# Modelling students' effort using behavioral data

**Barbara Moissa, Geoffray Bonnin, Sylvain Castagnos and Anne Boyer**
Université de Lorraine – LORIA, Nancy, France
{barbara.moissa, geoffray.bonnin, sylvain.castagnos, anne.boyer}@loria.fr

**ABSTRACT**: Students' effort is often considered a key factor for students' success. It has several related definitions, none of which is widely adopted. In this paper, we define students' effort as the experienced cognitive load, which is the total amount of cognitive resources used during the execution of a given task. We propose an effort model to quantify students' effort based on this construct. Our approach uses behavioral measures (i.e., interaction and eye gaze data). Our preliminary results show that the eye gaze measures have an intermediary relationship with effort, while the interaction measures have a weak relationship with effort and seem slightly complementary to eye gaze measures.

**Keywords**: students' effort, cognitive load, descriptive analytics, eye gaze data, interaction data

## 1 INTRODUCTION

Decades of studies have shown that student's success is strongly dependent on their effort (Hill, 1990; Swinton, 2010; Scariot et al., 2016). Being able to accurately measure students' effort can therefore lead to a better understanding of its relationship with learning outcomes, and to the design of new tools to help teachers identify students who are struggling or not truly engaged in their learning. However, measuring the effort is a particularly difficult task. One difficulty is that despite all the interest raised by this concept, student effort has no widely adopted definition (Meltzer et al., 2001). For instance, some define it as just "the amount of studying" (Schuman, 2001), while others define it in a more specific manner, e.g., "the amount of time and energy that students expend in meeting the formal academic requirements established by their teacher and/or school" (Carbonaro, 2005).

Not surprisingly, researchers have used a variety of approaches to measure students' effort. These approaches include the time spent on learning tasks ( Schuman et al., 1985; Hill, 1990) and grades assigned by teachers (Nagy, 2016; Swinton, 2010). Despite being easy to acquire, the time spent on learning tasks can be considered as unreliable, as a student may spend a long time on some activity precisely because he is not making much effort. Grades given by teachers (and even by students themselves) are time consuming, and cannot be automated.

Other approaches are related to the exploitation of behavioral data. These include the work of (Scariot et al., 2016), who proposed the modeling of students' effort using Moodle's log data as a behavioral measure. Their assumption was that greater students' participation on Moodle means greater students' effort. This assumption makes sense, since having a good attendance, delivering learning tasks on time (or not), and other actions (Carbonaro, 2005), are good attitudes expected from students. Another related study is the one from Huptych et al. (2017) which uses the total number of clicks given by a student in each activity to measure the effort. One downside of this approach is that they are only able to measure students' effort on activities that require students' to interact by clicking.

In this paper, our aim is to continue this last line of research and to propose a model that exploits behavioral data to quantify students' effort. We rely on the Cognitive Load Theory and on its methods for measuring the cognitive load. More precisely, we study the ability of different behavioral measurements to measure the effort, and propose different combination approaches to enhance the accuracy of the measurements, while being applicable to different types of learning activities.

The reminder of this paper is as follows: Section 2 describe the Cognitive Load Theory and some approaches for measuring cognitive load. Section 3 describes the dataset we collected and Section 4 shows our results using this dataset. Finally, Section 5 closes the paper.

# 2 THE COGNITIVE LOAD THEORY

The Cognitive Load Theory (CLT) was proposed by Sweller (1988). According to this theory, learning is the development and automation of schemas in the working memory and the storage of those schemas in the long-term memory for easy access and use (Paas et al., 2003; Leppink, 2017). The theory states that the learning design must take into account the limitations of the working memory in order to avoid underload and overload, allowing the storage of the learning schemas (Leppink, 2017). In case of underload, i.e., if the student does not exert the proper amount of effort, no new information will be stored in the long-term memory. On the other hand, in case of overload, i.e., if the student exerts a great amount of effort, there will be no cognitive resources left to store new information in the long-term memory. Overload typically happens when a task is too hard for a given student. For instance, the task may require too much knowledge in order to be completed. Processes that do not allow the creation of learning schemas may also consume the available cognitive resources, e.g., when a student divides his attention between different information sources in different spaces or times.

Many researchers consider the cognitive load as a form of student effort (Paas & Van Merriënboer, 1993; Paas et al., 2003; Leppink, 2017). Moreover, the CLT gained lots of attention since it was first proposed, becoming a consolidated research field. One goal of this field is to accurately measure the cognitive load. Although it cannot be measured directly, it can be inferred through other measures that are believed to have a high correlation with it (Xie & Salvendy, 2000). Those measures can be classified in four categories: subjective, performance, physiological and behavioral measures (Chen et al., 2016).

*Subjective measures* are a popular way of measuring the cognitive load (Paas et al., 2003; Shi et al., 2007). This approach is based on the assumption that people are capable to introspect their cognitive processes and report the mental effort exerted (Leppink, 2017). It consists in asking the participants to self-assess their cognitive load in the middle of a task (Shi et al., 2007) or immediately after the task is over (Chen et al., 2016), being unsuitable for applications that require real-time data. This approach has been shown to be sensitive to small differences, valid, reliable and unobtrusive (Paas et al., 1994, Paas et al., 2003), and is often used as a baseline to measure the correlation of other measures with the cognitive load (Chen et al., 2016).

*Performance measures* are based on the assumption that the experienced cognitive load will reflect on task outcomes, and correspond to grades, the number of correct exercises, etc. Although in the educational context it is often assumed that the more effort students exert on their learning task, the higher their outcomes will be, it is possible for two students to achieve the same outcome for a given task and exert different levels of effort on it (Paas et al., 2003). This can be explained by the potential differences in previous knowledge of the students (Hau & Salili, 1996). Thus, Paas & Van Merriënboer (1993) argue that performance measures should be combined with other types of cognitive load measures to assess the efficiency of the instructional design.

*Physiological measures* are based on the assumption that the increase of the experienced cognitive load leads to physiological changes (Paas et al., 2003). These changes affect various body properties, such as temperature, heartbeat, pupil dilation, brain waves, etc. (Kramer, 1990). One advantage of the corresponding measures is that they can be captured at a high rate and with a high degree of sensitivity (Paas & Van Merriënboer, 1994), and can therefore capture variations of cognitive load over time. However, they typically require the use of specific technologies that can bias the learning experience.

*Behavioral measures* capture objectively and implicitly the subjects' actions (Chen et al., 2016). Examples of behavioral measures include eye activity such as blink frequency, fixation frequency and fixation duration (Beatty & Lucero-Wagoner, 2000), speech features (Khawaja et al., 2007; Yin et al., 2007), linguistic features (Khawaja et al., 2014), mouse usage (Arshad et al., 2013), digital pen input (Ruiz et al., 2007; Yu et al., 2011), gait patterns (Verrel et al., 2009), and head movements and mouth openness.

Usually, researchers only use these measurements to detect increases and decreases of the cognitive load, or to classify the cognitive load into categories such as low, medium and high. Another key point when measuring the cognitive load is that using several measures instead of just one in isolation usually increases the accuracy of the measurements by reducing noise and eventually overcoming the lack of data (Mulder, 1992; Chen et al., 2016).

## 3 DATASET

Our goal in this paper is to propose an effort model based on behavioral data. In order to acquire such data, we wanted to rely on a course material that would contain various contents (text, images, graphics, etc.) and a clearly defined educational objectives. We chose the context of language learning, because this field allows to acquire knowledge and to restore it in a relatively short time, since the encoding of all aspects of the language and its restitution are mainly based on verbal short-term memory and verbal working memory (Baddeley, 2003).

In order to avoid a bias related to a mastery of the language prior to our study, we chose the Esperanto language, which is in little use and not studied at school. After a comparative study of the different Esperanto online learning sites, we selected the iKurso website[1] for its simple and complete course, accompanied by various exercises.

We recruited participants through a university mailing list. 14 French volunteers passed the test, 8 students, 1 engineer, 2 researchers, 2 PhD students, and 1 post-doctoral student. Our panel is composed of 6 women and 8 men. 7 are between 18 and 25 years old, 3 between 26 and 30 years old, 2 between 31 and 35 years old, and 2 between 36 and 40 years old. None of them knew Esperanto prior to our study (as answered in a questionnaire). In order to maximize the engagement of the participants in the tasks, we decided to set up a lottery whose outcome was dependent on the scores obtained on the final evaluation.

When each subject arrived, we briefly presented the material (an eye-tracker and a computer) and calibrated the eye tracker. The subject was then faced with an instruction page after which the site opened, and the learning phase began. Each subject had the opportunity to browse the different pages of the course with no time limitations. Once the participant was finished with the learning phase, he was directed to an evaluation questionnaire with 21 questions (11 sentences to translate and 10 multiple-choice questions), and could not return to the course. The experiment ended when the subject submitted his answers. Table 1 provides a summary of the scores obtained by the participants.

**Table 1: Summary of users' scores on the evaluation questionnaire**

| | *Grammar score (/11)* | *Vocabulary score (/11)* | *Translation score (/11)* | *MCQ (/10)* | *Global score* |
|---|---|---|---|---|---|
| Mean | 4.28 | 2.92 | 1.92 | 8.78 | 10.23 |
| Median | 3.50 | 2.50 | 1.00 | 9.00 | 10.00 |
| Standard deviation | 3.66 | 2.33 | 2.58 | 1.42 | 3.53 |

During the learning phase, users' gaze data were retrieved using a Tobii X1 Light eye-tracker and the software Tobii Studio. We manually defined 336 areas of interest (AOIs) for all course elements required in the evaluation questionnaire. In total, for each subject, we extracted 18 characteristics (Marchal et al., 2016; Marchal et al., 2018): number of fixation points, cumulative duration, mean and standard deviation of fixations, cumulative length, mean and standard deviation of saccades, sum, mean, and standard deviation of the absolute and relative angles of the visual path; length of the first fixation relatively to the edge of the screen, number of dynamic AOIs obtained with the DBSCAN clustering algorithm, and entropy measures.

## 4 MODELLING STUDENTS' EFFORT

As mentioned in the previous section, none of the subjects had prior knowledge about Esperanto. We therefore assume their effort is the main factor influencing their outcome, and consider their score at the final questionnaire a reliable measure of effort in this case. In order to study how students' effort can be effectively measured and modeled using behavioral data, we now analyze the correlations of

[1] https://ikurso.esperanto-france.org

the different measurements available in the dataset with these scores, and rely on the Spearman's rank correlation coefficient.

### 4.1 Measuring the effort using raw indicators

We first focus on the correlations of the raw indicators extracted from the dataset. Specifically, we provide the values of the following indicators: time spent on task (TaskTime), total number of page views (#Hits), total number of clicks (#Clicks), total number of keystrokes (#Keystrokes), total time of fixations (FixationsTime), and total number of fixations (#Fixations). The correlations between these indicators and the global scores obtained by the participants are shown in Table 2. As can be seen, only small correlations were obtained. The strongest correlation we found is the one with the total number of page views (#Hits) with a coefficient of 0.28. Surprisingly, the weakest correlations are the eye gaze indicators (FixationsTime and #Fixations) with correlations coefficients near to zero.

**Table 2: Raw indicators correlation with scores**

| Type | Indicator | Correlation | Level | |
|---|---|---|---|---|
| Interaction | TaskTime | 0.21 | Low | |
| | #Hits | 0.28 | Low | * |
| | #Clicks | 0.14 | Low | |
| | #Keystrokes | 0.10 | Low | |
| Eye gaze | FixationsTime | -0.08 | Low | * |
| | #Fixations | 0.00 | Low | * |

### 4.2 Combining several interaction indicators

As mentioned previously, one means of increasing the accuracy of the measurements is to combine different measures. We thus created new indicators using the three following types of combinations:

1. *Average time between actions:* These combinations correspond to the total duration of the session divided by the number of occurrences of a type of action. These actions are pages views (#Hits), clicks (#Clicks) and keystrokes (#Keystrokes). We also combined all of them by adding them together (#Actions):

$$AvgTimeBetweenHits = TaskTime \div \#Hits \quad (1)$$

$$AvgTimeBetweenClicks = TaskTime \div \#Clicks \quad (2)$$

$$AvgTimeBetweenKeystrokes = TaskTime \div \#Keystrokes \quad (3)$$

$$AvgTimeBetweenActions = TaskTime \div \#Actions \quad (4)$$

2. *Weighted actions:* This indicator combines three types of actions (page views, clicks and keystrokes) in a different way. The number of clicks (#Clicks) and keystrokes (#Keystrokes) are used as a means to weight the importance of the page views (#Hits)[2]:

$$WeightedActions = \#Hits \times (\#Clicks + \#Keystrokes + 1) \quad (5)$$

3. *Eye gaze:* These indicators correspond to the average time (duration) of the fixations (Equation 6), and the average elapsed time between fixations (Equation 7):

$$AvgTimeFixations = FixationsTime \div \#Fixations \quad (6)$$

$$AvgTimeBetweenFixations = TaskTime \div \#Fixations \quad (7)$$

The correlations between the indicators described above and the global scores of the participants are presented in Table 3. Although still generally small, the resulting correlations are higher than the correlations coefficients shown in Table 2. Perhaps the most surprising outcome is the medium correlation (0.54) of the average duration of fixations (AvgTimeFixations). Moreover, the correlation for the weighted actions (WeightedActions) and for the average time between clicks (AvgTimeBetweenClicks) is 0.35, which is an improvement compared to the previous highest interaction correlation coefficient of 0.28 for the page views (#Hits).

**Table 3: Combined indicators correlation with scores**

| Type | Indicator | Correlation | Level | |
|---|---|---|---|---|
| Interaction | AvgTimeBetweenHits | 0.05 | Low | |
| | AvgTimeBetweenClicks | 0.35 | Low | * |
| | AvgTimeBetweenKeystrokes | 0 | Low | |
| | AvgTimeBetweenActions | -0.18 | Low | |
| | WeightedActions | 0.35 | Low | * |
| Eye gaze | AvgTimeBetweenFixations | 0.36 | Low | |
| | AvgTimeFixations | 0.54 | Medium | * |

## 4.3 Towards an effort model

We finally proceeded to create the effort model by choosing one interaction and one eye gaze indicator to combine, expecting them to be complementary and therefore able to increase the correlation with the scores when combined. Beside the average time between fixations, two previous interaction indicators had the same highest correlation values (highlighted in Table 3). Of course, the

[2] This equation contains an addition of one to avoid having a value of zero when the task does not require any clicks or keystrokes to be completed (e.g., reading a text).

chosen eye gaze indicator is the average duration of fixations (AvgTimeFixations) as it is the highest correlation (also highlighted in Table 3).

Both types of indicators have different ranges and distributions. Thus, before combining these indicators, we normalized them by using an exponential function with and upper limit of one (Equation 8). We then combined the interaction indicators (AvgTimeBetweenClicks and WeightedActions) and the eye gaze indicator (AvgTimeFixations) into the model as shown in Equation 9[3].

$$norm(x) = 1 - e^{-\alpha x} \tag{8}$$

$$Effort = w_1 \times (norm(WeightedActions)) + w_2 \times (norm(AvgTimeFixations)) \tag{9}$$

The correlations results for these combinations are shown in Table 4. As can be seen, the highest correlation is 0.59 for the combination of WeightedActions ($w_1$ = 0.7) with AvgTimeFixations ($w_2$ = 0.3). Overall, we were thus able to increase the correlation by more than 100% compared to the initial correlations.

The other combination (AvgTimeBetweenClicks and AvgTimeFixations) did not improve the correlation (i.e., the highest correlation still equal to the correlation of the eye gaze indicator alone), which suggests that the average time between clicks, which is not as flexible to the type of task as the WeightedActions indicator, is not complementary to the eye gaze indicator. This is not surprising, as people usually watch the screen when using the mouse.

**Table 4. Comparison of the combined indicators**

| | |
|---|---|
| AvgTimeBetweenClicks and AvgTimeFixations | 0.54 |
| WeightedActions and AvgTimeFixations | 0.59 |

## 5 CONCLUSION

A number of definitions have been proposed for the concept of students' effort. Usually, these definitions are related to various students' behaviors such as attending classes, delivering assignments on time, participating in class, etc. The Cognitive Load Theory allows to understand the effort as the cognitive load experienced by students. The theory also explains that learning occurs when we create and automate schemas in the working memory and then store them in the long-term memory, stating that the limitations of the working memory should be respected in order to allow learning.

This theory gained lots of attention and several researchers are now looking for new and more effective ways of measuring the imposed cognitive load (i.e., effort) in order to detect and avoid underload and overload. Different means of acquiring measurements of the cognitive load exist, and

[3] We empirically determined the value of the parameters that led to the higher correlation coefficients. We did however not systematically optimize these values.

they can be classified into subjective, performance, physiological and behavioral measures. However, those measurements only allow to classify or identify an increase or decrease in the cognitive load.

In this paper, we proposed a new model that allows to quantify students' effort, as opposed to identified or classified as shown in the cognitive load research. Despite the limitations of our dataset (i.e., absence of effort ratings and number of subjects) and the need of further investigation, our study adopts some new approaches that can offer a better understanding of students' effort. First, it relies on the Cognitive Load Theory that explains how learning occurs, the limitations that should be considered while designing learning activities (i.e., the working memory has a limit that should be respected), and also the role of effort in achieving learning outcomes. Second, it uses interaction and eye gaze data which can reduce noise and account for missing data (Mulder, 1992; Chen et al., 2016). Third, the model can still be used if only part of the data is available, even though it would result in a smaller correlation (e.g., if only the interaction data are available, the correlation coefficient would be 0.35 instead of 0.59).

Our results show that our approach is able to provide measurements that have a medium correlation with effort. Especially, our proposed combination increased the correlation score by more than 100% compared to raw indicators. This approach can be exploited to develop effort-based educational tools and help both teacher and students. For instance, with proper dashboards, teachers could be able to identify students who are not well engaged into learning (i.e., not exerting enough effort on their learning tasks) and students who are struggling (i.e., exerting too much effort on their learning tasks, possibly meaning that they lack previous knowledge to handle the proposed tasks). Similarly, students could identify if they are just not engaged enough or if they are struggling with the proposed learning tasks, leading them to seek help from their teachers and classmates. The proposed model can also be used in fully automated tools. For instance, recommendation systems could identify how much effort a student should exert in a given moment and recommend the ideal learning tasks, with the goal of promoting his engagement.

In the future, we intend to create a new dataset and run the model to see how it performs and how it can be further enhanced. Later, we want to add new behavioral (e.g., movements) and physiological measurements (e.g., pupil dilation, and skin temperature). Our ultimate goal is to be able to exploit our effort model to provide engaging recommendations to the students.

## REFERENCES

Arshad, S., Wang, Y., & Chen, F. (2013). Analysing mouse activity for cognitive load detection. In *Proc. of the 25th Australian Computer-Human Interaction Conf.: Augmentation, Application, Innovation, Collaboration* (pp. 115–118). ACM.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience* 4, 829–839.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In *Handbook of Psychophysiology* (pp. 142–162). Hillsdale, NJ: G. Berntson & L. G. Tassinar.

Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education*, *78*(1), 27–49.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust Multimodal Cognitive Load Measurement*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-31700-7

Hau, K.-T., & Salili, F. (1996). Prediction of academic performance among Chinese students: Effort can compensate for lack of ability. *Organizational Behavior and Human Decision Processes*, *65*(2), 83–94.

Hill, L. (1990). Effort and reward in college: A replication of some puzzling findings. *Journal of Social Behavior & Personality*, *5*(4), 151–161.

Huptych, M., Bohuslavek, M., Hlosta, M., & Zdrahal, Z. (2017). Measures for recommendations based on past students' activity (pp. 404–408). ACM Press. https://doi.org/10.1145/3027385.3027426

Khawaja, M. A., Chen, F., & Marcus, N. (2014). Measuring Cognitive Load Using Linguistic Features: Implications for Usability Evaluation and Adaptive Interaction Design. *Int. Journal of Human-Computer Interaction*, *30*(5), 343–368.

Khawaja, M. A., Ruiz, N., & Chen, F. (2007). Potential speech features for cognitive load measurement. In *Proc. of the 19th Australasian Conf. on Computer-Human Interaction: Entertaining User Interfaces* (pp. 57–60). ACM.

Kramer, A. F. (1990). *Physiological Metrics of Mental Workload: A Review of Recent Progress*.

Leppink, J. (2017). Cognitive load theory: Practical implications and an important challenge. *Journal of Taibah University Medical Sciences*, *12*(5), 385–391.

Marchal, F., Castagnos, S., & Boyer, A. (2016). A First Step toward Recommendations Based on the Memory of Users. In *Proc. of the 28th IEEE Int. Conf. on Tools with Artificial Intelligence (ICTAI)*. IEEE.

Marchal, F., Castagnos, S., & Boyer, A. (2018). First Attempt to Predict User Memory from Gaze Data. *Int. Journal on Artificial Intelligence Tools*.

Meltzer, L., Katzir-Cohen, T., Miller, L., & Roditi, B. (2001). The Impact of Effort and Strategy Use on Academic Performance: Student and Teacher Perceptions. *Learning Disability Quarterly*, *24*(2), 85–98.

Mulder, L. J. M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology*, *34*(2–3), 205–236.

Nagy, R. (2016). Tracking and Visualizing Student Effort: Evolution of a Practical Analytics Tool for Staff and Student Engagement. *Journal of Learning Analytics*, *3*(2), 165–193.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, *38*(1), 63–71.

Paas, F., Van Merriënboer, J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, *79*(1), 419–430.

Paas, F., & Van Merriënboer, J. J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, *35*(4), 737–743.

Paas, F., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*(4), 351–371.

Ruiz, N., Taib, R., Shi, Y. D., Choi, E., & Chen, F. (2007). Using pen input features as indices of cognitive load. In *Proc. of the 9th Int. Conf. on Multimodal Interfaces* (pp. 315–318). ACM.

Scariot, A. P., Andrade, F. G., Silva, J. M. C. da, & Imran, H. (2016). Students Effort vs. Outcome: Analysis Through Moodle Logs. In *IEEE 16th Int. Conf. on Advanced Learning Technologies* (pp. 371–372). Austin, TX, USA: IEEE.

Schuman, H. (2001). Comment: Students' Effort and Reward in College Settings. *Sociology of Education*, *74*(1), 73–74.

Schuman, H., Walsh, E., Olson, C., & Etheridge, B. (1985). Effort and reward: The assumption that college grades are affected by quantity of study. *Social Forces*, *63*(4), 945–966.

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems* (pp. 2651–2656). ACM.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Swinton, O. H. (2010). The effect of effort grading on learning. *Economics of Education Review*, *29*(6), 1176–1182. https://doi.org/10.1016/j.econedurev.2010.06.014

Verrel, J., Lövdén, M., Schellenbach, M., Schaefer, S., & Lindenberger, U. (2009). Interacting effects of cognitive load and adult age on the regularity of whole-body motion during treadmill walking. *Psychology and Aging*, *24*(1), 75–81.

Xie, B., & Salvendy, G. (2000). Prediction of Mental Workload in Single and Multiple Tasks Environments. *Int. Journal of Cognitive Ergonomics*, *4*(3), 213–242.

Yin, B., Ruiz, N., Chen, F., & Khawaja, M. A. (2007). Automatic cognitive load detection from speech features. In *Proc. of the 19th Australasian Conf. on Computer-Human Interaction: Entertaining User Interfaces* (pp. 249–255). ACM.

Yu, K., Epps, J., & Chen, F. (2011). Cognitive load evaluation of handwriting using stroke-level features. In *Proc. of the 16th Int. Conf. on Intelligent User Interfaces* (pp. 423–426). ACM.