Confusion Detection Dataset of Mouse and Eye Movements

Michal Hucko
Slovak University of Technology in Bratislava, UXI Research Center, Slovakia
michal.hucko@stuba.sk

Robert Moro
Slovak University of Technology in Bratislava, UXI Research Center, Slovakia
robert.moro@stuba.sk

Maria Bielikova
Slovak University of Technology in Bratislava, UXI Research Center, Slovakia
maria.bielikova@stuba.sk

ABSTRACT

Since emotion detection mostly employs supervised machine learning, big labeled datasets are needed to train accurate detectors. Currently, there is a lack of the open datasets, especially in the domain of confusion detection on the web. In this paper, we introduce a confusion detection dataset comprising of two modalities – the mouse movements and the eye movements of the users. The dataset was gathered during a quantitative controlled user study with 60 participants. We chose a travel agency web application for the study, where we carefully designed six tasks reflecting the common behavior and the problems of the day-to-day users. In the paper, we also discuss the issue of labeling emotional data during the study and provide exploratory analysis of the dataset and insights into the confused users' behavior.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; *Web-based interaction*; • **Mathematics of computing** → Exploratory data analysis.

KEYWORDS

confusion detection, affective computing, labeled dataset, mouse movements data, eye movements data, emotions labeling

ACM Reference Format:
Michal Hucko, Robert Moro, and Maria Bielikova. 2020. Confusion Detection Dataset of Mouse and Eye Movements. In *28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), July 14–17, 2020, Genoa, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3386392.3399289

1 INTRODUCTION

Although emotions influence our behavior and decisions and are our natural reactions to events that we observe or that have an effect on us, the machines with which we interact remain largely ignorant of them. This shortcoming is addressed by the *affective computing* discipline [24]. One of the emotions relevant for machines to observe in humans is *confusion*, which typically indicates that a user has a problem solving his or her task. By detecting the exact moment of user confusion, the application interface can be adjusted to increase the user experience. Important for automatic emotion (confusion) detection is the fact that it can be inferred by observing the bodily feedback [19], such as facial expressions, pupil dilation, posture, heart rate, galvanic skin response, etc.

In our work, we focus on the issue of user *confusion detection* on the web. Although there are some available datasets for emotion detection in general which we summarize in Section 2, we can observe that they often lack more readily available modalities. While many datasets contain facial expressions of the users, the information on the users' behavior during a natural interaction with an application (e.g., their mouse movements) is largely missing. This is even more pronounced in case of confusion detection. There is a lack of open datasets for this issue, which means in practice that it is almost impossible to benchmark and compare confusion detectors which hinders further advancement in the field.

To bridge this gap, we introduce in this paper a novel labeled dataset dedicated to confusion detection in a web application. The main novelty of the presented dataset lies in the provided modalities, which (as far as we know) are not used together in any existing emotional dataset. Namely, it comprises *mouse movement* and *eye movement* data of users during their interaction with a web application. The dataset was obtained during a quantitative user study with 60 participants in a controlled environment of UXI Research Center [2] at the Slovak University of Technology in Bratislava. The study was already briefly described in [12], but here we provide its detailed methodology, publish the resulting dataset and discuss the results of its exploratory analysis.

Besides the raw data, the dataset contains labels of the moments of the user confusion, which were provided by the users by clicking on a confusion button integrated in the application interface. The presented dataset can be used to train detectors to detect confusion in the *real time* (i.e., in the order of tens or hundreds of milliseconds) as opposed to the approaches detecting the confusion after longer time periods (e.g., at the end of a session).

We believe the presented dataset can be used by a wider affective computing and user modeling community to train and compare confusion detectors. The provided modalities (especially mouse movements) are generally available also in other application domains and thus present an interesting extension of approaches typically based on facial expressions.

2 RELATED WORK

There is a comprehensive recent survey article by Pooria [25] which summarises available datasets for emotion detection used in the literature. Based on the summary, we can see that most of these datasets are suited for emotion and sentiment recognition from

UMAP '20 Adjunct, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-7950-2/20/07…$15.00
https://doi.org/10.1145/3386392.3399289

facial expressions or speech. They provide audio and video modalities as is the case, e.g., for *HUMAINE* dataset [9], Belfast database [8], *SEMAINE* dataset [18], *IEMOCAP* dataset [4], or *eNTERFACE* dataset [17]. Some datasets also provide text modality, e.g., ICT-MMMO [30] or MOUD [23]. However, there is a lack of datasets for real-time emotion detection from natural users' interaction with a computer, i.e., datasets that would include mouse interaction (or eye movement) data. At the same time, these modalities would be useful to have, since they are readily available and easy to collect (in case of mouse interaction) and have been shown to convey information on users' confusion [15, 21]. As to the confusion detection itself, the researchers are left to collect and label their own datasets, which hampers progress in the field and makes it very hard to compare the existing approaches.

Detecting user emotions—such as moments of user confusion—in the real-time has a potential to improve user experience on the web. If the users are confused (i.e., they do not know what do next), it might discourage them to use the application at all. If they do not have a choice (e.g., in case of the applications used by employees in the companies), low user experience and the resulting user confusion badly reflects on the time needed to train new employees, on their efficiency and their motivation to continue with such a job.

One of the most popular applications of emotion detection is the educational (or technology-enhanced learning) domain. Santos reviewed 26 different works of emotion and personality detection in this domain [26]. Although probably most approaches employ facial expression recognition [1, 20], there are some that utilize also data from other sensors, e.g., EEG [16] or keyboard [14].

Working specifically on confusion detection, Pentel proposed a method based on the mouse moves in a computer game [21, 22]. The game was an advent calendar where players clicked on the numbered rectangles (from 1 to 24) in ascending order as fast as possible with respect to accuracy. Pentel used a concurrent think-aloud protocol for labeling the confused moves and then used the data to train the classifier.

Other example of confusion detection approach is the work of Lallé et al. [15]. They tracked users' eye movements in a web application for house price visualisation using an eye tracker. To train the real time confusion detector, they used dynamical features of the eye movements called fixations and saccades [10] together with the pupil dilation features and the head distance from a screen.

Even though both works (that of Pentel and Lallé) report accuracy of their proposed approaches, a direct comparison (and consequent assessment of generalizability and predictive power of the used mouse and gaze features) is very hard because of different applications used for data collection.

The two approaches also differ in the emotion labelling approach they adopted. Labelling of the collected data is a serious challenge in training emotion detection models, since majority (if not all) approaches rely on the supervised classification methods. According to [6], we can distinguish three major groups of labeling approaches: verbal approaches, questionnaire approaches, and approaches leveraging the software elements.

Verbal approaches use either concurrent or post hoc verbal protocols. *Concurrent verbal protocols* employ speaking aloud about the emotion a participant is feeling while solving a task. It was used in a variety of emotion detection works, e.g., [7, 21]; however, it is problematic to employ this procedure in studies dealing with tens or hundreds of participants. Moreover, verbalizing user's emotion can influence the gaze patterns [6]. As there is a close link between the eyes' gaze and the mouse movements [5], this effect can have the same negative impact on the mouse logs. It can be alleviated by using *post hoc verbal protocols* that involve a retrospective inspection of participants' behavior by the participants themselves *after* the task. It was used by, e.g., Pentel for confusion labeling [22].

Another approach of obtaining the ground truth labels—using a questionnaire—can also be divided into concurrent and post hoc protocols. An example of the concurrent protocols is a popup question that appears based on a certain condition [27, 29]. The post hoc protocols are more often used with the emotion recognition approaches. Some works employ the Self-Assessment Manikin (SAM) [28, 31] devised in the work of Bradley [3]. It measures the levels of arousal and valence by choosing a picture. There are other approaches employing Likert scale-like questions about the target emotion [7, 21].

The emotion labeling approach, which is the most inspiring for our work, was presented in the work of Conati et al. [6]. The authors classify this type of gathering the ground truth labels into the category of *labels via interface input*. The method was used in the work of Lallé [15]. During the experiment, a software button was placed in the top right corner of the screen with the text *"I am confused"*, which the participants were instructed to click on when they felt confused. Leveraging this approach, the authors were able to identify the exact moment in time when the emotion occurred. To obviate the misclicks, after each session the authors also leveraged the post hoc verbal protocol.

Although this labelling approach may lead to collecting a smaller amount of labeled data [6] (if the users do not click on the button), we decided to use it to label moments of users' confusion as we describe in more detail in Section 3.3.

3 STUDY METHODOLOGY

Since we were interested in collecting users' interaction with (and moments of users' confusion in) a real-world web application, we used a web application of a popular Slovak travel guide agency *FiroTour*[1] for the study. We selected it for multiple reasons: (*i*) interface of the application is similar to many other e-commerce applications, so the findings should be representative enough also of other web applications in this or similar domains, (*ii*) it has thousands of daily visitors and (*iii*) although it provides a production-tested environment with a high level of user experience, the customer center at *FiroTour* nevertheless receives daily tens of customer requests regarding the complexity of their web application.

3.1 Study apparatus

The user study took place at the User Experience and Interactions (UXI) Research Center at the Slovak University of Technology in Bratislava[2] [2]. The laboratory consisted of 20 computers equipped with 60Hz Tobii X2-60 screen-based eye trackers mounted at the bottom of the 24-inch screens The computers were also equipped with a standard keyboard and a corded mouse Logitech M90 with 1000 DPI (dots per inch).

[1] https://www.firotour.sk
[2] http://uxi.sk

For logging the mouse activity within the web application, we used a custom implementation of the mouse logger with the sampling frequency of 60Hz. We used a *JavaScript* implementation of the logger which listens for all mouse activity in the web application [13], which was directly integrated into the source code of the *FiroTour* application.

The study was set up in *iMotions*[3], which handled the communication with *Tobii* eye trackers and provided a fixation filter.

3.2 Participants

There were 60 participants (students of the Slovak University of Technology in Bratislava) who took part in the study. They were all students of the *Human-Computer Interaction* course in the second year of their undergraduate study and they were awarded extra credit for their participation. They were all between 20 to 23 years old and had the same expertise in the field of computer science. There were 50 male and 10 female participants and all had a previous experience with the tested web application. They were randomly assigned to 4 separate runs which took place over 2 days with maximum of 16 participants being recorded at the same time.

3.3 Collecting confusion labels

To record the moments of users' confusion, we leveraged the *labels via interface protocol* discussed in the work of Conati et al. [6]. We adjusted the implementation of a confusion button introduced in the work of Lallé [15] where they used a button with the text *"I am confused"*. We carried out a small pilot study with just a few participants to test the infrastructure, the study setup and the users' perception of the button. We kept the button position in the top right corner of the viewport (similar to [15]), but we changed its label from *"I am confused"* to just the exclamation mark (see Figure 1). The change was made after repeated discomfort of the participants to report a negative emotion during the pilot study (we hypothesize that it was the result of a social desirability bias problem already reported in [11] to cause the participants to be less likely to report negative affective states). From a practical point of view the confusion button was directly integrated into the *FiroTour*'s web application with a custom *JavaScript* snippet.

3.4 Study scenario

Upon entering the laboratory, the participants were seated in front of the computers equipped with the eye trackers and running the *iMotions* software. They were asked to sign the informed consent including GDPR (General Data Protection Regulation) agreement. The study began with a short presentation describing the study and the function of the confusion button. We used the instruction similar to the one used in the work of Lallé [15] translated into the Slovak language and enriched with the examples taken from the *FiroTour* application[4].

[3]https://imotions.com

[4]In English, it reads as follows: "You will see a button labeled with the exclamation mark in the top right corner of the screen. We kindly ask you to click on the button every time when you would need someone to help you with the task or the things go in a different way than expected. You should click on the button also when you would like to leave the web page because you cannot find the information needed for completing the task. You can navigate anywhere in the *FiroTour* domain. You are expected to click on the button also when you would like to use a different application (other than *FiroTour*) to solve the task (e.g., Google). After clicking on the button, there will be no reaction of the system. You can continue solving the task or go to a next one by clicking on F10."

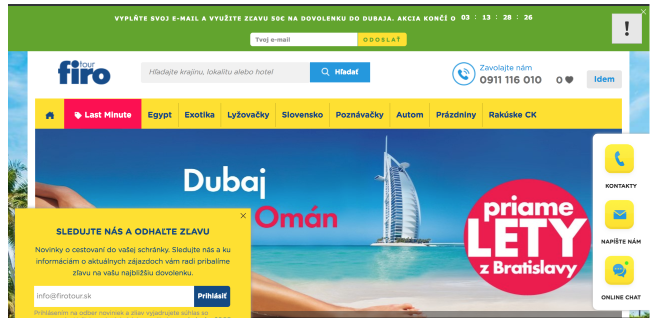

Figure 1: The landing page of the *FiroTour* application (in Slovak). The confusion button (the exclamation mark) for gathering the ground truth confusion labels is placed in the top right corner of the screen. It was designed not to unnecessarily attract users' attention, but to still be well visible.

Next, the participants started the study set up in *iMotions*. It consisted of a pre-study questionnaire, eye tracker calibration and six blocks (each containing one task). The pre-study questionnaire consisted of questions about the participants' gender and age and an open answer field for the user id which we later used for mapping the mouse logs to the eye tracker logs. For the eye tracker calibration, we used a standard 9-point calibration sequence [10]. After the calibration, a sequence of six blocks (tasks) followed.

Each block consisted of 3 elements (instruction, url, and survey). The first element was the *instruction element* with the task description in the Slovak language. After reading the instructions, the participants could move on to the next element by pressing the continue button in the middle of the screen or by pressing the F10 key. The second element was the *url element* which represented the actual task. It began by the Chrome browser being opened in the incognito mode to ensure the browser uniformity across the devices (empty cache, cookies, local storage, etc.). While solving the tasks, the participants were free to surf the internet as they needed and there were no time limit on the task.

Upon finishing the task, the participants pressed the F10 key to move on to the last element of the block – the *survey element*. This element consisted of three questions. The first one was an open question for the task answer. The participants were asked to fill in the desired information for each task. The second question was a binary question about pressing the confusion button during the task. The last question was an open answer question about reasons for pressing the confusion button. The order of the tasks (blocks) was randomized. The whole study was one hour long on average.

3.5 Study tasks

We consulted the study task design with the *FiroTour*'s customers service to create tasks that would reflect normal customer behavior (typical use cases of the users of the web application) as well as tasks known to be problematic and tasks requiring to use newly-added application features. The exact wording of the tasks was discussed with the customer service and piloted with a group of participants.

We designed six tasks, which required participants to find certain information or offer on the web page. Each participant started each task at the landing page of the application. This enables us to analyse travel paths within the application more easily and it also reflects common user behavior. Following task descriptions were shown to the participants in the instruction elements (before the start of each task); they were also printed out on a sheet of paper and available to the participants at all time during the study:

(1) Find out whether it is possible to pay with a credit card.
(2) Find out the title and the price of the cheapest offer within the special *Firáčik* program[5].
(3) Find out the name of the newest travelling blog post author.
(4) Find out the price of the cheapest vacation in Turkey with departure from Vienna.
(5) Find out how far *the Port Ghalib Resort* is from the beach.
(6) Find out in which year the *FiroTour* travel agency was founded.

From these, tasks 4 and 5 reflected common user behavior, tasks 1 and 3 problematic behavior and tasks 2 and 6 a newly-added functionality. Our hypothesis was that the problematic and the newly-added tasks would lead to more confusion button presses as the routine (common) tasks.

4 CONFUSION DATASET ANALYSIS

The collected confusion dataset consists of two main folders:

(1) *Mouse data* (`MouseTrackingData`). It contains complete mouse logs for each task of the study for each participant. It also contains `confusion` events logged when the participants pressed the confusion button.
(2) *Eye tracking data* (`EyeTrackingData`). It contains complete eye position logs for all parts of the study of each participant.

The dataset is available online[6]. It is important to note that we provide mouse data only from the web application (the actual tasks), but the provided eye tracking data cover all parts of the *iMotions* session (including instructions, surveys, calibration, and the pre-study questionnaire). The data from the two modalities can be joined using `userId` field. At the same time, it is the only (randomly assigned) identifier of a user. The dataset does not provide any further information about the participants based on which their identify could be identified.

During the study, we noticed some technical issues with some of the computers in the laboratory. For this reason, from the 60 original participant recordings, we ended up with 57 valid mouse recordings and 56 eye tracker recordings. Together, there are 54 participant recordings with both the mouse and the eye tracker data, which we publish in the dataset. We applied some minor processing on the raw data gathered during the study, in which we unified the timestamp format used in the mouse and the eye tracker logs (we used the YYYYmmdd HH:MM:SS.fff format with a millisecond precision).

4.1 Mouse and eye movements data

Mouse data folder labeled as `MouseTrackingData` consists of 57 `csv` files (one for each user) labeled with the unique `userId`. Each row of the `csv` files consists of a timestamp, user and task id, mouse event (one of the `mousemove`, `scroll`, `mousedown`, `mouseup`, and `confusion`), coordinates of the mouse cursor position ant the *HTML* element underneath the mouse cursor (identified by its Xpath).

The folder labeled as `EyeTrackingData` consists of 56 `csv` files (one for each user) labeled with a unique `userId`. Each row of the `csv` files consists of a timestamp, user and task id, user age and gender and eye tracker data (fixation coordinates and duration, pupil diameter of both eyes, and user's distance from the screen).

It is important to note that the `taskId` field differs from the one mentioned in the mouse data. In case of eye tracker data, we tracked the gaze also during the instruction (`taskId` prefixed with `Instruction`) and survey (`taskId` prefixed with `Survey`) part of the study. The gaze activity within the actual mouse tracked task is prefixed by `Url`. For example, `Url_2` can be mapped to the task with `taskId` 2 in the mouse logs.

Also the mouse coordinates and the eye gaze coordinates, even if seemingly similar, are two slightly different measures. X and Y coordinates of the mouse represent the position of the cursor within the rendered document object model (*DOM*), while the gaze position represents the information about the location of the gaze within the bounding box of the eye tracker. The coordinates match if a web page is not scrolled down. However, if a user scrolled down the page, the mouse coordinates take this into account, while the gaze coordinates do not. However, since the mouse data log contains scroll events of the users, it is possible to adjust the gaze coordinates to take scrolling into account (in which case they would have the same meaning as the mouse coordinates).

4.2 Confusion log

To log the presence of confusion within the dataset, we created a special mouse event marked as `confusion`. The event is stored only within the mouse activity data and from the log structure it only carries the information about the URL where the event occurred, the time when the confusion button was clicked, and the identification number of the task, during which it occurred. It is important to mention that the participants were allowed to click as many times as they needed on the confusion button during the study. For this reason, it is possible to find multiple occurrences of the confusion event within one task.

Overall, there are 108 confusion button presses in the dataset. There were situations with multiple confusion button clicks during a task. Namely, there were 9 cases when the user pressed the confusion button twice within a task and two cases when the button was pressed three times. In each of these cases the time period between two consecutive clicks was longer than two seconds, which means these were not double clicks (or misclicks), but the users' intention was to communicate that they are confused.

A complete summary of all button clicks across all study tasks and users is provided in Figure 2. For the sake of completeness we also provide the distribution of multiple clicks (two or three clicks) within the tasks. As we can see, the most problematic was the task 2 which contains a newly-added feature to the application. The tasks 1 and 4 were the second (and third) most confusing tasks. One of these tasks was a problematic task for general *FiroTour* users (task number 1) and the other one was a common task in the *FiroTour*

[5]The Firáčik program contains special offers for families with children.
[6]https://bit.ly/2UCOwX3

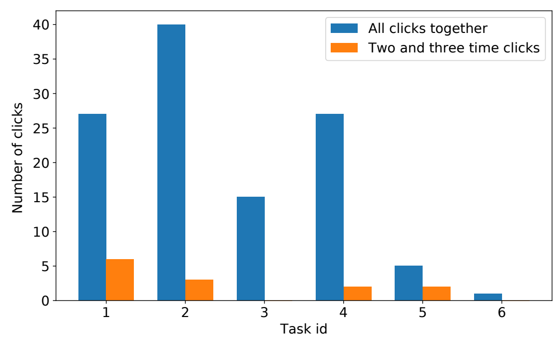

Figure 2: Distribution of the aggregated number of confusion button clicks across the tasks. We also provide the column representing number of multiple (two or three) clicks on the confusion button across the tasks.

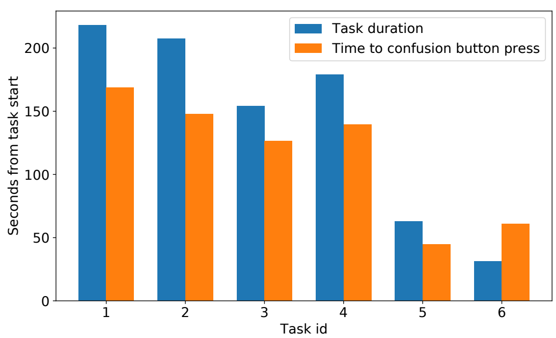

Figure 3: Average number of seconds to the first confusion button click for each task compared to the average total duration of the task. The time to the confusion button press is larger than the average total duration time of the task 6, since there was only one user reporting confusion who took longer than the average users to finish the task.

application (based on the *FiroTours*'s customer service experience). Surprising was that the task 6, which also contained a newly-added feature and therefore could pose some difficulty to the users, turned out to be a trivial task for the participants based on the number of confusion button clicks.

Next, we looked at the times within the study tasks when the users started to get confused. We provide a detailed graph in Figure 3, in which we compare this time to the average total duration of each task. We can see a common trend across the tasks that the users tend to press the button 80% into the task's total duration.

We also analysed the total task duration for users based on whether or not they got confused during the tasks. In Figure 4 we can see a comparison of the task times of the confused users and those who did not report confusion. Although the confused users took overall longer to solve the tasks on average, based on the means and the standard error, the only significant difference seems

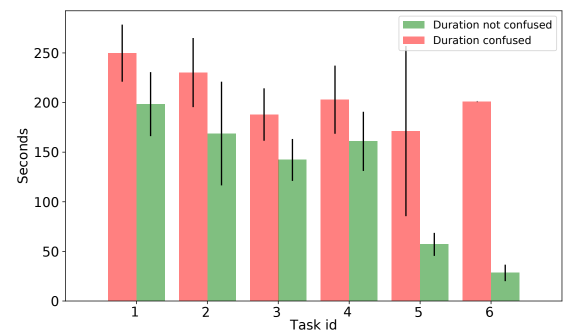

Figure 4: The comparison of the task duration (means with $\pm 2 \times SE$ error bars) across the tasks based on the confusion occurrence. The users were marked as confused during the task if they pressed the confusion button at least once. In case of task 6, we lack the confused data for the standard error to be computed (there was only one confused user).

to be in tasks 3 and 5, one of which (task 3) tested a problematic application feature and the other (task 5) a standard user interaction within the application. The lack of a significant difference in the task time between the confused and not confused users in other tasks can be due to multiple reasons: the users take different time to solve the tasks, so someone can solve the task in less time even though he or she had to overcome a moment of confusion than someone else who was not confused at all. Also, if someone gets confused and gives up, it might show as a shorter task duration, because the task was not completed. In any way, it suggests that simple time tracking is not enough to distinguish the confused users and more sophisticated features (derived from the mouse or eye movement data provided in the dataset) are necessary.

4.3 Post-task survey

After each task, the users were asked to fill in a post-task survey. We asked them whether they pressed the confusion button during the task – their answers matched with the occurrence of a logged `confusion` event. In the open answer field regarding reasons for confusion button press we found mostly expected answers like *"I was confused"*, *"I did not know where to click next"*, etc. which suggests that the participants correctly interpreted and used the confusion button.

We also verified whether the participants did not use the confusion button as a sort of *I give up* button. This does not seem to be the case, since although the majority of the participants did not solve the task after pressing the button, there were many instances of them solving the task even after a button press. This can often occur in a real-life scenario as well; the confused users may try to overcome their confusion and finish the task in the end, even more so if the task is important for them or they do not have an option of using a different application (e.g., in their work).

5 CONCLUSIONS

In this paper we introduced a confusion detection dataset, which we made available online. Compared to the existing affective datasets, it is unique in several ways: (*i*) it captures users' natural interaction (even though in a controlled environment) with a real-world web application, (*ii*) it contains two modalities (the mouse movements and the eye movements) which are not common in other datasets, even though the former is readily available in most computer-human interaction scenarios, and (*iii*) it is annotated by the users themselves at the moments of confusion occurrence leveraging labels via interface protocol. In addition, while other existing datasets are concerned with different types of emotions which are more easily discerned from the facial expressions, our presented dataset captures confusion, which can occur in many application scenarios and has a direct connection to the usability problems. On the other hand, a limitation of our dataset is that there may be many types of confusion manifesting differently in different domains, so our dataset might have captured only some of these types.

The dataset can serve as a benchmark for comparison of confusion detection algorithms using either mouse or eye movement data or both. The first results using solely mouse movement data were presented in [12]. The algorithms can aim to detect confusion at the task granularity or in the real time based on the user activity in a short time interval. The dataset provides many opportunities for feature engineering. Various features can be derived from mouse actions (clicks, scrolls) and movements (velocity, acceleration, etc.) as well as from fixations and pupil diameters in the gaze data. It is also possible to analyze the sequence of *HTML* elements that the users interacted with and mine their interaction patterns. Besides confusion detection, the dataset could be used to train and evaluate prediction of a next user action, user intention or task success and be thus of value to a wider user modeling community.

ACKNOWLEDGMENTS

This work was partially supported by the Slovak Research and Development Agency under the contracts No. APVV-15-0508, APVV-17-0267 and by the Scientific Grant Agency of the Slovak Republic, grants No. VG 1/0667/18 and VG 1/0725/19. The authors would also like to thank the FiroTour company for their help in study task design and for allowing data collection in their web application.

REFERENCES

[1] Kiavash Bahreini, Rob Nadolski, and Wim Westera. 2016. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments* 24, 3 (2016), 590–605. https://doi.org/10.1080/10494820.2014.908927

[2] Maria Bielikova, Martin Konopka, Jakub Simko, Robert Moro, Jozef Tvarozek, Patrik Hlavac, and Eduard Kuric. 2018. Eye-tracking en masse: Group user studies, lab infrastructure, and practices. *Journal of Eye Movement Research* 11, 3 (Aug. 2018). https://doi.org/10.16910/jemr.11.3.6

[3] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[5] Mon Chu Chen, John R Anderson, and Myeong Ho Sohn. 2001. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI'01 extended abstracts on Human factors in computing systems*. ACM, 281–282.

[6] Cristina Conati, Enamul Hoque, Dereck Toker, and Ben Steichen. 2013. When to Adapt: Detecting User's Confusion During Visualization Processing. In *UMAP '13 Extended Proceedings*.

[7] Sidney K D'Mello, Scotty D Craig, Jeremiah Sullins, and Arthur C Graesser. 2006. Predicting affective states expressed through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education* 16, 1 (2006), 3–28.

[8] Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. 2000. A new emotion database: considerations, sources and scope. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

[9] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. 2007. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction*. Springer, 488–500.

[10] Andrew T Duchowski. 2007. Eye tracking methodology. *Theory and practice* 328 (2007).

[11] Allen L Edwards. 1957. The social desirability variable in personality assessment and research. (1957).

[12] Michal Hucko, Ladislav Gazo, Peter Simun, Matej Valky, Robert Moro, Jakub Simko, and Maria Bielikova. 2019. YesElf: Personalized Onboarding for Web Applications. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP'19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 39–44. https://doi.org/10.1145/3314183.3324978

[13] Michal Hucko, Robert Moro, and Maria Bielikova. 2020. Scalable Real-Time Confusion Detection for Personalized Onboarding Guides. In *Proceedings of the 20th Int. Conf. on Web Engineering - ICWE'20*. Springer, to appear.

[14] Teoh Kung-Keat and Jasmine Ng. 2016. Confused, bored, excited? An emotion based approach to the design of online learning systems. In *7th Int. Conf. on University Learning and Teaching (InCULT 2014) Proceedings*. Springer, 221–233.

[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] Fu-Ren Lin and Chien-Min Kao. 2018. Mental effort detection using EEG data in E-learning contexts. *Computers & Education* 122 (2018), 63 – 79. https://doi.org/10.1016/j.compedu.2018.03.020

[17] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. 2006. The eNTERFACE'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 8–8.

[18] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing* 3, 1 (2011), 5–17.

[19] Paula M Niedenthal. 2007. Embodying emotion. *science* 316, 5827 (2007), 1002–1005.

[20] C. M. Paxiuba, J. Calado, C. P. Lima, and J. Sarraipa. 2018. CADAP: A student's emotion monitoring solution for e-learning performance analysis. In *2018 International Conference on Intelligent Systems (IS)*. 776–783. https://doi.org/10.1109/IS.2018.8710542

[21] Avar Pentel. 2015. Employing think-aloud protocol to connect user emotions and mouse movements. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 1–5.

[22] Avar Pentel. 2015. Patterns of Confusion: Using Mouse Logs to Predict User's Emotional State. In *UMAP Workshops*.

[23] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proc. of the 51st Annual Meeting of the Assoc. for Computational Linguistics*. 973–982.

[24] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[25] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[26] Olga C. Santos. 2016. *Emotions and Personality in Adaptive e-Learning Systems: An Affective Computing Perspective*. Springer Int. Publishing, Cham, 263–285. https://doi.org/10.1007/978-3-319-31413-6_13

[27] Philip Schmidt, Robert Dürichen, Attila Reiss, Kristof Van Laerhoven, and Thomas Plötz. 2019. Multi-target affect detection in the wild: an exploratory study. In *Proc. of the 23rd Int. Symposium on Wearable Computers*. ACM, 211–219.

[28] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proc. of the 2018 on Int. Conf. on Multimodal Interaction*. ACM, 400–408.

[29] Georgios Tsoulouhas, Dimitrios Georgiou, and Alexandros Karakos. 2011. Detection of learner's affective state based on mouse movements. *Journal of Computing* 3, 11 (2011), 9–18.

[30] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. Morency. 2013. YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. *IEEE Intelligent Systems* 28, 3 (May 2013), 46–53. https://doi.org/10.1109/MIS.2013.34

[31] Philippe Zimmermann, Sissel Guttormsen, Brigitta Danuser, and Patrick Gomez. 2003. Affective computing—a rationale for measuring mood with mouse and keyboard. *Int. journal of occupational safety and ergonomics* 9, 4 (2003), 539–551.