Differentiating Surgeons' Expertise solely by Eye Movement Features

Benedikt Hosp*
Human-Computer Interaction,
University of Tübingen
Tübingen, Germany
benedikt.hosp@uni-tuebingen.de

Myat Su Yin
MIRU, Mahidol University
Nakhon Pathom, Thailand
myatsu.yin@mahidol.ac.th

Peter Haddawy
Faculty of ICT, Mahidol University
Nakhon Pathom, Thailand
peter.had@mahidol.ac.th

Ratthapoom Watcharopas
Faculty of Medicine, Mahidol
University
Bangkok, Thailand
poom911@hotmail.com

Paphon Sa-ngasoongsong
Faculty of medicine, Mahidol
University, Ramathibodi Hospital
Bangkok, Thailand
paphonortho@gmail.com

Enkelejda Kasneci
Human-Computer Interaction,
University of Tübingen
Tübingen, Germany
enkelejda.kasneci@uni-tuebingen.de

ABSTRACT

Medical schools are increasingly seeking to use objective measures to assess surgical skills. This extends even to perceptual skills, which are particularly important in minimally invasive surgery. Eye tracking provides a promising approach to obtaining such objective metrics of visual perception. In this work, we report on results of a cadaveric study of visual perception during shoulder arthroscopy. We present a model for classifying surgeons into three levels of expertise using only eye movements. The model achieves a classification accuracy of 84.44% using only a small set of selected features. We also examine and characterize the changes in visual perception metrics between the different levels of expertise, forming a basis for development of a system for objective assessment.

CCS CONCEPTS

• **Social and professional topics** → **Medical technologies**; • **Computing methodologies** → **Supervised learning**; **Biometrics**.

KEYWORDS

surgeon, eye, tracking, diagnostic, model, machine learning

ACM Reference Format:
Benedikt Hosp, Myat Su Yin, Peter Haddawy, Ratthapoom Watcharopas, Paphon Sa-ngasoongsong, and Enkelejda Kasneci. 2021. Differentiating Surgeons' Expertise solely by Eye Movement Features. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3461615.3485437

ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8471-1/21/10…$15.00
https://doi.org/10.1145/3461615.3485437

1 INTRODUCTION

Arthroscopy is a popular minimally invasive surgical procedure that improves patient outcomes while at the same time conserving hospital resources. According to Monson et al. [9], patients experience less pain, have fewer complications and recover faster than with traditional open surgery. However, a surgeon needs advanced technical skills for this type of operation [5]. Arthroscopy involves inserting instruments and a scope into the joint (e.g. shoulder or knee) through small incisions. A key capability in performing arthroscopic surgery is the ability to use the scope to navigate through complex anatomy of the joint for inspection, diagnosis, and to locate the surgical site. The scope can rotate in multiple dimensions and casts its image on a screen placed next to the patient, which surgeons largely rely upon during surgery. Navigation is challenging due to complex anatomy, limited field of view, projection of the 3D space onto the 2D monitor, and the rotation of the monitor from the instrument plane.

Due to these technical challenges, there is growing interest within the medical community to optimize training, including having objective measures of performance for tasks like navigation. Since navigation is a psychomotor task in which visual perception plays a crucial role, it is natural to look to eye tracking for such a measure. Indeed, the role of eye movements is increasingly being investigated in surgery [5]. In particular, the role of eye movements is increasingly being investigated (for an overview see [5]). To determine whether eye tracking can serve as a basis for an objective measure in arthroscopy, first it must be determined whether, and to what extent, differences in surgeons' expertise are reflected by their eye movements. The findings from this study are significant for the design of adequate training and evaluation scenarios for perceptual-cognitive diagnostic and training systems.

In this work, we consider the perception of surgeons using eye movement patterns from three expertise levels in a human cadaveric study of diagnostic arthroscopy of the shoulder. We selected this task since it focuses on navigation skill in which perception plays a major role. We use stimulus-independent eye movement patterns to develop a model to classify the subjects into the three levels of expertise. Using only a small number of selected features, our model achieves a classification accuracy of over 82%. We further investigate differences in eye movement patterns among the

three classes in order to understand how these patterns evolve with increasing levels of expertise. We hope that such an understanding can assist in developing specialized training to provide the appropriate support to surgeons at different expertise levels.

2 RELATED WORK

In eye tracking studies, using artificial forms of presentation like virtual reality (VR) [8, 15] or images [4, 11] could omit important perceptual details requiring the participants to fill in through inference which often subsequently leads to the higher levels of frustration [15]. To provide a presentation mode that is as natural as possible, we use so called soft cadavers that provide natural tactile sensation while maintaining the naturalness of the scene. Although remote eye trackers are commonly used in lab studies [8], as soon as the participant changes to another direction (e.g. down at the cadaver), they can no longer capture the gaze signal. To allow the participant to use normal gaze behaviour and move freely without data lost, we use a head-mounted eye tracker in combination with a 4k-screen. This setup supports natural gaze behaviour as well as high control of the stimulus allowing us to capture highly detailed information of the tissue on a screen with high resolution and gaze signals on the cadaver, both with the same field camera. Eye tracking studies in surgery are differed in how they evaluated the gaze signal. The gaze signal on the stimulus was considered, i.e. target gaze behavior, switching behavior (alternating gaze between target and instrument), or following behaviour (eye following the instrument) [8]. Other studies focused on quiet eye periods [13]. However, there are also studies that have gained insights at the feature level. For example, Kocak et al.[7] used stimulus-independent eye features in their binary classification and found significantly lower saccade rates, as well as significantly higher peak velocities for experts, which was confirmed by other studies [5]. Tien et al. [12] found a higher fixation rate in experts. Eivazi et al. [4] show differences in time to first fixation and mean fixation duration. However, theses differences were not confirmed by Sondergren et al. [11], as in both studies fixation durations are analyzed differently and the choice of regions of interest plays an important role. These results show that eye movements can be used to assess the surgical expertise and to define differences between groups.

Many studies have focused on the detection of differences in expertise between experts and novices [13, 15] and only few studies have focused on the development of eye movements. Studies focusing on development have used mostly simulators [7] or images [11]. Hidden Markov models (HMM) used in the latter study reveal differences in eye movement patterns between high and low performers. So far, several algorithms have been introduced to eye tracking including supervised methods like support vector machines [2, 6] and neural networks [3]. Ahmidi et al. [1] mixed instrument movements with eye movement data and achieved a binary classification accuracy of 82.5% for skill level classification. All these studies show that eye movement data can be used to differentiate between experts and novices and that it is not necessary to determine exactly where the surgeons were looking to measure their skill accurately.

3 PARTICIPANTS AND METHODS

3.1 Procedure

This work makes use of the eye tracking dataset from the work of Yin et al. [14]. Their dataset contains eye movement data for three classes of surgeons: 3rd year residents (R3), 4th year residents (R4), and fellows. Each class consists of five (n=5) participants, equally. We even use the data of the two participants that were left out in their study because of a gaze signal offset. Since we only use relative features, we can use the data of these participants too. In their study, participants were placed in front of the cadaver and four feet away from the 4k, 52-inch screen where the output of the arthoscope was displayed (Figure 1). Each participant was familiarized with the setup and asked to navigate and diagnose 12 anatomical landmarks in the shoulder, while wearing a Tobii Pro Glasses 2 eye tacker. The gaze was recorded with Tobii software.

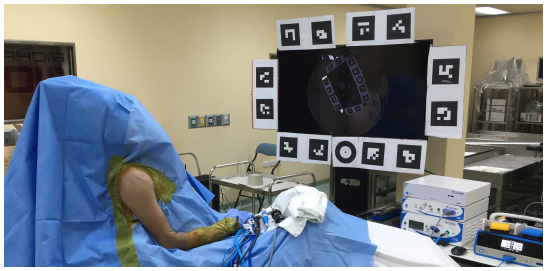

Figure 1: Experimental setup showing cadaver, arthroscopic equipment, and 4k monitor with ARUCO markers.

3.2 Data preparation

The Tobii Glasses 2 were set to a frame rate of 100 Hz , thus a gaze sample is available every 10 ms and saved with a timestamp, x-, and y-coordinates. The samples are used to calculate fixations and saccades metrics using the Tobii Fixation Filter, with a sliding window averaging method and the feature classification algorithm of Olson [10] with a default velocity threshold of 0.7 pixels/ms. The raw eye tracking data, as well as the fixations and saccade metrics, are exported from the Tobii Studio software. From the fixations, velocity of the saccades, saccade duration, values of the gyroscope (yaw, pitch and roll) as well as the amplitude of saccades, we use the person-specific average, minimum, maximum and standard deviation as features. While Tobii provides metrics about the first saccade and first fixation too, we did not include them. Since our participants were familiarized with the glasses for different lengths of time when the trials started, we end up with chaotic first saccades, which have no informative character.

As our aim is to infer which features contribute to expertise differences, we first used all the exported features from the Tobii Studio Software and added common metrics to them. Subsequently, we evaluated their frequency in the model building process and rated the most frequent used features to build a model with this subset of features for expertise acquisition. To incorporate uncertainties, we trained the model 150 times and calculated the most frequently used features by taking the features with the maximum number of occurrences in the training process. We added certain typical eye movement features which we calculated by ourselves.

The fixations were split into small fixations and smooth pursuit fixations. As the Tobii Software does not provide calculations of smooth pursuits, which are assumed to help differentiating different expertise classes, the smooth pursuit events were encoded in the fixations. We therefore treated fixations with a dispersion over 30 pixels as smooth pursuit fixations. This threshold was empirically defined during data analysis. The set of features was:

- Saccade duration (average, min, max. std. dev.)
- Fixation duration (average, min, max. std. dev.)
- Smooth pursuit dispersion (average, min, max, std. dev.)
- Fixation frequency
- Saccade frequency
- Pupil diameter (average, min, max. std. dev.)
- Gyroscope X,Y,Z (average, min, max. std. dev.)

We decided to include the gyroscope values because they could provide information about head movement between screen and cadaver The integration of pupil diameter features is based on the assumption that experts may have less fluctuating pupil diameter since their mental effort is considered to be smaller. Vice versa, the pupil diameter of intermediates and novices may reveal expertise differences by such effects.

3.3 Machine learning model

We used all 38 features to build a support vector machine (SVM) model in 150 independent runs. On each of the 150 runs we keep out one participant (leave-one-out validation). This participant is our test set and has never been seen by the model (of the current run) before. Therefore, in each run we take all data of the remaining 14 participants to train the model and test it with the unseen data of the test set participant. While the training algorithm iterates over the same procedure it changes the participant for the test set (sequentially iterating over the participant numbers from 1 to 15) 150 times. Thus, each participant is used as test set 10-times in total. By having 10 runs for each participant, we are taking statistical fluctuations into account. To ensure independency between runs, we train a new model on every run and report the cumulated accuracy values of the 150 runs. Thus, in each run, the model is trained with 14 participants and tested with the test set data of one participant, which is unseen by the current model. A strict separation of data in a participant-wise manner is very important, as mixing up samples of one person into training and testing data would allow the model to remember person-specific (idiosyncratic) features and restrict a real expertise learning process.

On each run, the data of the 14 participants of the training set is split into 10-folds. This is called a 10-fold cross-validation. The cross-validation is important to protect the model against over fitting. In each fold, $\lceil \frac{1}{10} \rceil$ of the 14 participants that belong to training set is used to validate the model that is trained with $\lfloor \frac{9}{10} \rfloor$ of 14 participants. Which participant belongs to training or validation set, is decided randomly. However, the split is always done participant-wise to prevent an idiosyncratic learning behavior of the model.

In a first model, we use all 38 features to check the classifiability of the data set and afterwards reduce the amount by taking the four most frequent features of 150 runs. The most frequent features are features that have the highest importance values for a single model prediction. In each run we built a queue of all 38 features sorted by importance for the current model. Subsequently, we computed their overall frequency over all models.

4 RESULTS

Our first classification model shows promising results with an average accuracy of 60%. As a system that would simply guess the class, would only reach a chance-level of 33.33%, the all feature model can already be considered as well-performing. But as we want to specify the results to allow a precise statement about a high performing classification with the least amount of features, we continued by collecting all features and their importance values on 150 runs of the all feature model and took the most frequent features (MFF) as a new set. With this subset of four features, shown in Table 1, earlier counteracting features may be avoided and a precise statement about the differences of the groups can be stated. The final SVM model with the four MFF uses a linear kernel and a box constraint of 11.0174. We adopted one-vs-all approach for multi-class classification with the kernel scale remains 1. Before training, we standardized the data. Training took about 56.03 sec.

4.1 Performance metrics

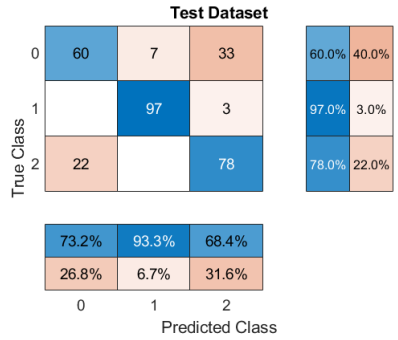

Figure 2: Performance values on 100 runs.

With an accuracy of 82.33% the model improved over 20 percentage points, compared to the all feature model. Figure 2 shows the confusion matrix after 100 runs. 7 samples of the novice class were classified as intermediate and 33 as expert. This results in a class accuracy of 60%. The classification of the intermediates peaked at 97%, as only 3 samples were classified as experts and non as novices. This is especially interesting since the intermediates are in between the other classes and are therefore more likely to spread to both sides. The expert samples were with 78 samples correctly and 22 samples as novice samples, the second best classified class.

Table 1: The most important and frequent features on 150 runs.

Feature	derivation
Peak velocity of saccades	standard deviation
Amplitude of saccades	minimum
Total amplitude of saccades	sum
Saccade duration	standard deviation

The average recall is with 95.43% extremely high which is confirmed by the average miss rates. Only 4.57% of samples were misclassified. For the SVM model we achieve an area under the curve (AUC) of 0.91.

4.2 Feature evolution

As we consider there is a cognitive process going on, forming the optimal gaze behavior from novice to expert, we have a look at the evolution of the gaze features between the classes to describe such process as good as possible. To analyze these evolutionary steps, we have a look at the single feature characteristics separately. We do that with the four MFF from Table 1. Table 2 contains the characteristics of the four MFF features.

The table shows that experts have a smaller standard deviation of the peak velocity of the saccades (93.26 °/s). This feature is hard to interpret, but one assumption may be that experts have a more uniform distribution of saccade velocities. This means they do more saccades at the same speed, in a structured and planned way, compared to intermediates and novices. Interestingly, intermediates as the middle class between expert and novice show a much more diverse saccade peak velocity behaviour (121.72 °/s). Novices are in the middle between experts and intermediates. A higher value for the standard deviation of the saccade peak velocities could be an indicator for a more chaotic gaze behaviour, but it is hard to draw a conclusion about such a feature. When having a look at the minimum saccade amplitudes, we can see the same differences. The experts have on average a larger minimal saccade of length 0.86 °, compared to the intermediates with 0.40°and the novices with 0.64°. Again, we can see that the novices are in between the experts and intermediates. Only the total amplitude of all saccades shows a uniform evolution. The experts do a total of 481.32°of saccade length, where intermediates do more than twice the experts (1120.74°) and novices (1956.21°) even more than five time the experts and nearly double the intermediates. Another interesting feature evolution can be seen in the standard deviation of the saccade durations. This feature is also hard to interpret, but one possibility could be that experts with 18.96 ms and intermediates with 16.58 ms have slightly more order in their saccades than novices. Though the differences are very small and should be confirmed with more data.

Table 2: Average feature evolution between classes.

Feature	Fellow	R4	R3
Saccade peak velocity (STD)	93.26 °/s	121.72 °/ s	117.45 °/s
Saccade amplitude (min)	0.86 °	0.40 °	0.64 °
Total saccade amplitude	481.32 °	1120.74 °	1956.21 °
Saccade duration (std. dev.)	18.96 ms °	16.58 ms °	23.54 ms °

5 DISCUSSION

In this work we developed a model with supervised machine learning techniques that is able to distinguish three levels of expertise solely on the basis of eye movements during an arthroscopic surgery of the shoulder. With an accuracy of 82.33% the model can be considered as performing well on this 3-class problem. Thus, it can be stated that expertise differences between three different groups of expertise are reflected by their eye movements.

To further understand the differences between the three levels of expertise, we had a look at the four most frequent features of the model and analyzed the evolution of the characteristics between the groups. Except for the total amount of saccade amplitudes, the remaining three of the four most frequent features show a uniform evolution. First, novices tend to have a more chaotic gaze behavior and distribute their gaze over a larger portion of the scene by making many different saccades with different speed. They also tend to look more at the outside than the center. The evolution to intermediates shows an atypical behavior, as they tend to still gaze over a larger area of the scene than the experts, but do smaller saccades with a still diverse velocity. This might indicate, that they try to focus on more specific visual clues and start to concentrate on the center of the scene. In the next evolution step, the saccade velocities shrink significantly, which signifies a more planned scanning behavior, with somewhat longer saccades, concentrated more on specific areas. To summarize our findings, one can state that the evolution of novices to intermediates first tends to lead to a partly more chaotic gaze behavior, then turning to be more precise. With the investigations on the evolutionary steps, we can also define class dependent weak-spots in perception for each class. An evolution between the single classes is clearly recognizable. Thus, opening the way to a class-specific training system that is optimized for different steps in perceptional evolution. We also showed that for a high accuracy classification there are not many features needed. A subset of four features describing the gaze behavior is already enough to distinguish different classes. Luckily, the four features are easily calculated, which would allow the usage of the classification as an online classification system. Though, the classification would need to be done segment-wise after a certain period of time.

Further steps are to add more participants to each class, and refine the number of classes. This would allow a much finer classification and therefore a better understanding of the differences between the levels of expertise. A finer classification is important to robust assumptions made by the model about gaze behavior and optimize the recognition of class-specific weak-spots to be used in a training system.

REFERENCES

[1] Narges Ahmidi, Gregory D Hager, Lisa Ishii, Gabor Fichtinger, Gary L Gallia, and Masaru Ishii. 2010. 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, 295–302.

[2] 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*. 1–9.

[3] Nora Castner, Thomas C Kuebler, 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 *ACM Symposium on Eye Tracking Research and Applications*. 1–10.

[4] Shahram Eivazi, Roman Bednarik, Markku Tukiainen, Mikael von und zu Fraunberg, Ville Leinonen, and Juha E Jääskeläinen. 2012. Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 377–380.

[5] Frouke Hermens, Rhona Flin, and Irfan Ahmed. 2013. Eye movements in surgery: A literature review. (2013).

[6] Benedikt Hosp, Florian Schultz, Enkelejda Kasneci, and Oliver Höner. 2020. Eye Movement Feature Classification for Soccer Expertise Identification in Virtual

Reality. *arXiv preprint arXiv:2009.11676* (sep 2020).

[7] Ergun Kocak, Jan Ober, Necip Berme, and W Scott Melvin. 2005. Eye motion parameters correlate with level of experience in video-assisted surgery: objective testing of three tasks. *Journal of Laparoendoscopic & Advanced Surgical Techniques* 15, 6 (2005), 575–580.

[8] 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*. 41–48.

[9] JR Monson. 1993. Advanced techniques in abdominal surgery. *British Medical Journal* 307, 6915 (1993), 1346–1350.

[10] Pontus Olsson. 2007. Real-time and offline filters for eye tracking.

[11] Mikael Hans Sodergren, Felipe Orihuela-Espina, James Clark, Ara Darzi, and Guang-Zhong Yang. 2010. A hidden markov model-based analysis framework using eye-tracking data to characterise re-orientation strategies in minimally invasive surgery. *Cognitive processing* 11, 3 (2010), 275–283.

[12] Geoffrey Tien, Bin Zheng, and M Stella Atkins. 2011. Quantifying surgeons' vigilance during laparoscopic operations using eyegaze tracking.. In *MMVR*. 658–662.

[13] Mark R Wilson, John S McGrath, Samuel J Vine, James Brewer, David Defriend, and Richard SW Masters. 2011. Perceptual impairment and psychomotor control in virtual laparoscopic surgery. *Surgical endoscopy* 25, 7 (2011), 2268–2274.

[14] 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.

[15] Bin Zheng, Geoffrey Tien, Stella M Atkins, Colin Swindells, Homa Tanin, Adam Meneghetti, Karim A Qayumi, and O Neely M Panton. 2011. Surgeon's vigilance in the operating room. *The American Journal of Surgery* 201, 5 (2011), 673–677.