Characterizing Collaboration in the Pair Program Tracing and Debugging Eye-Tracking Experiment: A Preliminary Analysis

Maureen M. Villamor
Ateneo de Davao University, Quezon City Philippines
University of Southeastern Philippines, Davao City, Philippines
maui@usep.edu.ph

Ma. Mercedes T. Rodrigo
Ateneo de Davao University
Quezon City
mrodrigo@ateneo.edu

ABSTRACT
This paper characterized the extent of collaboration of pairs of novice programmers as they traced and debugged fragments of code using cross-recurrence quantification analysis (CRQA). This was a preliminary analysis that specifically aimed to compare and assess the collaboration of pairs consisting of two individuals who may have different or same level of prior knowledge given a task. We performed a CRQA to build cross-recurrence plots using eye tracking data and computed for the CRQA metrics, such as recurrence rate (RR), determinism (DET), average diagonal length (L), longest diagonal length (LMAX), entropy (ENTR), and laminarity (LAM) using the CRP toolbox for MATLAB. Results showed that low prior knowledge pairs (BL) collaborated better compared to high prior knowledge pairs (BH) and mixed prior knowledge (M) pairs because of its high RR and DET implying that they had more recurrent fixations and matching scanpaths. However, the BL pairs' high ENTR and LAM could mean that they seemed to have more difficulty in understanding and debugging the programs. All pairs regardless of category had more or less exerted the same level of attunement when asked to debug the programs as evident in their L values. The mixed pairs seemed to have struggled with eye coordination the most as it had the most incidences of low LMAX.

Keywords
Eye-tracking, Collaboration, Cross-recurrence quantification

1. INTRODUCTION
Eye gaze plays an essential role in social interaction processes. In Computer-Supported Collaborative Learning (CSCL), eye-tracking had been used in previous works to study joint attention in collaborative learning situations [9][16]. Two eye-trackers, for instance, can be synchronized for studying the gaze of two persons collaborating in order to solve a problem and for understanding how gaze and speech are coupled [11-13].

The use of gaze coupling was first proposed in [11] to study conversation coordination. In this study, they defined gaze coupling as episodes when participants are looking at the same target. Their results showed that the coupling of eye gaze between collaborating partners may be an indicator of quality interaction and better comprehension. In the domain of pair programming, Pietinen et al. [10] suggested that gaze closeness could reflect tightness of collaboration. More prior studies [1][11-13] have shown that the coupling of eye gaze between collaborating partners may be an indicator of quality interaction and better comprehension and that joint attention, and more generally, synchronization between individuals is essential for an effective collaboration.

Cross-recurrence quantification analysis or CRQA, introduced in [18], is an extension of Recurrence Quantification Analysis (RQA) [7] that is used to quantify how frequently two systems exhibit similar patterns of change or movement in time. It takes two different trajectories of the same information as input and tests between all points of the first trajectory with all points of the second trajectory forming a cross-recurrence plot (CRP). The CRP permits visualization and quantification of recurrent state patterns between two time series. Analysis using CRP's has been proposed as a generalized method to unveil the interlocking of two interacting people [2]. It has been used to analyze the coordination of gaze patterns between individuals and has been used to determine how closely two collaborators' gaze follow each other. In the scientific literature, a cross-recurrence gaze plot is considered as the standard way of representing social eye-tracking data [16].

CRQA was used in [11], which provided the first quantification of gaze coordination in their monologue data to analyze the relation between eye movements of the speaker and the listener. The analysis revealed that the coupling between speaker and listener eye-movements predicted how well the listener understood what was said. They extended their findings in their succeeding studies [12-13] and results revealed that eye movement coupling found in monologue indeed extends to dialogues.

In the context of pair programming, Jermann et al. [5] used synchronized eye-trackers to assess how programmers collaboratively worked on a segment of code, and they also contrasted a "good" and a "bad" pair using cross-recurrence plots. Results showed that high gaze recurrence seems to be typical of a "good" pair where the flow of interaction is smooth and where partners sustain each other's understanding. A dual eye-tracking study was also conducted that demonstrated the effect of sharing selection among collaborators in a remote pair-programming scenario [4]. They used gaze cross-recurrence analysis to measure the coupling of the programmers' focus of attention. Their

findings showed that pairs who used text selection to perform collaborative references have high levels of gaze cross-recurrence.

This paper aimed to use CRQA to characterize collaboration of pairs of novice programmers in the act of tracing fragments of code and debugging. Specifically, this was a preliminary study that attempted to answer the following research question: Using CRQA, what characterizes collaboration of pairs consisting of (a) both high prior knowledge students, (b) both low prior knowledge students, and (c) high- and low-prior knowledge students?

Although the use of CRQA as an approach to assess collaboration between participants in a pair programming eye-tracking experiment is not an entirely novel approach, the main contribution of this study was the inclusion of the composition of the pairs in terms of expertise levels. Previous studies did not characterize the pairs based on prior knowledge in programming or level of expertise.

2. METHODS

2.1 Participants

The study was conducted in two private universities in the Philippines. Students who had taken the college-level fundamental programming course were recruited to participate in this study. Since the study is not finished yet and is still on-going, we recruited only 16 pairs of participants as of writing of this paper.

2.2 Structure of the Study

A screening questionnaire was distributed to student volunteers, to determine their eligibility to take part in this study (e.g. no cataracts, no implants, etc.), and they were required to undergo an eye-tracking calibration test. Participants who passed both screenings were given consent letters to fill up and sign. They were then asked to take a written program comprehension test (20 minutes) to determine their level of programming knowledge and skills. The actual eye-tracking experiment followed which was designed for 60 minutes at the maximum. Two Gazepoint eye-trackers were used to collect the pairs' eye-tracking data. The pairs were shown 12 programs with known bugs and were asked to mark the location of the bugs with an oval. There was no need to correct the errors.

A slide sorter program with "Previous", "Reset", "Finish" and "Next" buttons was created to display the program specifications followed by the buggy programs. The participants were free to click any of the buttons as they liked and were free to navigate the slides. No scrolling was needed. When the participant finds a bug, he/she clicks on the location of the bug and the software then draws an oval to mark it. Figure 1 is an excerpt from a specific slide in the slide sorter program showing the ovals.

The pairs were told to work with their partner on the problems and should collaborate using a chat program. All communications with their partner was via chat. The participants were seated together in the same room but were spaced far enough to ensure that all communication with their partners was via chat only. After the actual eye-tracking experiment, the pairs were asked to fill up a post-test questionnaire privately to assess how well they knew each other, how well they thought they collaborated, and how they felt about their partner. This study limits its analysis to the results of the programming comprehension test and the eye gaze data.

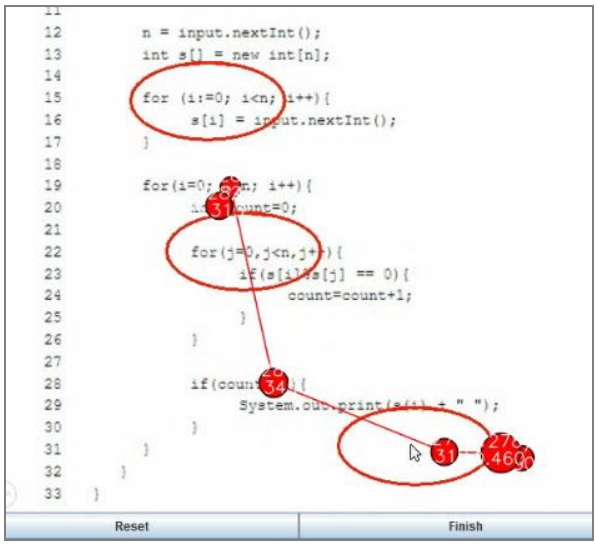

Figure 1. An excerpt from the slide sorter program showing the ovals after marking.

2.3 Constructing a Cross-Recurrence Plot

To conduct a cross-recurrence analysis, an N x N matrix called cross-recurrence plot is built, which is essentially a representation of the time coupling between two time series. The horizontal axis represents time for the first collaborator (C1) and the vertical axis represents time for the second collaborator (C2). Given two fixation sequences of the collaborators, f_i and g_i, i – 1… N, we define the cross- recurrence as r_{ij} = 1 if d (f_i, g_j) ≤ ρ, and 0, otherwise [7].

Recurrence occurs when two fixations from different sequences land within a given radius ρ of each other, where d is some distance metric (e.g., Euclidean distance). Cross-recurrence points are represented as a black point (pixel) in the plot (see Figure 2). For a pixel to be colored, the distance between the fixations of the two collaborators has to be lower than a given threshold. If two collaborators uninterruptedly looked at two different spots on the screen for the entire interaction, the resulting CRP would be completely blank (white space in Figure 2). On the contrary, if the two collaborators looked at the same spot on the screen continuously, the plot would show only a dark line on the diagonal. Points exactly on the diagonal of the plot correspond to synchronous recurrence, such as, collaborators look at the same target at exactly the same time. Points above the diagonal correspond to fixations of *C2* that happen after *C1* has fixated the element. Points below the diagonal correspond to *C2*'s gaze leading *C1*'s. Asymmetries above and below the diagonal line could therefore be indicative of leading and following behaviors.

2.4 CRQA Metrics

CRQA defines several measures that can be assessed along the diagonal and vertical dimensions. For the diagonal dimension, we have: recurrence rate, determinism, average and longest length of diagonal structures, entropy, and diagonal recurrence profile. For the vertical dimension, we have: laminarity and trapping time. The definitions that follow are taken from [7].

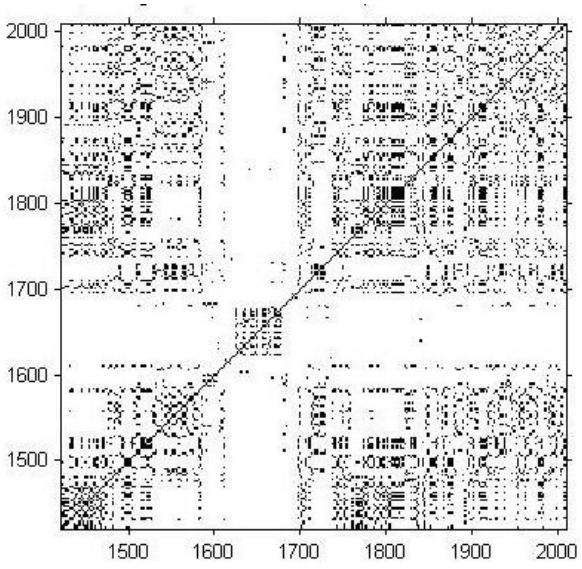

Figure 2. Example of a Cross-Recurrence Plot

Cross-Recurrence Rate (RR) represents the "raw" amount of similarities between the trajectories of two systems, which refers to the degree to which they tend to visit similar state. In eye-tracking data, this represents the percentage of cross-recurrent fixations. The more closely coupled the two systems are, in terms of sharing the same paths, the more recurrences will be formed along the diagonal lines. Hence, a high density of recurrence points in a diagonal results in a high value of RR.

Determinism (DET) is the proportion of recurrence points forming long diagonal structures of all recurrence points. Relative to eye-tracking data, this refers to the percentage of identical scanpath segments of a given minimal length in the two scanpaths.

The average diagonal length (L) reports the duration that both systems stay attuned. High coincidences of both systems increase the length of these diagonals. High values of DET and L represent a long time span of the occurrence of similar dynamics in both trajectories.

The longest diagonal length (LMAX) on a recurrence plot denotes the longest uninterrupted period of time that both systems are in concurrence, which can be seen as an indicator of stability of the coordination.

Entropy (ENTR) measures the complexity of the attunement between systems. In eye-tracking, this represents the complexity of the relation between scanpaths of the two eye-movement data. ENTR is low if the diagonal lines tend to all have the same length, signifying that the attunement is regular; otherwise, ENTR is high if the attunement is complex.

Using the diagonal recurrence profile (DiagProfile) offers the possibility of observing the direction of the coordination, that is, if there is an asymmetry with one interlocutor leading the other.

Vertical structures in a CRP quantify the tendency of the trajectories to stay in the same region. The laminarity (LAM) of the interaction refers to the percentage of recurrence points forming vertical lines, whereas trapping time (TT) represents the average time two trajectories stay in the same region.

2.4 Data Preparation and Measures

Results of the written program comprehension test, post-test and the number of bugs identified were recorded. The program comprehension results were used to categorize the students as having high or low prior knowledge. A student was considered to have high prior knowledge if his/her program comprehension score was equal to or greater than the median score. Otherwise, the student has low prior knowledge.

The fixation data was cleaned first by removing fixations less than 100 milliseconds [8]. The number of fixations per slide that contained the actual program were segregated and saved on separate files. Hence, each participant has at most 12 fixation files. Fixation alignment was performed in case of uneven number of fixations per program file. Fixation files with sequences less than 20 were discarded because it usually returned a NaN value when the CRQA was performed using the CRP toolbox for MATLAB [7].

Given 16 pairs and 12 programs, there should have been 16×12 = 192 cases, but we only had 179 cases for the analysis since some pairs did not finish all 12 programs and some fixations sequences were discarded. A cross-recurrence plot was then constructed for each pair for every program, and the cross-recurrence analysis was performed to get the RR, DET, LMAX, L, ENTR, and LAM.

The challenge of using CRQA is finding optimal parameters for *delay*, *embed*, and *radius* [7]. An optimal delay can be identified when mutual information drops and starts to level off. The embedding dimension can be determined using false nearest neighbors and checking when there is no information gain in adding more dimensions. For this experimental data, however, no further embedding was done [3]. With an embedding dimension of one, delay was also set equal to one since no points were time delayed [17]. For this experimental data, the radius, which is the threshold that determines if two fixation points are recurrent, was set to 5% of the maximal phase space diameter [15] to avoid subjective biases when looking at recurrent patterns.

3. RESULTS AND DISCUSSION

Of the 16 pairs, there were three (3) both high prior knowledge pairs, five (5) both low prior knowledge pairs, and eight (8) mixed prior knowledge pairs. The remainder of the text will refer to these categories as BH, BL, and M respectively. The CRQA metrics per program according to these relationships were averaged separately to get the aggregated CRQA metrics.

The aggregated results were examined to find differences among the categories, which entailed looking at incidences of high and low values of the CRQA metrics. A value was considered high if it was equal to or greater than the mean plus one standard deviation and low if it was equal to or less than the mean minus one standard deviation. Table 1 shows the descriptive values of all aggregated CRQA metrics per program. No further statistical measures were performed since there were not too many pairs to consider and this was only for hypothesis generation purposes.

Findings showed that the BH pairs only had incidences of low to average RR's and BL pairs only had incidences of average to high RR's. The M pairs had a mix of high, low, and average RR's. See Table 1 for high and low RR. Figure 3 shows the boxplots of RR in these categories.

Table 1. Descriptive values of the CRQA metric per program

CRQA Metric	Mean	SD	Min	Max	Low <=	High >=
RR	0.13	0.05	0.06	0.27	0.08	0.17
DET	0.42	0.09	0.25	0.67	0.33	0.51
L	3.50	1.31	1.94	7.12	2.19	4.81
LMAX	39.59	22.45	9.33	82.57	17.14	62.05
ENTR	0.76	0.20	0.44	1.34	0.57	0.96
LAM	0.50	0.13	0.26	0.78	0.38	0.64

This could possibly mean that the BL pairs collaborated better than BH and M pairs due to its incidences of higher recurrent fixations. However, it could also mean that the high RR's found in BL pairs was because of the BL pairs' greater number of fixation points, implying that the BL pairs had spent more time comprehending the program flow and finding the errors in the program. More time spent could have resulted to more chances of having more recurrent fixations. BH and M pairs exhibited the same degree of collaboration based on their comparable average RR's with M only slightly higher than BH. It can also be noted that the high RR's observed in all categories were all found in the middle programs, possibly indicating that the middle programs required more concentration compared to other programs.

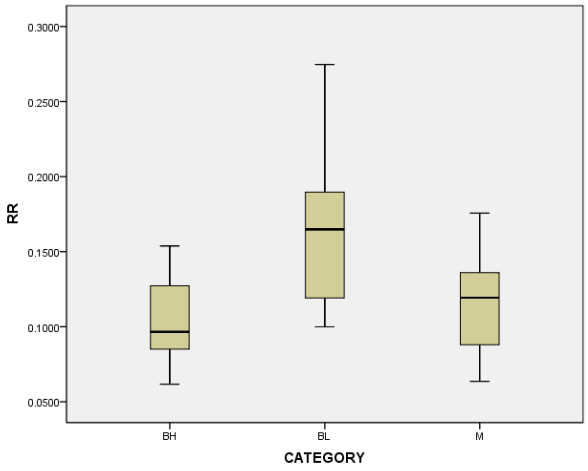

Figure 3. Boxplots of RR in All Catgories

The BL pairs only had average to high DET values, whereas BH and M pairs both only had low to average DET values. See Table 1 for high and low DET values. Figure 4 shows the boxplots of DET in all categories. The greater number of high DET values found in BL pairs could possibly mean that the BL pairs had shared more identical scanpaths compared to BH and M pairs. Also, since the BL pairs had more occurrences of high RR's and seemed to have spent longer durations in the task; this might have resulted to more matching scanpaths compared to BH and M pairs. As with RR, BH and M pairs' average DET were nearly the same, indicating the same degree of collaboration as assessed through their percentage of identical scanpaths.

Upon examination of their L values, results showed the BL pairs neither had high nor low L values. All but two of their L values were below the mean. The M pairs had few occurrences of high L values whereas BH pairs had one incidence each of high and low L values. Hence, a large majority of their L values were average. See Table 1 for high and low L values. Figure 5 shows the boxplots of the L values in all categories. These results implied that all of the pairs regardless of their expertise level or prior knowledge had more or less concentrated and exerted the same level of attunement on the given task. However, the M pairs possibly exhibited frequent longer durations where the pairs stay attuned compared to BH and BL pairs. BL pairs, on the other hand, had exhibited frequent shorter durations of attunement.

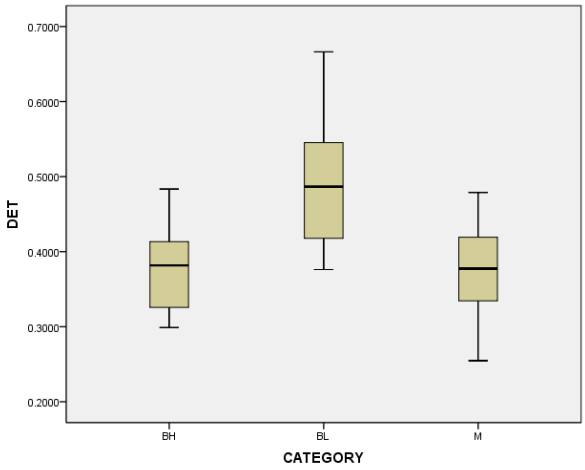

Figure 4. Boxplots of DET in All Catgories

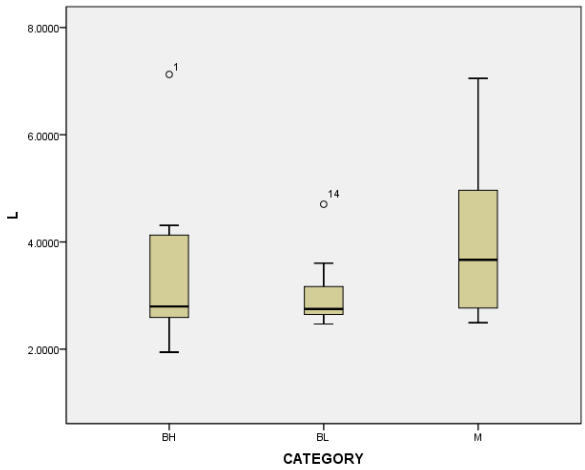

Figure 5. Boxplots of L in All Catgories

As for LMAX, BL pairs seemed to have exhibited better stability in terms of eye coordination particularly in the middle programs since they had more occurrences of high LMAX values. M pairs seemed to have struggled with eye coordination the most because of more incidences of low LMAX values. However, the average LMAX values of BH and M pairs were comparable, possibly indicating that the BH pairs' eye coordination stability was almost the same as M pairs. See Table 1 for high and low LMAX values. Figure 6 shows the boxplots of LMAX in all categories.

The same pattern in DET can also be observed in ENTR in terms of the incidences of high and low ENTR. The BL pairs had average to high ENTR values, whereas both BH and M pairs only had low to average ENTR, with M pairs having more low ENTR values than the BH pairs. See Table 1 for high and low ENTR values. Figure 7 shows the boxplots of ENTR in all categories. These findings imply that the BL pairs seemed to have more complex scanpaths in looking for bugs compared to BH and M pairs particularly in the middle programs. The BH pairs had the least complicated and, hence, more predictable scanpaths but their average ENTR was comparable to M pairs' average ENTR indicating that their scanpaths when looking for bugs were almost identical.

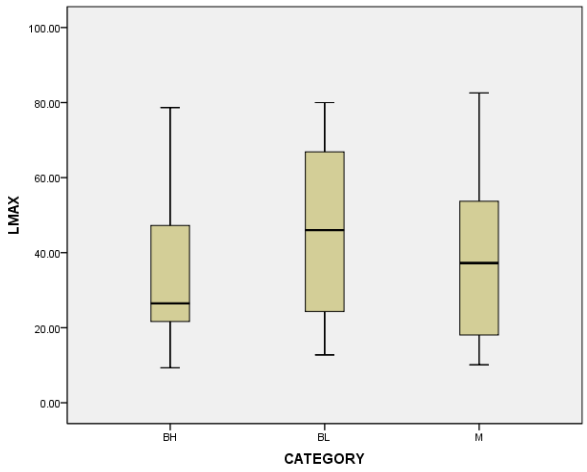

Figure 6. Boxplots of LMAX in All Catgories

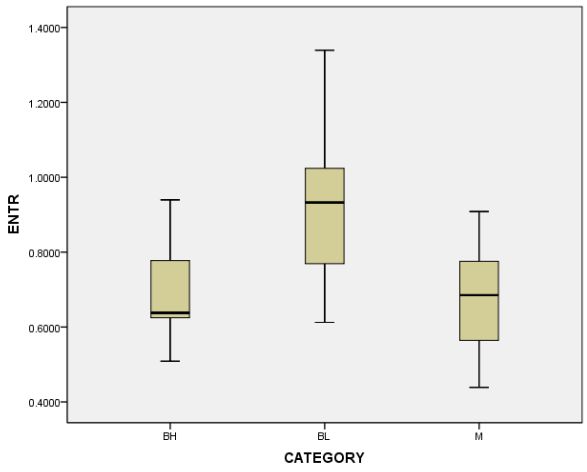

Figure 7. Boxplots of ENTR in All Catgories

As with DET and ENTR, the BL pairs only had average to high LAM values, whereas both BH and M pairs only had low to average LAM values. See Table 1 for high and low LAM values and Figure 8 for the boxplots. This could imply that the BL pairs seemed to have encountered more problems in understanding the program and, hence, tended to spend more time in certain regions of the code. BH and M pairs, on the other hand, seemed to have struggled less in understanding and debugging the programs. .

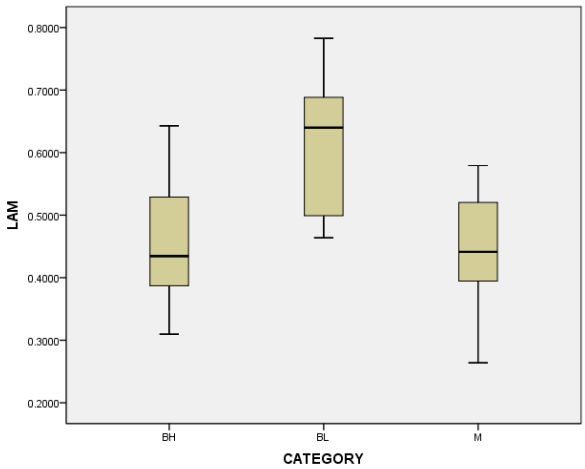

Figure 8. Boxplots of LAM in All Catgories

We also examined the number of slide switches between the program specification and the buggy program. We observed that the BL pairs had the least average number of slide switches among the pairs, but with the highest LAM values. This could mean that BL pairs tended to spend more time focusing on the actual program finding for bugs and switched less frequently between the program specification and the buggy program compared to other categories. BH and M pairs had higher frsequency of slide switches but with the lowest LAM values. BH and M pairs probably switched between slides more frequently because they just read the program specification to quickly check and recheck what the program does and were fast in terms of inspecting what was wrong in the actual program. BL pairs probably did not mind the program specification too much and just focused on the actual program locating bugs for the most part of the task.

Overall, it can be noted that for all the pairs, more evidences of collaboration and concentration happened in the middle part of the task. Perhaps, all the pairs perceived the middle programs the most difficult to debug.

4. SUMMARY AND CONCLUSION

The goal of this paper was to characterize the collaboration between pairs of novice programmers in the act of tracing and debugging a program in an attempt to understand the collaborative relationship of two individuals on a given task. Their collaboration was assessed through their CRQA metric results.

Findings showed that BL pairs are characterized with high RR, high DET, high ENTR and high LAM. Their high RR and DET signify that BL pairs are inclined to collaborate with their peers more compared to BH and M pairs. However, their high ENTR may signify complicated scanpaths in looking for bugs and their high LAM imply tendencies to stay in same regions of the code, which implies further that they frequently have difficulties in understanding and debugging the programs.

All pairs regardless of category tend to exhibit the same level of attunement in debugging as evident in their L values. The M pairs, however, are characterized as having more incidences of LMAX values, which could mean that they tend to struggle with

eye coordination the most. Overall, BH and M pairs are comparable in terms of collaboration as assessed through their CRQA results. We hypothesized, therefore, that the presence of a participant with high prior knowledge in M pairs may have contributed to the similarity between BH and M pairs

5. ACKNOWLEDGMENTS

The authors would like to thank Ateneo de Davao University and Ateneo de Naga University for allowing us to conduct the eye-tracking experiment. Many thanks also to Japheth Duane Samaco, Joanna Feliz Cortez, and Joshua Martinez for facilitating the data collection. We would like to thank also Dr. Norbert Marwan for giving us the permission to use the CRP toolbox for MATLAB. Lastly, thank you to Private Education Assistance Committee of the Fund for Assistance to Private Education for the grant entitled "Analysis of Novice Programmer Tracing and Debugging Skills using Eye Tracking Data."

6. REFERENCES

[1] Cherubini, M., Nüssli, M.A. and Dillenbourg, P., 2008, March. Deixis and gaze in collaborative work at a distance (over a shared map): a computational model to detect misunderstandings. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 173-180). ACM.

[2] Dale, R., Warlaumont, A.S. and Richardson, D.C., 2011. Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *International Journal of Bifurcation and Chaos*, *21*(04), pp.1153-1161.

[3] Iwanski, J.S. and Bradley, E., 1998. Recurrence plots of experimental data: To embed or not to embed?. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *8*(4), pp.861-871.

[4] Jermann, P. and Nüssli, M.A., 2012, February. Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 1125-1134). ACM.

[5] Jermann, P., Mullins, D., Nüssli, M.A. and Dillenbourg, P., 2011. Collaborative gaze footprints: Correlates of interaction quality. In *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings.* (Vol. 1, No. EPFL-CONF-170043, pp. 184-191). International Society of the Learning Sciences.

[6] Jermann, P., Nüssli, M.A. and Li, W., 2010, September. Using dual eye-tracking to unveil coordination and expertise in collaborative Tetris. In *Proceedings of the 24th BCS Interaction Specialist Group Conference* (pp. 36-44). British Computer Society.

[7] Marwan, N. and Kurths, J., 2002. Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, *302*(5), pp.299-307.

[8] Matos, R., 2010. Designing eye tracking experiments to measure human behavior. *Merriënboer, JJG van, and Sweller, J.(2005). Cognitive load theory and complex learning: Recent developments and future directions. Educational Psychology Review*, *17*.

[9] Pietinen, S., Bednarik, R. and Tukiainen, M., 2009. An exploration of shared visual attention in collaborative programming. In *21st Annual Psychology of Programming Interest Group Conference, PPIG*.

[10] Pietinen, S., Bednarik, R., Glotova, T., Tenhunen, V. and Tukiainen, M., 2008, March. A method to study visual attention aspects of collaboration: eye-tracking pair programmers simultaneously. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 39-42). ACM.

[11] Richardson, D.C. and Dale, R., 2005. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science*, *29*(6), pp.1045-1060.

[12] Richardson, D.C. and Dale, R., 2006. Grounding dialogue: eye movements reveal the coordination of attention during conversation and the effects of common ground. In *Proceedings of the 28th Annual Cognitive Science Society Conference*.

[13] Richardson, D.C., Dale, R. and Kirkham, N.Z., 2007. The art of conversation is coordination common ground and the coupling of eye movements during dialogue. *Psychological science*, *18*(5), pp.407-413.

[15] Schinkel, S., Dimigen, O. and Marwan, N., 2008. Selection of recurrence threshold for signal detection. *The European Physical Journal-Special Topics*, *164*(1), pp.45-53.

[16] Schneider, B., Abu-El-Haija, S., Reesman, J. and Pea, R., 2013, April. Toward collaboration sensing: applying network analysis techniques to collaborative eye-tracking data. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 107-111). ACM.

[17] Webber Jr, C.L. and Zbilut, J.P., 2005. Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials in contemporary nonlinear methods for the behavioral sciences*, pp.26-94.

[18] Zbilut, J.P., Giuliani, A. and Webber, C.L., 1998. Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters A*, *246*(1), pp.122-128.