LSTMs can distinguish dental expert saccade behavior with high "plaque-urracy"

Nora Castner
Jonas Frankemölle
Enkelejda Kasneci
nora.castner@uni-tuebingen.de
jonas.frankemoelle@student.uni-tuebingen.de
enkelejda.kasneci@uni-tuebingen.de
Human-Computer Interaction,
University of Tübingen
Tübingen, Germany

Constanze Keutel
Department of Radiology, Center of Dentistry, Oral Medicine and Maxillofacial Surgery,University Hospital Tübingen
Tübingen, Germany
constanze.keutel@med.uni-tuebingen.de

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

ABSTRACT

Much of the current expertise literature has found that domain specific tasks evoke different eye movements. However, research has yet to predict optimal image exploration using saccadic information and to identify and quantify differences in the search strategies between learners, intermediates, and expert practitioners. By employing LSTMs for scanpath classification, we found saccade features over time could distinguish all groups at high accuracy. The most distinguishing features were saccade velocity peak (72%), length (70%), and velocity average (68%). These findings promote the holistic theory of expert visual exploration that experts can quickly process the whole scene using longer and more rapid saccade behavior initially. The potential to integrate expertise model development from saccadic scanpath features into intelligent tutoring systems is the ultimate inspiration for our research. Additionally, this model is not confined to visual exploration in dental xrays, rather it can extend to other medical domains.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Computing methodologies** → **Neural networks**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Eye Tracking, Expertise, Scanpath analysis, Medical image interpretation, Deep Learning, intelligent tutoring systems

ACM Reference Format:
Nora Castner, Jonas Frankemölle, Enkelejda Kasneci, Constanze Keutel, and Fabian Hüttig. 2022. LSTMs can distinguish dental expert saccade behavior with high "plaque-urracy". In *2022 Symposium on Eye Tracking Research and Applications (ETRA '22), June 8–11, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3517031.3529631

ETRA '22, June 8–11, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9252-5/22/06…$15.00
https://doi.org/10.1145/3517031.3529631

1 INTRODUCTION

When we observe an expert in their element, we generally recognize that they are indeed experts by attributes such as performance and speed. But the underlying mechanics – the cognitive and procedural – that dictate success are less tangible. In an attempt to explain experts' superior performance in terms of cognitive ability, features such as intuition and effortlessness are put forth [Chi 2006; Dreyfus and Dreyfus 1986; Ericsson and Lehmann 1996; Polanyi 1962; Shanteau 1992]. The extent of an expert's intuition is manifested in their visual behavior, and many expert domains rely on effective visual processing [Brams et al. 2019; Gegenfurtner et al. 2011; Van der Gijp et al. 2017].

Specifically in medicine, visual inspection of medical images requires sensitivity to even the slightest feature aberration. Effective visual processing of medical images is asserted by the fixation behavior of experts, where fixation durations are shorter for experts compared to novices [Assaf et al. 2016; Fox and Faulkner-Jones 2017; Ganesan et al. 2018; Gegenfurtner et al. 2011; Van der Gijp et al. 2017]. Spending less time on the task as well as more rapid attention to relevant information (less fixations to obtain the desired result) are also characteristics of experts [Brams et al. 2019; Cooper et al. 2010; Krupinski 1996; Kundel et al. 2007; Mallett et al. 2014; Manning et al. 2006; Nodine et al. 1996; Van der Gijp et al. 2017; Warren et al. 2018]. This suggests that their experience and knowledge provides shortcuts that are more sophisticated than novice comprehension. For example, experts can view a radiograph for just a few milliseconds and tell if there was an anomaly present with extremely high accuracy [Brunyé et al. 2021; Evans et al. 2013; Feltovich et al. 2006]. Informally, we would refer to this quick understanding of the scene content as *getting the gist.*

Yet, there is some reality behind the impression that experts quickly know what is in the image and what further to look for. The *Holistic image processing* theory [Kundel et al. 2007] states that experts initially form a brief global sense of the problem by scanning the whole image. Then, they hone in on areas that require deeper investigation. Support for this theory in radiology has additionally found that expert search strategies employ a global-to-focal

order [Brams et al. 2020; Ganesan et al. 2018; Kok et al. 2012; Kundel et al. 2007; Van der Gijp et al. 2017]. This means that experts scan the outer periphery and central areas with sweeping saccades and short fixations and then scrutinize features with longer fixations and decreased saccadic amplitudes [Pannasch et al. 2008]. This behavior has also been linked to effective visual search in general [Parkhurst et al. 2002]. Novices, on the other hand, tend to exhibit primarily focal search behavior [Gandomkar and Mello-Thoms 2019]. They attend to more central and salient regions as exemplified by shorter saccade lengths and longer and more frequent fixations [Ganesan et al. 2018; Koide et al. 2015; Kok et al. 2012; Van der Gijp et al. 2017].

The information gained from expert visual search can convey to novices effective scene processing and attention to relevant features. Thus, expertise recognition through scanpath analysis is a crucial step toward effective gaze training. Research on scanpath classification for medical expertise is becoming more realized as a viable approach for teaching interventions. Robust recognition of a student's level of understanding through gaze can then accommodate the appropriate level of content for that student. This solution has the ability to smooth the transition between residency and professional environments for students by minimizing the knowledge gap. In the present work, we (1) distinguish scanpath differences between experts, intermediates and novices using a state of the art deep-learning classification algorithm (2) and observe search phase patterns in the temporal saccade features of these expertise levels.

2 SCANPATH ANALYSIS IN MEDICAL EXPERTISE

Research on the scanning patterns of medical experts has been able to determine specific strategies for certain images. For example, a circular pattern is preferred for mammogram inspection [Krupinski 1996; Kundel et al. 2007], spiraling outward for hand x-ray inspection [Hu et al. 1994], or drilling downwards in 3D Chest CTs [Drew et al. 2013; Mercan et al. 2018]. For dental radiographs, tooth-by-tooth and circular viewing patterns were evident depending on the nature of anomalies present in periapical projections [Hermanson et al. 2018]. However, for OPTs it was observed that spiraling inward (periphery areas first, then dental areas) and circular (going back and forth between central and periphery) techniques were preferred by more experienced clinicians [Grünheid et al. 2013]. Scanpath analysis can quantify these observed differences in visual search strategies and even discriminate advanced patterns linked to phases in the search.

Using string alignment for similarity assessment, expertise as well as performance promote more scanpath similarity when inspecting chest x-rays, ECGs, and brain scans [Crowe et al. 2018; Davies et al. 2018; Kok et al. 2016]. Classification based on similarity has shown high accuracy in distinguishing medical professionals from novices [Kübler and Kasneci 2015; Kübler et al. 2017]. Specifically for radiography examination, [Li et al. 2019] showed that experts had more similar patterns than novices when clustering experts and novices using contrast mining with temporal binning to extract subsequences of attentional behavior indicative of different strategies employed by both groups. However, Gandomkar et al. [Gandomkar et al. 2017, 2018] classified novice and experts using an SVM on RQA features and found expertise corresponded with more unique scanpath dynamics during mammogram inspection compared to novices [Gandomkar and Mello-Thoms 2019], which was further corroborated in expert inspection of orthopedic radiographs [Assaf et al. 2016]. These approaches in similarity extraction focus on the spatial aspect of the gaze behavior, often alluding to what experts are looking at and, more important, in what relevant order.

Spatial information from fixations is an important feature in scanpath analysis of medical experts. However, much like the temporal understanding of fixational behavior, saccade behavior over time can recognize patterns related to key intervals in expert visual search. There are growing efforts to incorporate saccade features (e.g. saccade velocities) to distinguish medical professionals from novices [Hosp et al. 2021; Li et al. 2012, 2016; Yin et al. 2020]. Concerning temporal saccade behavior, Li et al [Li et al. 2012, 2016] found expert dermatologists had longer fixation durations coupled with decreasing saccade amplitudes as an effect of viewing time compared to novices. Yet, classifying medical experts from the temporal saccade behavior using state-of-the-art scanpath classifiers has not been heavily investigated. Using saccades over fixations offers characteristics not bound to AOIs and robust to spatial offsets [Jarodzka et al. 2010]. Outside the medical domain, saccade angle patterns have been used successfully for task and reading classification [French et al. 2017; Fuhl et al. 2019; Kelton et al. 2019; Kunze et al. 2013].

To our knowledge, only one study used saccade behavior over time to classify expert and novice radiologists. Yoon et al. [Yoon et al. 2018] trained a CNN using gaze velocity profiles and were able to classify subjects at roughly 70%. In general, deep learning models are giving more traction to medical expert gaze classification. Castner et al. [Castner et al. 2020] proposed an approach that used image patches at the fixation level as input for a CNN to extract similar features linked to subsequences in dental expert and novice scanpaths. Expert similarity in the gaze behavior was linked to semantic features rather than stimulus specific regional information and achieved 73% accuracy. Another approach that can handle more sequential information in the input is recurrent neural networks (RNNs). One type of RNN that has been used recently in scanpath classification is Long-Short Term Memory (LSTM) [Tao and Shyu 2019]. These models have exhibited aptitude for handling time series data and forecasting [Shao and Soong 2016; Wang et al. 2017]. Their architecture better handles learning relevant information from long-term dependencies [Hochreiter and Schmidhuber 1997]. Sodoké et al. [Sodoké et al. 2020] used eye movement sequences related to AOIs during an intubation simulation as input for their CNN-LSTM for expert novice classification and achieved 84% accuracy.

3 PROPOSED APPROACH

3.1 Scanpath Data

The data is a subsample of participants from a larger cohort study that investigated the visual search strategies involved in dental OPT inspection and anomaly detection. Participants were sixth (n=58) and tenth (n=54) semester dental students, plus experienced dentists (n=26). We consider the sixth semester students as *novices*,

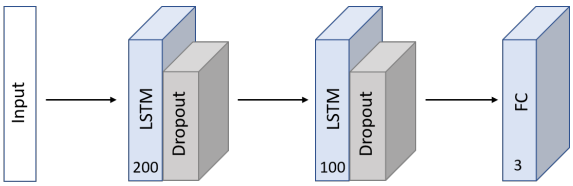


Figure 1: The architecture of the classification model.

as they have not yet received any explicit OPT analysis training. However, tenth semester students are in their final semester, thus we consider them as *intermediates*. The expert dentists were from the University Hospital with an average of 10 years of professional experience.

Participants were presented OPTs and asked to perform a visual search task to determine any anomalies. Further details of the experimental paradigm can be found in [Castner et al. 2018a, 2020]. Sixth and tenth semester students saw the images for 90 seconds and experts for 45 seconds. This shortened duration was determined because much of the previous literature has shown that experts are faster at scanning radiographs [Gegenfurtner et al. 2011; Manning et al. 2006; Nodine and Mello-Thoms 2000; Turgeon and Lam 2016; Van der Gijp et al. 2017]. Students were shown 20 OPTs. Experts were presented 15 OPTs, 10 being the same ones the students viewed. For compatibility, we evaluated gaze data for the 10 OPTs all groups viewed. Additionally, to control for effect of scanpath length, we took the first 45 seconds of the students' viewing time, in line with the experts' total viewing time.

Eye tracking data was collected with an SMI RED250 remote eye tracker with 250 Hz sampling frequency. A multiple-point calibration was performed prior to presentation with a validation criteria of deviations less than 1 degree. The SMI software's standard implementation of the I-VT algorithm was used for fixation and saccade detection.

One dataset is defined as the temporal saccade events of one subject looking at one image. In total, 1159 datasets were used for the analysis. This is because we removed datasets with low signal quality (see [Castner et al. 2018b, 2020] for further pre-processing details).

3.2 LSTM Architecture

We chose to predict three levels of expertise (novice, intermediate, and expert) from the temporal saccade features using an LSTM network architecture (illustrated in figure 1). Since LSTM models require a fixed input sequence length, we empirically determined this length to be 300. Zero-padding at the end was used for shorter sequences, i.e., a dataset with 250 saccade events would receive an additional zero-padding of 50. This approach was suggested in [Goodfellow et al. 2016] and provided the best results. The network contains two LSTM cells with *tanh* activation functions, which are both followed by a dropout layer with a dropout rate of $p = 0.2$ to reduce overfitting. The first LSTM cell contains 200 units, the second contains 100. The data is finally passed to a fully connected layer with 3 units and $softmax$ activation. The network was implemented and trained in keras. It was trained for 125 epochs using the Adam optimizer with the default keras parameters ($learningrate = 0.001$) and categorical crossentropy loss was used. For training, we upsampled the minority classes, specifically the novice and expert classes, to achieve a balanced data set.

The input sequences are the timecourse of visual inspection as determined by a saccade feature. To determine which features (e.g. length, velocity, peak, etc.) produced the best distinction of expertise levels, we evaluated them individually.

4 EVALUATION

4.1 Leave-One-Out Cross-Validation

To see whether we can predict a participant's expertise using saccade features, we performed a Leave-One-Out Cross-Validation (LOOCV). This cross-validation approach uses each participant once as the validation set while the remaining participants form the training set. Although being computationally expensive, LOOCV provides a reliable way to measure our model's ability to classify unseen gaze behavior using saccade features. This method ensures that the model is trained on most of the data set, thereby reducing its bias. Additionally, we enforce that the scanpaths of one participant are not part of both the training and validation set.

Table **??** shows the performance results for each saccade feature. The highest overall accuracy was the saccade velocity peak (0.72), then the saccade length (0.70) and saccade velocity average (0.68). Oddly enough, saccade amplitude produced the lowest accuracy (0.43).

Regarding expertise levels, the recall for all features promotes that experts (0.71) and novices (0.68) were easily distinguishable; yet, intermediates were harder to distinguish (0.51). Saccade length was the most accurate at recognizing novices (0.77) and experts (0.79) and, to some extent, intermediates (0.59). Figure 2 shows the confusion matrix of predicted and actual expertise levels for the saccade length. Here it is apparent that intermediates were more often misclassified as experts (0.28).

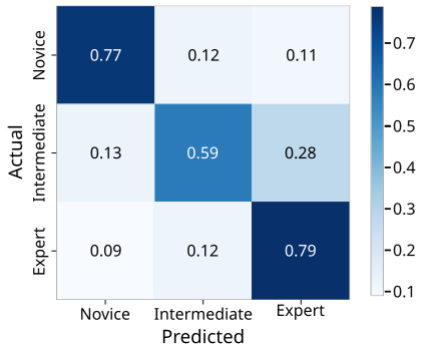


Figure 2: Confusion matrix representing the performance for saccade length (px) as feature input for LSTM.

Overall, our LSTM network is able to classify the expertise level from the saccade behavior above chance level (0.33) for all features and with high accuracy for all features except the saccade amplitude.

Table 1: Precision and recall, calculated for each class.

	Total	*Novice*		*Intermediate*		*Expert*	
Saccade Feature	Accuracy	Recall	Prec.	Recall	Prec.	Recall	Prec.
Saccade Length [px]	0.70	0.77	0.78	0.59	0.70	0.79	0.67
Saccade Amplitude [°]	0.43	0.42	0.47	0.38	0.42	0.52	0.42
Saccade Acceleration Average [$°/s^2$]	0.64	0.72	0.70	0.49	0.64	0.76	0.64
Saccade Acceleration Peak [$°/s^2$]	0.60	0.67	0.68	0.46	0.54	0.70	0.61
Saccade Deceleration Peak [$°/s^2$]	0.61	0.71	0.70	0.43	0.60	0.73	0.58
Saccade Velocity Average [$°/s$]	0.68	0.76	0.74	0.59	0.61	0.70	0.69
Saccade Velocity Peak [$°/s$]	0.72	0.76	0.71	0.65	0.67	0.75	0.78
Position of Saccade Peak Velocity [%]	0.62	0.66	0.70	0.49	0.60	0.76	0.61

4.2 Timecourse of saccade features visualized

In an effort to determine how these features distinguish the expertise levels, we plotted the features overtime. Each expertise level is depicted by the average of all datasets within 250ms windows (the eye tracker sampling frequency) with noise filtered out using a simple exponential smoothing function S_t:

$$S_t = \alpha x_t + (1 - \alpha) S_{t-1}, \tag{1}$$

where x_t is the values of the feature, x, at the current time, t. we used α value of 0.3. For initialization of S_t, we took the average of the first 5 values of x. This smoothing acts as a low pass filter so big jumps are maintained potentially indicating more global scanning behavior (i.e., longer saccade length, faster saccade velocity). Figure 3 shows the best performing saccade features from the LSTM.

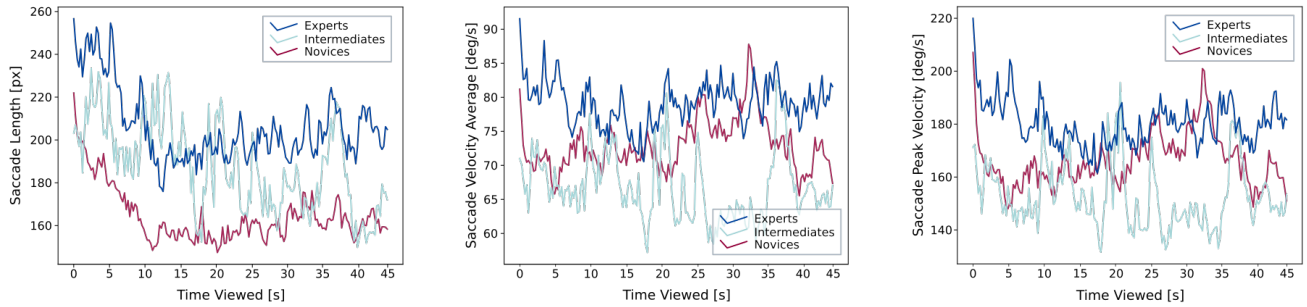
For these features, clear differences are apparent in the saccade behavior of all expertise levels. The saccade lengths overtime (figure 3a) show the clearest distinctions between the expertise levels. Where experts exhibit larger saccade lengths over the entirety of image viewing compared to novices. However, both novices and experts seem to employ a similar strategy of larger saccades over roughly the inital 10 seconds of image viewing, then afterwards employ shorter saccades. This behavior could indicate that novices also have an initial search phase that covers a larger span of the image; though in that same phase, experts cover a larger span of the image. After this initial search phase, both groups employ more local investigation of regions, though, again, experts cover a larger local span compared to novices. Intermediates did not exactly follow this strategy. Rather, their saccade lengths fluctuate back and forth over the whole time course. This behavior could possibly be indicative of a large jump to a new region, local inspection of that region, and a large jump to another new region.

Both average (figure 3b) and peak (figure 3c) saccade velocities have similar behaviors overtime as a peak for the event is the local maxima above a specified threshold [SensoMotoric Instruments 2017] and the average is the average velocities of that event. Again, it is interesting to see that up until 10 seconds of viewing time there are large differences in the saccade velocities of experts and novices, where experts exhibit faster velocities than novices initially. After approximately 10 seconds, the velocities converge and fewer differences in behavior is apparent. One interpretation of this behavior could be that experts initially employ faster saccades than novices; more interesting, however, novices produce faster saccades overtime, becoming more comparable to experts. Whereas, intermediates exhibit velocities that appear similar to novices initially (also up to 10 seconds), but then proceed to have highly fluctuating behavior between slower and faster velocities.

5 DISCUSSION

Our LSTMs using temporal saccade features were highly capable of distinguishing experts, intermediates, and novices with saccade velocity and saccade length resulting in the highest accuracy. Expert behavior differed the most from novices, which is comparable to other research in deep learning models for medical expert classification (84%) [Sodoké et al. 2020] as well as for dental expert classification(73% for LOOCV) [Castner et al. 2020]. Furthermore, we were able to show that intermediates could be accurately classified against their expert or novice counterparts.

The strength of these models lies in the behavioral aspect of the saccades and is completely data driven. Saccade behavior reflects aspects of the visual search strategy such that, from a physiological perspective, the cost of triggering a larger saccade needs to be outweighed by the certainty of a detected target [Stevenson et al. 1986]. Thus, expert heuristics of the task dictate their saccade behavior, promoting larger and faster saccades [Evans et al. 2013; Van der Gijp et al. 2017]. In general, saccade amplitudes have been shown to decrease over time as a product of search time (i.e. ambient towards focal processing) [Buswell 1935; Pannasch et al. 2008; Uemura et al. 2014] and has also been linked to task efficiency [Parkhurst et al. 2002; Wolfe et al. 2021]. Li et al. [Li et al. 2012, 2016] found experts and intermediates had longer fixation durations coupled with decreasing saccade amplitudes as an effect of viewing time. Saccade velocities related to more distant targets exhibit higher peak velocities when falling short of the target [Munoz et al. 1996]. Hosp et al. [Hosp et al. 2021] found that experts performing laproscopic surgery had more uniform saccade velocities compared to intermediates and novices. Above all, saccade features appear to be robust against image semantics, which has been found to highly affect the fixation behavior [Castner et al. 2018b; Donovan and Litchfield 2013; Kok et al. 2012; Turgeon and Lam 2016; Wood et al. 2013]. Thus, the saccade features are ideal for classification models. Our choice of LSTM is further supported as an appropriate demonstration for patterns linked to the timecourse. In general, deep learning



(a) Saccade Length (in pixels) **(b) Saccade Average Velocity** **(c) Saccade Peak Velocity**

Figure 3: Timecourse of best performing saccade features for expertise levels. Average values for all groups for 250ms time windows then smoothed

networks are advantageous because they can handle high dimensional data and are able to easily recognize patterns [Arulkumaran et al. 2017]. However, they are often a black box where learning happens through the weighted connections throughout hidden layers creating a situation where the patterns recognized by the machine may not be obvious to the researcher.

5.1 Proponent for Holistic Processing

In an effort to provide explainability to the model performance based on saccade behavior, we visualized the features over time (see figure 3) and found clear differences between experts, novices, and intermediates. Especially for saccade length and for the velocity metrics, we can see an initial interval of highly differing saccade behavior. For our task, this initial phase was roughly 10 seconds and depicts experts having longer saccade lengths and higher velocities than novices. The saccade peak velocity, which was the best performing feature for the LSTM, and the average velocity also showed clear distinctions in the expertise levels (figures 3b and 3c). These higher velocities in experts support the theory of holistic processing [Kundel et al. 2007]. This initial phase could be the *getting-the-gist* phase. In contrast, novices have slightly lower velocities that increase towards similar behavior of experts, thus more local inspection initially. But after this initial inspection phase, expert and novice velocities become less distinguishable. Yet, the saccade length in this initial inspection phase shows an offset between experts and novices – novices having shorter saccade lengths – but both experts and novices have steadily decreasing saccade lengths in this initial inspection phase. This similar trend could indicate that novices also do their own global processing, it may just be on a smaller scale as compared to experts. Further research beyond the current qualitative investigation is necessary to solidify this novel assumption. Previous research in medical image inspection that could support this behavior has found that experts view the stimulus as more goal-driven and novices as more stimulus-driven due to inexperience [Al-Moteri et al. 2017; Fox and Faulkner-Jones 2017; Ganesan et al. 2018; Koide et al. 2015; Kok et al. 2012; Krupinski et al. 2006].

Intermediates, interestingly enough, show saccade lengths more similar to experts in this phase, but velocities more similar to novices. This could align with the literature that has found intermediate gaze behavior sometimes aligns with novices and sometimes aligns with experts [Brams et al. 2019; Gegenfurtner et al. 2011]. Currently, research regarding how intermediates relate to the holistic processing theory is under-explored [Brams et al. 2019; Gegenfurtner et al. 2011; Van der Gijp et al. 2017]. Although the LSTM could accurately detect intermediates, the recall was lower and, especially for saccade length, intermediates were slightly (0.28) misrepresented as experts. Li et al. [Li et al. 2012, 2016] also found similarities between expert and intermediate dermatologists regarding saccade amplitude decreasing over time, but we have only found this in our initial inspection phase; Afterwards, the saccade behavior (velocities and lengths) of intermediates fluctuates greatly. We can only theorize that the visual inspection of intermediates could be more of a global - focal - global- focal. If this is the case, these finding are highly interesting to the order in which relevant information is attended to and processed in intermediates. Moreover, it offers crucial understanding of this in-between stage and how to develop appropriate learning interventions.

5.2 Limitations and Future Research

It should be noted that we are only investigating the first 45 seconds of the novices and the intermediates (both groups being students who investigated OPTs for 90 seconds) in order to control for the scanpath length affecting the outcome of the classifier as well as to observe potential patterns the model could have detected among the expertise levels. It is a valid concern that students need more time to appropriately inspect and diagnose a medical image, and future research should address the gaze behavior over a longer time period to establish other prominent phases in the visual search strategy.

Another potential confound to saccade behavior related to expertise could be the age of experts. Generally, experts are older than their novice counterparts – who are often students – and age can contribute to variations in the saccade behavior [Crabb et al. 2014; Irving et al. 2006; Munoz et al. 1998]. However, we found intermediates were, to a small extent, misclassified as experts. For the current investigation, our intermediate sample were advanced dental students in their last semester before beginning their residency. Even

though they are still considered students, it is interesting that their search strategies are starting to exhibit similar patterns to those of experts. Even though we found similar patterns among younger intermediates and older experts, further investigation is necessary to better rule out age effects by examining these expertise levels in an age controlled population.

6 CONCLUSION

To date, intermediate data in conjunction with expert and novice data has not been investigated enough. The current work addresses this concern by providing an LSTM network that is capable of classifying experts, intermediates, and novice dentists with high accuracy. Our approach uses the temporal saccade features, which proved to be a highly feasible, data-driven approach that does not rely on areas of interest, but patterns related to search strategy. Saccade length and velocity information were the best performing features and their behavior time was visualized for the expertise levels. When visualizing the saccade features over time, the behavior follows the holistic theory of expert visual exploration that experts can quickly process the whole scene using longer and more rapid saccade behavior initially. However, further investigation is pivotal to understanding intermediate medical image inspection, especially regarding which saccadic features are more expert-like and which are more novice-like. This work shows that saccadic scanpath features are a viable input for intelligent tutoring systems. Not only can the findings of this work be targeted toward dental students, but also in other medical domains.

ACKNOWLEDGMENTS

Enkelejda Kasneci is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

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.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* 34, 6 (2017), 26–38.

Dan Assaf, Eyal Amar, Norbert Marwan, Yair Neuman, Moshe Salai, and Ehud Rath. 2016. Dynamic patterns of expertise: the case of orthopedic medical diagnosis. *Plos One* 11, 7 (2016), e0158820.

Stephanie Brams, Gal Ziv, Ignace TC Hooge, Oron Levin, Thomas De Brouwere, Johny Verschakelen, Siska Dauwe, A Mark Williams, Johan Wagemans, and Werner F Helsen. 2020. Focal lung pathology detection in radiology: Is there an effect of experience on visual search behavior? *Attention, Perception, & Psychophysics* (2020), 1–14.

Stephanie Brams, Gal Ziv, Oron Levin, Jochim Spitz, Johan Wagemans, A Mark Williams, and Werner F Helsen. 2019. The relationship between gaze behavior, expertise, and performance: A systematic review. *Psychological Bulletin* 145, 10 (2019), 980.

Tad T Brunyé, Trafton Drew, Manob Jyoti Saikia, Kathleen F Kerr, Megan M Eguchi, Annie C Lee, Caitlin May, David E Elder, and Joann G Elmore. 2021. Melanoma in the blink of an eye: Pathologists' rapid detection, classification, and localization of skin abnormalities. *Visual Cognition* (2021), 1–15.

Guy Thomas Buswell. 1935. *How people look at pictures: a study of the psychology and perception in art.* Univ. Chicago Press.

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.

Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Therese Eder, Fabian Huettig, Constanze Keutel, and Enkelejda Kasneci. 2020. Deep Semantic Gaze Embedding and Scanpath Comparison for Expertise Classification during OPT Viewing. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Full Papers)*. Association for Computing Machinery, New York, NY, USA, Article 18, 10 pages. https://doi.org/10.1145/3379155.3391320

Michelene TH Chi. 2006. Two approaches to the study of expert' characteristics. Cambridge University Press, 21–30.

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? 7627 (2010), 76270B.

David P Crabb, Nicholas D Smith, and Haogang Zhu. 2014. What's on TV? Detecting age-related neurodegenerative eye disease using eye movement scanpaths. *Frontiers in Aging Neuroscience* 6 (2014), 312.

Emily M Crowe, Iain D Gilchrist, and Christopher Kent. 2018. New approaches to the analysis of eye movement behaviour across expertise while viewing brain MRIs. *Cognitive Research: Principles and Implications* 3, 1 (2018), 12.

Alan Davies, Markel Vigo, Simon Harper, and Caroline Jay. 2018. Using simultaneous scanpath visualization to investigate the influence of visual behaviour on medical image interpretation. *Journal of Eye Movement Research* 10, 5 (Feb. 2018). https://doi.org/10.16910/jemr.10.5.11

Tim Donovan and Damien Litchfield. 2013. Looking for cancer: Expertise related differences in searching and decision making. *Applied Cognitive Psychology* 27, 1 (2013), 43–49.

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.

Hubert L Dreyfus and Stuart E Dreyfus. 1986. The power of human intuition and expertise in the era of the computer. *Mind Over Machine. Nueva York: the Free Press* (1986).

K Anders Ericsson and Andreas C Lehmann. 1996. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology* 47, 1 (1996), 273–305.

Karla K Evans, Diane Georgian-Smith, Rosemary Tambouret, Robyn L Birdwell, and Jeremy M Wolfe. 2013. The gist of the abnormal: Above-chance medical decision making in the blink of an eye. *Psychonomic Bulletin & Review* 20, 6 (2013), 1170–1175.

Paul J Feltovich, Michael J Prietula, and K Anders Ericsson. 2006. Studies of expertise from psychological perspectives. In *The Cambridge Handbook of Expertise and Expert Performance*, K Anders Ericsson, Robert R Hoffman, and Aaron Kozbelt (Eds.). Cambridge University Press, 41–67.

Sharon E Fox and Beverly E Faulkner-Jones. 2017. Eye-Tracking in the Study of Visual Expertise: Methodology and Approaches in Medicine. *Frontline Learning Research* 5, 3 (2017), 29–40.

Robert M French, Yannick Glady, and Jean-Pierre Thibaut. 2017. An evaluation of scanpath-comparison and machine-learning classification algorithms used to study the dynamics of analogy making. *Behavior Research Methods* 49, 4 (2017), 1291–1302.

Wolfgang Fuhl, Nora Castner, Thomas Kübler, Alexander Lotz, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2019. Ferns for Area of Interest Free Scanpath Classification. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 39, 5 pages. https://doi.org/10.1145/3314111.3319826

Ziba Gandomkar and Claudia Mello-Thoms. 2019. Visual search in breast imaging. *the British Journal of Radiology* 92, 1102 (2019), 20190057.

Ziba Gandomkar, Kevin Tay, Patrick C Brennan, and Claudia Mello-Thoms. 2017. A model based on temporal dynamics of fixations for distinguishing expert radiologists' scanpaths. In *Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment*, Vol. 10136. International Society for Optics and Photonics, 1013606.

Ziba Gandomkar, Kevin Tay, Patrick C Brennan, and Claudia Mello-Thoms. 2018. Recurrence quantification analysis of radiologists' scanpaths when interpreting mammograms. *Medical Physics* 45, 7 (2018), 3052–3062.

Aarthi Ganesan, Maram Alakhras, Patrick C Brennan, and Claudia Mello-Thoms. 2018. A review of factors influencing radiologists' visual search behaviour. *Journal of Medical Imaging and Radiation Oncology* 62, 6 (2018), 747–757.

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.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning, Chapter 10: Sequence Modeling: Recurrent and Recursive Nets.* MIT Press. http://www.deeplearningbook.org.

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.

Brian P Hermanson, Grant C Burgdorf, John F Hatton, Darrin M Speegle, and Karl F Woodmansey. 2018. Visual fixation and scan patterns of dentists viewing dental periapical radiographs: an eye tracking pilot study. *Journal of Endodontics* 44, 5 (2018), 722–727.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

Benedikt Hosp, Myat Su Yin, Peter Haddawy, Ratthapoom Watcharopas, Paphon Sangasoongsong, and Enkelejda Kasneci. 2021. Differentiating Surgeons' Expertise solely by Eye Movement Features. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*. 371–375.

Chartene H Hu, Harold L Kundel, Calvin F Nodine, Elizabeth A Krupinski, and Lawrence C Toto. 1994. Searching for bone fractures: a comparison with pulmonary nodule search. *Academic Radiology* 1, 1 (1994), 25–32.

Elizabeth L Irving, Martin J Steinbach, Linda Lillakas, Raiju J Babu, and Natalie Hutchings. 2006. Horizontal saccade dynamics across the human life span. *Investigative ophthalmology & visual science* 47, 6 (2006), 2478–2484.

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

Conor Kelton, Zijun Wei, Seoyoung Ahn, Aruna Balasubramanian, Samir R Das, Dimitris Samaras, and Gregory Zelinsky. 2019. Reading detection in real-time. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*. 1–5.

Naoko Koide, Takatomi Kubo, Satoshi Nishida, Tomohiro Shibata, and Kazushi Ikeda. 2015. Art expertise reduces influence of visual salience on fixation in viewing abstract-paintings. *Plos One* 10, 2 (2015), e0117696.

Ellen M Kok, Anique BH De Bruin, Simon GF Robben, and Jeroen JG Van Merrienboer. 2012. Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. *Applied Cognitive Psychology* 26, 6 (2012), 854–862.

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. 1996. Visual scanning patterns of radiologists searching mammograms. *Academic Radiology* 3, 2 (1996), 137–144.

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 C Kübler and Enkelejda Kasneci. 2015. Automated Comparison of Scanpaths in Dynamic Scenes. In *SAGA-International Workshop on Solutions for Automatic Gaze Data Analysis: Proceedings*. 1–3.

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, 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.

Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. 2013. I know what you are reading: recognition of document types using mobile eye tracking. In *Proceedings of the 2013 International Symposium on Wearable Computers*. 113–116.

Rui Li, Jeff Pelz, Pengcheng Shi, Cecilia Ovesdotter Alm, and Anne R Haake. 2012. Learning eye movement patterns for characterization of perceptual expertise. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*. 393–396.

Rui Li, Pengcheng Shi, Jeff Pelz, Cecilia O Alm, and Anne R Haake. 2016. Modeling eye movement patterns to characterize perceptual skill in image-based diagnostic reasoning processes. *Computer Vision and Image Understanding* 151 (2016), 138–152.

Yu Li, Carla Allen, and Chi-Ren Shyu. 2019. Quantifying and understanding the differences in visual activities with contrast subsequences. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*. 1–5.

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.

David Manning, Susan Ethell, Tim Donovan, and Trevor Crawford. 2006. How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography* 12, 2 (2006), 134–142.

Ezgi Mercan, Linda G Shapiro, Tad T Brunyé, Donald L Weaver, and Joann G Elmore. 2018. Characterizing diagnostic search patterns in digital breast pathology: Scanners and drillers. *Journal of Digital Imaging* 31, 1 (2018), 32–41.

DP Munoz, JR Broughton, JE Goldring, and IT Armstrong. 1998. Age-related performance of human subjects on saccadic eye movement tasks. *Experimental brain research* 121, 4 (1998), 391–400.

DOUGLAS P Munoz, DAVID M Waitzman, and ROBERT H Wurtz. 1996. Activity of neurons in monkey superior colliculus during interrupted saccades. *Journal of Neurophysiology* 75, 6 (1996), 2562–2580.

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.

Calvin F Nodine and Claudia Mello-Thoms. 2000. The nature of expertise in radiology. *Handbook of Medical Imaging. SPIE* (2000), 859–895.

Sebastian Pannasch, Jens R Helmert, Katharina Roth, Ann-Katrin Herbold, and Henrik Walter. 2008. Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research* 2, 2 (2008).

Derrick Parkhurst, Klinton Law, and Ernst Niebur. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Research* 42, 1 (2002), 107–123.

Michael Polanyi. 1962. Personal knowledge: towards a post-critical. *Philosophy* (1962).

SensoMotoric Instruments 2017. *BeGaze Manual* (version 3.7 ed.). SensoMotoric Instruments.

James Shanteau. 1992. Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes* 53, 2 (1992), 252–266.

Hongxin Shao and Boon-Hee Soong. 2016. Traffic flow prediction with long short-term memory networks (LSTMs). In *2016 IEEE Region 10 Conference (TENCON)*. IEEE, 2986–2989.

Komi Sodoké, Roger Nkambou, Aude Dufresne, and Issam Tanoubi. 2020. Toward a deep convolutional LSTM for eye gaze spatiotemporal data sequence classification. In *Proceedings of the 13th International Conference on Educational Data Mining*. 672–676.

SB Stevenson, FC Volkmann, JP Kelly, and Lorrin A Riggs. 1986. Dependence of visual suppression on the amplitudes of saccades and blinks. *Vision research* 26, 11 (1986), 1815–1824.

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.

Munenori Uemura, Morimasa Tomikawa, Ryuichi Kumashiro, Tiejun Miao, Ryota Souzaki, Satoshi Ieiri, Kenoki Ohuchida, Alan T Lefor, and Makoto Hashizume. 2014. Analysis of hand motion differentiates expert and novice surgeons. *Journal of Surgical Research* 188, 1 (2014), 8–13.

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.

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and S Yu Philip. 2017. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Advances in Neural Information Processing Systems*. 879–888.

Amy L Warren, Tyrone L Donnon, Catherine R Wagg, Heather Priest, and Nicole J Fernandez. 2018. Quantifying Novice and Expert Differences in Visual Diagnostic Reasoning in Veterinary Pathology Using Eye-Tracking Technology. *Journal of Veterinary Medical Education* 45, 3 (2018), 295–306.

Jeremy M Wolfe, Chia-Chien Wu, Jonathan Li, and Sneha B Suresh. 2021. What do experts look at and what do experts find when reading mammograms? *Journal of Medical Imaging* 8, 4 (2021), 045501.

Greg Wood, Karen M Knapp, Benjamin Rock, Chris Cousens, Carl Roobottom, and Mark R Wilson. 2013. Visual expertise in detecting and diagnosing skeletal fractures. *Skeletal Radiology* 42, 2 (2013), 165–172.

Myat Su Yin, Peter Haddawy, Benedikt Hosp, Paphon Sa-ngasoongsong, Thanwarat Tanprathumwong, Madereen Sayo, Supawit Yangyuenpradorn, and Akara Supratak. 2020. A Study of Expert/Novice Perception in Arthroscopic Shoulder Surgery. In *Proceedings of the 4th International Conference on Medical and Health Informatics*. 71–77.

Hong-Jun Yoon, Folami Alamudun, Kathy Hudson, Garnetta Morin-Ducote, and Georgia Tourassi. 2018. Deep gaze velocity analysis during mammographic reading for biometric identification of radiologists. *Journal of Human Performance in Extreme Environments* 14, 1 (2018), 3.