly types than intraoral (periapical) radiographs, making correct detection more difficult [14, 15]. Therefore, less sensitive imagery of an anomaly can evoke higher gradation of difficulty for its accurate interpretation. Further understanding of both expert and novice OPT examination is necessary to effectively improve the training of medical image interpretation. Previous research has only scratched the surface of the cognitive processes during visual inspection of radiological images and the dichotomy between experts and novices. For this reason, our work goes one step further by examining the adaptability of cognitive processes during visual inspection of multiple features in decision making.

Background: Characterizing expertise

Expertise lies in the mind. The theory that expert aptitude develops a more structured long-term memory designated for domain-specific tasks [16] offers insight into experts' faster and more accurate abilities [5]. *Long-term working memory*, proposed by Ericsson and Kintsch [16], offers this explanation for how experts seemingly effortlessly handle their domain-specific tasks. Their memory structuring facilitates their ability to maintain working memory at optimal capacity, avoiding overload, which affects productivity and performance.

Generally, working memory is understood as temporary storage for processing readily available information [17]. Long-term working memory relates to the structuring available to the larger, long-lasting storage and is of interest in skill learning [16]. For instance, chess players employ memory chunking that enables them to quickly recognize favorable positions and movements with less focus on single pieces [18]. Athletes show faster reaction to attentional cues, especially in interceptive sports, (e.g. basketball), indicating more rapid mental processing [19]. Also, medical professionals have been thought to proficiently employ heuristics in their decision-making strategies, i.e. visual search of radiographs [20] and diagnostic reasoning in case examinations [21, 22].

Developing new skills and the related memory structures for a specific discipline rely heavily on the capacity of working memory. According to Just and Carpenter [23], when the working memory demands exceed available capacity, comprehension is inhibited, leading to negative effects on performance. Effective comprehension then relies on *resource allocation*

[23]. Optimal resource allocation supports rapid convergence to the most appropriate task-solution. Experts can filter out irrelevant information, which is evident in gaze behavior; they focus more on areas relevant to the task solution and less on areas that are irrelevant [5, 24, 25]. For instance, expert radiologists devote more fixations to anomaly-prone areas [26, 27] and devote shorter fixation time to an anomaly in detection tasks [20, 28]. Dental students' gaze behavior has also been shown to be an effective feature to classify level of conceptual knowledge [29].

Additionally, when the task becomes too difficult or is perceived as such, there is more demand on working memory [30]. Sweller points out that the means-to-an-end problem solving strategies that novices employ can overload working memory [31]. And though perceived task-difficulty is influenced by acquired knowledge [32], even experts can face challenging problems that could evoke more load on working memory [33, 34]. Cognitive load, or more specifically intrinsic cognitive load [35], is the effect of "heavy use of limited cognitive-processing capability" [31]. For more information, see review by Paas and Ayres [36]. High cognitive load has been shown to have negative effects on performance [30] and effective learning in general [37].

One way to assess levels of cognitive load is the pupillary response [38–40], where pupil size has been shown to increase as a response to memory capacity limits [41, 42] as well as when the task becomes too difficult [37, 43]. Accordingly, experts have a higher threshold for what is difficult compared to their novice counterparts, which is evident in the pupil response. Therefore, we are interested in expert and novice dentists when interpreting anomalies of varying degree of difficulty in panoramic radiographs. More important, our aim is to further understand experts' perception of difficulty in their domain-specific tasks and whether this affects cognitive load.

Pupil diameter as a measure of cognitive load

Not only does visual search strategy reflect cognitive processes [44–46], but pupil diameter has also been shown to be a robust, non-invasive measurement of cognitive load [37–39, 41–43, 47–52]. Hence, with an increase in task difficulty, the diameter increases, otherwise known as task-evoked pupillary response. Originally, Kahneman and Beatty [47] linked pupil response to attentional differences. Then, the link between attention and capacity was promoted [43]; where higher load on the working memory showed a larger change in pupil dilation. Additionally, pupillary response has been found to be an indicator of learning [37], where pupil diameter decreased with more experience in a task.

Much of the early research in processing capacity and cognitive load has found that pupil activity correlates to workload during a variety of tasks [41–43, 53]. Specifically for visual search tasks, cognitive load has also been measured by pupil activity. For instance, more distractors make the paradigm more difficult, affecting the pupil diameter increase [54]. Also, monochrome displays evoked longer search time and more pupil dilation than colored displays for both object counting and target finding tasks [55]. Regarding uncertainty, an increase in pupil diameter was associated with response time and uncertainty of target selection [56]. One of the more important takeaways from the visual search literature is the interplay of selective attention, increasing task demand, and the mental effort evoked. Moreover, this interplay is apparent in medical professionals and their diagnostic interpretation of radiographs. Students may not be as exposed to such tasks of varying difficulties, but accumulate more experiences overtime, which can reduce cognitive load. Regarding learning, pupil dilation decreases as an effect of training over time [57].

Though it is apparent that pupillary response is a product of cognitive load, other factors have been shown to effect pupil size, e.g. fatigue [58, 59], caffeine consumption [60], etc. [59, 61]. Most important to this work is changes in luminance in the environment, which result in the physiological response of constriction or dilation [52]. Age difference has also been shown to affect pupil size differences, where overall pupil size in older adults is smaller than younger adults, though variance between subjects in similar age groups is also quite high [48, 52]. With these factors in mind, studies on pupil diameter and load recommend a task-to-baseline comparison in luminance-controlled environments [37–39, 41–43, 47, 50, 54, 56, 62, 63]. Therefore, when measuring pupillary response in relation to cognitive load, these factors should be controlled in order to avoid such confounds.

Previous research

Only a few studies have comprehensively addressed cognitive load and medical expertise, and even fewer have addressed cognitive load during visual search. Trained physicians showed more accurate performance and smaller pupillary response during clinical multiple-choice questions compared to novices, and this effect was larger for more difficult questions [50]. Expert surgeons' pupil diameter increased as a result of increasing task difficulty during laparoscopic procedures [64]. Additionally, Tien et al. [65] found that junior surgeons exhibited larger pupil sizes than experts during a surgical procedure. More important, they found that specific tasks affected junior surgeons' pupillary response to a higher degree. For more references highlighting lower pupillary response as an effect of medical expertise (e.g. surgeons, anesthesiologists, physicians), see Szulewski et al. [66].

Regarding specifically medical image interpretation, Brunyé and colleagues [49] found pupil diameter increases as an effect of difficulty in diagnostic decision making, more so for cases that were accurately diagnosed. They further highlight the prospects that pupillary response in combination with gaze behavior has in understanding uncertainty in medical decision making [67]. Specifically for dental expertise and OPT interpretation, experts' gaze behavior (e.g. fixations) was highly distinguishing of difficult and obvious images, where students' gaze behavior was not [68, 69]. Castner et al. [13] found that fixation behavior changed with respect to differing anomalies. Therefore, the degree of difficulty in accurate pathology detection can affect gaze behavior, which can be indicative of the reasoning strategies used.

With this intention in mind, we looked at expert and novice dentists' pupillary response while fixating on anomalies of varying difficulty in panoramic radiographs. To our knowledge, we are the first to apply differentiable pupillometry to the dental imagery visual search domain. Not only do these OPTs have multiple anomalies, but also within one OPT, varying difficulties can be present. Therefore, we are not analyzing an overall impression of easy or difficult image. Rather, through the course of the search strategy, we are extracting when dentists spot an anomaly and consequently mental processing at that moment. We propose the degree of anomaly interpretation difficulty can be indicated by changes in the pupillary response; where a larger response is more representative of harder to interpret anomalies. We also hypothesize to find a difference in the pupillary response between experts and novices, as established by prior research; where baseline-related pupil difference, as a measure of cognitive load, is sensitive to experts' processing of anomalies of varying degree of difficulty. Additionally, we report that students, after acquiring the appropriate training to inspect OPTs, have higher cognitive load compared to experts. More interesting is whether students are attuned to the varying gradations of the anomalies.

Materials and methods

Participants

Data collection took place in the context of a larger project performed over multiple semesters from 2017 to 2019. Dentistry students from semesters six through ten were recorded during an OPT inspection task. For reference, sixth semester students are in the second half of their third year and the tenth semester is in the fifth year of their studies, being the last semester before they continue on to the equivalent of a residency.

The sixth semester students were evaluated three times in each period of data collection due to their curriculum requirement of an OPT interpretation training course. For the purpose of the present paper, we chose to only evaluate the sixth semester students after this course (N_{sixth} = 50). They have the necessary knowledge to perform the OPT task as it is intended (i.e. they know what they have to look for), without having yet acquired the routine skills.

Table 1 details both the student and expert data. Experts ($N_{experts}$ = 28) from the University clinic volunteered their expertise for the same task that students performed. Experience was defined as professional years working as a dentist and ranged from 1 to 43 years (M_{years} = 9.88). 50% of experts reported seeing between 11 and 30 patients on a typical work day and the remainder saw less than 10 patients a day. All experts had the necessary qualifications to practice dentistry and or any other dental related specialty: e.g. Prosthodontics, Orthodontics, Endodontics, etc. Due to technical difficulties, eye tracking data was lost for two participants, leaving $N_{experts}$ = 26 participants for the eye tracking analysis.

The Ethical Review Board of the Leibniz-Institut für Wissensmedien Tübingen approved the student cohort of the study with the project number LEK 2017/016. All participants were informed in written form and consented in written form that their pseudonymous data can be analyzed and published. Due to a self-constructed pseudonym, they had the option to revoke this consent until the date of anonymization of the data after data collection is finished. The Independent Ethics Committee of the Medical Faculty and University Hospital Tübingen approved the expert cohort of the study with the project number 394/2017BO2. All participants were informed in written form and consented verbally that their anonymous data can be analyzed and published. Due to a self-constructed pseudonym, they had the option to revoke this consent at any time.

Experimental paradigm

The experimental protocol for the students consisted of an initial calibration, task instruction, then two image phases: Interpretation and Marking. The details of the experimental protocol are found in Fig 1. Prior to the interpretation, a two second fixation cross was presented: This served as baseline for our analysis. Then, an OPT was presented in the interpretation phase for

Table 1. Participant data overview.

	Students	Experts
N	50	26
$N_{glasses}$	12*	9
OPTs viewed/person	20	15
Total Datasets	750	390
Poor Tracking Ratio**	14.3%	14.3%

* data regarding glasses for one collection is unknown

** Percentage of poor data quality. Proportion of valid gaze points less than 80%.

https://doi.org/10.1371/journal.pone.0223941.t001

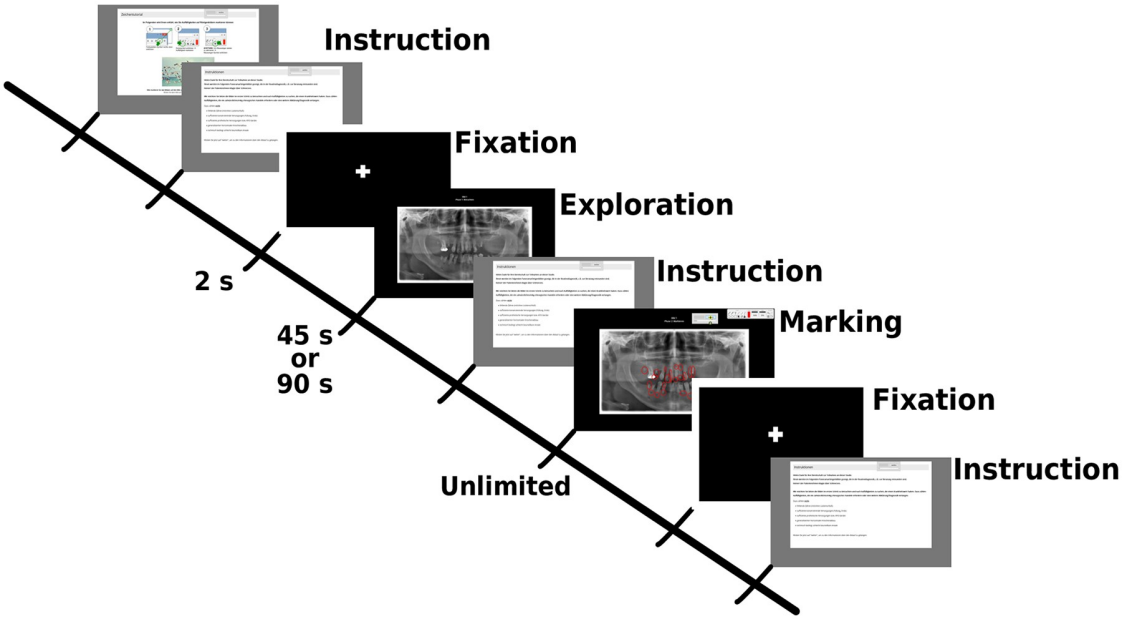


Fig 1. Outline of experimental session. Initially, there was a calibration and procedural instructions. Then for each image, there is a fixation cross for baseline data, the exploration phase (45s duration for experts and 90s for students), instructions for the marking phase, and the marking phase (unlimited time). Students received two sets of 10 OPTs with a break in between and experts received one set of 15 OPTs with a break after the first seven.

https://doi.org/10.1371/journal.pone.0223941.g001

90 seconds and the participant was instructed to only search for areas indicative of any pathologies in need of further intervention. The marking phase came after interpretation; where the same OPT was shown with the instruction to only mark the anomalies found in the interpretation phase using an on-screen drawing tool. There was unlimited time for the marking phase and participants could continue with a button click. This procedure was repeated for all OPTs. In total, the students viewed 20 OPTs with a short break after the first ten.

The diagnostic task for the expert group was highly similar to that of the students. However, it was determined that 90 seconds is too long of a duration for the experts, since much of the previous literature has shown experts are faster at scanning radiographs [5, 20, 26, 27, 68, 70–72]. Therefore, the exploration phase was shortened to a duration of 45 seconds. Additionally, due their busy schedules, experts only viewed 15 OPTs, with a short pause after the first seven.

Both students and experts were unrestrained during the experiment, although they were instructed to move their head as little as possible. Further details of one of the student data collections can be found in Castner et al. [29] and expert data collections can be found in Castner et al. [13].

Stimuli

OPT images. The 15 OPTs viewed by both the experts and the post-training course sixth semester students were used for the current analysis to avoid effects from unseen images. The OPTs were chosen from the university clinic database by the two expert dentists involved in this research project and were determined to have no artifacts and technological errors. Both dentists independently examined the OPTs and the patient workups and further consolidated together to determine ground truths for each image. Two OPTS were negative (no anomalies) controls.

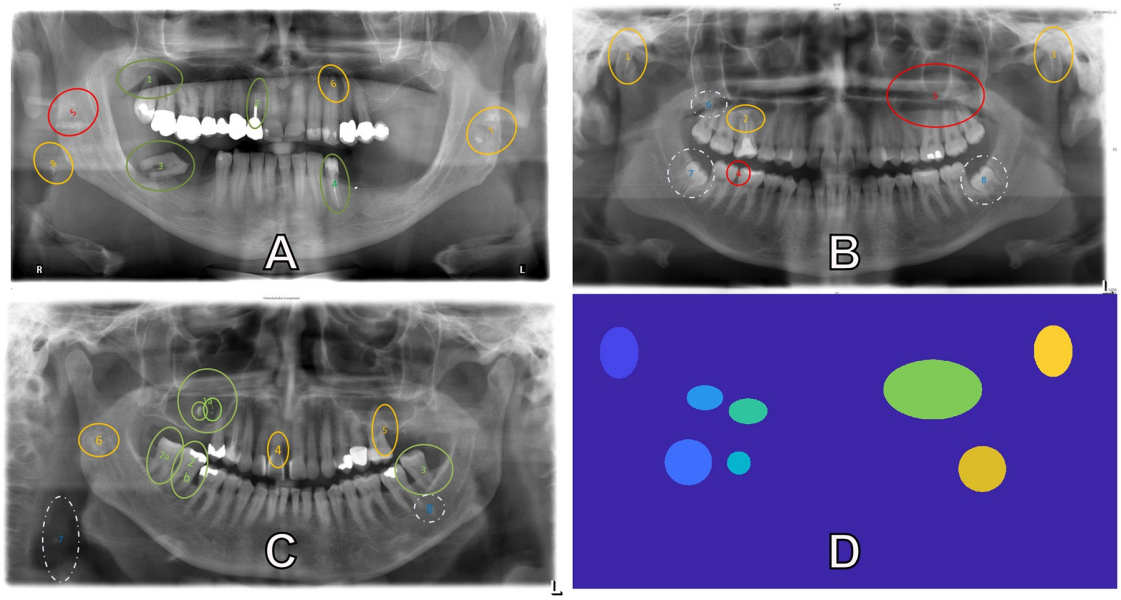


Fig 2. OPTs with pre-determined ground truth. Example of the OPTs used in the experiment. Pre-determined ground truths are indicated by the ellipses and their colors indicate the level of difficulty each anomaly is: Green (least difficult), yellow (intermediary), red (most difficult) and white (nature of difficulty unclear). Image (D) is the ground truth map for image (B). Each anomaly is segmented and given a distinguishing integer.

https://doi.org/10.1371/journal.pone.0223941.g002

Additionally, the level of difficulty for each anomaly was pre-determined. Fig 2 shows three OPT images viewed in the experiment. Anomalies are illustrated in green, yellow, and red, and represent easy, medium, and difficult, respectively. This classification was set up in a blinded review and the consent process of two senior dentists (6th and 7th authors). For example, the green anomalies in Fig 2A are dental cyst (1) and insufficient root canal fillings. (2a,b) in Fig 2C are an example of elongated lower molars due to missing antagonists. The yellow anomalies in Fig 2B are irregular forms of the mandibular condyle (1,3) and (2) is an apical translucency indicative of inflammation due to a contagious (bacterially colonized) root canal filling. The red anomalies in this image are approximal caries (4) and a maxillary sinus mass. Anomalies indicated by the white dashed circles were determined as ambiguous, e.g. the nature of their difficulty and or pathology is unclear. For example, in Fig 2B (7,8) are impacted wisdom teeth, though it is uncertain whether this will become a problem for the patient and therefore is regarded as potentially pathologic. (6) is an apical translucency at the mesial root apex and it is unclear whether it is indicative of an inflammation. Therefore, they were kept in this analysis even though the nature of their difficulty is unclear.

Ground truth maps. We created maps for the 15 OPTs evaluated (See Fig 2C) using Matlab 2018. As input, all OPTs were loaded as .png files with their respective anomalies—all colored red. Thresholding for red values was performed to automatically get the pixel coordinates of the ellipse edges. Then, the ellipses were filled with the `poly2mask()` function. Anomalies automatically extracted from this process were double checked for overlapping and had their boundaries corrected. Similar anomalies inside of another, such as (2a,b) in Fig 2C, were grouped together as one anomaly. Other anomalies too close together and too different in pathology, such as (3,8) in Fig 2C, were excluded from the analysis, due to possible spatial accuracy errors in the gaze. Similarly, anomalies that were denoted by too small of an ellipse were padded to have a larger pixel area, e.g. (4) in Fig 2B, to account for the spatial accuracy

errors in the gaze. Each segmented anomaly is given a distinguishing integer for its respective pixels. Raw gaze points from the left eye are then mapped to the map and gaze coordinates receive the corresponding integer value.

Data acquisition

Environment. Data collection for students took place in a digital classroom equipped with 30 remote eye trackers attached to laptops with 17inch HD display screens running at full brightness. This setup allows for data collection of up to 30 participants simultaneously, minimizing the overall time needed for collection. For this study, verbal instructions were given en masse pertaining to a brief overview of the protocol and an explanation of eye tracking, then individual calibrations were performed with a supervised quality check; students could then run the experiment self-paced.

Data collection for the experts took place in the university hospital so the experts could conveniently participate during work hours. There, the room used for data collection was dedicated for radiograph reading. The same model remote eye tracker was used for expert data collection and was run with the same sampling frequency on a laptop with 17inch HD display screen running at full brightness.

More important to the current study, both data collection environments had the room illumination levels controlled with no effects from sunlight or other outdoor light. The standard maintained illuminance for experimental sessions was between 10 to 50 lux, measured with a lux sensor (Gossen Mavo-Max illuminance sensor, MC Technologies, Hannover, Germany). It is advised that environment illumination during radiograph reading should be ambient (25–50 lux) for the best viewing practices [73] and to optimize contrast perception in radiographs [74–76]. Therefore, with room illumination controlled, we can evaluate pupillary response independent of environmental illumination changes.

Laptops. Regarding the screen display, radiograph reading is not affected by the luminance of the display [75]. However, both the laptop models used for the experimental sessions abided by the multiple medical and radiology commission standards [72, 73, 77]. The HP Z Book 15 (for students) has screen brightness averages approx. $300 cd/m^2$ [78]. The Dell Precision m4800 (for experts) averages approx. $380 cd/m^2$ [79]. While the screen luminance was also controlled and followed the standard protocols for viewing radiographs, the exact effect of the screen brightness on the pupillary response is out of the scope of this work; rather the pupillary response dependent on mental load during these reading task is the focus.

Eye tracker. The SMI RED250 remote eye tracker is a commercial eye tracker with 250Hz sampling frequency and used for gaze data collection. We used the included software for both the experiment design (*Experiment Center*) and event analysis (*BeGaze*). Since the eye tracker has a high sampling frequency, both stable (fixations) and rapid (saccadic) eye movements for static stimuli can be measured. Analysis was performed on the raw gaze data output from the eye tracker: x and y coordinates with timestamps mapped to the screen dimensions. The raw data points also have pupil diameter output in millimeters [80]. Although the data is raw and has not been run through event detection algorithms, raw gaze points are labeled as fixation, saccade, or blink.

Calibration was performed for all participants. A validation also was performed as a quality check to measure the gaze deviation for both eyes from a calibration point: A deviation larger than one degree constituted recalibration. Calibrations were performed prior to the experiments as well as one or two times during the experimental session, depending on how many images were presented.

Data preprocessing

Quality of raw data. Only gaze data from the exploration phase was of interest to this work since gaze data from the marking phase was affected by the use of the screen drawing-tool. Initially, the raw gaze data was examined for signal quality. The eye tracker reports proportion of valid gaze signal to stimulus time as the tracking ratio. Therefore, if a participant's tracking ratio for an OPT was deemed insufficient—less than 80%—we omitted his or her data for this OPT. If overall, a participant had poor tracking ratios for more than three of OPTs he or she viewed, all gaze data for that participant was removed. This preprocessing stage can assure that errors (e.g. post-calibration shifts, poor signal due to glasses) in the gaze data are substantially minimized. Table 1 gives the distribution of participants and the percent of datasets excluded due to low tracking ratio (last row). We started with 1140 data sets, but 199 datasets were initially excluded on the grounds of poor data quality.

Blink removal. The SMI-reported tracking ratio does not take into account when the eye tracker detects a blink [80]. Nevertheless, inaccurately detected blinks created an alarming number of cases with acceptable tracking ratios even though there was an inordinate amount of undetected gaze. Fig 3a shows an example of a participant's pupil size samples over time for the left and right eye for an OPT presentation. This participant had a reported tracking ratio of 98%, but a large portion of the left eye gaze signal– approximately 33.5 seconds out of 90 seconds—could be signal loss labeled as a blink. In contrast, Fig 3b shows a participant who also has a high tracking ratio, though the data appears to be acceptable with typical blink durations detected and little signal loss.

Consequently, the main issue stems from the apparent lack of a maximum blink duration threshold. Extra criteria were necessary to further detect and exclude datasets with pupil signal loss mislabeled as a blink. We overestimated the threshold for atypical blink durations, setting this value to 5000 ms, to account for situations where a participant could possibly be rubbing his or her eye/s or even closing the eye shortly. This threshold optimally maintains an acceptable amount of pupil data for the entire stimulus presentation (90 or 45 seconds). Since baseline data was sampled during the two seconds the fixation cross was displayed, we set the threshold blink duration to 500 ms and added an extra criterion of a minimum 200 pupil samples to effectively extract enough samples for an acceptable pupil diameter baseline. Therefore, 570 datasets from 72 participants (48 students, 24 experts) were used for the final analysis.

Pupil diameter measurement. Data analysis was done for the left eye. For further signal processing, we removed gaze coordinates and pupil data for the raw data points labeled as

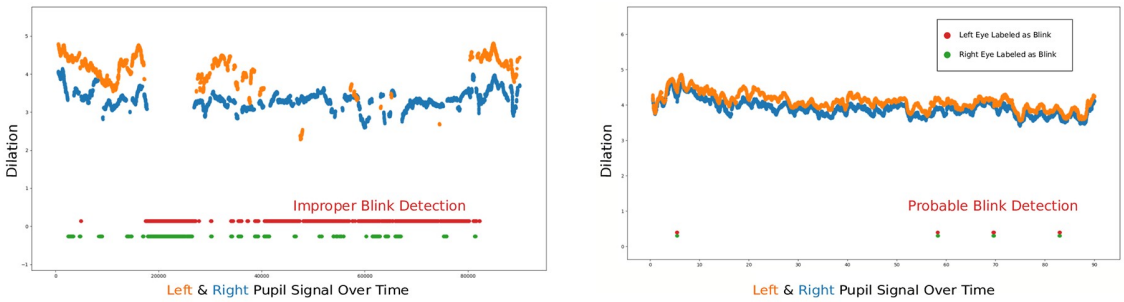


Fig 3. Blink detection in the raw gaze data. (**a**) Low Data Quality Example (**b**) High Data Quality Example. The raw pupil signal of the left and right eye (orange and blue dots) over the course of image presentation. Red and green dots in the lower part show when the eye tracker labels the data point as a blink for the left and right eye, respectively. The particular subject in 3a had a high tracking ratio, though many data samples could be incorrectly labeled as blinks. The participant in 3b also has a high tracking ratio and his or her data appears to be acceptable quality.

https://doi.org/10.1371/journal.pone.0223941.g003

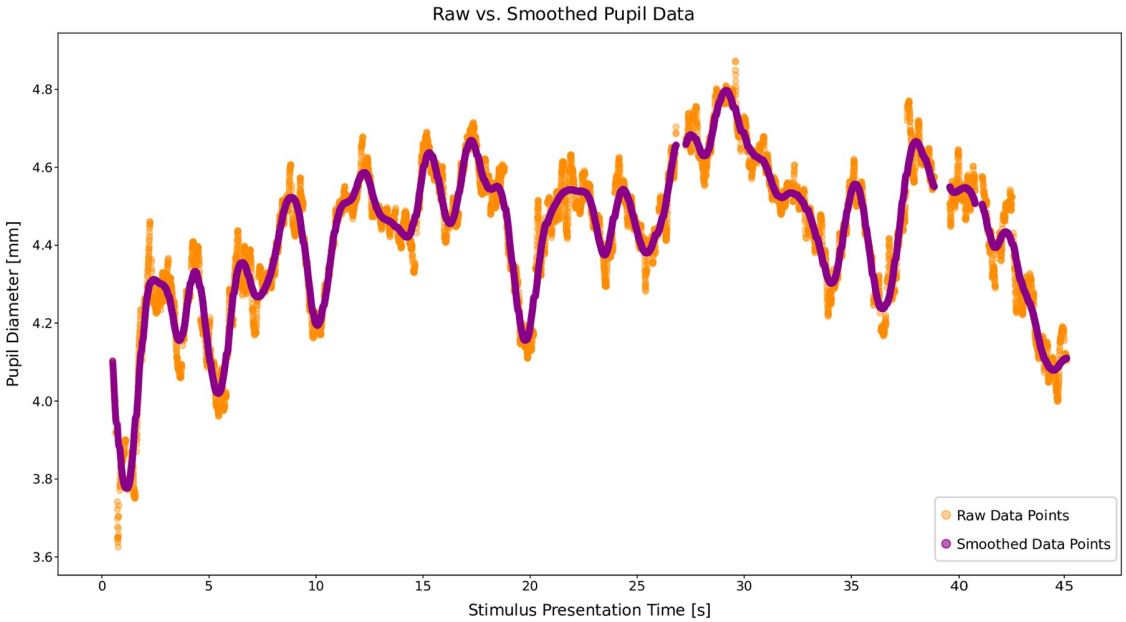


Fig 4. Smoothed pupil signal. Raw signal from the left eye (orange) and the smoothed signal (purple) with a Butterworth filter with 2Hz cutoff.

https://doi.org/10.1371/journal.pone.0223941.g004

saccades (since visual input is not perceived during rapid eye movements [3]). Data points with a pupil diameter of zero or labeled as a blink were also removed. Additionally, data points 100 ms before and after blinks were removed, due to pupil size distortions from partial eye-lid occlusion. Lastly, the first and last 125 data points in the stimulus presentation were removed due to stimulus flickering [81–83]. The remaining data was smoothed with a third order low-pass Butterworth filter with a 2Hz cutoff as illustrated by the purple data points in Fig 4.

Gaze hit mapping. For both students and experts, we plotted the raw gaze points that landed in each anomaly and extracted its level of difficulty. For simplicity, we will refer to them as gaze hits. For all hits on an anomaly for a participant, we calculated the median pupil diameter. The median pupil diameter for each anomaly was then subtracted from the respective baseline data for that image. We performed subtractive baseline correction because it has been found to be a more robust metric and have higher statistical power [63]. Therefore, the difference from baseline could indicate diameter increase (positive value) or diameter decrease (negative value) compared to baseline.

With the gaze hits on anomalies of varying difficulties, we can evaluate the pupillary response of both experts and students during anomaly fixations. The pupillary response, as measured by change from baseline, can then provide insight into the mental/cognitive load both groups are undergoing while interpreting the anomalies.

Results

Overall change from baseline

Independent of gaze on anomaly difficulty, we looked at participants' median pupil diameter for each image compared to baseline median pupil diameters. We favored the median over the mean because it has greater robustness towards noise and outliers. Fig 5a shows the average of the median pupillary response from baseline for both students and experts. Overall, students

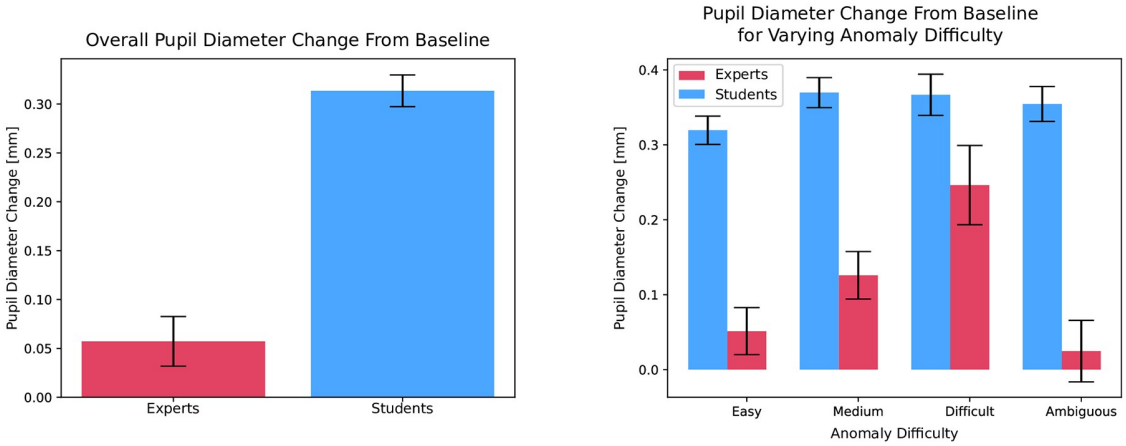


Fig 5. Pupillary response of experts and novices during visual Inspection. (**a**) Median Pupil Change From Baseline for Experts and Novices. (**b**) Median Pupil Change From Baseline for Gaze on Anomalies. The median pupil diameter change from baseline for students (blue bars) and experts (red bars) for the overall image behavior (5a) and when gazing on anomalies of varying difficulty (5b). Standard errors are indicated in black. Students had larger pupillary response from baseline compared to experts, but this effect was homogeneous for the differing anomalies. Whereas experts showed an increased pupillary response behavior as an effect of increasing difficulty.

https://doi.org/10.1371/journal.pone.0223941.g005

(M = 0.314, SD = 0.315) had a larger increase from baseline than experts (M = 0.057, SD = 0.353: $t(568) = -8.824$, $p < 0.001$). We also performed a supplementary analysis to rule out any effects that fatigue could have on the pupillary response (see S1 Fig).

Gaze on anomalies

To evaluate whether anomaly difficulty had an effect of student and expert pupillary response, we ran a 2 × 4 factor ANOVA to test for expertise and anomaly difficulty interactions. There was a main effect for expertise ($F(1, 1388) = 161.68$, $p < 0.001$) indicating that students had a larger increase from baseline than experts. There was also an effect for anomaly difficulty ($F(3, 1388) = 3.87$, $p = 0.009$) indicating that there was a larger increase in pupil size from baseline for more difficult anomalies. There was a significant interaction between expertise and anomaly difficulty ($F(3, 1388) = 2.76$, $p = 0.041$). There were no significant effects of anomaly difficulty on student pupillary response. However, there were significant effects of anomaly difficulty on expert pupillary response. Fig 5b details the pupillary response of experts and novices on the varying anomaly difficulties.

Post hoc analyses with Bonferroni correction for anomaly difficulty on the expert data revealed significant differences for the more difficult anomalies (M = 0.246, SD = 0.370) compared to least difficult (M = 0.0514, SD = 0.396, $t(207) = -3.0582$, $p = 0.003$) and ambiguous ($t(150) = 3.1796$, $p = 0.002$). There were no significant differences for medium anomalies (M = 0.1259, $SD = 0.3904$) compared to the difficult ($t(200) = 1.8989$, $p = 0.059$). Meaning, experts had the largest pupil size change from baseline for more difficult anomalies, especially compared to least difficult and ambiguous anomalies.

Discussion

Students showed larger and more homogenous pupil size change from baseline for all anomaly gradations compared to experts. Thus for students, pupillary response was independent of whether an anomaly was easy or difficult to interpret. This effect was also found during visual inspection of the whole image (Fig 5a), where students had overall greater change from

baseline compared to experts. Pupillary response differences between students and experts have been supported by the previous literature [49, 50, 65–67, 84]. However, the more interesting takeaway from this work is the lack of influence of anomaly gradation on student cognitive processing. One would imagine that even the most pronounced of anomalies would make the recognition process easier. Our findings from student pupillary response indicate that, regardless of how conspicuous, the level of mental workload remains constant.

Conversely, experts showed a strong pupillary response to anomaly gradation. The least difficult to interpret anomalies showed less change from baseline, then the intermediary anomalies, and finally the largest response was for the most difficult anomalies (Fig 5b). Meaning, as the gradation of difficulty increases so does the pupillary response. This behavior, however, was not evident for the ambiguous anomalies, which showed the smallest response change from baseline. This effect may lie in the nature of the uncertainty of these anomalies. As determined by the two experts involved in the project, this category was a mixture of potential areas that may or may not have included an anomaly: Or even an anomaly, but with no cause for alarm. Therefore, it is uncertain how difficult, easy, or even existing these anomalies were.

Cognitive load is often used to explain findings regarding learning [23, 31, 36, 62]. For instance, Tien et al. [65] found that novices reported higher memory load compared to experts performing the same task. This behavior can be likened to students' lack of conceptual knowledge and experience, producing them to "think harder" [85, 86] to interpret these images. Furthermore, large pupil size can be reflective of learning during the task [23, 37, 41, 43, 47, 82]. During learning, students are developing the proper memory structures as theorized by Ericsson and Kintsch [16] and Sweller [31]. Additionally, their pupillary response could reflect that they have not yet developed the conceptual knowledge to quickly recognize the image features indicative of the specific anomalies or how to interpret their underlying pathologies. Even for easy anomalies, they may be unsure of whether they accurately interpreted it or not. Uncertainty as well as perceived task difficulty have been found to affect the pupillary response, and acquired knowledge has been shown to reduce uncertainty and perceived difficulty [32, 56]. Moreover, prior knowledge to a problem has been shown to reduce cognitive load [31, 36, 41, 50].

Cognitive load can also be indicative of inefficient reasoning strategies. Efficient reasoning strategies reduce load on working memory, in turn enhancing performance [30]. Patel et al. [33] found that when novices interpreted clinical case examinations, they tended to employ reasoning strategies that have been known to elicit higher workload. Our findings also suggest that students may employ similar cognitive strategies that evoke higher load for all anomaly gradations. Comparatively, experts employ more efficient strategies; however, they are more sensitive to task features.

In general, as task difficulty increases, so does the workload [64] and correspondingly, the pupil dilation [30, 43, 87, 88]. With increasingly difficult stimuli, Duchowski et al. [89] also showed increased cognitive load via microsaccade rates during decision making. However, Patel et al. [34] found more cognitive load in physicians when examining more complicated case examinations. When expert dentists perform a visual inspection of an OPT, they gaze in many areas that potentially have a multitude of differing pathologies or even positional and summation errors. Depending on the gradation of the area they are focusing on, proper interpretation may need to evoke adaptations in the decision-making strategies. Our findings show that experts dentists are capable of this adaptability during the course of visual inspection of OPTs.

Gaze behavior in expert dentists was also shown to change with difficult images [13, 68]. The current work went one step further and found changes within the visual search of an OPT in contrast to the overall response to image interpretation. Kok et al. [46] found that expertise

reflected visual search strategies employed. *Top-down* strategies that experts generally employ use acquired knowledge and understanding of the current problem to focus on the relevant aspects of an image to quickly and more accurately process it [24, 90, 91]. Whereas *bottom-up* strategies that student generally employ is less efficient, as focus is on salient, noticeable images features, regardless of relevancy [20, 46, 91]. Furthermore, systematic search (inspecting all features of an image in a pre-determined orders) evokes more load on the working memory [20, 27]. However, students are generally trained to perform this type of search when they first get exposed to these images [72, 90].

An expert generally knows in what areas of the OPT anomalies are prevalent and how they are illustrated in the image features. Therefore, an expert can quickly recognize an image feature as a specific anomaly. In contrast to overall visual inspection—where experts showed low pupillary response compared to students—when inspecting specific areas, pupil dilation fluctuation can be indicative to changes in their cognitive processes to accommodate more complex features. Naturally, interpretation of medical images is not trivial and certain image or pathology features can avert the true diagnosis. Experts are more robust at determining more difficult or subtle anomalies [11, 27, 68, 72, 92]. Although when anomalies become harder to interpret, experts evoke pupillary response indicative of increasing task-difficulty, leading to behavior that is likely of more thorough inspection.

Limitations and future work

It should be noted that there were age differences between the two groups. Due to the sensitivity of the expert demographic data, we did not record their ages; but we expect them to be older than their student counterparts. Age has been found to have an effect on the average pupil size [48, 52]. For this reason, we measured a change from baseline to control such for age effects. Additionally, Van Gerven et al. [51] found that pupillary response to workload in older adults (early seventies) is not as pronounced as in younger adults (early twenties). Though we cannot say exactly how old our expert population was, they were all still working in the clinic and therefore more than likely to be younger than early seventies. Also, their years of experience in the clinic (average of 10 years) suggests they were more middle aged (30 to 45 years old). Further research is needed to better address this limitation and control for possible age difference effects on pupillary response.

Another limitation to this work could be the technical problems associated with the eye tracker data collection. We systematically removed data sets determined as poor quality; however, spatial resolution errors can accumulate within an experimental session if a participant moves too much. Then, the gaze appears to have a shifted offset, which would affect precision. Multiple calibrations during collection help with precision. We also increased the areas of smaller ground-truth anomalies and excluded anomalies that were too close and too different in nature. The total gaze hits on each type of anomaly were not evenly distributed, with more gaze hits on easier and intermediary anomalies (See S1 Table in Supporting Information). Students used more total gaze hits due to longer OPT presentation time, but the distributions were highly similar to experts. Future research could further untangle the differences in gaze hits on easier and difficult anomalies, while controlling for presentation time differences.

The temporal scanpath information is also an interesting direction for future research, i.e. systematic search in students and its effect on workload and pupillary response. For example, how often do "look backs" on anomaly areas occur and does the pupil dilation increase with each look back. Also, whether easy or more conspicuous anomalies are viewed at first and how the pupillary response in students incorporates this initial information. Following up on the

understanding that systematic search produces more memory load as measured by pupil dilation [93], would also be interesting to replicate with temporal information from our findings.

Conclusion

We measured pupil diameter change from baseline when gazing on anomalies of varying difficulty during visual search of dental panoramic radiographs. We found that the gradation of anomalies in these images had an effect on expert pupillary response. Anomaly gradation did not have an effect on student pupillary response, which suggests higher workload and less sensitivity to complex features compared to experts. Experts are able to selectively allocate their attention to relevant information and is evident in the pupillary response. However, selective attention coupled with focus on features perceived as challenging can increase the pupil dilation as we found in our investigation. Although a majority of expert studies have established that experts are more robust at accurately solving their domain-specific tasks than their student counterparts [5, 16, 24, 91], increased pupillary response during difficult anomaly inspection supports adaptable processing strategies.

With more insight into expert decision-making processes during visual search or medical images, appropriate learning interventions can be developed. These interventions can incorporate not only the scanpath behavior, but also the cognitive load during appropriate detection of pathologies. From this combination, image semantics can be better conveyed to the learner. Training sessions that convey the appropriate information through adaptive gaze interventions based on cognitive load detection via the pupillary response offers a promising direction in medical education.

Supporting information

S1 Fig. Pupillary response over course of experiment. The average pupillary response from baseline for students (blue bars, 20 images total) and experts (red bars, 15 images total) during the first set of OPTs presented and the second set of OPTs presented. Their is no effect in the pupillary response that could be attributed to fatigue during the experiment.
(PDF)

S1 Table. Table of Expert and Student Gaze Counts. Shows the gaze hits on each anomaly type for both students and experts. For both levels of expertise, the least difficult and intermediate have the most gaze hits. The following are the ambiguous and the most difficult anomalies. Students had overall more gaze hits than experts; however, this may be attributed to the 90 second viewing time they had in comparison to the 45 second viewing time that the experts had.
(PDF)

Author Contributions

Conceptualization: Katharina Scheiter, Constanze Keutel, Fabian Hüttig, Enkelejda Kasneci.

Data curation: Nora Castner, Thérése Eder, Juliane Richter.

Formal analysis: Tobias Appel.

Investigation: Nora Castner.

Supervision: Nora Castner, Tobias Appel, Thérése Eder, Juliane Richter, Katharina Scheiter, Constanze Keutel, Fabian Hüttig, Andrew Duchowski, Enkelejda Kasneci.

Writing – original draft: Nora Castner.

Writing – review & editing: Nora Castner, Tobias Appel, Thérése Eder, Juliane Richter, Katharina Scheiter, Constanze Keutel, Fabian Hüttig, Andrew Duchowski, Enkelejda Kasneci.

References

1. Qvarfordt P, Biehl JT, Golovchinsky G, Dunningan T. Understanding the Benefits of Gaze Enhanced Visual Search. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. ETRA'10. New York, NY: ACM; 2010. p. 283–290. Available from: http://doi.acm.org/10.1145/1743666.1743733.
2. Otto K, Castner N, Geisler D, Kasneci E. Development and evaluation of a gaze feedback system integrated into eyetrace. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications; 2018. p. 1–5.
3. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Van de Weijer J. Eye tracking: A comprehensive guide to methods and measures. OUP Oxford; 2011.
4. Tafaj E, Kasneci G, Rosenstiel W, Bogdan M. Bayesian online clustering of eye movement data. In: Proceedings of the symposium on eye tracking research and applications; 2012. p. 285–288.
5. Gegenfurtner A, Lehtinen E, Säljö R. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. Educational Psychology Review. 2011; 23(4):523–552. https://doi.org/10.1007/s10648-011-9174-7
6. Moacdieh N, Sarter N. Display Clutter: A Review of Definitions and Measurement Techniques. Human Factors. 2015; 57(1):61–100. https://doi.org/10.1177/0018720814541145 PMID: 25790571
7. Gwizdka J. Distribution of Cognitive Load in Web Search. J Am Soc Inf Sci Technol. 2010; 61 (11):2167–2187. https://doi.org/10.1002/asi.21385
8. Huettig F, Axmann D. Reporting of dental status from full-arch radiographs: Descriptive analysis and methodological aspects. World Journal of Clinical Cases: WJCC. 2014; 2(10):552. https://doi.org/10.12998/wjcc.v2.i10.552 PMID: 25325067
9. Baghdady MT, Pharoah MJ, Regehr G, Lam EW, Woods NN. The role of basic sciences in diagnostic oral radiology. Journal of dental education. 2009; 73(10):1187–1193. PMID: 19805783
10. Baghdady MT, Carnahan H, Lam EW, Woods NN. Dental and dental hygiene students' diagnostic accuracy in oral radiology: effect of diagnostic strategy and instructional method. Journal of dental education. 2014; 78(9):1279–1285. PMID: 25179924
11. Diniz MB, Rodrigues JA, Neuhaus K, Cordeiro RCL, Lussi A. Influence of Examiners' Clinical Experience on the Reproducibility and Validity of Radiographic Examination in Detecting Occlusal Caries. Clinical Oral Investigations. 2010; 14(5):515–523. https://doi.org/10.1007/s00784-009-0323-z PMID: 19669175
12. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics. 2015; 35(6):1668–1676. https://doi.org/10.1148/rg.2015150023 PMID: 26466178
13. Castner N, Klepper S, Kopnarski L, Hüttig F, Keutel C, Scheiter K, et al. Overlooking: the nature of gaze behavior and anomaly detection in expert dentists. In: Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data. ACM; 2018. p. 8.
14. Douglass CW, Valachovic RW, Wijesinha A, Chauncey HH, Kapur KK, McNeil BJ. Clinical efficacy of dental radiography in the detection of dental caries and periodontal diseases. Oral Surgery, Oral Medicine, Oral Pathology. 1986; 62(3):330–339. https://doi.org/10.1016/0030-4220(86)90017-4 PMID: 3462638
15. Akarslan Z, Akdevelioglu M, Gungor K, Erten H. A comparison of the diagnostic accuracy of bitewing, periapical, unfiltered and filtered digital panoramic images for approximal caries detection in posterior teeth. Dentomaxillofacial Radiology. 2008; 37(8):458–463. https://doi.org/10.1259/dmfr/84698143 PMID: 19033431
16. Ericsson KA, Kintsch W. Long-term working memory. Psychological review. 1995; 102(2):211. https://doi.org/10.1037/0033-295X.102.2.211 PMID: 7740089
17. Cowan N. What are the differences between long-term, short-term, and working memory? Progress in brain research. 2008; 169:323–338. https://doi.org/10.1016/S0079-6123(07)00020-9 PMID: 18394484
18. Charness N, Reingold EM, Pomplun M, Stampe DM. The perceptual aspect of skilled performance in chess: Evidence from eye movements. Memory & cognition. 2001; 29(8):1146–1152. https://doi.org/10.3758/BF03206384
19. Voss MW, Kramer AF, Basak C, Prakash RS, Roberts B. Are expert athletes 'expert'in the cognitive laboratory? A meta-analytic review of cognition and sport expertise. Applied Cognitive Psychology. 2010; 24(6):812–826. https://doi.org/10.1002/acp.1588

20. Van der Gijp A, Ravesloot C, Jarodzka H, van der Schaaf M, van der Schaaf I, van Schaik JP, et al. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Advances in Health Sciences Education. 2017; 22(3):765–787. https://doi.org/10.1007/s10459-016-9698-1 PMID: 27436353

21. Coderre S, Mandin H, Harasym PH, Fick GH. Diagnostic reasoning strategies and diagnostic success. Medical education. 2003; 37(8):695–703. https://doi.org/10.1046/j.1365-2923.2003.01577.x PMID: 12895249

22. Schmidt HG, Boshuizen HP. On acquiring expertise in medicine. Educational psychology review. 1993; 5(3):205–221. https://doi.org/10.1007/BF01323044

23. Just MA, Carpenter PA. A capacity theory of comprehension: individual differences in working memory. Psychological review. 1992; 99(1):122. https://doi.org/10.1037/0033-295X.99.1.122 PMID: 1546114

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

25. Eivazi S, Hafez A, Fuhl W, Afkari H, Kasneci E, Lehecka M, et al. Optimal eye movement strategies: a comparison of neurosurgeons gaze patterns when using a surgical microscope. Acta Neurochirurgica. 2017; 159(6):959–966. https://doi.org/10.1007/s00701-017-3185-1 PMID: 28424915

26. Nodine CF, Mello-Thoms C. The nature of expertise in radiology. Handbook of Medical Imaging SPIE. 2000.

27. Kok EM, Jarodzka H, de Bruin AB, BinAmir HA, Robben SG, van Merriënboer JJ. Systematic viewing in radiology: seeing more, missing less? Advances in Health Sciences Education. 2016; 21(1):189–205. https://doi.org/10.1007/s10459-015-9624-y PMID: 26228704

28. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Academic radiology. 1996; 3(12):1000–1006. https://doi.org/10.1016/S1076-6332(96)80032-8 PMID: 9017014

29. Castner N, Kasneci E, Kübler T, Scheiter K, Richter J, Eder T, et al. 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; 2018. p. 39.

30. Veltman J, Gaillard A. Physiological workload reactions to increasing levels of task difficulty. Ergonomics. 1998; 41(5):656–669. https://doi.org/10.1080/001401398186829 PMID: 9613226

31. Sweller J. Cognitive load during problem solving: Effects on learning. Cognitive science. 1988; 12 (2):257–285. https://doi.org/10.1207/s15516709cog1202_4

32. Lance CE, Hedge JW, Alley WE. Joint relationships of task proficiency with aptitude, experience, and task difficulty: A cross-level, interactional study. Human Performance. 1989; 2(4):249–272. https://doi.org/10.1207/s15327043hup0204_2

33. Patel VL, Groen GJ. Knowledge based solution strategies in medical reasoning. Cognitive science. 1986; 10(1):91–116. https://doi.org/10.1207/s15516709cog1001_4

34. Patel VL, Groen GJ, Arocha JF. Medical expertise asa function of task difficulty. Memory & cognition. 1990; 18(4):394–406. https://doi.org/10.3758/BF03197128

35. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE Guide No. 86. Medical teacher. 2014; 36(5):371–384. https://doi.org/10.3109/0142159X.2014.889290 PMID: 24593808

36. Paas F, Ayres P. Cognitive load theory: A broader view on the role of memory in learning and education. Educational Psychology Review. 2014; 26(2):191–195. https://doi.org/10.1007/s10648-014-9263-5

37. Sibley C, Coyne J, Baldwin C. Pupil dilation as an index of learning. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 55. SAGE Publications Sage CA: Los Angeles, CA; 2011. p. 237–241.

38. Paas F, Tuovinen JE, Tabbers H, Van Gerven PW. Cognitive load measurement as a means to advance cognitive load theory. Educational psychologist. 2003; 38(1):63–71. https://doi.org/10.1207/S15326985EP3801_8

39. Appel T, Scharinger C, Gerjets P, Kasneci E. Cross-subject workload classification using pupil-related measures. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications. ACM; 2018. p. 4.

40. Appel T, Sevcenko N, Wortha F, Tsarava K, Moeller K, Ninaus M, et al. Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures. In: 2019 International Conference on Multimodal Interaction; 2019. p. 154–163.

41. Granholm E, Asarnow RF, Sarkin AJ, Dykes KL. Pupillary responses index cognitive resource limitations. Psychophysiology. 1996; 33(4):457–461. https://doi.org/10.1111/j.1469-8986.1996.tb01071.x PMID: 8753946

42. Hyönä J, Tommola J, Alaja AM. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. The Quarterly Journal of Experimental Psychology. 1995; 48(3):598–612. https://doi.org/10.1080/14640749508401407 PMID: 7568993

43. Beatty J. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological bulletin. 1982; 91(2):276. https://doi.org/10.1037/0033-2909.91.2.276 PMID: 7071262

44. Unema PJA, Pannasch S, Joos M, Velichkovsky BM. Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. Visual Cognition. 2005; 12(3):473–494. https://doi.org/10.1080/13506280444000409

45. Velichkovsky BM, Joos M, Helmert JR, Pannasch S. Two Visual Systems and their Eye Movements: Evidence from Static and Dynamic Scene Perception. In: CogSci 2005: Proceedings of the XXVII Conference of the Cognitive Science Society. Stresa, Italy; 2005. p. 2283–2288.

46. Kok EM, De Bruin AB, Robben SG, Van Merrienboer JJ. Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. Applied Cognitive Psychology. 2012; 26(6):854–862. https://doi.org/10.1002/acp.2886

47. Kahneman D, Beatty J. Pupil diameter and load on memory. Science. 1966; 154(3756):1583–1585. https://doi.org/10.1126/science.154.3756.1583 PMID: 5924930

48. Birren JE, Casperson RC, Botwinick J. Age changes in pupil size. Journal of Gerontology. 1950; 5 (3):216–221. https://doi.org/10.1093/geronj/5.3.216 PMID: 15428618

49. Brunyé TT, Eddy MD, Mercan E, Allison KH, Weaver DL, Elmore JG. Pupil diameter changes reflect difficulty and diagnostic accuracy during medical image interpretation. BMC medical informatics and decision making. 2016; 16(1):77. https://doi.org/10.1186/s12911-016-0322-3 PMID: 27378371

50. Szulewski A, Roth N, Howes D. The use of task-evoked pupillary response as an objective measure of cognitive load in novices and trained physicians: a new tool for the assessment of expertise. Academic Medicine. 2015; 90(7):981–987. https://doi.org/10.1097/ACM.0000000000000677 PMID: 25738386

51. Van Gerven PW, Paas F, Van Merriënboer JJ, Schmidt HG. Memory load and the cognitive pupillary response in aging. Psychophysiology. 2004; 41(2):167–174. https://doi.org/10.1111/j.1469-8986.2003.00148.x PMID: 15032982

52. Winn B, Whitaker D, Elliott DB, Phillips NJ. Factors affecting light-adapted pupil size in normal human subjects. Investigative ophthalmology & visual science. 1994; 35(3):1132–1137.

53. van der Wel P, van Steenbergen H. Pupil dilation as an index of effort in cognitive control tasks: A review. Psychonomic bulletin & review. 2018; 25(6):2005–2015. https://doi.org/10.3758/s13423-018-1432-y

54. Porter G, Troscianko T, Gilchrist ID. Effort during visual search and counting: Insights from pupillometry. The Quarterly Journal of Experimental Psychology. 2007; 60(2):211–229. https://doi.org/10.1080/17470210600673818 PMID: 17455055

55. Backs RW, Walrath LC. Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. Applied ergonomics. 1992; 23(4):243–254. https://doi.org/10.1016/0003-6870(92)90152-L PMID: 15676872

56. Geng JJ, Blumenfeld Z, Tyson TL, Minzenberg MJ. Pupil diameter reflects uncertainty in attentional selection during visual search. Frontiers in human neuroscience. 2015; 9:435. https://doi.org/10.3389/fnhum.2015.00435 PMID: 26300759

57. Takeuchi T, Puntous T, Tuladhar A, Yoshimoto S, Shirama A. Estimation of mental effort in learning visual search by measuring pupil response. PloS one. 2011; 6(7):e21973. https://doi.org/10.1371/journal.pone.0021973 PMID: 21760936

58. Lowenstein O, Feinberg R, Loewenfeld IE. Pupillary movements during acute and chronic fatigue: A new test for the objective evaluation of tiredness. Investigative Ophthalmology & Visual Science. 1963; 2(2):138–157.

59. Murata A. Assessment of fatigue by pupillary response. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences. 1997; 80(7):1318–1323.

60. Abokyi S, Owusu-Mensah J, Osei K. Caffeine intake is associated with pupil dilation and enhanced accommodation. Eye. 2017; 31(4):615. https://doi.org/10.1038/eye.2016.288 PMID: 27983733

61. Bradley MM, Miccoli L, Escrig MA, Lang PJ. The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology. 2008; 45(4):602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x PMID: 18282202

62. Gerjets P, Scheiter K, Catrambone R. Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures. Instructional Science. 2004; 32(1-2):33–58. https://doi.org/10.1023/B:TRUC.0000021809.10236.71

63. Mathôt S, Fabius J, Van Heusden E, Van der Stigchel S. Safe and sensible preprocessing and baseline correction of pupil-size data. Behavior research methods. 2018; 50(1):94–106. https://doi.org/10.3758/s13428-017-1007-2 PMID: 29330763

64. Zheng B, Jiang X, Atkins MS. Detection of changes in surgical difficulty: evidence from pupil responses. Surgical innovation. 2015; 22(6):629–635. https://doi.org/10.1177/1553350615573582 PMID: 25759398

65. Tien T, Pucher PH, Sodergren MH, Sriskandarajah K, Yang GZ, Darzi A. Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair. Surgical endoscopy. 2015; 29 (2):405–413. https://doi.org/10.1007/s00464-014-3683-7 PMID: 25125094

66. Szulewski A, Kelton D, Howes D. Pupillometry as a Tool to Study Expertise in Medicine. Frontline Learning Research. 2017; 5(3):53–63. https://doi.org/10.14786/flr.v5i3.256

67. Brunyé TT, Drew T, Weaver DL, Elmore JG. A review of eye tracking for understanding and improving diagnostic interpretation. Cognitive research: principles and implications. 2019; 4(1):7.

68. Turgeon DP, Lam EW. Influence of experience and training on dental students' examination performance regarding panoramic images. Journal of dental education. 2016; 80(2):156–164. PMID: 26834133

69. Castner N, Kübler TC, Richter J, Eder T, Huettig F, Keutel C, et al. Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing. In: Eye Tracking Research and Applications. ACM; 2020.

70. Nodine CF, Kundel HL, Mello-Thoms C, Weinstein SP, Orel SG, Sullivan DC, et al. How experience and training influence mammography expertise. Academic radiology. 1999; 6(10):575–585. https://doi.org/10.1016/S1076-6332(99)80252-9 PMID: 10516859

71. Manning D, Ethell S, Donovan T, Crawford T. How do radiologists do it? The influence of experience and training on searching for chest nodules. Radiography. 2006; 12(2):134–142. https://doi.org/10.1016/j.radi.2005.02.003

72. Krupinski EA, Tillack AA, Richter L, Henderson JT, Bhattacharyya AK, Scott KM, et al. Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. Human pathology. 2006; 37(12):1543–1556. https://doi.org/10.1016/j.humpath.2006.08.024 PMID: 17129792

73. American College of Radiology. ACR technical standard for digital image data management. Reston (VA): American College of Radiology. 2002; p. 811–9.

74. Brennan PC, McEntee M, Evanoff M, Phillips P, O'Connor WT, Manning DJ. Ambient lighting: effect of illumination on soft-copy viewing of radiographs of the wrist. American Journal of Roentgenology. 2007; 188(2):W177–W180. https://doi.org/10.2214/AJR.05.2048 PMID: 17242225

75. Goo JM, Choi JY, Im JG, Lee HJ, Chung MJ, Han D, et al. Effect of monitor luminance and ambient light on observer performance in soft-copy reading of digital chest radiographs. Radiology. 2004; 232 (3):762–766. https://doi.org/10.1148/radiol.2323030628 PMID: 15273338

76. Pollard BJ, Samei E, Chawla AS, Beam C, Heyneman LE, Koweek LMH, et al. The effects of ambient lighting in chest radiology reading rooms. Journal of digital imaging. 2012; 25(4):520–526. https://doi.org/10.1007/s10278-012-9459-5 PMID: 22349990

77. Kagadis GC, Walz-Flannigan A, Krupinski EA, Nagy PG, Katsanos K, Diamantopoulos A, et al. Medical imaging displays and their use in image interpretation. Radiographics. 2013; 33(1):275–290. https://doi.org/10.1148/rg.331125096 PMID: 23322841

78. Winkler T. NotebookCheck Review Dell Precision M4800 Notebook; 2013. Available from: https://www.notebookcheck.net/Review-Dell-Precision-M4800-Notebook.104416.0.html.

79. Ngo A. NotebookCheck Review HP ZBook 15 Workstation; 2014. Available from: https://www.notebookcheck.net/Review-HP-ZBook-15-Workstation.108229.0.html.

80. SensoMotoric Instruments. BeGaze Manual; 2017.

81. Hoeks B, Levelt WJ. Pupillary dilation as a measure of attention: A quantitative system analysis. Behavior Research Methods, Instruments, & Computers. 1993; 25(1):16–26. https://doi.org/10.3758/BF03204445

82. Klingner J, Kumar R, Hanrahan P. Measuring the task-evoked pupillary response with a remote eye tracker. In: Proceedings of the 2008 symposium on Eye tracking research & applications. ACM; 2008. p. 69–72.

83. Kiefer P, Giannopoulos I, Duchowski A, Raubal M. Measuring cognitive load for map tasks through pupil diameter. In: The Annual International Conference on Geographic Information Science. Springer; 2016. p. 323–337.

84. Brunyé TT, Gardony AL. Eye tracking measures of uncertainty during perceptual decision making. International Journal of Psychophysiology. 2017; 120:60–68. https://doi.org/10.1016/j.ijpsycho.2017.07.008 PMID: 28732659

85. Ahern S, Beatty J. Pupillary responses during information processing vary with Scholastic Aptitude Test scores. Science. 1979; 205(4412):1289–1292. https://doi.org/10.1126/science.472746 PMID: 472746

86. Verney SP, Granholm E, Marshall SP. Pupillary responses on the visual backward masking task reflect general cognitive ability. International Journal of Psychophysiology. 2004; 52(1):23–36. https://doi.org/10.1016/j.ijpsycho.2003.12.003 PMID: 15003370

87. Chen S, Epps J. Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. Human–Computer Interaction. 2014; 29(4):390–413. https://doi.org/10.1080/07370024.2014.892428

88. Duchowski AT, Krejtz K, Krejtz I, Biele C, Niedzielska A, Kiefer P, et al. The Index of Pupillary Activity: Measuring Cognitive Load Vis-à-vis Task Difficulty with Pupil Oscillation. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI'18. New York, NY: ACM; 2018. p. 282:1–282:13. Available from: http://doi.acm.org/10.1145/3173574.3173856.

89. Duchowski A, Krejtz K, Żurawska J, House D. Using Microsaccades to Estimate Task Difficulty During Visual Search of Layered Surfaces. IEEE Transactions on Visualization and Computer Graphics. 2019; 213.

90. Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Investigative radiology. 1978; 13(3):175–181. https://doi.org/10.1097/00004424-197805000-00001 PMID: 711391

91. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology. 2007; 242(2):396–402. https://doi.org/10.1148/radiol.2422051997 PMID: 17255410

92. Kundel HL, Nodine CF, Krupinski EA, Mello-Thoms C. Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. Academic radiology. 2008; 15(7):881–886. https://doi.org/10.1016/j.acra.2008.01.023 PMID: 18572124

93. Attar N, Schneps MH, Pomplun M. Working memory load predicts visual search efficiency: Evidence from a novel pupillary response paradigm. Memory & cognition. 2016; 44(7):1038–1049. https://doi.org/10.3758/s13421-016-0617-8

Supporting information

S1 Fig. Pupillary Response over Course of Experiment. The average pupillary response from baseline for students (blue bars, 20 images total) and experts (red bars, 15 images total) during the first set of OPTs presented and the second set of OPTs presented. Their is no effect in the pupillary response that could be attributed to fatigue during the experiment.

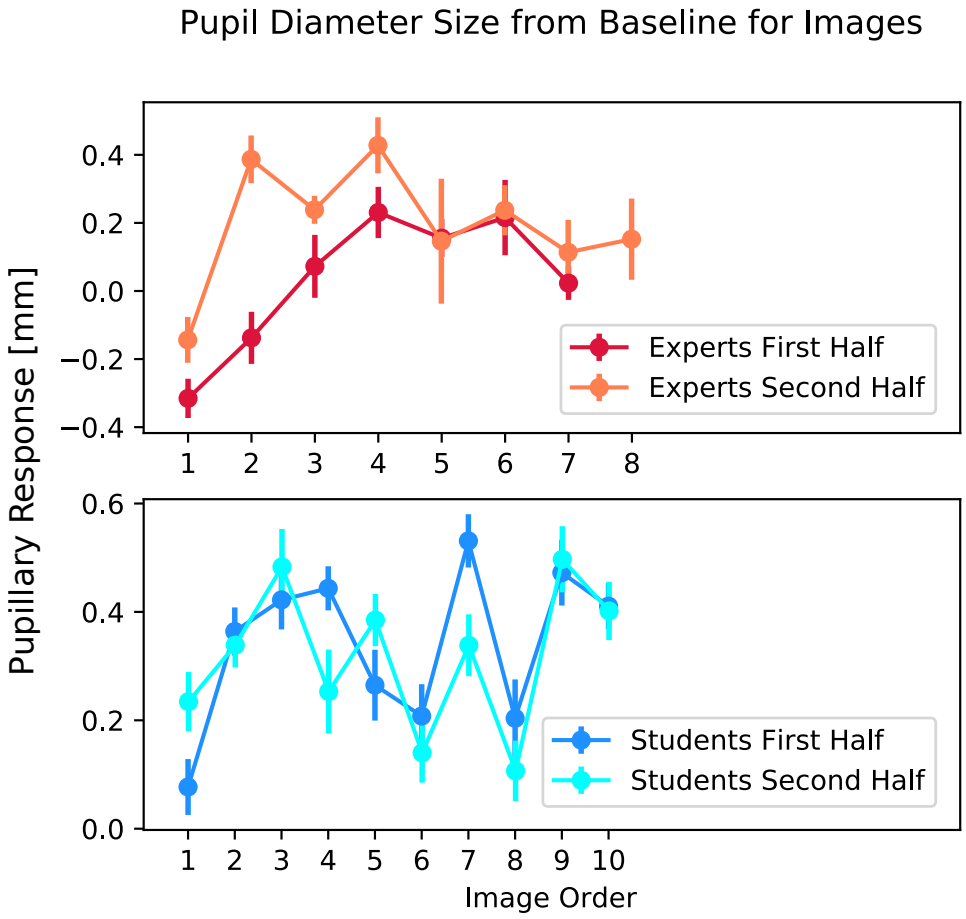


Figure 1: Pupillary Response for first and sectond set of images.

S1 Table. Table of Expert and Student Gaze Counts. shows the gaze hits on each anomaly type for both students and experts. For both levels of expertise, the least difficult and intermediate have the most gaze hits. The following are the ambiguous and the most difficult anomalies. Students had overall more gaze hits than experts; however, this may be attributed to the 90 second viewing time they had in comparison to the 45 second viewing time that the experts had.

Table 1: **Raw Gaze Count on Anomaly.**

Anomaly Type	Less Difficult	Intermediate	More Difficult	Ambiguous
Total	471	448	173	304
Student	312	296	124	202
Expert	159	152	49	102

Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing

Nora Castner
Perception Engineering, University of Tübingen
Tübingen, Germany
nora.castner@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

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

ABSTRACT

Modeling eye movement indicative of expertise behavior is decisive in user evaluation. However, it is indisputable that task semantics affect gaze behavior. We present a novel approach to gaze scanpath comparison that incorporates convolutional neural networks (CNN) to process scene information at the fixation level. Image patches linked to respective fixations are used as input for a CNN and the resulting feature vectors provide the temporal and spatial gaze information necessary for scanpath similarity comparison. We evaluated our proposed approach on gaze data from expert and novice dentists interpreting dental radiographs using a local alignment similarity score. Our approach was capable of distinguishing experts from novices with 93% accuracy while incorporating the image semantics. Moreover, our scanpath comparison using image patch features has the potential to incorporate task semantics from a variety of tasks.

*This work was sponsored by the Federal Ministry of Education and Research Germany
†Department of Prosthodontics
‡Department of Radiology, Center of Dentistry, Oral Medicine and Maxillofacial Surgery

ETRA '20 Full Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7133-9/20/06…$15.00
https://doi.org/10.1145/3379155.3391320

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Human-centered computing** → Human computer interaction (HCI); • **Applied computing** → **Psychology**.

KEYWORDS

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

ACM Reference Format:
Nora Castner, Thomas Kübler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, Constanze Keutel, and Enkelejda Kasneci. 2020. Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers), June 2–5, 2020, Stuttgart, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3379155.3391320

1 INTRODUCTION

Through eye movements, our thoughts, motivations, and expertise can be distinguished. We can accurately classify what someone is looking at and, more important, in what context they are looking at it, simply from the patterns in our gaze behavior. Eye-tracking data is, however, still subject to large intra- and inter-individual variance. Neither two subjects are likely to look at a given stimulus in an identical way, nor is the same person likely to exhibit the identical gaze sequence when looking at the same stimulus twice. This variability becomes non-trivial when developing online systems that can recognize specific groups: e.g., distinguish experts from novices or doing performance prediction.

We measure these distinct gaze patterns as a scanpath: Areas of focus (fixations) where the eye behavior remains relatively still before moving to another area via a rapid eye movement (saccade) [Holmqvist et al. 2011]. Discriminating scanpaths necessitates

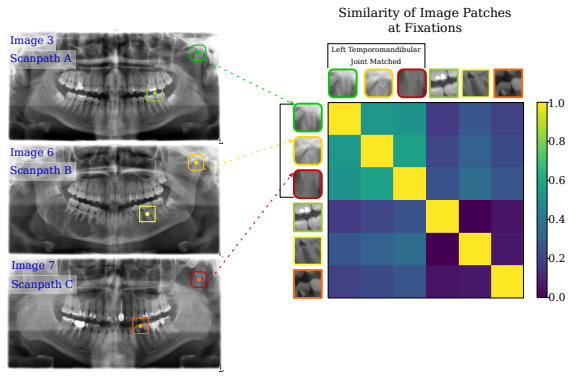


Figure 1: Matching image patch descriptors are recognized as similar across stimuli. When three different participants fixate on the left temporomandibular joint, the feature descriptors from DeepScan value them as similar. In contrast to when these participants fixate elsewhere, e.g. teeth, roots, etc.

effective ways of aggregating and averaging gaze data over multiple trials to achieve converging summarizations of representative scanpaths (e.g. attention density maps [Le Meur and Baccino 2013]).

Such aggregation techniques are simple to implement as long as subjects view the same stimulus from the same perspective, e.g., an image on a computer screen. Although, when either aggregation over a range of different stimuli or dynamic stimuli is required, analysis becomes challenging. For instance, semantically identical regions – also known as areas of interest (AOIs) – with regard to the studied task have to be identified and annotated. Once annotated, the sequence of AOIs visited by gaze can be analyzed as a proxy representation of the scanpath.

However, even though it is apparent that task and subject differences affect scanpaths, often accurate prediction is still elusive. Greene et al. [Greene et al. 2012] failed to predict an observers' task from their gaze behavior using sequence information from manually defined AOIs. Additionally, when aggregating the scanpath data, [Borji and Itti 2014] found they still could not accurately classify the task. Prediction increased in [Kanan et al. 2014] when considering the scanpath as a collection of features representing a fixations position and duration. Finally, the largest improvement in prediction performance was found when training an HMM model per stimulus [Haji-Abolhassani and Clark 2014]. Although it was accurate and incorporated spatial information, it has constrained applicability across stimuli.

In order to apply task or subject prediction from scanpath information, conventional approaches that handle one image, one subject, or both are not feasible. One realm in particular that has shown promising potential for gaze behavior is training of medical personnel [Gegenfurtner et al. 2011; Van der Gijp et al. 2017; Waite et al. 2019]. For instance, gaze analysis has often been proposed as a measure for adaptive training systems (i.e. searching radiographs for pathologies [Jarodzka et al. 2012, 2010b], practicing surgery or laproscopy in VR/AR [Law et al. 2004]). However, actually working training procedures are still scarce. Massed practice approaches, i.e., lengthy viewing of hundreds of radiographs, is still common educational practice [Rozenshtein et al. 2016]. Even though it has been available for decades, as of now eye-tracking has yet to deliver the promises for adaptive training. The challenge of expediting a novice to expert solely through training gaze behavior has yet to be fully operational [Van der Gijp et al. 2017].

In this work, we show how to incorporate high-level, deep neural network-generated image patch representations into classical scanpath comparison measures. We apply our method DeepScan to expertise classification on an eye movement dataset of expert and student dentists. Dentistry, in particular, relies heavily on effective visual inspection and interpretation of radiographs [Huettig and Axmann 2014]. Even then, panoramic dental radiographs are highly susceptible to diagnostic error [Akarslan et al. 2008; Bruno et al. 2015; Douglass et al. 1986; Gelfand et al. 1983]. We demonstrate our method by decoding expertise from eye movements during dental radiograph inspection, which is a crucial first step towards adaptive learning procedures. It is worth noting, this metric is not confined to dental expertise recognition, rather developed with the intention for various use cases. It offers the future potential to assess student's learning progress in real-time and to adapt stimulus material based on current aptitude, while not being restricted to the stimulus material used during creation of the classifier.

2 REVISITING VISUAL SCANPATH COMPARISON

2.1 Traditional Approach: String Alignment

One of the most common and traditional approaches to scanpath comparison is extraction of a similarity score by representing a scanpath as a sequence of symbols and comparing the resulting string to one another [Anderson et al. 2015]. AOIs on a given stimuli can be semantically or structurally linked to a symbol [Cristino et al. 2010; Goldberg and Helfman 2010; Jarodzka et al. 2010a; Kübler et al. 2014]. Thus, coded strings provide information on the temporal and spatial order of the user's gaze behavior. Temporal resolution (i.e. fixation duration) can also be factored into the sequence [Cristino et al. 2010].

The output of such a comparison – the similarity score – is based on a total derived from rewarding matches and penalizing mismatches or gaps[1]. A scoring matrix can be used to represent the relative similarity of characters to one another [Baichoo and Ouzounis 2017; Day 2010; Goldberg and Helfman 2010]. A positive matching score represents similar regions and a negative score mismatches. Gaps are inserted in order to make neighboring characters match and to compensate small shifts of highly similar segments between the sequences.

Global sequence alignment with a notion of AOI similarity can be performed via the Needleman-Wunsch algorithm [Anderson et al. 2015; Needleman and Wunsch 1970]. Global sequence alignment determines the most optimal alignment for the entirety of two sequences. It has been shown to be a robust metric in scanpath comparison, e.g. in ScanMatch [Cristino et al. 2010], classification

[1] inserting a space into one of the sequences.

of attentional disorder [Galgani et al. 2009], multiple scanpath sequence alignment [Burch et al. 2018], and expert and novice programmer classification [Busjahn et al. 2015]. Castner et al. [Castner et al. 2018a] found incoming dental students with no prior training in radiograph interpretation could be classified from later semester students with 80% accuracy from Needleman-Wunsch similarity scores.

Similarly, scanpath similarity from local sequence alignment has often been used as a robust classifier. Rather than deal with the entirety of two sequences, local alignment determines the most optimal aligned subsequence between the two. Local alignment compensates to a greater degree for sequences of differing lengths and is not as strongly influenced by differences in the beginning or end of the sequences [Khedher et al. 2018]. For example, [Khedher et al. 2018] used the Smith-Waterman algorithm [Smith et al. 1981] for local alignment of medical undergrads' scanpaths during a clinical reasoning task. They found similarly well performing students had highly correlative scores. Similarly, [Çöltekin et al. 2010] found high comprehension and scanpath similarity of local subsequences in reading interactive map displays.

Determining the optimal alignment between two sequences is computationally costly. Additionally, though commonly used, these methods suffer from a severe drawback: The manual selection of AOIs. This process is subjective, not only in which AOIs are considered relevant for the analysis, but also with regard to their size [Cristino et al. 2010; Jarodzka et al. 2010a]. For instance, Deitelhof et al. [Deitelhoff et al. 2019] found that scanpath transitional information can be highly impacted by the AOI size and padding, which can affect validity. Moreover, some measures (e.g. Levenshtein distance) only rate exact matches and mismatches and do not consider any potential AOI similarity - and the ability of an algorithm to include this introduces the additional hard problem of judging AOI similarity objectively.

Much of the prior literature on scanpath comparison using sequence alignment have employed manual AOI definitions. However, these approaches suffer errors in spatial resolution or require task-subjective AOI labels [Cristino et al. 2010; Jarodzka et al. 2010a]. Kübler et al. [Kübler et al. 2014] developed a method –SubsMatch– for sequence comparison without AOI definitions, which uses a bag-of-words model and looks at the transitional behavior of subsequences. Castner et al. [Castner et al. 2018a] used these subsquence transitions from SubsMatch with an SVM Classifier [Kübler et al. 2017] and found comparable results to sequence alignment with grid AOIs.

However, these automatic approaches lack any notion of what is actually being looked at. Therefore, they usually perform excellent when subjects view the exact same stimulus (because then location identity corresponds to semantic identity to some extent). But when performing cross-stimulus analysis or the stimulus is subject to noise, performance drops significantly.

As of now, gaze pattern comparison is based either only on gaze location – not on the semantic object that is being looked at – or relies on human annotation to determine the semantics. Yet, scene semantics are absolutely critical for judging gaze behavior. For larger experiments and *in the wild* head-mounted eye tracker data [Pfeiffer et al. 2016; Wan et al. 2018], manual annotation is unfeasible. We propose a method that combines the traditional approach of sequence alignment with deep learning for fixation target understanding. Combining these methods enables us to understand (and automatically analyze) the semantics behind a scanpath.

2.2 Current Directions: Deep Learning

Convolutional neural networks (CNNs) can provide information of image semantics that can be used for segmentation [Chen et al. 2017; Long et al. 2015] or classification [Krizhevsky et al. 2012] and saliency prediction [Hong et al. 2015; Huang et al. 2015], and many other applications. In the field of eye tracking research, they have also provided robust performance in eye movement behavior and scanpath generation [Assens Reina et al. 2017; Liu et al. 2015]. For instance, methods using probabilistic models and deep learning techniques coupled with ground truth gaze behavior have been shown to predict fixation behavior [Kümmerer et al. 2015; Wang et al. 2015].

Concerning human scanpath classification, [Fuhl et al. 2019] encoded gaze data as a compact image with the spatial, temporal, and connectivity represented as pixel values in the red, green, and blue channels respectively. These images were input for a CNN classifier, which showed high accuracy in classifying task-based gaze behavior. Mishra et al. [Mishra and Bhattacharyya 2018] followed a similar approach of depicting scanpath information as an image for a CNN sarcasm detector.

Tao and Shyu [Tao and Shyu 2019] offer an approach similar to our proposed approach. They developed a CNN-Long Short Term Memory (LSTM) network that runs on scanpath-based patches from a saliency-predicted map[2] and classifies typical/autism spectrum disorder gaze behavior [Tao and Shyu 2019]. Square patches are defined based on fixation positions as they occur in the scanpath. Then, each patch is run through a shallow CNN, and the patch feature vector with the duration information provides an LSTM network input for classification from a dense layer from each patch [Tao and Shyu 2019]. Most notable, they maintain the sequential information of the scanpath.

We utilize powerful Deep Neural Network(DNN)-based feature descriptors to represent the semantics of a gaze sequence (scanpath). Our proposed approach follows a similar idea of incorporating the sequential fixation information in conjunction with visual features using a CNN. However, we extract scanpath similarity from the culmination of image patch features using the traditional approach of sequence alignment. For the current work, we chose local alignment in order to focus on common subsequences that could be indicative of expertise. Then, we cluster the scanpaths based on this similarity. Subsequently, we evaluate our proposed approach on detecting expert and student dentists' scanpaths when inspecting dental radiographs.

3 PROPOSED APPROACH

3.1 Image Features at the Fixation Level

Each individual fixation corresponds to a visual intake of a certain stimulus region. We then encode each fixation location on the specific stimulus image by a vector that describes the local image

[2] ASD specific saliency prediction from the Saliency4ASD challenge.

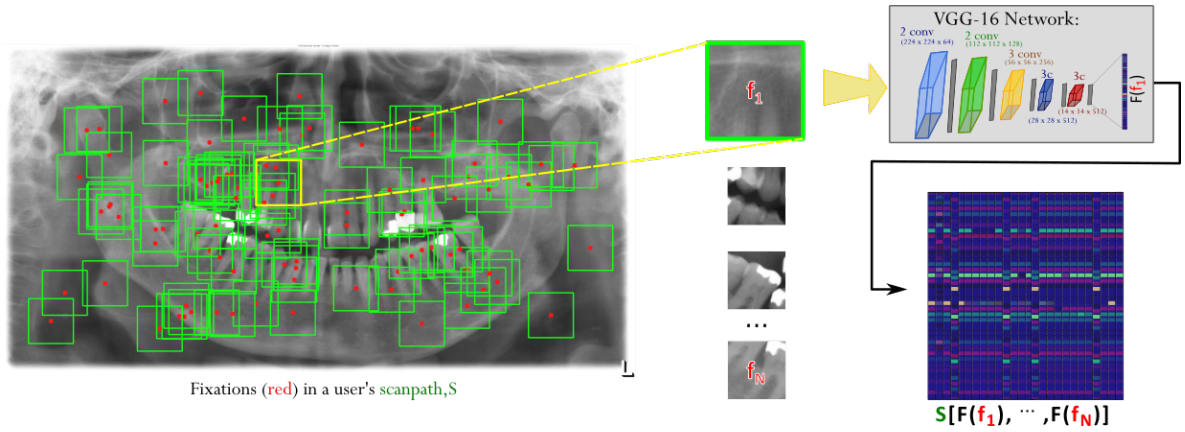


Figure 2: Proposed Model: DeepScan. For a scanpath, we extract the fixation locations and, using the VGG-16 CNN architecture, we create a feature corresponding to an image patch relative to the ith fixation $\mathbf{F}(f_i)$. The resulting vector illustrating the scanpath S can then be compared to another scanpath vector. In our work, we compared scanpaths via local alignment similarity. The pre-trained VGG-16 network consist of 5 blocks of convolutions with ReLus with max-pooling between each layer.

region. We generate such encodings via the output from the VGG-16 architecture [Simonyan and Zisserman 2014]. Accordingly, for each fixation location on the stimulus image, we extract a patch of 100×100 pixels as input to the network. This step is relatively similar to [Tao and Shyu 2019], although we determined that using a fixed size bounding box is adequate for our stimuli. The fixation coordinates indicate the center of the bounding box of the image patch, unless a fixation is too close to the stimulus borders. Then, appropriate shifting of the box along the x or y axis is necessary.

The architecture we employed for patch processing originally takes 224×224 RGB input images. For the current evaluation on experts and students searching dental radiographs, our stimuli were grayscale with pixel dimensions 1920×1080. In development, we determined that patch sizes of 224×224 for our stimuli were too large (e.g. four or more teeth would be in this sized area). Smaller patches were more preferable so that enough information from an entity is extracted. Therefore, we rescaled the 100×100 image patches to the desired input size for the network, and replicated the one channel image information to get three channels that can utilize the weights pre-trained on ImageNet [Deng et al. 2009].

However, image patch input size and channels could be adapted for other stimuli or any other preferred network for the fixation encodings. The takeaway from this image patch approach is that through only the gaze: 1) we map the image features of interest in temporal order, and 2) we can extract the semantics from these features.

3.2 CNN Architecture

For patch descriptor extraction, we employed a VGG-16 network [Simonyan and Zisserman 2014] as implemented in keras[3] and pre-trained on ImageNet. Figure 2 shows the network: Consisting of five blocks of convolutions, with each block followed by ReLUs and max-pooling.

[3]Python 3.6 with GPU compatibility.

Since we are only interested in the features, we omit the fully-connected and prediction output layers of the model and use the output after max-pooling, which has $7 \times 7 \times 512$ dimensions, and flatten it to a $1 \times 1 \times 25088$ vector. This feature description from the final convolutional layer, $\mathbf{F}(f_i)$, represents the image patch at the ith fixation f_i.

The feature descriptors provide the semantic information for each fixation in a user's scanpath and are the equivalent to a symbol representation in the traditional string-sequence representation. In the following, we discuss the changes required in the alignment algorithm in order to work with alignment scores generated by comparing these image features to each other. Figure 1 shows an example of how similar features can compare to each other.

We chose the VGG-16, in contrast to a network pre-trained on radiology images since it is more generalizable to a variety of tasks and stimuli. Additionally, it is publicly available and easily integrated for replication purposes. Pre-trained networks for medical images are often not publicly available due to the data sensitivity and protection, and any existing architectures for these images are not yet up to par with the generic image trained architectures. Choosing a network that is trained for a specific stimulus category, e.g., panoramic radiographs or other X-Rays, might improve results. However, it introduces the risk of limiting data analysis to specific elements, which is comparable to manual AOI selection. Ultimately, though our approach is evaluated on medical image expertise, we developed it for generalizability in multiple applications.

3.3 Local Alignment

Once we have descriptors for each fixation, we assemble them as a scanpath. The resulting matrix of image features at each fixation creates a scanpath matrix. $S_A = (F_{f_1}, F_{f_2}, \cdots, F_{f_N})$. With this matrix representation, we can compare its similarity to the matrix representing another scanpath.

For scanpath comparison, we perform local alignment using a variant of the Smith-Waterman Algorithm. We preferred local alignment scoring over global alignment due to its ability to find similar subsequences, even if the scanpaths may otherwise be highly varying [Khedher et al. 2018]. Moreover, we did not want to enforce strict global alignment due to different viewing times required by students and experts. In sequence alignment, the penalty system can have a major effect on values in the scoring matrix, and therefore, the similarity score [Baichoo and Ouzounis 2017]. Our scoring choice prioritizes finding long rather than short similar subsequences by accumulating scores. Equation 1 details the scoring system used for the current evaluation:

$$M_{ij} = \max \begin{cases} M_{i-1,j-1} + \; c - \sum_{i,j} |A_{:F_j} - B_{:F_i}|, & \text{Match} \\ M_{i,j-1} - gap, & \text{Gap in A} \\ M_{i-1,j} - gap, & \text{Gap in B} \\ 0 & \text{No Similarity.} \end{cases} \quad (1)$$

Where M is the scoring matrix of size $(n+1) \times (m+1)$ for two scanpaths A and B with n and m fixations respectively. Element $M_{i,j}$ takes the maximum value based on if there is a match between the values at index j of scanpath A and index i of scanpath B. The original algorithm scores matches as the score value added to the value at the previous indices: $M_{i-1,j-1} + score(a_j, b_i)$. Then, if there is no match, it determines whether the value of a gap penalty (gap) in either scanpath, or no similarity (0) are more optimal for the score.

The interesting part of our approach is contained in the calculation of the match score. Since it is highly unlikely that two features will be exactly the same, we cannot explicitly match or mismatch. Therefore, we calculate the score by taking the sum of absolute differences in feature descriptor j of scanpath A and descriptor i of scanpath B as shown in the first line of equation 1. This is simple to implement and cheap to compute, but other metrics such as cosine or Euclidean distance could also be used. This procedure leads to a dissimilarity score between the image patches. The more dissimilar the image patches, the larger the scoring value.

In order to convert it to a similarity score, we can subtract the dissimilarity score from a constant c. We calculated c in equation 1 by averaging the sum of the differences for all features between all scanpaths of one random image in the dataset. Therefore, c was 21,049 in the evaluation of our proposed approach. This constant affects highly similar image patches positively, but highly dissimilar image patches are penalized negatively with the same weight. Meaning it functions similar to a match/mismatch threshold. Additionally, we set our gap penalty (lines 2 and 3 in eq.1) to 42,000 to highly penalize gaps, therefore almost double c.

This choice of c makes the algorithm consider about half of the image patches relatively dissimilar to each other. Furthermore, gaps are penalized quite strongly, resulting in compact alignments that are not drastically influenced by large differences in sequence lengths. Figure 3 shows an example of the similarity matrix created from the local alignment performed for two scanpaths. The maximum value in the matrix is the similarity score [Smith et al. 1981]. In figure 3, the highest yellow color indicates the final similarity score and backtracing from this index till 0 will give the optimal local alignment of both sequences.

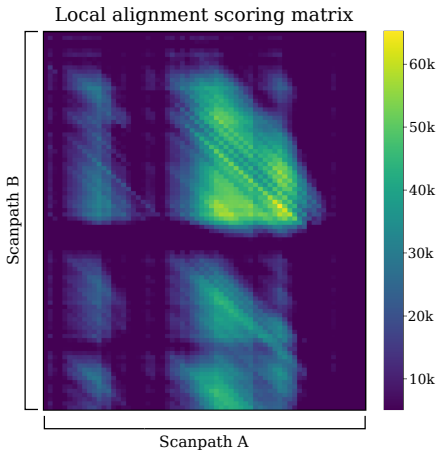


Figure 3: Scoring matrix of the local alignment. Backtracing from the index with the highest value (yellow) will give you the optimal local alignment of two scanpaths.

The resulting similarity score for the two scanpaths is $\max(M)$. Then, we normalize this score based on the length of the shorter scanpath, thus:

$$similarity = \frac{\max(M)}{\min(|S_A|, |S_B|)}. \quad (2)$$

We compared the performance of our DeepScan method to Smith-Waterman local alignment of hand-labeled semantic AOIs (the gold standard in adding semantic information to image patches, see Supplementary Material Figure 1). These AOIs indicate specific anatomical structures and regions across the dental radiographs and provide the paramount in semantic information that can be represented in a scanpath. For scoring the semantic scanpath comparisons, we used a simple, standard scoring system: 1 for matches, −1 for mismatches, and −2 for gaps.

4 EVALUATION

4.1 Scanpath Data of Dentists

Students (n=57) were incoming dental students (sixth semester) from their initial pre-med studies. They had no prior training in dental radiograph interpretation, but basic conceptual knowledge in general medical concepts. Experts (n=30, average 10.16 years experience) were dentists working in the local university clinic with all the proper qualifications and some had further licensing for other particular specializations (e.g. Endontology, Prosthetics, Orthodontology, etc.). Diagnostic performance results from both groups indicated that the experts had 79.91% higher pathology detection accuracy than students [4].

Both students and experts were asked to perform a visual search task of panoramic dental radiographs (OPTs); then following image inspection, indicate any areas indicative of pathologies. Students

[4]Performance metrics and expert results can further be found in [Castner et al. 2018b]

had 90 seconds to inspect each OPT, where experts had 45 seconds to inspect each OPT. This shortened duration was due to the research indicating that experts are much faster when visually inspecting radiographs [Gegenfurtner et al. 2011; Turgeon and Lam 2016]. Students inspected two blocks of 10 OPTs in one experimental run and experts – due to their hard-pressed schedules – inspected 15 OPTs.

All eye tracking data was collected with SMI RED250 remote eye trackers sampling at 250Hz attached to laptops with FullHD displays. A quality assessed calibration[5] was performed for each participant before and during data collection. Gaze data, i.e. fixations, were determined using a velocity based metric provided by the eye tracker's software. Further details of the data collection and pre-processing can be found in [Castner et al. 2018a,b].

For compatibility, we chose to evaluate gaze data from the first 45 seconds of each student participant, in line with the experts' total viewing time. Additionally, our model is only evaluated on gaze data for the 10 OPTs that both groups viewed. Gaze data was lost for two expert participants due to software failure. Also, 5 participants were excluded due to having high data loss (under 80% tracking ratio[6] and 3 or more low signal quality images) leaving 25 experts and 54 students for the final analysis. The resulting total for all participants for all images was 733 scanpaths.

4.2 Similarity Scoring

We performed local alignment of the scanpath vectors with patch features for each participant for all images. In order to get the scanpath behavior representative of each participant, we averaged a participants' similarity output for all images. Figure 4 shows the similarity scores from DeepScan of each participants' scanpath behavior over the images viewed in pairwise comparison to other participants. The diagonal of the matrix indicates the highest similarity value, which is a participants' gaze behavior compared to his or herself.

From the similarity matrix, a trend is apparent where experts (labeled green in figure 4) show higher similarity scores among each other, as visible by the more yellow values. Conversely, students' gaze behavior shows less similarity among each other, especially when compared to experts.

4.3 Hierarchical Clustering

We clustered the similarity scores of all participants using agglomerative hierarchical clustering [Corpet 1988; Johnson 1967; West et al. 2006]. As the similarity matrix can easily be inverted to a distance matrix, the unsupervised clustering approach was straight forward; however one could introduce additional weighting factors or more complex classification methods on top as well. This approach evaluates the distance between data points and links closer in distance clusters until one cluster remains [Johnson 1967]. Partitioning the clusters then is determined by the linkage distance. We used Ward's [Johnson 1967] method for proximity definition, which minimizes the sum of the squared distances of points from the cluster centroid.

[5]less than one degree average deviation from a four point validation.
[6]A metric reported from SMI indicating proportion of valid gaze signals.

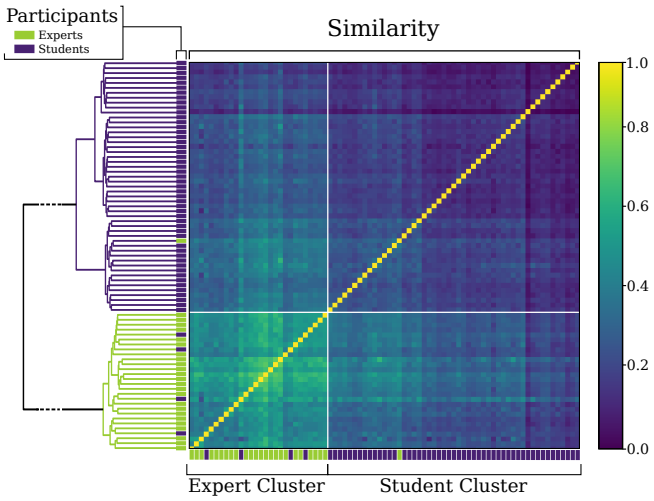


Figure 4: Similarity matrix of subjects' average scanpath behavior. Purple labels indicate students' gaze behavior. Green labels indicate experts' gaze behavior. Values closer to yellow indicate higher similarity, Where the diagonal is a participant compared against themselves. Values shown on the diagonal are rescaled relative to values off-diagonal solely for perceivability. On the y-axis is the resulting clustering of the dendrogram, which recognized 2 clusters. On cluster (purple) with mainly students and the other cluster (green) with mainly experts.

Average Gaze Behavior of Each Subject. For the scores of each student and expert summed over all images, the resulting dendrogram (2-dimensional tree view of the nested clusters) is shown on the y-axis in figure 4.

The clustering seen in figure 4 recognizes two main clusters evident in the gaze behavior with the majority of students in one cluster (purple cluster, purple labels) and the majority of experts (green cluster, green labels) in the other. Table 1 calculates the true positive rate (TPR) when utilizing the clustering as a classification for both students and experts as well as the overall accuracy. We achieved 93.7% accuracy. We also found two clusters evident in the traditional local alignment with manual AOIs; however more students were misplaced in the expert cluster (as seen in table 1), resulting in an overall accuracy of 85%.

Gaze Behavior on the Image Level. We then ran the hierarchical clustering for participants' gaze at the image level (over all 733 datasets and not the average similarities for each participant as above). The dendrogram also recognized two clusters, therefore we calculated the number of experts in one cluster and the number of students in the other. The achieved accuracy for our approach was 68.62%: Experts had 85.65% TPR and students had 61.18% TPR. The achieved accuracy for the traditional, semantic approach was 64.39%: Experts had 51.76% TPR and students had 93.27% TPR. This slight dip in performance could be attributed to pathology differences in the stimuli. Previous literature has also found that

Table 1: Performance of linkage clustering for our approach (*Feature*) and Semantic AOIs as measured by the True Positive Rate (TPR). Two main clusters were found based upon the gaze behavior for both approaches.

	Student		*Expert*		***Accuracy***	
	Feature	*Semantic*	*Feature*	*Semantic*	*Feature*	*Semantic*
Student	50	44	1	1		
Expert	4	10	24	24		
TPR	92.5 %	81.5 %	96.0 %	96.0 %	**93.7 %**	86.06 %

gaze behavior of expert and novice dentists can be highly stimulus dependent, where dental radiographs considered easy to interpret evoke similar gaze behavior in experts and novices [Grünheid et al. 2013; Turgeon and Lam 2016].

4.4 Cross-Image Classification

To further see whether we could predict classification performance on an image level, we performed a leave one subject and one image out cross-validation using the similarity scores from DeepScan. We performed classification to 1) see whether we could predict a participant's expertise from their scanpath on a new image, not contained in the set that we compare to. 2) to confirm that certain stimuli may affect the similarities more than others. For each subject, we ran a 3-nearest neighbor classifier, trained on the remaining subjects and images. Table 2 shows the performance for each image. Here, it is clear that for some images, distinguishing expert and student scanpaths becomes more difficult. For instance, image 1 shows a heavy tendency to classify all participants' scanpaths for that image as experts, and image 3 shows a tendency to over-classify as students. Nevertheless, five images allowed us to determine expertise of a new subject on a new stimulus that were not contained in the data we used for the classification. Especially, image 8 shows the highest accuracy in classifying level of expertise, meaning this OPT and its semantics can possibly trigger experts to inspect the image in a distinctive way.

The cross-validation for the traditional local alignment scoring for the scanpaths with manual AOIs, showed better performance on the image level than DeepScan, and slightly better overall (77% versus 73% respectively). Thereby, it is possible that we cannot yet utilize the full potential of semantic encoding using the feature approach. However, given that DeepScan is purely data driven, its results are comparable and relegates the tedious process of manual AOI labeling. Retraining the network on OPT data might help the encoding to come closer to manually-defined semantic labels.

Additionally, we sorted the similarity scores of all scanpaths from DeepScan to isolate those that expose especially high similarity values to many other scanpaths. We hoped to extract archetype-scanpaths this way. The histogram in figure 5 shows that two expert scanpaths had the highest similarity scores to the most other scanpaths. Interestingly enough, both these scanpaths and a number of the other high similarity scanpaths are for image 1. Thus from the local alignment similarity, certain scanpaths from image 1 offer highly similar subsequences to other scanpaths regardless of image. Image 1 was one of the stimuli that made a distinction between

Table 2: Performance of kNN classifier when one image is left out and each participants' expertise for that image is predicted. Note that chance level is not 50%, therefore we provide Cohen's Kappa (κ) as an indicator of performance, with bold text indicating fair performance.

	Expert TPR		*Student TPR*		***Accuracy***			
	Feature	*Semantic*	*Feature*	*Semantic*	*Feature*		*Semantic*	
Chance:	32 %		68 %		*Overall*	κ	*Overall*	κ
Image 1	100 %	75 %	20.4 %	76.6 %	44.9 %	0.14	**78.2 %**	0.52
Image 2	59.1 %	68.2 %	83.3 %	85.4 %	**75.7 %**	0.43	**80 %**	0.54
Image 3	28.6 %	66.7 %	93.5 %	80.4 %	73.1 %	0.26	**76.1 %**	0.46
Image 4	52.4 %	57.1 %	89.8 %	83.7 %	**78.6 %**	0.45	**75.7 %**	0.41
Image 5	76.2 %	53.4 %	68.6 %	88.2 %	70.8 %	0.39	**77.8 %**	0.43
Image 6	66.7 %	75 %	67.9 %	81.1 %	65.5 %	0.31	**79.2 %**	0.54
Image 7	60.9 %	30.4 %	86.5 %	90.4 %	**78.7 %**	0.49	72 %	0.24
Image 8	73.9 %	91.3 %	88.2 %	68.6 %	**83.8 %**	0.62	**75.7 %**	0.51
Image 9	45.8 %	58.3 %	92.6 %	96.3 %	**78.2 %**	0.43	**84.6 %**	0.60
Image 10	30 %	80 %	96.2 %	65.4 %	77.8 %	0.32	69.4 %	0.37
Overall	60.1 %	65.5 %	78.2 %	82 %	72.7 %	0.37	**76.9 %**	0.46

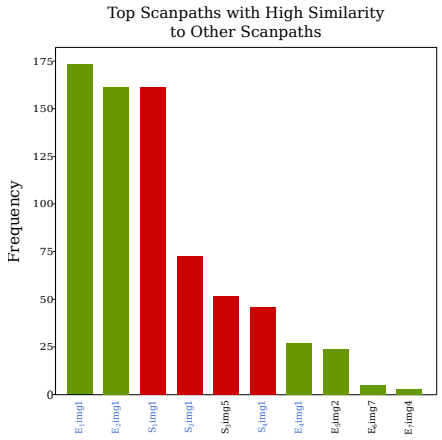


Figure 5: The top scanpaths who have the highest frequencies of similarities to other scanpaths; With experts indicated in green and students indicated in red. The majority of theses scanpaths are for image 1, as indicated by the blue text.

expertise levels hard. It might therefore represent a standard scanpath for checking OPTs that abstracts over special attributes of individual stimuli.

The two experts scanpaths (illustrated by their image patches) with the most highest similarities to each other and many other subjects' scanpaths are shown in figure 2 in the Supplementary Material.

5 DISCUSSION

We were able to successfully extract similarities in the scanpath behaviors between experts and the differences towards students gaze behavior while interpreting panoramic dental radiographs.

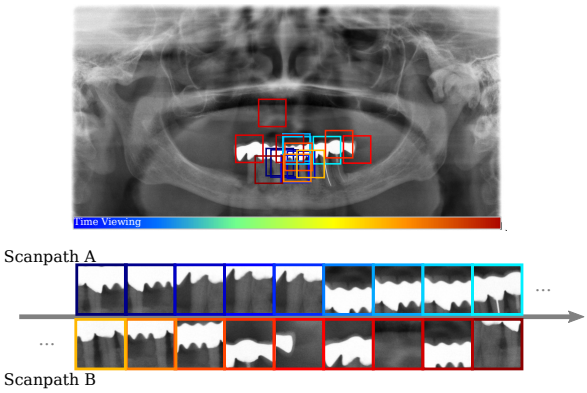


Figure 6: Two relatively dissimilar scanpaths from students. The local alignment finds the optimal matching subsequence starting in scanpath A at the twentieth fixation (far left top) and in scanpath B at the fiftieth fixation (far left bottom).

Our developed scanpath comparison approach uses temporal scanpath information to extract image features at the fixation level. The resulting similarity comparison of scanpaths therefore incorporates this image information into the traditional approach of sequence alignment to detect patterns between the behaviors.

From traditional local alignment techniques using image features, we found that experts showed highly similar behavior to each other and therefore, were more likely to be clustered together. More interesting, students' similarity scores indicated that their scanpaths were not highly similar to those of experts, but also there was no distinct homogeneity among themselves (see figure 6). One possible reason for their low similarity to each other could be that they are incoming students with some conceptual background; however, they had no training on radiograph interpretation. Previous research has found that students evoke more systematic search strategies after training, resulting in more similar gaze behaviors [Kok et al. 2016; Van der Gijp et al. 2017]. Additionally, the heterogeneity of background and training can affect scanpath similarity [Davies et al. 2016]. Possibly students have varying levels of conceptual knowledge or familiarity with radiographs before entering their first year of dental studies.

Our algorithm was able to accurately classify unseen scanpaths given scanpaths from other participants and other images. Although we found that, depending on the image, it could be easier or harder to differentiate the levels of expertise from the scanpath similarities. This finding is, however, in alignment with previous studies specifically on dentists and dental radiograph examination. For instance, [Turgeon and Lam 2016] found that radiographs defined as easy to interpret offered no differences in the gaze behavior of experts and novices. Castner et al. [Castner et al. 2018b] also found that even among experts some images evoked highly differing gaze behavior to achieve accurate anomaly detections.

With the system at hand, we could classify expertise of dentist students in an adaptive feedback setting from viewing just a single stimulus (with decent accuracy), even if the stimulus itself is an arbitrary OPT that is unknown to the classifier. This could be used to guide students through the learning process and to adapt the difficulty of stimulus material to their current knowledge level. When viewing multiple stimuli (which students do in the current mass practice approach), classification accuracy can be increased.

Futhermore, we observed that some stimuli allowed for a classification of expertise, while others did not. We could utilize this information as a hint on which stimuli are likely to induce a training effect and to differentiate them from stimuli that are too easy (for the current student).

Moreover, we designed DeepScan to handle image variability. One image feature descriptor of a patch in one image can match to similar patches in other images (see figure 1); This way, scanpaths can be more easily compared cross-stimuli, but this process also replaces a manual AOI-annotation. By the assumption that similar semantic meaning in a visual task corresponds to similar looking features in the stimulus, we have introduced a notion of stimulus semantics into the automated scanpath interpretation. A similar workflow could be used to compare data where the annotation of dynamic AOIs is usually unfeasible, e.g., recordings of mobile eye-tracking devices to each other. Furthermore, we do not restrict the algorithm to individual annotated AOIs, but represent each fixation by its feature descriptor, no matter whether a data analyst would deem it relevant for the analysis at hand or not.

One limitation for the current work could be the methodological confound of the viewing time differences in the expert and student paradigms. Since a consistently longer viewing time for the students would heavily affect the similarity scoring regardless of normalization, we took the first 45 seconds of the students, so that our similarity scores would be less biased by their longer scanpaths.

6 CONCLUSION

Our proposed model for scanpath classification, DeepScan, is capable of extracting gaze behavior indicative of expertise in dental radiograph inspection. More important, this approach employs deep learning to extract image features. Consequently, human expert gaze behavior coupled with relevant image semantic extraction offers an accurate approach to automated scanpath classification. However, the motivation for this model does not finish here. Rather, it was developed for applicability not only in the medical expertise domain, but also for scenarios with dynamic, semantically varying tasks (i.e. Training in VR, real world scenarios with mobile eye tracking).

Future directions of the proposed approach optimization for online classification of scanpaths. We chose a local alignment evaluation as a traditional approach to scanpath comparison, since it provides for a standard and robust evaluation of the scanpath feature matrix created. DeepScan has the potential for online use and further evaluation are therefore necessary for working towards integrating this model into adaptive feedback scenarios.

REFERENCES

ZZ Akarslan, M Akdevelioglu, K Gungor, and H Erten. 2008. A comparison of the diagnostic accuracy of bitewing, periapical, unfiltered and filtered digital panoramic images for approximal caries detection in posterior teeth. *Dentomaxillofacial Radiology* 37, 8 (2008), 458–463.

Nicola C Anderson, Fraser Anderson, Alan Kingstone, and Walter F Bischof. 2015. A comparison of scanpath comparison methods. *Behavior research methods* 47, 4 (2015), 1377–1392.

Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. 2017. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision*. 2331–2338.

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

Ali Borji and Laurent Itti. 2014. Defending Yarbus: Eye movements reveal observers' task. *Journal of vision* 14, 3 (2014), 29–29.

Michael A Bruno, Eric A Walker, and Hani H Abujudeh. 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics* 35, 6 (2015), 1668–1676.

Michael Burch, Kuno Kurzhals, Niklas Kleinhans, and Daniel Weiskopf. 2018. EyeMSA: exploring eye movement data with pairwise and multiple sequence alignment. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ACM, 52.

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.

Nora Castner, Enkelejda Kasneci, Thomas Kübler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, and Constanze Keutel. 2018a. 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.

Nora Castner, Solveig Klepper, Lena Kopnarski, Fabian Hüttig, Constanze Keutel, Katharina Scheiter, Juliane Richter, Thérése Eder, and Enkelejda Kasneci. 2018b. Overlooking: the nature of gaze behavior and anomaly detection in expert dentists. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*. ACM, 8.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

Arzu Çöltekin, Sara Irina Fabrikant, and Martin Lacayo. 2010. Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. *International Journal of Geographical Information Science* 24, 10 (2010), 1559–1575.

Florence Corpet. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic acids research* 16, 22 (1988), 10881–10890.

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.

Alan Davies, Gavin Brown, Markel Vigo, Simon Harper, Laura Horseman, Bruno Splendiani, Elspeth Hill, and Caroline Jay. 2016. Exploring the relationship between eye movements and electrocardiogram interpretation accuracy. *Scientific reports* 6 (2016), 38227.

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.

Fabian Deitelhoff, Andreas Harrer, and Andrea Kienle. 2019. The influence of different AOI models in source code comprehension analysis. In *Proceedings of the 6th International Workshop on Eye Movements in Programming*. IEEE Press, 10–17.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Chester W Douglass, Richard W Valachovic, Anila Wijesinha, Howard H Chauncey, Krishan K Kapur, and Barbara J McNeil. 1986. Clinical efficacy of dental radiography in the detection of dental caries and periodontal diseases. *Oral Surgery, Oral Medicine, Oral Pathology* 62, 3 (1986), 330–339.

Wolfgang Fuhl, Efe Bozkir, Benedikt Hosp, Nora Castner, David Geisler, Thiago C Santini, and Enkelejda Kasneci. 2019. Encodji: encoding gaze data into emoji space for an amusing scanpath classification approach. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. ACM, 64.

Filippo Galgani, Yiwen Sun, Pier Luca Lanzi, and Jason Leigh. 2009. Automatic analysis of eye tracking data for medical diagnosis. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, 195–202.

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.

Martin Gelfand, Eric J Sunderman, and Melvin Goldman. 1983. Reliability of radiographical interpretations. *Journal of endodontics* 9, 2 (1983), 71–75.

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.

Michelle R Greene, Tommy Liu, and Jeremy M Wolfe. 2012. Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision research* 62 (2012), 1–8.

Thorsten Grünheid, Dustin A Hollevoet, James R Miller, and Brent E Larson. 2013. Visual scan behavior of new and experienced clinicians assessing panoramic radiographs. *Journal of the World Federation of Orthodontists* 2, 1 (2013), e3–e7.

Amin Haji-Abolhassani and James J Clark. 2014. An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision research* 103 (2014), 127–142.

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.

Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. 2015. Online tracking by learning discriminative saliency map with convolutional neural network. In *International conference on machine learning*. 597–606.

Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 262–270.

Fabian Huettig and Detlef Axmann. 2014. Reporting of dental status from full-arch radiographs: Descriptive analysis and methodological aspects. *World Journal of Clinical Cases: WJCC* 2, 10 (2014), 552.

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.

Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32, 3 (1967), 241–254.

Christopher Kanan, Nicholas A Ray, Dina NF Bseiso, Janet H Hsiao, and Garrison W Cottrell. 2014. Predicting an observer's task using multi-fixation pattern analysis. In *Proceedings of the symposium on eye tracking research and applications*. ACM, 287–290.

A Ben Khedher, Imène Jraidi, and Claude Frasson. 2018. Local sequence alignment for scan path similarity assessment. *International Journal of Information and Education Technology* 8, 7 (2018).

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.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

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.

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.

Benjamin Law, M Stella Atkins, Arthur E Kirkpatrick, and Alan J Lomax. 2004. Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In *Proceedings of the 2004 symposium on Eye tracking research & applications*. ACM, 41–48.

Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods* 45, 1 (2013), 251–266.

Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. 2015. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 362–370.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

Abhijit Mishra and Pushpak Bhattacharyya. 2018. Automatic extraction of cognitive features from gaze data. In *Cognitively Inspired Natural Language Processing*. Springer, 153–169.

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.

Thies Pfeiffer, Patrick Renner, and Nadine Pfeiffer-Lessmann. 2016. EyeSee3D 2.0: Model-based real-time analysis of mobile eye-tracking in static and dynamic three-dimensional scenes. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications*. ACM, 189–196.

Anna Rozenshtein, Gregory DN Pearson, Sherry X Yan, Andrew Z Liu, and Dennis Toy. 2016. Effect of massed versus interleaved teaching method on performance

of students in radiology. *Journal of the American College of Radiology* 13, 8 (2016), 979–984.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Temple F Smith, Michael S Waterman, et al. 1981. Identification of common molecular subsequences. *Journal of molecular biology* 147, 1 (1981), 195–197.

Y. Tao and M. Shyu. 2019. SP-ASDNet: CNN-LSTM Based ASD Classification Model using Observer ScanPaths. In *2019 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*. 641–646. https://doi.org/10.1109/ICMEW.2019.00124

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.

Stephen Anthony Waite, Arkadij Grigorian, Robert G Alexander, Stephen Louis Macknik, Marisa Carrasco, David Heeger, and Susana Martinez-Conde. 2019. Analysis of perceptual expertise in radiology–Current knowledge and a new perspective. *Frontiers in human neuroscience* 13 (2019), 213.

Qianwen Wan, Srijith Rajeev, Aleksandra Kaszowska, Karen Panetta, Holly A Taylor, and Sos Agaian. 2018. Fixation oriented object segmentation using mobile eye tracker. In *Mobile Multimedia/Image Processing, Security, and Applications 2018*, Vol. 10668. International Society for Optics and Photonics, 106680D.

Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2015. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3183–3192.

Julia M West, Anne R Haake, Evelyn P Rozanski, and Keith S Karn. 2006. eyePatterns: software for identifying patterns and similarities across fixation sequences. In *Proceedings of the 2006 symposium on Eye tracking research & applications*. ACM, 149–154.

Supplemental Material

Supplementary Figures

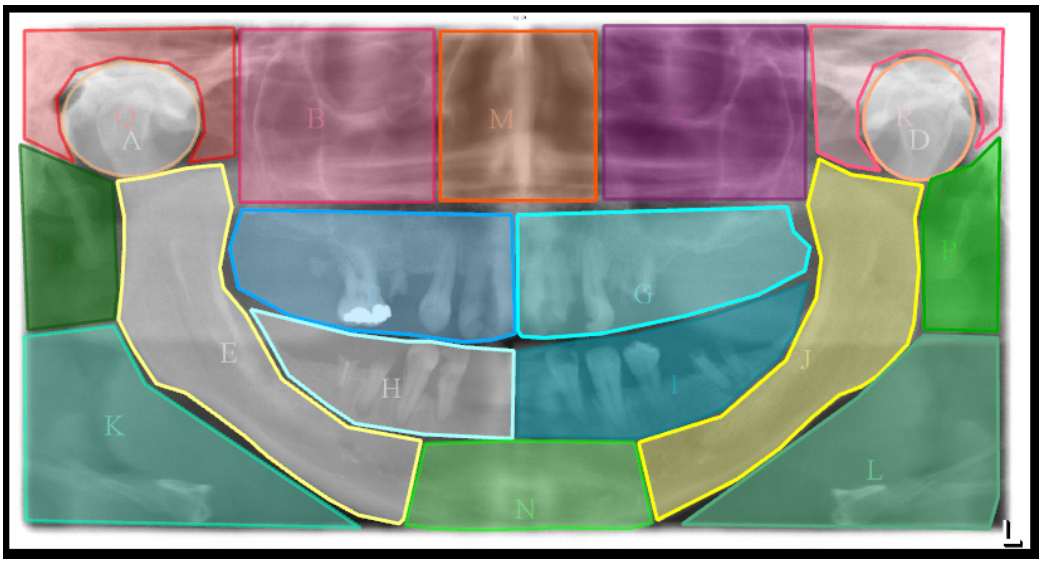


Figure 1: Manually defined Semantic AOIs for one OPT. AOI names are the same across all ten OPTs and correspond to a respective structure or region. Fixations in AOIs were then encoded to strings and used as input for the traditional Smith-Waterman sequence alignment.

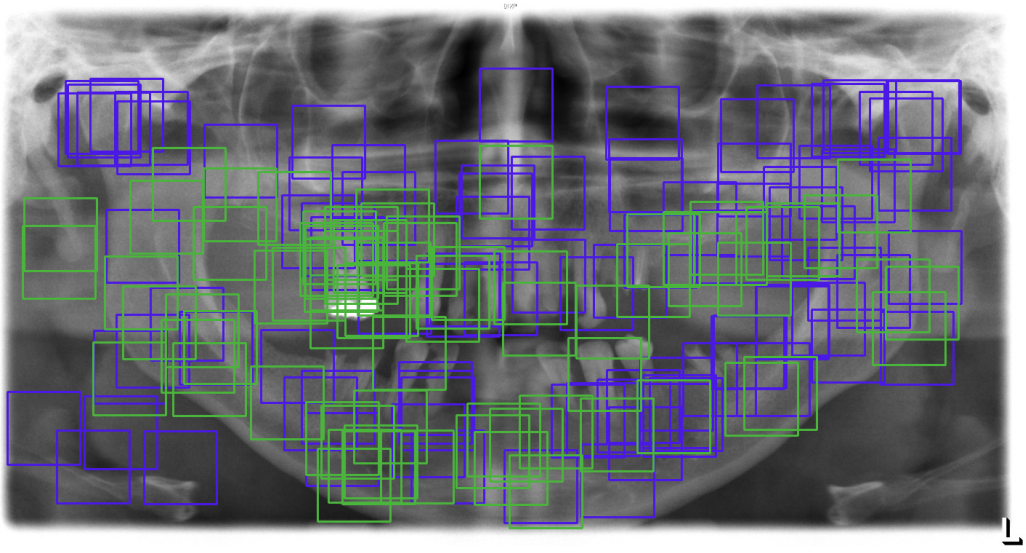


Figure 2: Two experts scanpath on image 1 (one scanpath in green, the other scanpath in blue) with highly similar scanpaths to themselves as well as many other subjects.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2020) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Towards expert gaze modeling and recognition of a user's attention in realtime

Nora Castner[a,b,*], Lea Geßler[b], David Geisler[a,b], Fabian Hüttig[c], Enkelejda Kasneci[a,b]

[a] *Human-Computer Interaction, University of Tübingen, Germany*
[b] *Institute of Computer Science, University of Tübingen, Germany*
[c] *Department of Prosthodontics, University Hospital Tübingen, Germany*

Abstract

One of the appealing areas of expertise research is devoted to measuring the effectiveness of training programs for novices. With recent progress in eye tracking, gaze-based interaction systems recognize a user's attention and can direct it accordingly. Moreover, dynamic visualization of an expert gaze model facilitates novice training by guiding the gaze to relevant areas. In addition, the system should be aware of realtime attention to remove an overlay that could occlude relevant information. We use an implementation of subtle gaze direction (SGD) and the simplified scanpath of a dentist to train naive participants in finding anomalies in dental radiographs. We were able to effectively direct user gaze to relevant image features without occluding the area when attention was recognized. Additionally, participants reported that the intervention was helpful for image inspection. The results of the model intervention show minimal improvements in anomaly detection, which is expected of naive subjects. We advocate that the system has the potential to be highly effective for advanced students and trainees with a certain foundation of conceptual knowledge.

Keywords: Eye Tracking; Gaze-based Interaction; Attention Guiding; Expertise; Learning

1. Introduction

Pervasive eye tracking provides a rich source of input to systems regarding a user's attention [6, 43]. As the interpretation of attention is often open to debate, we restrict the current work's definition to the visual attention aspect, as measured by the gaze locations over time for a given stimuli. Thus, gaze aware systems can detect user attention to certain areas at a given time and give customized support for the current task in a way that exploits natural human behavior, e.g. scanning a scene [29, 45]. One area that has shown promising applicability of these systems is intelligent tutoring systems. They cater to the user by offering adaptability and personalized feedback.

* Corresponding author. Tel.: +49-(0)7071-29-70492.
E-mail address: nora.castner@uni-tuebingen.de

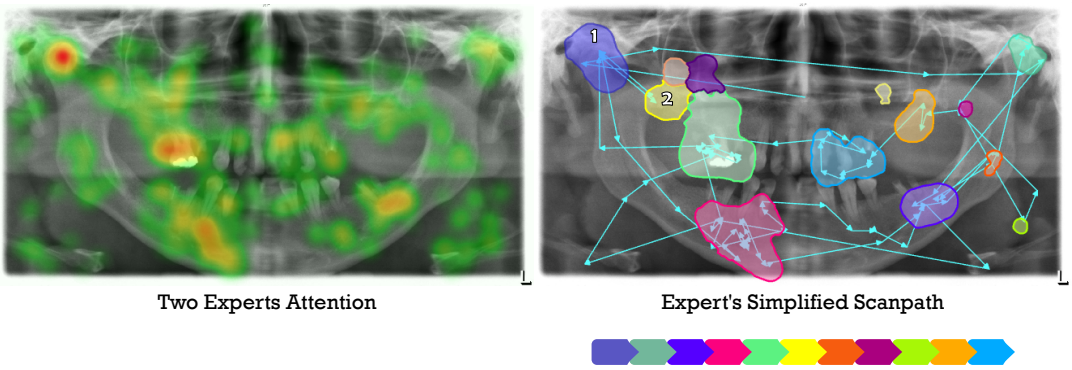


Fig. 1: Gaze guiding through experts' attention. AOIs are calculated from the heatmap. Then, the simplified transitional behavior becomes the order of presentation.

Using students' realtime gaze behavior to register attentional information has shown to successfully serve as input to adapt training systems in online learning portals [44, 40, 3, 7]. More important, they can augment traditional teaching approaches for effective visual inspection, by better breaking down complex imagery based on the realtime attentional information [41, 31]. One field that can highly benefit from gaze-aware tutoring is medical image inspection.

Medical experts are accountable for a high degree of sensitivity and specificity in diagnoses, since proper patient care is at stake. Visual search strategies can exemplify their perceptual expertise, where successful verification can rely on the slightest changes in the image features [39, 26, 20, 25, 47, 21]. In diagnostic radiology, experts initially perceive abnormalities faster and can also better discriminate what is irrelevant and what can indicate a pathology compared to novices [26, 22, 48].

Training perceptual expertise in radiology is a key component in novice training. For radiograph images, directed training in high volume has shown increased perceptual sensitivity in low-contrast target recognition [46] and semantic target recognition [11]. However, this approach becomes comparable to traditional massed practice approaches that require time and numerous images. Research has drawn attention to using gaze models for students or trainees [31, 30, 17, 16] to improve their perception to relevant features and thus streamline the learning process. Though proficient performance can be obtained, it has also been stressed that conceptual knowledge accelerates proper diagnostic interpretation [19, 12, 30].

Our work combines domains that have been previously running in parallel: Expert gaze modelling for learning and user-attention awareness. We designed a framework for gaze guiding based on expert viewing behavior on dental radiographs while recognizing a user's real-time gaze. Our interests are two-fold, 1) whether we can effectively guide a user's gaze to relevant regions of an image without occluding any information and 2) whether expert gaze guiding can improve perceptibility of anomaly features for non-experts. We present an exploratory evaluation of the intervention design with naive participants and assess its efficacy by its ability to guide the gaze unobtrusively and from usability feedback. Additionally, we look at detected anomaly features; however, we are aware that diagnostic performance would be more appropriately evaluated with students and advanced trainees, who have a more appropriate skill set for pathology interpretation.

2. Related Work

Gaze-based systems can offer an array of methods to visualize either a user's gaze in realtime or a gaze guiding model. Gaze contingency is visualizing a user's gaze, e.g. spotlighting, unmasking or unblurring, etc. [15, 38, 37]. For example, [28] used a white ring to indicate expert dentists' online gaze while viewing periapical (One tooth/ region of teeth) radiographs. Recognizing areas that were previously attended to and occluding them was shown to reduce workload during target detection [41]. Conversely, [14] found no effect on target detection accuracy or response time in a search task when users could see their own fixated or "yet to be fixated" regions overlayed in a colored translucent grid form. Moreover, they conclude that their protocol assumes a target will be detected if it is

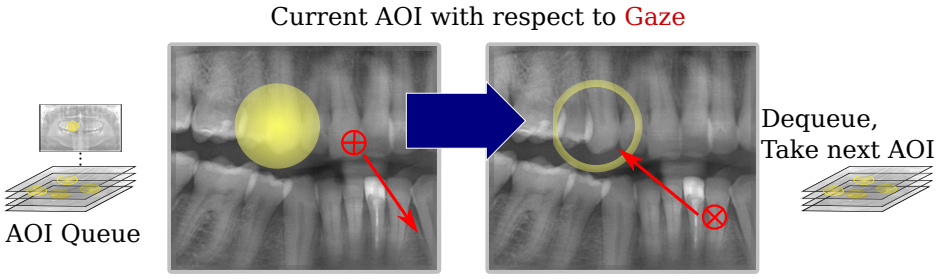


Fig. 2: Illustration of feedback animation. When the gaze attention (red cross-hair) is not directed towards the AOI, it pops up as a semi-translucent yellow circle (left image). When the gaze attention goes towards or is in the AOI, it presents the feedback as a translucent yellow ring (right image).

fixated on; it does not account for errors in target interpretation [14]. Regarding interpretation, tasks involving domain expertise can lead to recognition or decision errors (e.g. false negatives), where novices were more prone to these errors [34, 5, 1]. Concerning novice training, [42] found that blurring novice basketball players' periphery improved their decision making performance and that performance was even stable longitudinally. Specifically in medical image inspection, [35] found that gaze contingent windows on a chest radiograph acted as a guide for foveal vision towards inconspicuous nodules in the periphery, improving time to detection in experts. To our knowledge, gaze contingent feedback in novice medical image inspection has yet to be investigated.

Illustrating an expert's gaze as opaque yellow dots guided students to relevant features of the task [31]. Similarly, blurring areas where experts did not attend guides students' gaze without occluding relevant features [30]. Showing an experts' gaze dynamically as red dots coupled with think-aloud protocols was used for training both novice and experts in radiograph interpretation [19]. However, these aforementioned systems passively display an expert model and lack the user awareness and interaction.

Gaze-guiding systems that adapt to a user's gaze online have also been shown to be effective [38, 18, 41]. However, illustrating the user's gaze or a region of interest in a salient fashion can lead to feature occlusion that could hinder decision making in certain detection tasks. In order to not occlude potentially relevant regions, [2] proposed subtle gaze direction (SGD). It guides gaze during visual search by manipulating either the luminance or color and present it in the subject's periphery. When a saccade is detected towards the area of interest, the masking is removed [2]. Similarly, SGD with flickering in the periphery has been successful in gaze-guiding [49].

We employ a version of the SGD using areas of expert attention when examining panoramic dental radiographs (OPTs). We chose expert attentional areas because efficient anomaly detection is apparent in their search strategies. Among experts, scanpath variability can be high as they tend to employ their own "short cuts" [33, 32, 36] However, it has been found that relativity similar scanpaths can indicate correct medical image interpretation [13]. Therefore, our model is an expert's scanpath with respect to areas of attention of theirself and another highly similar expert scanpath. We present our SGD implementation with this expert model to non-experts.

3. Methods

Participants. We recruited 27 (20 male, M_{age} = 28.4) participants. Their backgrounds were mainly computer science (13). However, one was a medical assistant and another was a paleo-anthropologist. Both had more experience with general human anatomy and some radiology, though not specifically dentistry. 11 participants wore glasses during the experiment.

Experimental Paradigm. Prior to the experiment, all participants were debriefed regarding eye-tracking and the general protocol and signed a consent form. At the end, they were asked to fill out a brief questionnaire regarding the task difficulty, the gaze feedback, usability etc.

A five-point calibration with four point validation was performed for each subject at the beginning and in the middle of the experiment after a short pause. We followed the same experimental paradigm that can be found in [8]. Participants saw ten panoramic dental radiographs (OPTs). Each OPT was presented twice subsequently: First for

90 seconds, where they were instructed to inspect the image and then again where they could mark any areas they perceived as an anomaly. For each participant, we randomly determined which five OPTs would show the gaze-guiding; the other OPTs provided no feedback. This way we could compare within-subjects, whether the feedback had an effect. The second presentation of the OPTs had unlimited time for participants to mark detected anomalies at their own pace. A chin rest was used to assure stable gaze signal.

Expert Ground Truths. The OPTs were taken from Castner et al. [9, 8] and had pre-determined ground-truth anomaly information from two dentists involved in the project. The ground truth data was used to calculate the anomaly detection performance of the current participants. An anomaly was labeled as detected (true positive) if the participant marked the respective area of a ground truth anomaly. False negatives and false positives were if the participant did not mark a specific ground truth anomaly area or marked an area where no anomaly was present, respectively (see Castner et al. [9] for further details on the detection performance protocol).

To create the areas of interest (AOIs), we chose gaze data from two experts from a previous data collection with expert OPT inspection. Experts from this data collection had an average of 10 years of experience. Through similarity clustering, two experts were found to have scanpaths highly similar to all other experts' scanpath (see [10] for further details); their data was chosen to develop the expert model. From their heatmap, areas with higher concentration of gaze are segmented as illustrated in the right image in figure 1. We chose the scanpath of the more accurate (higher detected anomalies) of the two experts to provide transitional behavior. We preferred a simplified version of the transition, denoting the first glance into an AOI and not revisits, since it was determined that revisits would be too hard to follow. An example of a simplified scanpath is also found in figure 1: The first blue AOI is looked at (1) then transitions to four other AOIs were made before going back to the first AOI, we omit the revisit and set the next transition to the yellow AOI (2). Without revisits, scanpaths ranged from 9 to 23 transitions, and with revisits, they ranged from 88 to 175 transitions.

Software. We based the experiment software off the experiment designer and gaze-contingent feedback developed in [38]. This software already has the usability for presenting image stimuli for either a set time or key-press interrupt. We added an on-screen drawing tool, so we could gather the anomaly detection recall and precision of the non-experts. We also added the ability to upload customized feedbacks with AOI positions as csv-files.

We incorporated the AOIs and the ability to recognize attention towards them; Our method is based off the subtle gaze direction (SGD) method from [2]. We added a short delay of 5 seconds, before the first AOI pops up, so participants could scan the image shortly.

AOIs for a certain feedback are placed into a queue. Upon an animation timer timeout, the current AOI is dequed and painted over the stimulus. For this work, we set the timer to timeout every 3.8 seconds so participants would not feel rushed, as they were non-experts. The AOI is initially illustrated as yellow ($RGB : 252, 252, 103$) with a translucent radial gradient (left image in figure 2). We chose this color as we felt it would be salient against our grayscale stimuli.

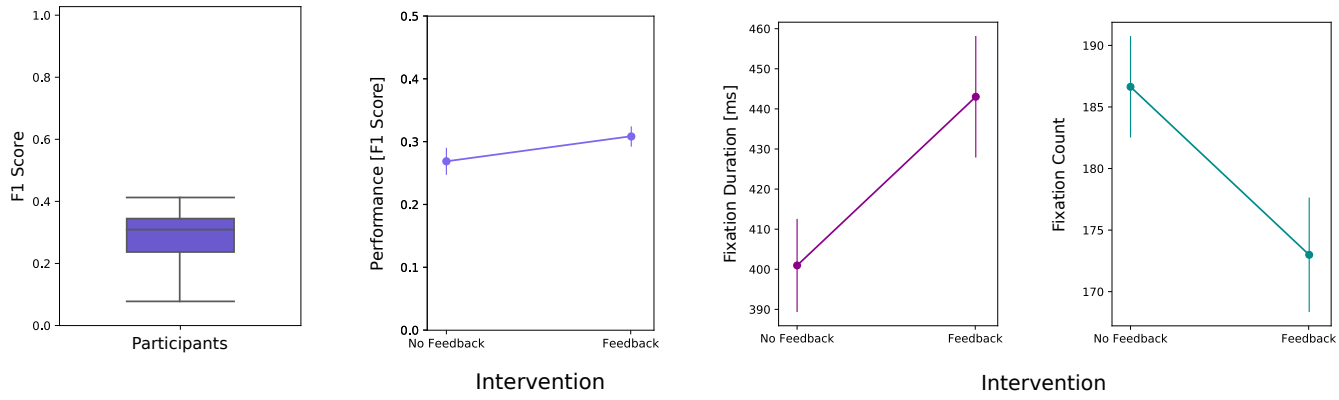
In order to avoid occlusion of important image features, we repaint the AOI area with a translucent yellow ring (right image in figure 2), when our SGD implementation detects the gaze angle as going towards the AOI. Where the angle, α, is calculated as follows:

$$\alpha = \cos^{-1}\left(\frac{\vec{v}\cdot\vec{t}}{|\vec{v}|\cdot|\vec{t}|}\right), \tag{1}$$

where $\vec{v}$ indicates the vector from the previous gaze point to the current gaze point and $\vec{t}$ indicates the vector from the previous gaze point to the target AOI. We calculate α five times using equation 1: with one $\vec{t}$ to center coordinates of the AOI and then $\vec{t}$ for each of the corner coordinates of its bounding box. We calculate the previous gaze as the average of the last two gaze coordinates stored in a buffer. We take the minimum of the five angles and subtract it from 360° if it is larger than 180°.

Then, if α is between 0 and 10°, the AOI updates from the circle to the ring. This threshold was used in [2], and was determined stable when testing our implementation. For gaze input, we used the SMI RED250 remote eye tracker running at 60Hz.

4. Results



(a) Non-Expert Performance (b) Performance with Intervention (c) Gaze Behavior with Intervention

Fig. 3: Performance as measured by the F1 Score (a) overall images (b) comparing the intervention of expert gaze feedback against no feedback, and (c) the gaze behavior with respect to feedback or no feedback. Means (circles) and standard errors (tails) are plotted for all figures.

Performance and Gaze. We calculated the sensitivity and precision of the participants over all images, then calculated the harmonic mean (F1 score) between the metrics. Sensitivity is the true positive rate (TPR), precision is the positive predictive value (PPV). The F1 score is $2 \cdot (PPV \cdot TPR/PPV + TPR)$. One participant's performance was omitted upon learning that they did not understand the instructions given at the beginning. As was expected with non experts, performance in OPT anomaly detection was relatively low: The average F1 score overall was $M = 28.42\%, SD = 8.45$. The distribution is shown in figure 3a.

To see if there were any effects of the expert gaze feedback intervention, we ran a repeated measures t-test on both the performance and the gaze behavior for "feedback" versus "no feedback" conditions. No major effect was found for the intervention on performance ($t(26) = -2.021, p = 0.054$), with the performance with the feedback was slightly better ($M = 30.80\%, SD = 8.23$) than without the feedback ($M = 26.85\%, SD = 10.97$.). Figure 3b shows the performance with respect to the intervention.

However, the intervention had a stronger effect on the gaze behavior. Average fixation durations were higher for the feedback condition ($M = 443.03, SD = 78.76$) compared to the no feedback condition ($M = 400.96ms, SD = 60.29$, $t(26) = -4.704, p < 0.0001$). Additionally, the average fixation count for the feedback condition was lower ($M = 173.0, SD = 24.15$) than the no feedback condition ($M = 186.64, SD = 21.41$, $t(26) = 4.502, p = 0.00012$).

Attention to AOIs. To assess whether the intervention successfully guided the gaze behavior, we looked at subjects' gaze behavior in relation to the AOIs as shown in figure 4. We ran repeated measures t-test for AOI glances and transition similarity.

We looked at the effect of the intervention on the AOI glances. We measure AOI glances as the proportion of a glance on an AOI in relation to the total AOIs from the expert model. We found that with the feedback, subjects had significantly higher proportion of glances ($M = 0.8359, SD = 0.0935$) than without the feedback ($M = 0.7060, SD = 0.0863$, $t(26) = -8.165, p < 0.0001$).

We looked at the effect of the intervention on the similarity of subject's AOI transitions to the expert's transition. Similarity was calculated with the levenshtein distance [24] for subjects' scanpaths compared to the expert's scanpath and normalized to the length of the longest scanpath. We found that with the feedback, subjects had significantly more similarity to the expert ($M = 0.7203, SD = 0.072$) than without the feedback ($M = 0.7937, SD = 0.0416$,

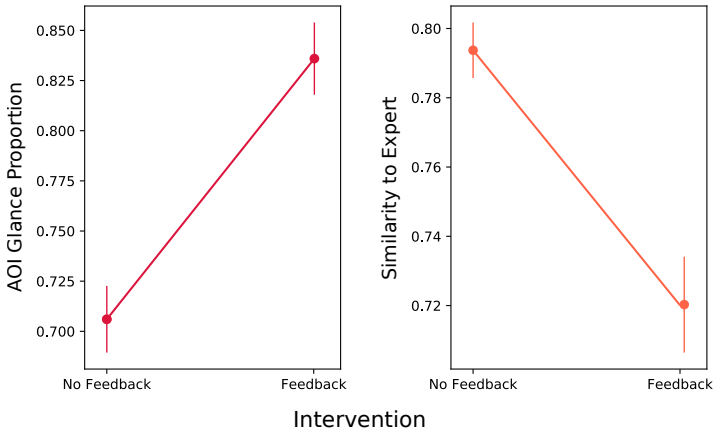


Fig. 4: Performance as measured by the F1 Score for each image with respect to intervention.

$t(26) = 4.791, p < 0.0001$). Figure 5 shows the transitional information for one image of subjects with (middle) and without (right) the intervention compared to the expert's gaze transitions relative to the AOIs (Left). Here, it is evident that the similarity is of the subjects who received the gaze feedback is closer to the expert's gaze behavior than the subjects who received no feedback: Note the transitions to (lines originating) and from (lines landing) AOI 5 (burgundy).

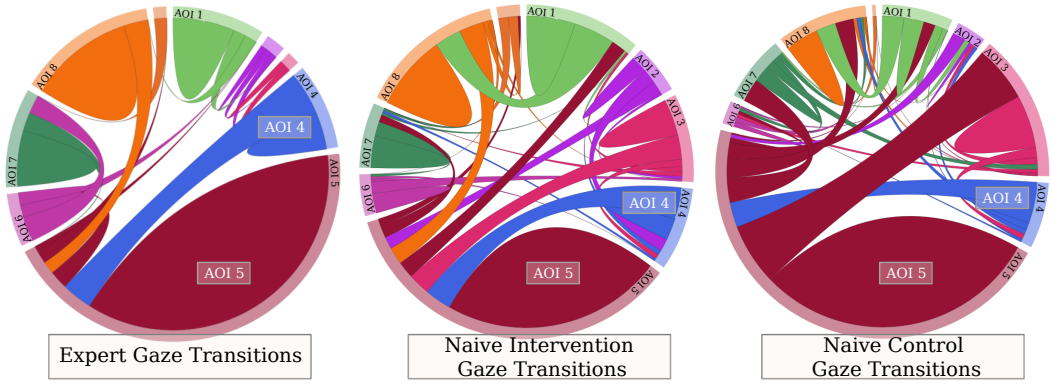


Fig. 5: Example of AOI transitions for one image. Where the left most diagram is the expert's transitional information and the middle is the transitional information of subjects who received the gaze intervention and the right most is the transitional information of subjects who received no gaze intervention .

User response. Regarding usability, we asked subjects to fill out a short questionnaire about the task and the gaze feedback. Average responses for the questions are plotted in figure 6. Overall, the subjects found the task difficult and were not confident in their performance. This could be expected as the nature of anomalies in these images are likely to be very subtle to the untrained eye. Moreover, they were overall positive regarding the intervention, finding it beneficial and depending on it to complete the task. Some participants made informal comments to the researchers that, after a few images with interventions, they started to recognize features (e.g. dark shadows in the gums), which they felt could be indicative of something abnormal (peridontitis). They did however find the task a bit too long and slightly rushed. These responses will be helpful for future testing and system development.

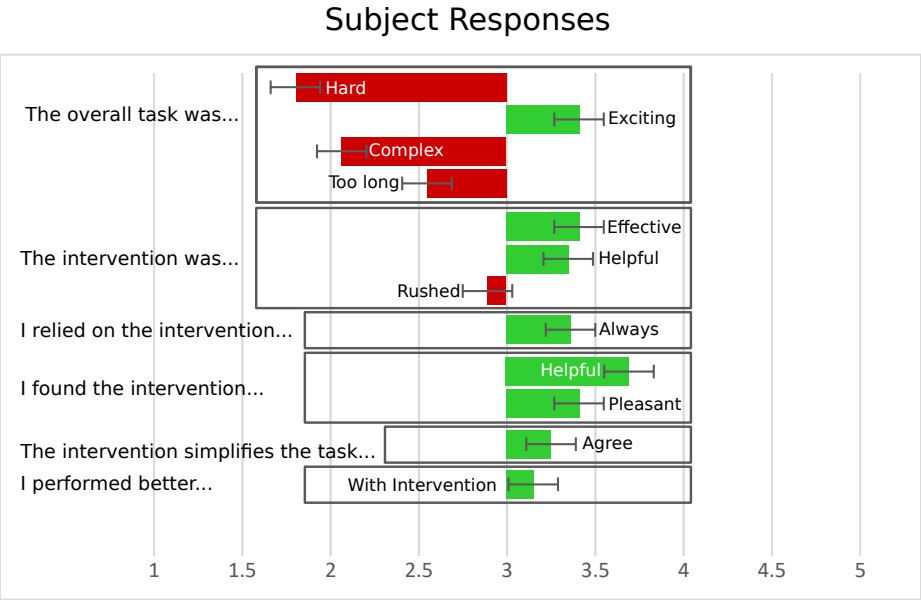


Fig. 6: Average responses for questionnaire regarding the task and the gaze intervention. The task was reported as difficult for the non-experts. They did report that the feedback was helpful and they used it.

5. Discussion

Overall, there were significant differences in the gaze behavior. When presented with the expert gaze model, participants exhibited fewer fixations, but longer fixation durations. This behavior could be indicative of more information processing and associated with novices [20, 32, 33, 27]. Additionally, the gaze model elicited a higher proportion of AOI glances. Therefore, there was more attention to relevant-areas of the image. However, subjects did not detect anomalies in dental radiographs with high accuracy. The expert gaze model intervention did not significantly improve performance compared to no intervention at all. One reason for this finding could be the low sample size. Additionally, this low sample size could explain the high variance in the gaze behavior for both the intervention and no intervention condition. Further research with an appropriate sample size to observe a significant difference is necessary.

Moreover, participants reported feeling more confident with the gaze intervention and relied on it to complete the task. They successfully followed the expert gaze model and were more similar to the expert's AOI transitional behavior. Although, they lacked the conceptual knowledge that facilitates the proper interpretation of the relevant features. Previous research has also indicated that search pattern training draws attention to relevant areas, but did not affect performance [33, 19, 23]. Waite et al. [48] highlights the reciprocity of perception and cognition in diagnostic performance: For instance, initial feature localization, then conceptual knowledge facilities the decision that this feature needs further inspection (e.g. difference in contrast, and area prone to anomalies, etc.) and whether it is recognized as a specific pathology or could be ruled out.

It should be noted that dental radiographs, as with all medical images, are highly complex in nature and require some form of conceptual knowledge to interpret reliably. Presenting only ten OPTs may not have been enough for a significant training effect. Considering the low number of OPTs, the naive participants seemed to recognize features the intervention highlights in later images as they reported. To get improved performance in naive observers, [11] used around 800 images to improve hip fracture detection. Further research is needed that addresses the optimal amount of images needed to improve interpretation, without inducing fatigue while still providing ample time to interact with the gaze model. In our study, we were limited to investigating short term effects of training naive participants. A longitudinal study regarding the gaze-aware feedback system on naive subjects' or novices' learning overtime would be an interesting aspect for further research. Furthermore, the notion of implicit feature learning is also interesting for future work. Beesley and colleagues [4] found gaze contingency aided in implicit rule learning. Staggering training sets of

certain types of anomalies and expert gaze behavior related to them may improve detecting the features indicative of these anomalies.

Moreover, we show a potentially effective learning intervention for either novices or more advanced dentists. Students undergo intense studying and exposure to get to the level of professional expertise that makes them successful later in their careers. More effective learning interventions can smooth the transition of students to residency and professional environments by minimizing the knowledge gap between each stage. With better preparation, less professional resources need to be expended on supervising incoming residents and early professionals. Even then, *expert* is never a final state, but should always be open for further learning and improving. Generally, it has been found that experts and more advanced trainees benefit highly from gaze interventions [19, 21]. Our implementation of the SGD with expert AOIs could also potentially be catered to advanced learners, in hopes to further fine-tune established skills.

6. Conclusion

We employ subtle gaze direction to present expert attention while examining panoramic dental radiographs. Our method does not occlude relevant areas in the foveal vison, as it recognizes when attention is directed towards the area. We could successfully guide the gaze to relevant image features and promoted further inspection. Our findings with naive participants showed that the gaze feedback could not develop successful dental radiograph diagnosis, but elicited gaze transitions similar to the expert model. They also felt more confident and that the framework helped them properly inspect radiographs. This aspect suggests further research to promote SGD as a suitable way to illustrate expert gaze behavior in learning interventions with students or advanced trainees.

Acknowledgements

Dr. med. Dr. med. dent. Constanze Keutel from the Department of Radiology, Center of Dentistry, Oral Medicine and Maxillofacial Surgery at the University Hospital Tübingen for their expertise and help overseeing the project with Dr. Hüttig.

References

[1] Baghdady, M.T., Carnahan, H., Lam, E.W., Woods, N.N., 2014. Dental and dental hygiene students' diagnostic accuracy in oral radiology: effect of diagnostic strategy and instructional method. Journal of dental education 78, 1279–1285.

[2] Bailey, R., McNamara, A., Sudarsanam, N., Grimm, C., 2009. Subtle gaze direction. ACM Transactions on Graphics (TOG) 28, 1–14.

[3] Barrios, V.M.G., Gütl, C., Preis, A.M., Andrews, K., Pivec, M., Mödritscher, F., Trummer, C., 2004. Adele: A framework for adaptive e-learning through eye tracking. Proceedings of IKNOW , 609–616.

[4] Beesley, T., Pearson, D., Le Pelley, M., 2015. Implicit learning of gaze-contingent events. Psychonomic bulletin & review 22, 800–807.

[5] Bruno, M.A., Walker, E.A., Abujudeh, H.H., 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 35, 1668–1676.

[6] Bulling, A., 2016. Pervasive attentive user interfaces. Computer , 94–98.

[7] Calvi, C., Porta, M., Sacchi, D., 2008. e5learning, an e-learning environment based on eye tracking, in: 2008 Eighth IEEE International Conference on Advanced Learning Technologies, IEEE. pp. 376–380.

[8] Castner, N., Kasneci, E., Kübler, T., Scheiter, K., Richter, J., Eder, T., Hüttig, F., Keutel, C., 2018a. 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. p. 39.

[9] Castner, N., Klepper, S., Kopnarski, L., Hüttig, F., Keutel, C., Scheiter, K., Richter, J., Eder, T., Kasneci, E., 2018b. Overlooking: the nature of gaze behavior and anomaly detection in expert dentists, in: Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, ACM. p. 8.

[10] Castner, N., Kübler, T.C., Scheiter, K., Richter, J., Eder, T., Hüttig, F., Keutel, C., Kasneci, E., 2020. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing, in: Eye Tracking Research and Applications, ACM.

[11] Chen, W., HolcDorf, D., McCusker, M.W., Gaillard, F., Howe, P.D., 2017. Perceptual training to improve hip fracture identification in conventional radiographs. PloS one 12.

[12] Coderre, S., Mandin, H., Harasym, P.H., Fick, G.H., 2003. Diagnostic reasoning strategies and diagnostic success. Medical education 37, 695–703.

[13] Davies, A., Brown, G., Vigo, M., Harper, S., Horseman, L., Splendiani, B., Hill, E., Jay, C., 2016. Exploring the relationship between eye movements and electrocardiogram interpretation accuracy. Scientific reports 6, 38227.

[14] Drew, T., Williams, L.H., 2017. Simple eye-movement feedback during visual search is not helpful. Cognitive Research: Principles and Implications 2, 1–8.
[15] Duchowski, A.T., Cournia, N., Murphy, H., 2004. Gaze-contingent displays: A review. CyberPsychology & Behavior 7, 621–634.
[16] Eder, T.F., Richter, J., Scheiter, K., Keutel, C., Castner, N., Kasneci, E., Huettig, F., 2020. How to support dental students in reading radiographs: effects of a gaze-based compare-and-contrast intervention. Advances in Health Sciences Education: Theory and Practice .
[17] Fichtel, E., Lau, N., Park, J., Parker, S.H., Ponnala, S., Fitzgibbons, S., Safford, S.D., 2019. Eye tracking in surgical education: gaze-based dynamic area of interest can discriminate adverse events and expertise. Surgical endoscopy 33, 2249–2256.
[18] Foulsham, T., Underwood, G., 2011. If visual saliency predicts search, then why? evidence from normal and gaze-contingent search tasks in natural scenes. Cognitive Computation 3, 48–63.
[19] Gegenfurtner, A., Lehtinen, E., Jarodzka, H., Säljö, R., 2017. Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis. Computers & Education 113, 212–225.
[20] Gegenfurtner, A., Lehtinen, E., Säljö, R., 2011. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. Educational Psychology Review 23, 523–552.
[21] Van der Gijp, A., Ravesloot, C., Jarodzka, H., van der Schaaf, M., van der Schaaf, I., van Schaik, J.P., Ten Cate, T.J., 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, 765–787.
[22] Van der Gijp, A., Van der Schaaf, M., Van der Schaaf, I., Huige, J., Ravesloot, C., Van Schaik, J., ten Cate, T.J., 2014. Interpretation of radiological images: towards a framework of knowledge and skills. Advances in Health Sciences Education 19, 565–580.
[23] van der Gijp, A., Vincken, K.L., Boscardin, C., Webb, E.M., Ten Cate, O.T.J., Naeger, D.M., 2017. The effect of teaching search strategies on perceptual performance. Academic radiology 24, 762–767.
[24] Goldberg, J.H., Helfman, J.I., 2010. Scanpath clustering and aggregation, in: Proceedings of the 2010 symposium on eye-tracking research & applications, pp. 227–234.
[25] Grünheid, T., Hollevoet, D.A., Miller, J.R., Larson, B.E., 2013. Visual scan behavior of new and experienced clinicians assessing panoramic radiographs. Journal of the World Federation of Orthodontists 2, e3–e7.
[26] Gunderman, R.B., Patel, P., 2019. Perception's crucial role in radiology education. Academic radiology 26, 141–143.
[27] Haider, H., Frensch, P.A., 1999. Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition 25, 172.
[28] Hermanson, B.P., Burgdorf, G.C., Hatton, J.F., Speegle, D.M., Woodmansey, K.F., 2018. Visual fixation and scan patterns of dentists viewing dental periapical radiographs: an eye tracking pilot study. Journal of endodontics 44, 722–727.
[29] Jacob, R.J., Karn, K.S., 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises, in: The mind's eye. Elsevier, pp. 573–605.
[30] Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., Eika, B., 2012. Conveying clinical reasoning based on visual observation via eye-movement modelling examples. Instructional Science 40, 813–827.
[31] Jarodzka, H., Scheiter, K., Gerjets, P., van Gog, T., Dorr, M., 2010. How to convey perceptual skills by displaying experts' gaze data, in: Proceedings of the 31st annual conference of the cognitive science society, pp. 2920–2925.
[32] Kok, E.M., De Bruin, A.B., Robben, S.G., Van Merrienboer, J.J., 2012. Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. Applied Cognitive Psychology 26, 854–862.
[33] Kok, E.M., Jarodzka, H., de Bruin, A.B., BinAmir, H.A., Robben, S.G., van Merriënboer, J.J., 2016. Systematic viewing in radiology: seeing more, missing less? Advances in Health Sciences Education 21, 189–205.
[34] Kundel, H.L., Nodine, C.F., Carmody, D., 1978. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Investigative radiology 13, 175–181.
[35] Kundel, H.L., Nodine, C.F., Toto, L., 1991. Searching for lung nodules. the guidance of visual scanning. Investigative radiology 26, 777–781.
[36] Nodine, C.F., Kundel, H.L., Mello-Thoms, C., Weinstein, S.P., Orel, S.G., Sullivan, D.C., Conant, E.F., 1999. How experience and training influence mammography expertise. Academic radiology 6, 575–585.
[37] Orlov, P.A., Bednarik, R., 2016. Screenmasker: An open-source gaze-contingent screen masking environment. Behavior research methods 48, 1145–1153.
[38] Otto, K., Castner, N., Geisler, D., Kasneci, E., 2018. Development and evaluation of a gaze feedback system integrated into eyetrace, in: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, pp. 1–5.
[39] Palmeri, T.J., Wong, A.C., Gauthier, I., 2004. Computational approaches to the development of perceptual expertise. Trends in cognitive sciences 8, 378–386.
[40] Parikh, S., Kalva, H., 2018. Eye gaze feature classification for predicting levels of learning, in: InProceedings of the 8th Workshop on Personalization Approaches in Learning Environments (PALE 2018).
[41] Qvarfordt, P., Biehl, J.T., Golovchinsky, G., Dunningan, T., 2010. Understanding the benefits of gaze enhanced visual search, in: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, pp. 283–290.
[42] Ryu, D., Mann, D.L., Abernethy, B., Poolton, J.M., 2016. Gaze-contingent training enhances perceptual skill acquisition. Journal of vision 16, 2–2.
[43] Santini, T., Brinkmann, H., Reitstätter, L., Leder, H., Rosenberg, R., Rosenstiel, W., Kasneci, E., 2018. The art of pervasive eye tracking: unconstrained eye tracking in the austrian gallery belvedere, in: Proceedings of the 7th workshop on pervasive eye tracking and mobile eye-based interaction, pp. 1–8.
[44] Sharma, K., Alavi, H.S., Jermann, P., Dillenbourg, P., 2016. A gaze-based learning analytics model: in-video visual feedback to improve learner's attention in moocs, in: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pp. 417–421.
[45] Sibert, L.E., Jacob, R.J., 2000. Evaluation of eye gaze interaction, in: Proceedings of the SIGCHI conference on Human Factors in Computing

Systems, pp. 281–288.
[46] Sowden, P.T., Davies, I.R., Roling, P., 2000. Perceptual learning of the detection of features in x-ray images: a functional role for improvements in adults' visual sensitivity? Journal of Experimental Psychology: Human perception and performance 26, 379.
[47] Turgeon, D.P., Lam, E.W., 2016. Influence of experience and training on dental students' examination performance regarding panoramic images. Journal of dental education 80, 156–164.
[48] Waite, S.A., Grigorian, A., Alexander, R.G., Macknik, S.L., Carrasco, M., Heeger, D., Martinez-Conde, S., 2019. Analysis of perceptual expertise in radiology–current knowledge and a new perspective. Frontiers in human neuroscience 13, 213.
[49] Waldin, N., Waldner, M., Viola, I., 2017. Flicker observer effect: Guiding attention through high frequency flicker in images, in: Computer Graphics Forum, Wiley Online Library. pp. 467–476.