Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Detecting Off-Task Behavior of Learners in Minecraft Using Exploration and Personalized Features

Maricel A. ESCLAMADO[a*] & Maria Mercedes T. RODRIGO[b]
[a]*University of Science and Technology of Southern Philippines, Philippines*
[b]*Ateneo Laboratory for the Learning Sciences, Ateneo de Manila University, Philippines*
*maricel.esclamado@ustp.edu.ph

Abstract: Off-task behavior refers to any action by a learner that is unrelated to the learning task, and it can have a negative effect on learning outcomes. Determining when a behavior is off task is challenging because these behaviors vary across different learning environments and goals. Off-task behavior in Minecraft might be more difficult to detect because of the open-ended nature of the game, which allows learners to explore the environment and complete learning tasks in various ways. This study aims to model off-task behavior of learners using the What-If Hypothetical Implementations using Minecraft (WHIMC). Detector of off-task behavior was developed using features from the interaction logs of the learners in WHIMC. Initially, the detector was constructed using the basic feature set as the baseline, that includes the time-based features, frequency-based features, and sequence features. The feature set was further expanded to include exploration features and personalized features, and the performance of the models using different sets of features were compared. This study found that the detector built with the addition of exploration features to the baseline feature set showed slightly higher performance compared to the detector built using the baseline feature set. Then, the detector built using the feature set that also included personalized features had better performance compared to the detector using the feature set with the exploration features. Some selected exploration and personalized features were also found to be more predictive of off-task behavior compared to some features in the baseline feature set.

Keywords: Minecraft, WHIMC, Off-task Behavior, exploration features, personalized features, Philippines.

1. Introduction

Off-task behavior can be defined as any behavior performed during a learning activity that does not involve the learning task or material (Rodrigo et al., 2013). This behavior has been found to have a negative effect on learning (Sabourin & Lester, 2014). Thus, it is important to detect these behaviors and be able to provide opportunities for some intervention, either by the teacher or the learning system, to try to bring the student back on track.

Determining when a behavior is off-task is challenging because these behaviors vary among learning environments and the learning goals they are trying to achieve. Off-task behavior in Minecraft might be more difficult to detect because of the nature of the game. Open-ended games like Minecraft provide players autonomy on how to do the tasks in the learning environment, thus there is no prescribed learning sequence and analysis of behavior becomes significantly important. With the open-ended nature of Minecraft, it is challenging to detect behavior patterns such as off-task behavior. This also makes it difficult to extract relevant features that can capture attributes of the student's learning which could help detect behavior such as being off task.

The focus of this study is on off-task behavior in Minecraft, a sandbox-type video game. In a sandbox game, players can create worlds and craft anything they want within those worlds. Minecraft is an open-ended game that focuses on exploration and building. Players have the freedom on how to explore the world making Minecraft an excellent domain in which to study player behaviors (Packard & Ontañón, 2015). There have been studies investigating player behavior in Minecraft, but no prior research has focused on detection of off-task behavior.

In this study, detectors of off-task behavior were developed using relevant features from the interaction logs of the learners in Minecraft, particularly in What-If Hypothetical Implementations using Minecraft (WHIMC; https://whimcproject.web.illinois.edu/), which is a set of simulations that learners can explore in order to learn more about science, technology, engineering, and mathematics (STEM). WHIMC uses Minecraft Java Edition as an interactive learning environment for students to explore and observe alternate versions of Earth through "what-if?" scenarios.

Initially, the detector was constructed using a basic feature set, as the baseline, that includes the time-based features, frequency-based features, and sequence features. The feature set was further expanded to include exploration features and personalized versions of the features, and the performance of the models using different sets of features were compared. This study sought to answer the following research questions:

RQ1: What set of basic features detect off-task behavior?
RQ2: What exploration features can be used in detecting off-task behavior?
RQ3: What personalized features can be used in detecting off-task behavior?
RQ4: How does the performance of detectors vary using different sets of features?

2. Prior Work on Off-task Detection

Prior studies attempted to build detectors of off-task behavior in various learning environments and explored different sets of features in detecting off-task behavior. Several studies explored the basic features such as time-based, performance-based, and frequency-based features (Wixon et al., 2012; Gobert et al., 2015; Jiang et al., 2018). Some studies also explored other types of features such as chat message features (Carpenter et al., 2020) and mouse movement features (Cetintas et al., 2009) which can possibly detect off-task behavior. These features are determined by the context of the learning goals, the nature of the learning environment, and the data logs that can be collected. For example, exploratory learning environments such as Minecraft provide exploration logs of the learners which can be used to track how they explore the environment (Lane et al., 2022), and features extracted from these logs can be useful in detecting off-task behavior. However, no prior studies had included exploration features in off-task behavior detection.

Another study (Cetintas et al., 2009) also considered personalized versions of the features to be included in the feature set. Personalized features refer to individual behavioral patterns established based on interactions with the system over time. A personalized feature taken at a specific time segment can be compared to that pattern to determine if the individual learner is deviating from the norm. Cetintas et al. (2009) identifies personalized version of a feature as the absolute value of a feature minus the average value of this feature on the same problem by the same student so far. Different students exhibit different behavioral patterns, thus suggesting that introducing personalized versions of each feature into off-task models makes the built models more flexible and adaptive to different students.

In this study, we attempt to build a model that detects off-task behavior in an open-ended learning environment built on Minecraft. By leveraging the exploration logs available in Minecraft, we incorporate exploration features into the off-task behavior detection model. Additionally, we explore the inclusion of personalized versions of features to make the model more flexible and adaptive to individual student behavior patterns.

3. Methods

Data was collected from a total of 79 students in the Philippines: 17 who participated in the WHIMC summer camp and 62 from a junior high school. The 17 participants of the WHIMC summer camp were from various schools in the Philippines and had varying ages between 10

– 14 years old. The 62 participants from the junior high school were composed of Grade 7 to Grade 10 students. Camp facilitators and partner teachers developed learning modules that explored the WHIMC worlds. In each world, the students can complete one or more quests. The students explored the assigned worlds synchronously. WHIMC is instrumented to collect data about learner actions as they use the game. Data that can be collected includes learners' positions, observations, and accessed science tools.

Log files were converted to text replays, which are easy-to-read versions of the log files and are effective for providing ground truth labels for behaviors (Wixon et al., 2012). In the text replay, clips were identified based on the quest duration. A clip represents a segment of the data starting from the initiation of a quest and concluding upon quest completion, change, or the end of the session without accepting a new quest. A total of 249 clips were generated from the logs of 79 participants from the summer camp and junior high school implementation. Two human coders practiced coding the clips in the text replay to determine whether the behavior in the clip was off-task or not. During the practice coding, the coders identified possible indicators of off-task behavior in the text replay, which included more time spent in irrelevant areas, more pauses longer than 30 seconds, more time spent on actions that were not part of the quest, and the quest not being completed. Most of these indicators of off-task behavior were based on prior studies, such as spending too much time in locations irrelevant to the task and doing activities unrelated to the task (Sabourin & Lester, 2014), and long pauses that may indicate off-task behavior (Cetintas et al., 2009). Next, the two coders separately coded all the clips. Based on the results, the two coders had a fair level of agreement with Cohen's Kappa of 0.26. Then, the coders checked the clips to confer with the disagreements. We found out that most of the disagreements were clips from the junior high school implementation. In their learning modules, aside from the tasks from the in-game quest, additional tasks were assigned by the teachers. An example is if most of the actions done by the student were for the additional tasks given by the teacher, the first coder had labeled this as on-task, and the second coder had labeled this as off-task since the coder only considered those actions from the in-game quests. Then, it was agreed to only consider the actions from the in-game quests since the off-task behavior detection was only based on the in-game data. The coders recoded these clips with disagreements based on the actions of in-game quests, resulting in an agreement for 247 clips out of 249. Two clips were excluded from the analysis because the coders had not reached an agreement on the coding of the two clips. Of the 247 clips that were coded, 163 (66%) were labeled as off-task and 84 (34%) were labeled as on-task.

A set of candidate features that could potentially detect off-task behavior in WHIMC based on the features of off-task behavior discussed in the literature review (Jiang et al., 2018) were generated from the data. A total of 28 features were distilled to compose the baseline feature set - basic feature set. This feature set includes time-based features, frequency-based features, and sequence features. Time-based features capture the time spent on user actions (Jiang et al., 2018). This includes idle time and time spent in irrelevant areas which are possible indicators of off-task behavior. Idle time is the time during which the learner stayed at a location point, with no change in distance traveled and no other activity (Esclamado et al., 2022). Time spent in irrelevant areas was computed by getting the amount of time the learner stayed in a location in which there were no nonplayer characters (NPCs) or the location was not related to the quest. Frequency-based features calculated the number of times each action is executed (Jiang et al., 2018). Sequence features calculated the frequency of common three-action sequences (Jiang et al., 2018). A total of 17 three-action sequences that occurred frequently were selected to be included in the feature set, and each action sequence was assigned to a feature variable. Quest completion or whether the quest was completed or not, was also included in the basic feature set since this is an important indicator of the performance of the learners in WHIMC.

The feature set was then expanded to include exploration features and the impact on the model's performance was evaluated. In WHIMC, the students are tasked to explore the environment and make observations. The data logs include the position where the action was taken in which considering exploration features may enhance the detection model. Table 1 shows the list of exploration features that were explored in this study.

Table 1. *List of Exploration features and their description*

Features	Description and how it was computed
distance	the total distance traveled by the learner. Successive pairs of location points in the logs were treated as line segments. Using the Euclidean distance, the total distance traveled by each learner was calculated by adding all distances of all line segments of the learner's path.
area	To get the value for the area covered by the learner, the smallest convex polygon that contains all locations that the learner visited, i.e. the convex hull, was determined using the Jarvis March algorithm. Based on these convex hulls, the area that each learner explored were computed.
ave_speed	To get the value for the average speed, the cumulative distances traveled by the learner within a clip in 3-second intervals were calculated. The speed rates between the 3-second intervals were calculated and then the average speed for the clip was calculated to determine how the learner explored the worlds.
num_worlds	the total number of worlds explored by the learner within a clip.
revisits	the number of times locations were revisited. To determine if a location was revisited, an approach based on Lane and colleagues (2022) was used. They divided the map into a 10 x 10 grid and if the player visited a cell in the grid, that cell was recorded as visited. In this study, a similar approach was used but aside from tagging a cell as visited, the cell was tagged as revisited if the learner visited the cell more than once.

Then, the feature set was further expanded by adding the personalized features and the impact on the model's performance was evaluated. In this study, a personalized version of a feature was identified by following a similar approach used by Cetintas et al. (2009). In the prior study, a personalized version of a feature can be identified by taking the absolute value of a feature and subtracting the average value of this feature on the same problem by the same student so far. However, since for every student there was only one clip for each quest unlike in the prior study where there were many attempts for each problem, the clustering was used to group similar student-quest data points in terms of behavior. This helped to identify a personalized feature, which is the value of a feature minus the average value of this feature for all student-quest data points on the same cluster. The personalized features of the basic and exploration features were generated, resulting in a total of 33 features. In this study, k-means clustering was used to group similar data points. Each data point is a unique combination of student and clip/quest.

Using the logistic regression classification algorithm, a detector of off-task behavior was built with the candidate features as independent variables and the ground truth labels obtained using text replay tagging, whether it is off-task or not, as the dependent variable. Forward selection was used to arrive at a parsimonious model. Feature selection was executed on training data only. Model performance was evaluated using 5-fold student-level cross-validation.

4. Results and Discussion

RQ1: What set of basic features detect off-task behavior?

Out of the 28 candidate features in the baseline feature set (basic features), 17 were selected after performing forward selection algorithm feature selection technique to arrive at a parsimonious model and student-level cross validation was used. The completion of the quest and the total time spent in the clip were found to be negatively related to off-task behavior. This means that not completing the quest and spending less time in the clip were indicative of being off-task. On the other hand, engaging in actions that were not part of the quest, spending more time in irrelevant areas, and having longer pauses contributed positively to off-task behavior. This indicated that spending too much time on unnecessary actions and irrelevant areas, and having more extended periods of inactivity were also indicative of being off-task.

RQ2: What exploration features can be used in detecting off-task behavior?

Out of the 5 candidate exploration features, 3 were selected along with the selected features from the baseline feature set after performing forward selection algorithm feature selection technique to arrive at a parsimonious model and student-level cross validation was used. The selected exploration features are the area covered, average speed, and number of worlds explored within the clip. These exploration features were more predictive of off-task behavior compared to some basic features. The area explored by the learner in the assigned world within a clip contributed negatively to off-task behavior, while the number of worlds explored by the learner within a clip contributed positively to off-task behavior. The more area covered in the assigned world, the less likely it was off-task, whereas visiting other worlds other than the assigned world, the more likely it was off-task. It shows that exploring less within the assigned world was indicative of off-task behavior. Furthermore, the average speed of the learners contributed positively to off-task behavior, which indicated that moving too fast around the WHIMC world, the more likely it was off-task.

RQ3: What personalized features can be used in detecting off-task behavior?

Out of the 33 candidate personalized features, 14 were selected along with the selected features from the second feature set (basic + exploration features) after performing forward selection algorithm feature selection technique to arrive at a parsimonious model, and student-level cross validation was used. Some selected personalized features were more predictive of off-task behavior compared to some basic and exploration features. These include the personalized versions of the number of observations, area covered, and time spent on unnecessary actions which were highly predictive of off-task behavior. The personalized version of the number of observations and the area covered by the student within the clip contributed negatively to off-task behavior. This indicated that making lesser observations and covering smaller area within a clip compared to the average value of similar clips in the same cluster are indicative of off-task behavior. On the other hand, the personalized version of time spent on unnecessary actions contributed positively to off-task behavior which means that spending too much time on actions not related to the quest compared to the average value of similar clips in the same cluster are indicative of off-task behavior.

RQ4: How does the performance of detectors vary using different sets of features?

The performance of the detectors using different sets of features using the logistic regression classification algorithm is shown in Table 2. The detector built using the second feature set (basic + exploration features) showed slightly higher cross-validated performance compared to the detector using the baseline feature set. The detector built using the feature set that included personalized features also had a higher cross-validated performance compared to the second feature set (basic + exploration features).

Table 2. *Cross-validated performance of OTB detector using different sets of features using Logistic Regression*

Feature Set	Accuracy	Kappa	AUC
Basic Features (baseline)	0.842	0.621	0.789
Basic Features + Exploration Features	0.846	0.622	0.783
Basic Features + Exploration Features + Personalized Features	0.846	0.63	0.795

5. Conclusion

This study investigated and modeled off-task behavior of learners using Minecraft, an open-ended learning environment, and compared the performance of detectors using different sets of features. Prior work had investigated detection of off-task behavior in game-based learning environments, but the set of features considered in these studies had typically been limited to time-based, frequency-based, and sequence features. This study developed these features, based on prior work (Jiang et al., 2018), that can be generated from WHIMC and was considered as the baseline feature set.

In exploratory learning environments such as Minecraft, in which learners are expected to explore the environment to complete tasks, the addition of exploration features to the baseline feature set was explored to determine the impact on the performance of the detector.

Based on the results, the detector built with the addition of exploration features to the baseline feature set showed slightly higher performance compared to the detector built using the baseline feature set. Among the five exploration features considered in this study, the area covered, average speed, and number of worlds visited were the selected features that best contributed to improving the performance of the detector. Also, these exploration features contribute more to predicting off-task behavior compared to some basic features. Additionally, the detector built using the feature set that included personalized features had better performance compared to the second feature set. This indicates that the addition of personalized features contributed to the improvement of the model's performance. Some selected personalized features were also more predictive of off-task behavior compared to some basic and exploration features.

The findings of the study, particularly on the features that are good predictors of off-task behavior, could provide insights to the teachers on what behavior could be indicative of off-task behavior such as spending less time on quests, interacting less with NPCs, pausing more frequently, exploring less area on the assigned worlds, moving too fast around the learning environment, or visiting other worlds other than the assigned world. They may use this information to suggest ways to help or guide students to better learning outcomes.

Acknowledgements

The authors thank H Chad Lane and Jeff Ginger for their enthusiastic collaboration, Dominique Marie Antoinette B. Manahan, Mikael Fuentes, Louise Marie C. Tulayba, Marlon C. Moralejo, and Ma. Rosario Madjos for their support, the Ateneo Laboratory for the Learning Sciences, and the Ateneo de Manila University.

References

Carpenter, D., Emerson, A., Mott, B.W., Saleh, A., Glazewski, K.D., Hmelo-Silver, C.E. and Lester, J.C. (2020). Detecting off-task behavior from student dialogue in game-based collaborative learning. In *Proceedings of the International Conference on Artificial Intelligence in Education*, Part I,55-66.

Cetintas, S., Si, L., Xin, Y.P.P. and Hord, C. (2009). Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Transactions on Learning Technologies, 3*(3), 228-236.

Esclamado, M. A., Rodrigo, M. M. T., & Casano, J. A. (2022). Achievement, Behaviors, and STEM Interest of Frustrated and Bored Learners Using Minecraft. *Proceedings of the 30th International Conference on Computers in Education, I*, 477-486.

Gobert, J.D., Baker, R.S., & Wixon, M.B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist, 50*(1), 43-57.

Jiang, Y., Bosch, N., Baker, R.S., Paquette, L., Ocumpaugh, J., Andres, J.M.A.L., Moore, A.L., & Biswas, G. (2018). Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection?. In *19th International Conference on Artificial Intelligence in Education (AIED 2018),* Part I, 19, 198-211.

Lane, H. C., Gadbury, M., Ginger, J., Yi, S., Comins, N., Henhapl, J., & Rivera-Rogers, A. (2022). Triggering stem interest with Minecraft in a hybrid summer camp. *Technology, Mind, and Behavior*, *3*(4). https://doi.org/10.1037/tmb0000077

Packard, B. and Ontañón, S. (2015). Learning behavior from demonstration in minecraft via symbolic similarity measures. In *Proceedings of the 10th International Conference on the Foundations of Digital Games*.

Rodrigo, M.M.T., Baker, R.S., & Rossi, L. (2013). Student off-task behavior in computer-based learning in the Philippines: comparison to prior research in the USA. *Teachers College Record, 115*(10), 1-27.

Sabourin, J. L., & Lester, J. C. (2014). Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing*, *5*(1), 45–56. https://doi.org/10.1109/t-affc.2013.27

WHIMC. (n.d.) What-If Hypothetical Implementations using Minecraft. https://whimcproject.web.illinois.edu/

Wixon, M., Baker, R.S.D., Gobert, J.D., Ocumpaugh, J., & Bachmann, M. (2012). WTF? detecting students who are conducting inquiry without thinking fastidiously. In *20th International Conference on User Modeling, Adaptation, and Personalization (UMAP 2012), Proceedings 20*, 286-296. https://doi.org/10.1007/978-3-642-31454-4_24