States of Confusion: Eye and Head Tracking Reveal Surgeons' Confusion during Arthroscopic Surgery

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

Myat Su Yin
Mahidol-Bremen Medical Informatics
Research Unit, Faculty of Information
and Communication Techology
Nakhon Pathom, Thailand
myat.su@mahidol.ac.th

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

Ratthapoom Watcharopas
Faculty of medicine, Ramathibodi
Hospital, Mahidol University
Bangkok, Thailand
poom911@hotmail.com

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

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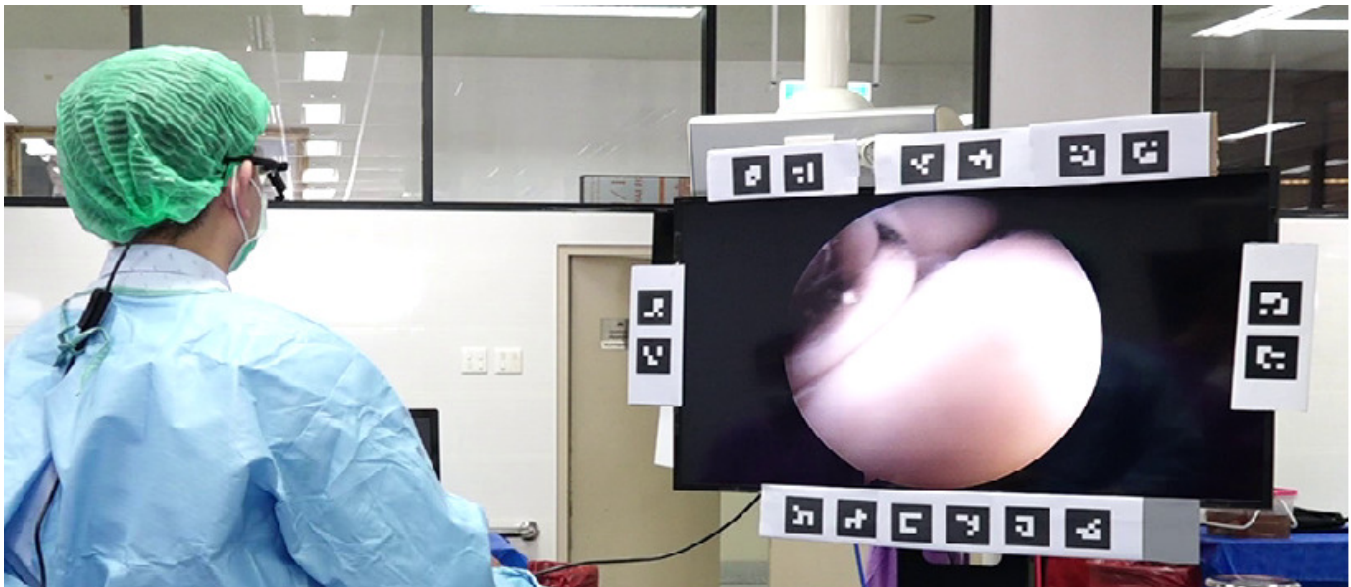enkelejda.kasneci@uni-tuebingen.de

Figure 1: Operating site with surgeon in front of 4k-screen, showing output of arthroscope.

ABSTRACT

During arthroscopic surgeries, surgeons are faced with challenges like cognitive re-projection of the 2D screen output into the 3D operating site or navigation through highly similar tissue. Training of these cognitive processes takes much time and effort for young surgeons, but is necessary and crucial for their education. In this study we want to show how to recognize states of confusion of young surgeons during an arthroscopic surgery, by looking at their eye and head movements and feeding them to a machine learning model. With an accuracy of over 94% and detection speed of 0.039 seconds, our model is a step towards online diagnostic and training systems for the perceptual-cognitive processes of surgeons during arthroscopic surgeries.

CCS CONCEPTS

• **Human-centered computing** → **User models**; *Ubiquitous and mobile computing systems and tools*; • **Applied computing** → **Health informatics**; • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

gaze, head, eye, tracking, machine learning, random forest, surgery, medicine, confusion

1 INTRODUCTION

Advancements in computer science have typically been a motor for new applications in fields like medicine. Next to classical imagery techniques like magnetic resonance imaging (MRI) [8] or arthroscopy [11], nowadays, the interaction between surgeons and their patients or instruments are increasingly being investigated. There are a lot of new sources of information, e.g. about the vital parameters of the patient or new perspectives/views of the operating site, which are shown to the physicist. They are all meant to improve the work of the surgeon. However, all these new advancements come with a certain level of complexity. Surgeons need to learn how to operate and benefit from these applications. For example, in arthroscopy, the surgeon needs to transfer the 2D image on the scope output into the 3D tissue of the patient. Information is shown on the screen, but navigation takes place on the operating site with a multidimensional instrument. This translation already poses a challenge.

Even in medical image reading, Brady et al. [3] estimated that the miss rate for interpreting the results correctly, may be up to 30% in some ares of radiology. For arthroscopy there is no such study, but arthroscopic surgery is a much more complex procedure than image reading, as surgeons are usually under time and success pressure, while working with patients and the stimulus is constantly and dynamically changing. Therefore, ways to teach surgeons to use these new technologies optimally, are as important as the developments of such. This is where human computer interaction comes into play. Methods of human-computer interaction find their way into the world of medicine. Indeed, there are multiple goals to pursue. Besides, i.e. touchless interaction techniques [20], the recognition of strategies of surgeons during an operation [24] are investigated. The recognition of skill [1, 25, 30, 31] or states of confusion [26, 33] of surgeons play a central role in interaction design, as they can help to draw a picture of a surgeons' skills and to find weak-spots which need to be focused on in training. This is done to maximize the output of surgeons and to improve their training. Along with confusion, often frustration or disengagement are involved, if the confusion lasts for too long [5]. Pachman et al. [21] summarized different approaches of the last few years and show that multiple ways of detection have been tried, e.g. facial expressions [19, 32] or learners' postures [4]. D'Mello et al. [4] postulated that models based on a single source had high error rates. Thus, later research focused on multiple sources to detect confusion but could not be fully automated, as external judges needed to be involved [6]. Lallé et al. [15] studied various combinations to predict occurrences of confusion. They reported 61% prediction rate with 193 features.

Thus, their system is hardly usable online since computation of these 193 features takes too much time. Similarly, the model of Shi et al. [23] is hardly usable in an online setting, too, as their model is too complex and thus needs too much computation power and time. Further, they use images on a display, which might cut off environmental influences, thus, preventing the application of natural gaze behavior. Conversely, we use a soft-cadaver in a real surgical setting, which allows the application of natural gaze behavior. We further use a simple but fast and robust model for classification, which allows the usage in an online fashion.

Most often, surgeons need both of their hands for the operation. So new information and interaction techniques need to focus on other modalities than the surgeons hands. One way to address this, is the use of eye tracking technology. This technique can either be directly used as interaction method [20] or as information provider about the skill or current state of the surgeons themselves. And as these devices are getting more ubiquitous, faster and more accurate, there are ever new possibilities to study the gaze behavior of the subject. Eye tracking can serve as a perceptual-cognitive diagnosis system. The interest in using eye tracking as a research method in medicine is growing rapidly (for an overview see Lévêque et al [16]).

There are even studies that focus on assessment of the impact of training with eye tracking, too [14, 28, 29]. Wilson et al. [29], i.e found significant differences in completion time when showing young surgeons a video with the gaze signal of an expert during laparoscopy, compared to only showing the plain video of the surgery or allowing a free viewing phase. There are plenty of such studies, showing that the findings of gaze behavior studies can even be used to optimize and/or shorten the training surgeons need to go through. While eye tracking devices are getting faster and ubiquitous, they produce more data, too. On the one hand more data means more usable information, but on the other hand there is a rise in complexity, too. With more data, there can be more inter-dependencies which are hard to understand and handle, especially with traditional techniques like AOI intersection counts [2, 12, 13, 17, 18]. To allow the analysis of such big data to be much more complex, there is another very important advancement in computer science that has an heavy impact on medicine. Artificial intelligence is applied in a variety of applications in medicine [7, 9, 22, 27]. The ever new potentials of machine learning and especially deep learning enabling even more complex tasks to be solved and more data to be analyzed.

In this work we are focusing on the analysis of data from 15 participants during arthroscopic surgery with so called soft-cadavers. During arthroscopy, the surgeon is mainly focusing on the output of the arthroscope, which shows a plane 2D view of the arthroscopic camera inside the portal hole of the patient. Surgeons need to rely on these images, while they navigate through tissue and bones. A young surgeon with low experience may get confused during navigation, since the structures look pretty similar for untrained surgeons. Expert surgeons can rely on their experience and know which visual clues they can use for navigation. In order to optimize the training of young surgeons, we introduce a real time ready confusion detection model, that recognizes states of confusion of surgeons during arthroscopic surgeries. With the combination of eye tracking, head tracking and machine learning methods, we present a highly accurate and fast classification model. Detections

of such model can be used to find weak-spots of surgeons in real time and signal assistive actions to be made.

2 METHOD

2.1 Data collection

We collected data of 15 surgeons who are all either members of the Orthopedic Department in the Faculty of Medicine from Mahi-dol University, Thailand or in the Orthopedics Surgery Residency Program. All subjects were wearing a TobiiGlasses 2 eye tracker (running at 100 Hz) during an arthroscopic surgery of the shoulder on a soft cadaver. The cadaver was placed in front of the surgeon and four feet further away we placed a 4k-screen which shows the output of the scope. During the navigation from the portal hole to the operating side, surgeons were telling verbally where they are and where they go to. They also told when they are confused. Which means they can either not tell their current position inside the joint or how to continue for sure. In relation to the beginning of the operation, we measured these points of time, where the surgeon told to be unsure/confused.

2.2 Feature space

At first we synchronized the eye tracking data with the timing data, by adjusting their timestamps to start at the same time relatively to the start of the surgery. This allows us to find the points of time of confusion inside the eye tracking data. In a next step we cut out a window around every confusion point (+/- one second before and after the event). These pieces of data are considered as "confusion event" samples and the remaining data with no confusion event as "no event" samples.

Each sample contains the following features:

- point of regard (x, y)
- pupil position (average of both eyes)
- pupil diameter (average of both eyes)
- gyroscope (x, y, z)
- accelerometer (x, y, z)

2.3 Classification

To build a random forest model, we split the samples into training and test data set. This is done in a participant wise manner, which means, if a subject is picked to belong to training set, all of their samples belong to training set. We need to do this, as the model would otherwise learn person-specific, so called idiosyncratic, features (for further information, see [10]). We followed two different approaches, for testing with unseen data.

The first approach follows a 2/3-strategy. We randomly pick 2/3 of the subjects for training and count the number of confusion event samples for each. Afterwards, we collect the same amount of "no event" samples from the same subjects. This means for our training set we have the same amount of confusion event samples as no event samples. This firstly leads to a balanced training set (50% confusion event samples and 50% no event samples) and secondly, to a chance-level of 50%, which allows easy interpretation of the results later.

In the second approach, we want to see whether a cross-validation during the training would optimize the results. Thus, we split the training set data by a 5-fold cross validation, which means in every run 1/5 of the data (of the training set) is picked to validate/optimize the model, while 4/5 of the data are used for training the model. After each run, we use the samples of the remaining 1/3 subjects (n=5) to test the classification performance with unseen data.

As we want to use our model in an online fashion, we need to test the classification accuracy (with unseen data) and the classification speed as well. We show the online computability by creating a queue, which consists of n=2000 samples. In our test we keep reading the gaze signal and add one sample to the queue in each step, while the oldest sample is kicked out of the queue. This means at every state the queue has a total of n=2000 samples. The average of each of the features of all samples inside the queue is now computed. These values are now representing the current content of the queue, which we call delta sample. This delta sample is now given to the trained random forest model and to classify it as "confusion sample" or "no confusion sample". To infer the average performance time, we measure the computation time of 100 single runs and calculate the average performance time.

3 RESULTS

Out of 1,266,758 samples, we have 7103 samples with an confusion event and 1,259,655 samples with no event. Out of these samples we collect 7103 confusion samples and 7103 no confusion samples. In every run, we randomly pick 1,000 samples of both to predict their class. The other samples are used for training.

We tested our approach - by randomly assigning training and testing data like aforementioned, in 100 times. The average accuracy of the random forest model is 94.2%. According to the accuracy, the average misclassification cost / loss is 0.0595. Figure 2 shows the development of the loss over all runs as a function of the number of trained trees. The differences are small but noticeable. The approach with test data set is performing a little bit better than the cross-validation approach. Test set approach reaches the best performance of the cross-validation approach (0.11) already with about 25-30 trees. The optimal loss value for the test approach is reached at 50 trees with a misclassification cost of 0.085.

Figure 3 shows the confusion matrix which contains the predictions of all 100 runs. In total we have 50,000 samples for each class. Of class 0 (no event), 47,016 samples out of 50,136 samples were predicted correctly and 3,120 as confusion event samples. Similarly, for class 1 (confusion event), the model predicted 47,023 samples correctly as confusion event and 2,841 samples wrongly as no event. This result is supported, by the average accuracy over all 100 runs of 94.2%.

To measure the performance speed of the model, we measured the computing time of every of the 100 runs. On average the prediction takes 0.039 seconds. This corresponds to a frame rate of 25 fps.

4 DISCUSSION

In this work we presented a random forest model that is able to classify states of confusions of surgeons during an arthroscopic surgery of the shoulder with an accuracy of over 94.2%, by taking only 9 features of eye and head movement into account. In our
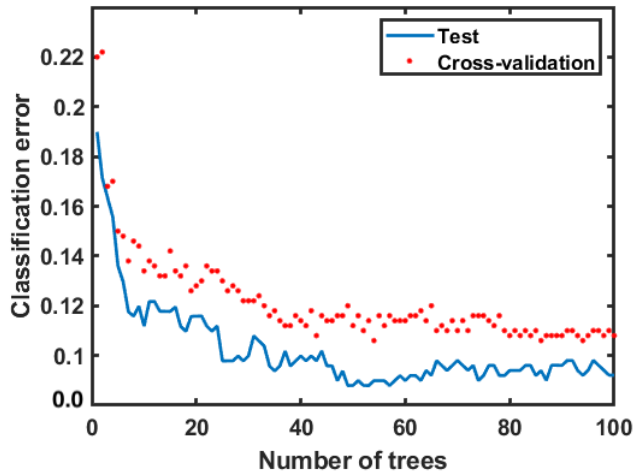
Figure 2: Validation with test data vs. validation with cross-validation as function of number of learners.

Figure 3: Confusion matrix showing number of correct and falsely predicted samples.

calculations, the model was able to provide a prediction of the content of a queue containing n=2000 samples (2 seconds of samples) in 0.039 seconds. This corresponds to the temporal resolution of common head mounted eye trackers which run at a frame rate between 25-30 fps. The speed may need to be optimized, to allow the application to higher paced field cameras. But in the scenario of a surgery, the speed is not a crucial part, rather, a high detection rate is important. With the detection of confusion states, one can help surgeons to proceed, either pointing out visual clues, which may be used by expert surgeons to navigate, or drawing arrows on the output of the arthroscope which tells the surgeon where to navigate next. Another possible usage of the knowledge of states of confusion can be to augment the whole output by describing the scene by segmenting and labeling each bone or tissue. Or simply name the shown parts in the output. There are multiple ways of supporting the confused surgeon. Depending on the state of expertise, the level of support may be chosen, to allow different
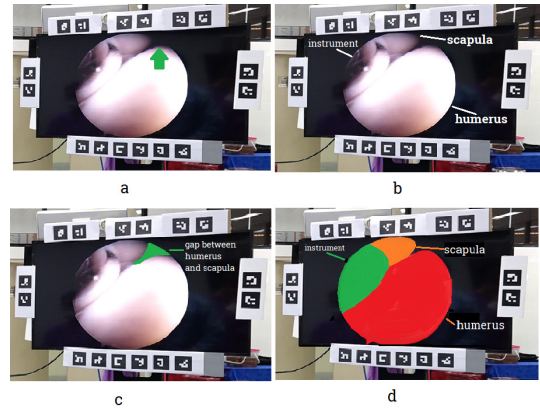

Figure 4: Different kinds of support for a confused surgeon.

skilled surgeons, to train their different weak spots. The different kinds of support can be seen on Figure 4. a) shows a simple arrow, which tells the surgeon where to go next with the arthroscope. b) shows more support by naming the single party of the output, so the surgeon knows which parts are involved and may remember how to proceed. Figure 4, c shows a similar output like a), but there are only visual clues highlighted, and d) this help would provide the most support, by segmenting and coloring the single parts in different colors and name them, accordingly.

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] Cristina Almansa, Muhammad W Shahid, Michael G Heckman, Susan Preissler, and Michael B Wallace. 2011. Association between visual gaze patterns and adenoma detection rate during colonoscopy: a preliminary investigation. *American Journal of Gastroenterology* 106, 6 (2011), 1070–1074.
[3] Adrian P Brady. 2017. Error and discrepancy in radiology: inevitable or avoidable? *Insights into imaging* 8, 1 (2017), 171–182.
[4] Sidney D'Mello and Art Graesser. 2009. Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence* 23, 2 (2009), 123–150.
[5] Sidney D'Mello and Art Graesser. 2014. Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta psychologica* 151 (2014), 106–116.
[6] Sidney K D'mello and Arthur Graesser. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction* 20, 2 (2010), 147–187.
[7] Pavel Hamet and Johanne Tremblay. 2017. Artificial intelligence in medicine. *Metabolism* 69 (2017), S36–S40.
[8] Karl H Höhne, Henry Fuchs, and Stephen M Pizer. 2012. *3D imaging in medicine: algorithms, systems, applications*. Vol. 60. Springer Science & Business Media.
[9] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4 (2019), e1312.
[10] Benedikt Hosp, Florian Schultz, Oliver Höner, and Enkelejda Kasneci. 2020. Eye Movement Feature Classification for Soccer Goalkeeper Expertise Identification in Virtual Reality. *arXiv preprint arXiv:2009.11676* (2020).
[11] Robert W Ike, William J Arnold, and Kenneth C Kalunian. 2021. Arthroscopy in rheumatology: where have we been? Where might we go? *Rheumatology* 60, 2 (2021), 518–528.
[12] Brendan S Kelly, Louise A Rainford, Sarah P Darcy, Eoin C Kavanagh, and Rachel J Toomey. 2016. The development of expertise in radiology: in chest radiograph interpretation,"expert" search pattern may predate "expert" levels of diagnostic accuracy for pneumothorax identification. *Radiology* 280, 1 (2016), 252–260.
[13] Ellen M Kok, Anique BH de Bruin, Jimmie Leppink, Jeroen JG van Merriënboer, and Simon GF Robben. 2015. Case comparisons: an efficient way of learning radiology. *Academic radiology* 22, 10 (2015), 1226–1235.

[14] Elizabeth A Krupinski, Anna R Graham, and Ronald S Weinstein. 2013. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Human pathology* 44, 3 (2013), 357–364.

[15] Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2016. Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data.. In *IJCAI*. 2529–2535.

[16] Lucie Lévêque, Hilde Bosmans, Lesley Cockmartin, and Hantao Liu. 2018. State of the art: Eye-tracking studies in medical imaging. *Ieee Access* 6 (2018), 37023–37034.

[17] Michael Mackert, Sara E Champlin, Keryn E Pasch, and Barry D Weiss. 2013. Understanding health literacy measurement through eye tracking. *Journal of health communication* 18, sup1 (2013), 185–196.

[18] 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.

[19] Bethany McDaniel, Sidney D'Mello, Brandon King, Patrick Chipman, Kristy Tapp, and Art Graesser. 2007. Facial features for affective state detection in learning environments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29.

[20] Andre Mewes, Bennet Hensen, Frank Wacker, and Christian Hansen. 2017. Touchless interaction with software in interventional radiology and surgery: a systematic literature review. *International journal of computer assisted radiology and surgery* 12, 2 (2017), 291–305.

[21] Mariya Pachman, Amaël Arguel, Lori Lockyer, Gregor Kennedy, and Jason Lodge. 2016. Eye tracking and early detection of confusion in digital learning environments: Proof of concept. *Australasian Journal of Educational Technology* 32, 6 (2016).

[22] AN Ramesh, Chandra Kambhampati, John RT Monson, and PJ Drew. 2004. Artificial intelligence in medicine. *Annals of the Royal College of Surgeons of England* 86, 5 (2004), 334.

[23] Qiyun Shi, Laura Warren, Gustavo Saposnik, and Joy C MacDermid. 2013. Confusion assessment method: a systematic review and meta-analysis of diagnostic accuracy. *Neuropsychiatric disease and treatment* 9 (2013), 1359.

[24] 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.

[25] Stefanie Speidel, Michael Delles, Carsten Gutt, and Rüdiger Dillmann. 2006. Tracking of instruments in minimally invasive surgery for surgical skill analysis. In *International Workshop on Medical Imaging and Virtual Reality*. Springer, 148–155.

[26] Mark J Stillman and Lisa A Rybicki. 2000. The bedside confusion scale: development of a portable bedside test for confusion and its application to the palliative medicine population. *Journal of palliative medicine* 3, 4 (2000), 449–456.

[27] Peter Szolovits. 2019. *Artificial intelligence in medicine*. Routledge.

[28] Samuel J Vine, Rich SW Masters, John S McGrath, Elizabeth Bright, and Mark R Wilson. 2012. Cheating experience: Guiding novices to adopt the gaze strategies of experts expedites the learning of technical laparoscopic skills. *Surgery* 152, 1 (2012), 32–40.

[29] Mark R Wilson, Samuel J Vine, Elizabeth Bright, Rich SW Masters, David Defriend, and John S McGrath. 2011. Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: a randomized, controlled study. *Surgical endoscopy* 25, 12 (2011), 3731–3739.

[30] Jie Ying Wu, Aniruddha Tamhane, Peter Kazanzides, and Mathias Unberath. 2021. Cross-modal self-supervised representation learning for gesture and skill recognition in robotic surgery. *International Journal of Computer Assisted Radiology and Surgery* (2021), 1–9.

[31] 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.

[32] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2008. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31, 1 (2008), 39–58.

[33] Yun Zhou, Tao Xu, Shiqian Li, and Shaoqi Li. 2018. Confusion State Induction and EEG-based Detection in Learning. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 3290–3293.