

Ateneo de Manila University

Archium Ateneo

Department of Information Systems &
Computer Science Faculty Publications

Department of Information Systems &
Computer Science

12-23-2019

Gaze collaboration patterns of successful and unsuccessful programming pairs using cross-recurrence quantification analysis

Maureen Villamor

Ma. Mercedes T. Rodrigo

Follow this and additional works at: <https://archium.ateneo.edu/discs-faculty-pubs>




Part of the [Computer Sciences Commons](#)

RESEARCH

Open Access



Gaze collaboration patterns of successful and unsuccessful programming pairs using cross-recurrence quantification analysis

Maureen M. Villamor^{1*}  and Ma. Mercedes T. Rodrigo²

*Correspondence:
maui@usep.edu.ph

¹College of Information and Computing University of Southeastern Philippines, Bo. Obrero, Davao City, Philippines
Full list of author information is available at the end of the article

Abstract

A dual eye tracking experiment was performed on pairs of novice programmers as they traced and debugged fragments of code. These programming pairs were categorized into successful and unsuccessful pairs based on their debugging scores.

Cross-recurrence quantification analysis (CRQA), an analysis using cross-recurrence plots (CRP), was used to determine whether there are significant differences in the gaze collaboration patterns between these pair categories. Results showed that successful and unsuccessful pairs can be characterized distinctively based on their CRPs and CRQA metrics. This study also attempted to interpret the CRQA metrics in relation to how the pairs collaborated in order to provide a somewhat clear picture of their relevance and meaning. The analysis results could serve as a precursor in helping us understand what makes a programming pair more successful over other pairs and what behaviors exhibited by unsuccessful pairs that should be avoided.

Keywords: Collaboration, Pair programming, Eye tracking, Cross-recurrence plots

Introduction

Eye tracking methodologies have been employed in collaborative tasks to describe how collaboration unfolds based on gaze patterns. For example, there had been studies that investigated how quickly a name target is visually fixated after it has been mentioned by a partner (Richardson and Dale 2005). This may be used as indication of how well the listener understood what has been said. For pairs in collaborative learning situations, joint attention has been of particular interest. This occurs when two mutually aware individuals look at something together (Schilbach 2015). This can be used as an indicator that informs researchers when do participants synchronize their gazes with their partners. The effect of joint attention in the quality of collaboration have already been explored and results suggest that productive collaboration is associated with more joint visual attention (Jermann et al. 2011; Schneider and Pea 2013).

A popular collaboration paradigm applied in teaching introductory programming courses is pair programming. Prior research have shown that pair programming is beneficial for students' learning and attitudes towards programming (Hannay et al. 2009; Murphy et al. 2010). These benefits include producing better quality of code, being more

confident with their solutions, increasing the likelihood of succeeding in their programming courses, finishing tasks faster, and attaining goals that would seem difficult or impossible to do if done individually.

In the context of pair programming, dual eye tracking has been explored to study the degree of joint attention of two individuals solving a problem together for the purpose of understanding how gaze and speech are coupled (Pietinen et al. 2008; Jermann et al. 2011; Olsen et al. 2015). These eye tracking studies on pair programming frequently employ the use of gaze coupling (Richardson and Dale 2005), which refer to moments when two individuals are looking at the same target. Prior research suggests that the level of gaze coupling is related to the quality of interaction and better comprehension (Richardson and Dale 2005), tightness of collaboration (Pietinen et al. 2008; Jermann et al. 2011), and quality of collaboration (Nüssli 2011).

One of the standard ways of representing social eye tracking data is using cross-recurrence plots (CRPs) (Schneider and Pea 2013). A CRP can be used to measure how much and when two subjects look at the same spot. Cross-recurrence in eye tracking is synonymous with gaze coupling (Nüssli 2011). An analysis on CRPs is called cross-recurrence quantification analysis (CRQA) (Zbilut et al. 1998). CRQA is used to quantify how often two systems display similar patterns of behavior in time, which involves taking two different trajectories of the same information as input and performing a test of “similarity” or “closeness” between the two trajectories.

Our previous studies on the use of CRQA characterized collaboration patterns according to participants’ prior knowledge (Villamor and Rodrigo 2017a), degree of acquaintanceship (Villamor and Rodrigo 2018a), both prior knowledge and degree of acquaintanceship (Villamor and Rodrigo 2017c), and determining leader-follower profiles Villamor and Rodrigo (2017b). Cross-recurrence was also found to be positively correlated to team performance (Cherubini et al. 2010; Zheng et al. 2016). The study of (Kuriyama et al. 2011) showed that cross-recurrence is higher in successful pairs than in successful pairs in solving tangram puzzles. In one study, CRPs were used to contrast a “good” and a “bad” pair that correlated to a good and bad collaboration quality (Jermann et al. 2011). In another study, eye movement patterns caused by collaboration were identified where it was observed that gaze coupling level is lower for a pair with a bad collaboration flow (Nüssli 2011). Hence, the concept of cross-recurrence is not entirely new.

This study would like to follow up on these previous works by investigating the coupling between the collaborators’ gazes quantified via CRQA to see whether the degree of coupling visualized through CRPs can be used to distinguish how successful and unsuccessful programming pairs collaborate. Specifically, this study seeks answers to the following questions: (1) Is there a significant difference on the CRQA results between successful and unsuccessful programming pairs, and (2) What characterizes the gaze collaboration patterns of successful and unsuccessful programming pairs based on the textures reflected on their respective CRPs and CRQA results? Whereas previous studies were able to establish the relationship between CRPs and collaboration quality, this study is different since it attempts to draw out the differences between the gaze collaboration patterns of the successful and unsuccessful programming pairs using CRPs. In addition, an attempt is also made to interpret the meaning of the CRQA results in relation to how the pairs collaborated.

Gaze cross-recurrence plot and CRQA

Cross-recurrence fixations and CRP

A cross-recurrence plot (CRP) is a matrix that visualizes the time coupling between two time series. It requires that the data should have the same unit and same phase reconstruction for the states of the two time series to be compared. Using eye tracking data, for instance, the two time series could be the fixation sequences of two collaborators containing the fixation x - and y -coordinates and the time when the fixations occur. Hence, given two fixation sequences f_i and g_i , $i = 1 \dots N$, a cross-recurrence is defined as $r_{ij} = 1$ if $d(f_i, g_j) \leq p$; and 0, otherwise (Marwan et al. 2007).

This means that a recurrence occurs when two fixations from different sequences settle within a given threshold or radius p of each other, where d is some distance metric (e.g., Euclidean distance). In Fig. 1, for example, assume that the numbered red and green dots are from the fixations sequences of collaborator 1 (A) and collaborator 2 (B), respectively. Given a certain radius bounded by the black bordered circle shown in the figure, which is considered the threshold whether the two fixations are recurrent or not, fixation pairs (1, 10) and (2, 10) are judged as recurrent since their distances fall within the radius of the circle.

If fixations i and j are recurrent (i.e., $r_{ij} = 1$), these fixations are shown as a black point or pixel on the plot. Figure 2 shows an example of a CRP. The labels along the horizontal and vertical axes refer to the fixation timelines of the first and second collaborators, respectively. Both collaborators in Fig. 2, for example, started at about the same time, 2250 s past the starting time of this particular session. A CRP, therefore, indicates fixations from different collaborators that are recurrent at their respective times.

Different types of small-scale structures called *textures* may be seen on the CRPs (Marwan et al. 2007). Table 1 lists down some of these noticeable textures and their corresponding meanings.

The snapshots in Fig. 3 showing one of the programs used as stimulus in this experiment can help us understand better the relationship between CRPs and collaborative eye tracking. The overlaid colored circles are the fixation points of the two collaborators in this pair. The aqua-colored circles on the left and the purple-colored circles on the right are for the first and second collaborators, respectively. At the top of these snapshots are the times (in seconds) past the starting time of this particular eye tracking session when

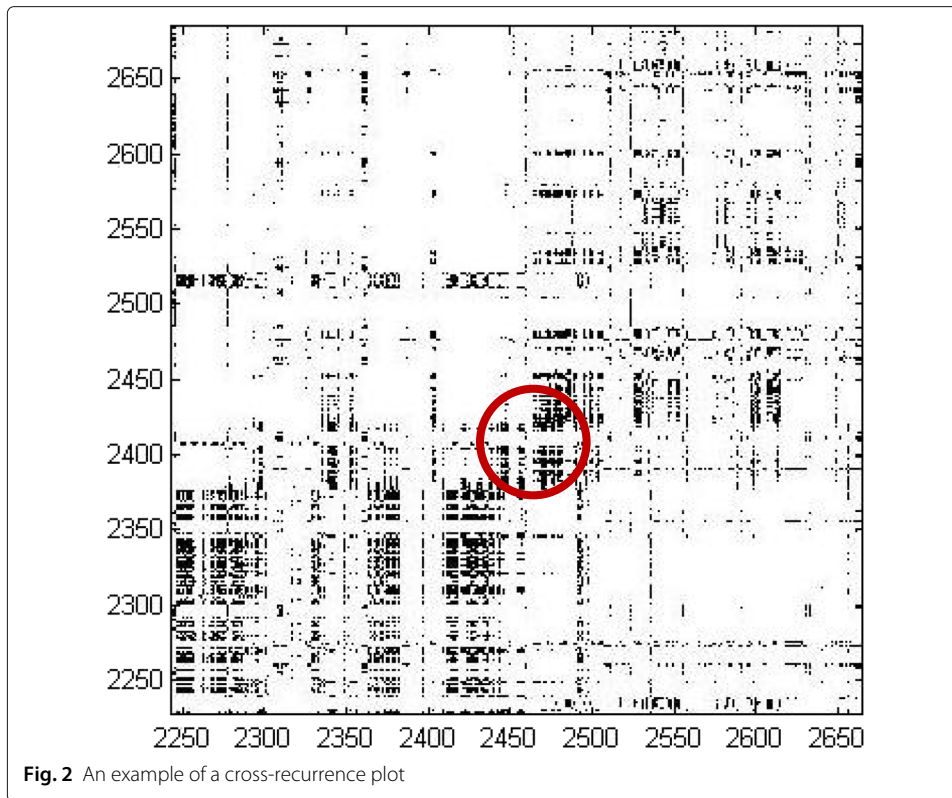
```

Scanner input = new Scanner(System.in);
String str;
int a;

str = input.nextLine();
for(a=str.length-1; a>0; a--){
    System.out.print(str.charAt(a-1));
}
}

```

Fig. 1 An illustration of recurrence fixations

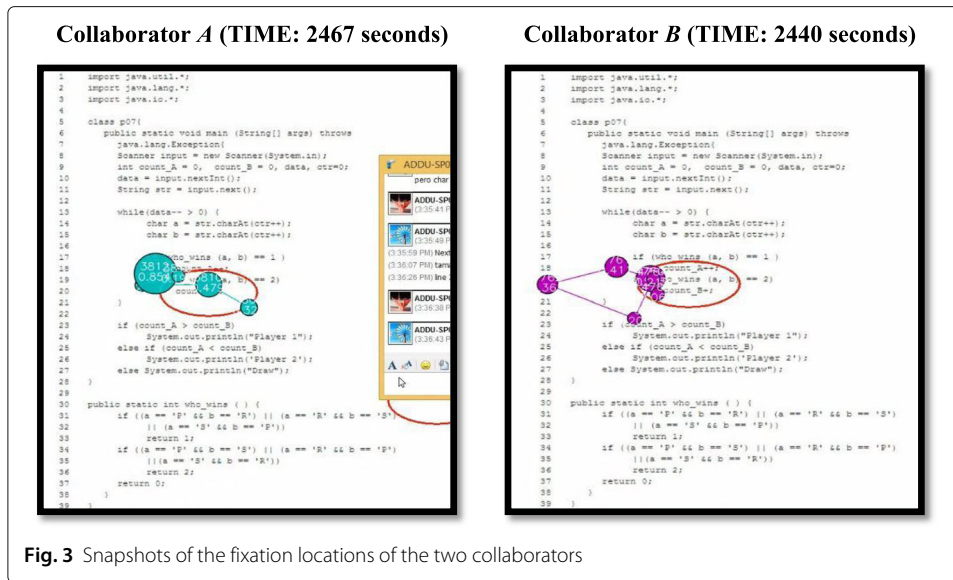


these fixations occurred. The fixations here are considered recurrent based on a given threshold since at these times the fixation points are about in the same location on the stimulus. These recurrent fixations are indicated by the pixelated regions enclosed in a red circle on the CRP in Fig. 2.

Figure 4 shows the corresponding scan pattern of the CRP in Fig. 2 using a line graph. The two subplots illustrate the side-by-side comparison of the fixation x -coordinates (top

Table 1 Some CRP textures and their meanings

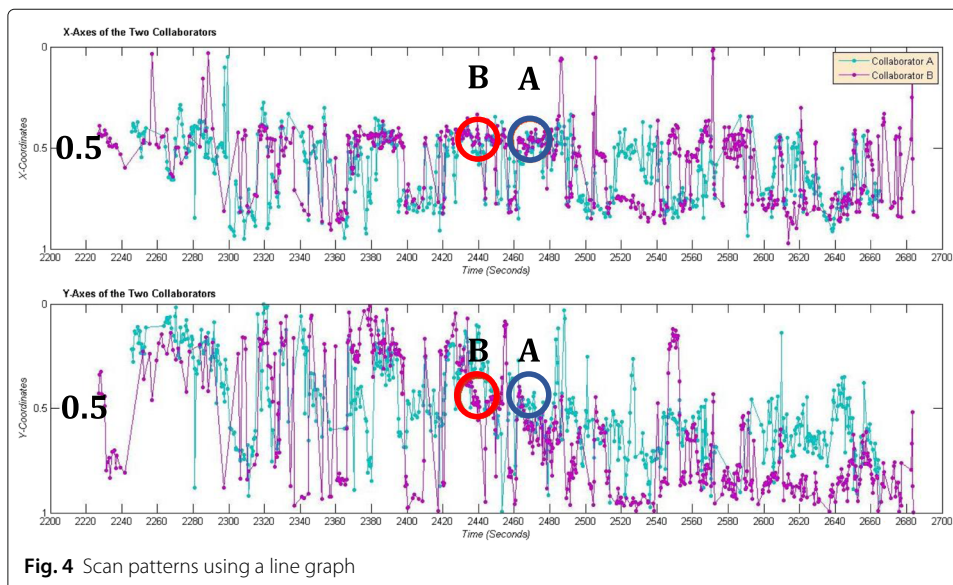
Texture	Meaning
Fading portions to the upper left and lower right corner	Data is non-stationary, i.e., data has slowly varying parameters
Single and isolated recurrence points	Reflect random and strong fluctuations in the data but are not considered a unique sign of chance or noise
Horizontal/vertical lines and rectangular clusters	Marks a time length which a state does not change or changes very slowly or the process is halted at a singularity in which the dynamic is stuck in paused states (called "laminar" or "trapped" states)
Bands of white space	Indicate abrupt changes or transitions that may reflect an underlying state changes
Empty regions	Two collaborators uninterruptedly looked at two different spots on the screen
Diagonal lines parallel to the main diagonal	The states within this period do not occur at any other times (e.g., states are unique or far from normal)
	Segment of one trajectory runs almost parallel to another segment (i.e., the trajectories visit the same region of the phase space at different times)
	Collaborators looked at the same spot on the screen continuously



subplot) and fixation y -coordinates (bottom subplot) of the two collaborators with the aqua and purple line graphs referring to the first and second collaborators, respectively. The x -axes represent the combined timelines of the two collaborators, while the y -axes denote the valid range of values for the fixations, which is between 0 and 1 in reverse order. The fixations of collaborator A that occurred after 2467 s can be found in the portion enclosed by the blue circles, while the fixations of collaborator B that occurred after 2440 s are bounded by the red circles in the figure. The positions of these fixation x - and y -coordinates, which is at about 0.5, suggest that the fixations points of the two collaborators are recurrent.

CRQA

Cross-recurrence quantification analysis (CRQA) is an analysis using CRPs (Marwan et al. 2007). It determines how frequently two systems exhibit similar patterns of behavior over



time by taking two different trajectories of the same information as input and performing a test of “closeness” between all points of the first trajectory with all points of the second trajectory.

Using CRQA, a number of measures can be extracted from the diagonal and vertical dimensions of the CRP (Marwan et al. 2007). For the diagonal dimensions, we have *recurrence rate*, *determinism*, and *average* and *longest diagonal length* and *entropy*. For the vertical dimension, we have *laminarity* and *trapping time*.

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 result in a high value of *RR*.

Determinism (*DET*) is the proportion of recurrence points forming long diagonal structures of all recurrence points. Stochastic and heavily fluctuating data cause none or only short diagonals, whereas deterministic systems cause longer diagonals. 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 line on a recurrence plot (*LMAX*) denotes the longest uninterrupted period of time that both systems stay attuned, which can be seen as an indicator of stability of the coordination. In the context of eye-tracking, it gives the longest time where both scanpaths of the two collaborators are synchronized.

The possibility of measuring the complexity of the attunement between systems can be done using entropy (*ENTR*), which is the Shannon entropy of the probability distribution of the diagonal line lengths $P(l)$. Using eye tracking data, 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. Sharma et al. (2018) also defined entropy as a measure of the level of uncertainty of a random variable, which refers to the objects looked at by the subjects where a high entropy indicates that the subjects looked at many objects while a low entropy indicates that the subjects mostly looked at few objects.

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. In eye tracking data, this represents the prolonged duration where the collaborators tend to focus on certain regions of the screen, either to denote increased concentration or problems in comprehension.

Methods

Participants

The dataset gathered for this study was from the six (6) universities spread across the Philippines. Students aged 18–23 years old who were in their second year to fourth year

level in college and had taken the college-level fundamental programming courses were recruited to participate in this study. A total of 84 participants were paired randomly irrespective of gender, degree of acquaintanceship, and programming experience producing 42 programming pairs.

Experimental procedures

Informed consent forms were distributed to the six (6) universities by the assigned local investigators. A screening questionnaire was distributed to student volunteers to determine their eligibility to take part in this study. The following were the exclusion criteria: (1) wearing bifocals, trifocals, layered, or regression lenses; (2) have difficulty reading a computer screen with contacts and/or eyeglass on; (3) have cataracts; (4) have eye implants; (5) have glaucoma; (6) using a screen reader or magnifier or other assistive technology to use the computer; and (7) if either of the pupils are permanently dilated.

Students who passed the initial screening were asked to take a written program comprehension test for 20 min and a self-efficacy survey for 5 min using the Computer Programming and Self-Efficacy Scale of Ramalingam and Wiedenbeck (1998). The program comprehension test scores were used to determine the participants' prior knowledge or proficiency level in programming and the self-efficacy rating results were used to determine their confidence level in programming. For the experiment proper, the participants were required first to undergo a nine-point eye tracking calibration test. The experiment was designed for 60 min.

Experimental tools

Two Gazepoint GP3 eye trackers were used to collect the pairs' eye movement data. These eye trackers are high-performance and easy-to-use eye trackers with 0.5–1 degree of visual angle accuracy, 60 Hz machine-vision camera, with allowable horizontal and vertical movement of 25 cm and 11 cm, respectively, and depth movement of ± 15 cm range. A slide sorter program with "Previous," "Reset," "Finish," and "Next" buttons was created to display the 12 erroneous programs preceded by a program specification. When a bug is found, the participant clicks on the location of the bug using a mouse. The software then draws an oval to mark it. There was no need to correct the errors. The participants were free to click any of the buttons and navigate to the next or previous slide at their own pace. No scrolling was needed.

The participants were informed how many bugs were there in each program. Each of the 12 programs contained a syntax, semantic, logic, or a combination of these types of errors. The programs were categorized as *easy*, *moderate*, and *hard* depending on the type of errors the program contained. The distribution of the programs based on difficulty was as follows: *easy* (programs 1–3 and 10), *moderate* (programs 4–6 and 11), and *hard* (programs 7–9 and 12). Programs 1–3 contained a single bug and the rest had three bugs. The total number of errors for the 12 programs was 30. Each error marked correctly was awarded with one point.

Because of the limitations of the eye trackers used in this experiment, the participants in the pair cannot see each other's work so if they intend to collaborate and work together on the same problem, a chat program was provided. A chat program was used so that the pairs would not be tempted to look away from their screens if they want to communicate

with their partners. The pairs were co-located but they were spaced far enough to ensure that all communication with their partner was via chat only.

Though the pairs were encouraged to work with their partner and use the chat program, they were not informed that this research was primarily about collaboration. No further instructions were given as to how to proceed with the task and which problems to solve first.

Data cleaning and segmentation

The slide sorter program generated log files for every participant, which contained a recording of buttons (“Next,” “Previous,” “Reset”) pressed and if the participant has already marked the location of the bug (“Mark”), the timestamps when these buttons were pressed, the slide numbers (e.g., slide 0 refers to the program specification of program 1 and slide 1 refers to the actual program), and the x and y screen coordinates of the ovals that appeared after the mouse click.

The fixation files produced by the Gazeport eye trackers were cleaned first by removing fixations with negative x - and y - gaze coordinates because the valid coordinates are those that fall between 0 and 1 only. The number of fixations per slide that contained the actual program were segmented with the help of the information contained in the slide sorter program log files. These segmented fixation files were saved on separate files. Hence, each participant had at most 12 fixation files. Some participants did not finish the 12 programs.

Data analysis

Pair success was measured in terms of debugging scores. The pair debugging score was computed by getting the average of the debugging scores of the individuals in the pairs. Two levels of granularity were used in the analysis: *pair-level* (average of all 12 programs) and *case-level* (individual programs under each pair). “Successful” and “unsuccessful” were defined in both pair-level and case-level to see if differences can be found in terms of their gaze collaboration patterns as they trace and debug programs. Both levels of granularity were used to verify whether the aggregate results (pair-level results) would be consistent when the programs are analyzed individually under each pair (case-level results).

For pair-level, a pair is successful if their average debugging score for the 12 programs is greater than or equal to the mean score ($M = 15.58$, $SD = 4.38$, $Min = 9$, $Max = 25.25$). Otherwise, the pair is unsuccessful. For the case-level, a case is successful when the individuals in the pairs correctly marked at least half of the errors in each program. A case is unsuccessful if both failed to mark all the errors or if only one of them marked at least half of the errors.

The CRQA metrics were derived for each of the 12 programs under each pair after constructing the CRPs. The process was done using the CRP toolbox for MATLAB (Marwan et al. 2007). The challenge of using CRQA is finding optimal parameters for *delay*, *embed*, and *threshold* or *radius*. However, for this experimental data, no further embedding was done, which was based on the work of Iwanski and Bradley (1998). With an embedding dimension of one, delay was also set equal to one since no points were time delayed (Webber Jr. and Zbilut 2005). The threshold was set to a default of 10% of the maximal phase space diameter (Schinkel et al. 2008). Because of varying fixation counts, threshold

adjustments were done as needed to ensure that the threshold was just the right size. If it is too small, the recurrence structure may not provide enough information. Otherwise, if it is too large, it could result to thicker and longer diagonal structures as they actually are since almost every point is a neighbor of every other point.

A *t* test for independent sample means at the 0.05 level of significance was performed to test for statistically significant differences on the CRQA results and other metrics between successful and unsuccessful pairs as well as successful and unsuccessful cases.

Results and discussion

Differences in CRQA results and CRPs

One pair was discarded from the total of 42 pairs because of extreme fixation count discrepancies, i.e., one collaborator in this pair had very high fixation counts and other one had very low fixation counts across 12 programs. Ideally, there should have been 41 pairs \times 12 programs = 492 cases but some pairs did not finish the 12 programs. Other cases also had very low fixation counts in all 12 programs and were not good candidates for CRQA and, thus, were not included. Hence for the case-level analysis, only 376 cases were used. Of the 41 pairs, 19 were successful and 22 were unsuccessful. Of the 376 cases remaining, 196 were successful and 180 were unsuccessful.

Pair-level *t* test results showed no significant differences on the aggregate CRQA results between successful and unsuccessful pairs. This confirms previous findings that cross-recurrence analysis does not support aggregating data from several pairs over long periods to dig out individual differences and discover generalizable patterns of interaction. Despite success in the study of gaze coordination, this kind of analysis is more suitable for examining data from short time windows and one pair at a time (Andrist et al. 2015).

Case-level *t* test results, however, showed significant differences on the CRQA results between successful and unsuccessful cases. This is shown in Table 2. The successful cases have significantly lower CRQA results than unsuccessful cases.

Incidences of high and low instances of each CRQA metric in the successful and unsuccessful cases were examined if notable differences could be found. A CRQA value is high if it is equal to or greater than the mean plus one standard deviation, and low if it is equal to or lesser than the mean minus one standard deviation. Table 3 shows the descriptive values of the CRQA metrics, which covers the mean, standard deviation, the minimum and maximum values, and bases for low and high CRQA values. The large majority of the CRQA values in both categories are average. However, the successful cases have more

Table 2 *T* test CRQA results in the case-level analysis ($N = 376$)

CRQA metric	Successful ($N = 196$)		Unsuccessful ($N = 180$)		<i>t</i> value	<i>p</i> value
	Mean	SD	Mean	SD		
RR	0.36	0.12	0.45	0.14	6.981	0.000
DET	0.75	0.13	0.81	0.11	5.476	0.000
<i>L</i>	3.49	0.82	4.03	1.14	5.378	0.000
LMAX	24.58	16.08	30.57	17.73	3.435	0.001
ENTR	1.57	0.42	1.81	0.44	5.314	0.000
LAM	0.84	0.09	0.89	0.08	5.342	0.000
TT	4.82	1.57	5.68	1.96	4.696	0.006

Table 3 Descriptive values of the CRQA metrics in the case-level analysis ($N = 376$)

CRQA Metric	Mean	Std. Dev.	Minimum	Maximum	Low \leq	High \geq
RR	0.40	0.13	0.04	0.76	0.27	0.54
DET	0.78	0.12	0.36	0.97	0.66	0.90
L	3.75	1.02	2.26	7.54	2.73	4.77
LMAX	27.44	17.13	5.00	111.00	10.31	44.58
ENTR	1.68	0.44	0.63	2.74	1.24	2.13
LAM	0.86	0.09	0.52	0.98	0.78	0.95
TT	5.23	1.82	2.34	5.23	3.41	7.04

instances of low RR, DET, L , LMAX, ENTR, LAM, and TT; and the unsuccessful cases have more occurrences of high RR, DET, L , LMAX, ENTR, LAM, and TT.

To determine further what factors could have contributed to the CRQA differences, all successful cases representing successful pairs that have all low CRQA metric values and all unsuccessful cases representing unsuccessful pairs with all high CRQA metric values from the new subset were extracted. Fifteen (15) successful cases and eleven (11) unsuccessful cases fit the CRQA criteria. The 15 successful cases have fixation count overall that are either low or below the fixation count midpoint. All, except one, in the 11 unsuccessful cases have fixation count overall that are either high or above the fixation count midpoint.

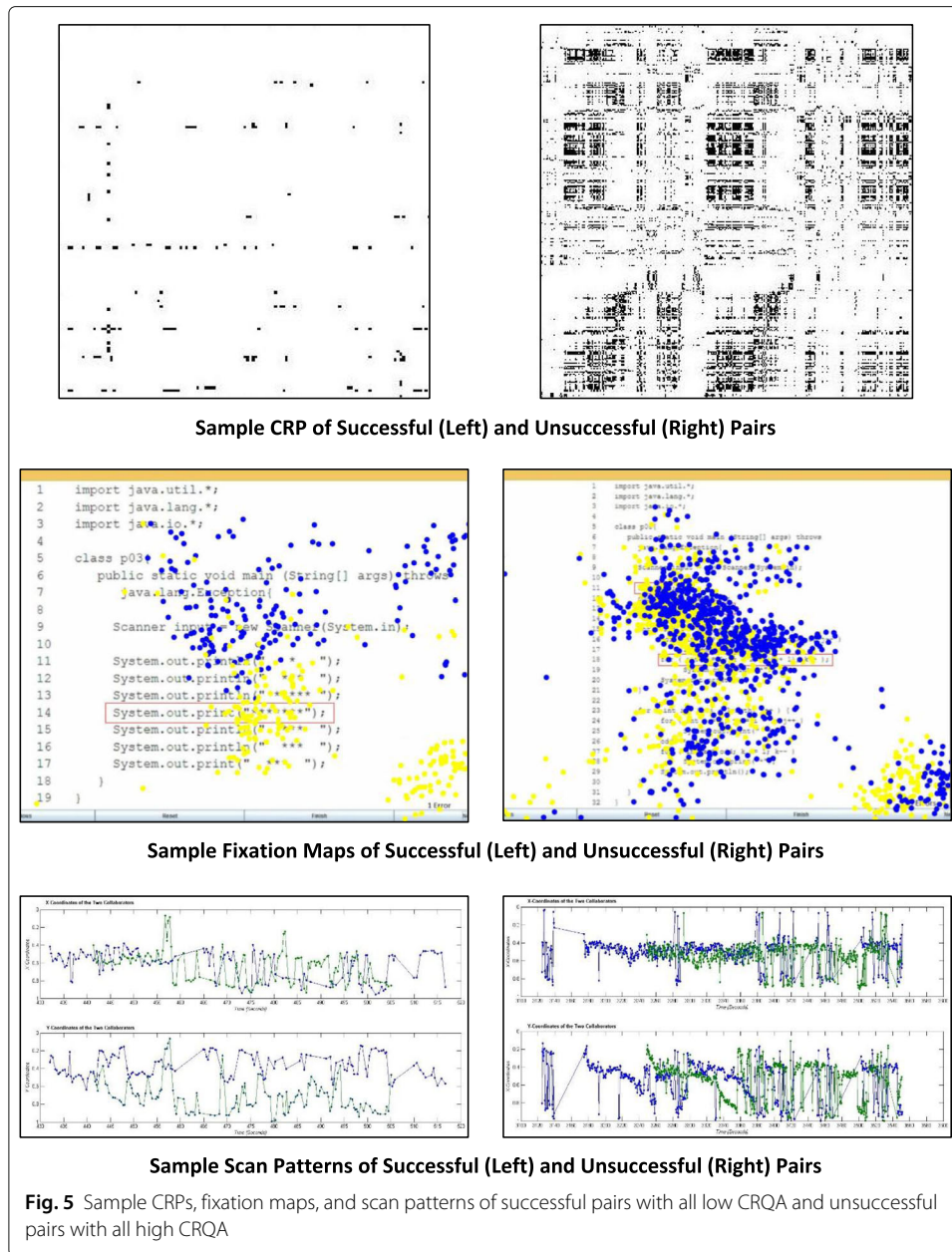
The CRPs of these pairs were examined next based on the CRP textures to extract observable differences that could potentially explain the CRQA characterizations between successful and unsuccessful pairs. The succeeding subsections present the differences between the CRPs, fixation maps, and scan patterns of the successful and unsuccessful pairs.

CRPs of successful and unsuccessful pairs

The fixation count overall of the successful pairs that are either low or below the midpoint are depicted on the 15 CRPs of the successful pairs with all low CRQA as having mostly single isolated points, more bands of white spaces and empty regions, and fewer rectangular segments of recurrence points. The corresponding fixation maps of these CRPs were also explored to see the fixation distribution of the successful pairs on the stimulus. The corresponding fixation maps show more dispersed fixations and the collaborators' fixations are mostly not clustered in similar areas. The corresponding scan patterns were also inspected and their scan patterns reflect more dissimilar patterns compared alongside the scan patterns of the unsuccessful pairs.

The fixation count overall of the unsuccessful pairs that are either high or above the midpoint are reflected on the 11 CRPs of the unsuccessful pairs with all high CRQA as having more rectangular or larger clusters of recurrence points and more visually recurring patterns. The fixation maps reveal that the fixations are heavily clustered on certain locations of the stimuli, and the scan patterns manifest more pronounced similarities compared to the scan patterns of the successful pairs. Figure 5 shows a comparison of the CRPs, fixation maps, and scan patterns representative of successful pairs with all low CRQA and unsuccessful pairs with all high CRQA.

Putting all these together, the lower number of fixation count overall, more dispersed fixations, and less similar scan patterns resulted to low CRQA in successful pairs. On the other hand, the higher number of fixation count overall, fixation cluster patterns, and more pronounced scan pattern similarities led to high CRQA in unsuccessful pairs.



The larger horizontal or vertical segments and clusters of recurrence points on the CRPs, called “laminar” states, also increased the values of LAM and TT.

Loci and sequence similarities

Two additional measures were computed to provide supporting evidence to the differences in RR and DET between successful and unsuccessful pairs. These measures are the *loci similarity* and *sequence similarity*. Loci similarity is defined as the percentage of locations where the two collaborations have looked at, independently of time and sequence. Sequence similarity is the percentage of similar scanpaths. Levenshtein distance, also known as the edit distance algorithm (Levenshtein 2002), was used in the computation of these two measures. This algorithm counts the number of operations (e.g., deletions,

insertions, and substitutions) needed to transform one string into to the other. This is accessible using the OGAMA software (Voßkühler et al. 2008) where a loci similarity value of 100% denotes that the fixations of the participants in the pair are in the same locations on the stimulus, and a sequence similarity value of 100% means that the participants have very identical scanpaths.

To do this, each program stimulus was divided into an $N \times N$ grid, where each cell in the grid was assigned a unique letter. Each participant's scanpath per program was built into a string using the letters of the cell that contained the current fixation location, and then Levenshtein distance was applied to the scanpaths of the pairs. Results showed that the successful pairs have lower loci similarity ($M_{\text{Successful}} = 61.30$, $SD_{\text{Successful}} = 9.79$) than the unsuccessful pairs ($M_{\text{Unsuccessful}} = 65.47$, $SD_{\text{Unsuccessful}} = 9.42$), and the difference is significant at ($t = 3.324$, $p = 0.001$). Likewise, the successful pairs have significantly lower sequence similarity than the unsuccessful pairs ($M_{\text{Successful}} = 11.63$, $SD_{\text{Successful}} = 4.17$; $M_{\text{Unsuccessful}} = 13.09$, $SD_{\text{Unsuccessful}} = 3.95$; $t = 2.753$, $p = 0.006$). This implies that the successful pairs had less incidences where they looked at the same parts of the program and their scanpaths were not identical most of the time. These results corroborate with the findings that successful pairs have significantly lower RR and DET than unsuccessful pairs.

Controlling for confounds

Is it possible that the number of errors in the program or program complexity and the proficiency level of the pairs confounded the CRQA results? Upon inspection, it was found that 11 and 4 out of the 15 successful cases representing successful pairs came from programs that were categorized as easy (with single error) and moderate (with three errors), respectively. Seven (7) and 4 out of the 11 unsuccessful cases representing unsuccessful pairs emanated from programs tagged as hard (three errors) and moderate (three errors), respectively. In short, none of the 15 were hard programs, and none of the 11 were easy programs. Hence, to control for these confounds, successful cases representative of the hard programs as well as unsuccessful cases representative of the easy programs containing a single error were selected.

Prior to examining their CRPs, fixation maps, and scan patterns, two separate t tests were performed to determine if there are significant differences on the CRQA values between successful and unsuccessful cases on easy and hard programs. The successful cases have significantly lower CRQA values than unsuccessful cases on the easy programs, except for LAM and TT. For the hard programs, only RR, DET, and LAM are significant.

Eight (8) CRPs each of the successful cases representative of hard programs and unsuccessful cases representative of easy programs with a single error were sampled. Single and isolated points and the presence of more bands of white spaces and empty regions are still evident on the CRPs of the successful pairs on hard programs. However, small clusters of points can already be seen forming mostly in vertical and horizontal patterns, and these clusters are manifested on their fixation maps. These incidences of laminar states increase the value of LAM and TT, suggesting that the successful pairs also need more time to locate all the errors in the hard programs. The scan patterns, however, remain to be more different than alike.

The clusters of points or laminar states are still larger and more prominent on most of the CRPs of the unsuccessful pairs even when working on easy programs, which are

also reflected on their fixation maps. This could explain why they have more occurrences of high LAM and TT, which could also be the reason for the visually recurring patterns found on their CRPs. Though there are also isolated incidences of recurrence points on their CRPs, the heavily clustered points are still more apparent. Their scan patterns also show more similarities compared to the scan patterns of the successful pairs.

The heavy cluster of points implies that given only a single error to find, most of the unsuccessful pairs would still have a hard time looking for it. Their more similar scan patterns also suggest that the individuals in the unsuccessful pairs tend to follow a certain pattern of scanning programs, which could either be linear scanning or alternately scanning from top to bottom.

One of the reasons also that could be a factor that makes some pairs successful over other pairs is the proficiency level of the pairs. Of the 125 successful cases from 19 successful pairs, 68 of these cases were from highly proficient pairs, 46 from mixed proficiency pairs, and only 11 cases from low proficiency pairs. Of the 112 unsuccessful cases from the 22 unsuccessful pairs, only 17 of these cases were from highly proficient pairs, 43 from mixed proficiency pairs, and 55 from low proficiency pairs. Prior research show that prior knowledge from previous courses can influence student achievement (Hailikari et al. 2008), so it is a known fact that students with high prior knowledge outperform students with low prior knowledge in problem solving tasks. Hence, it is also expected that when students with high proficiency levels are paired or group together in collaborative learning situations, they would perform better than pairs or groups with low proficiency levels.

In one of the exploratory analysis conducted using half of the dataset in this study, it was found that low proficiency pairs have significantly higher RR, DET, ENTR, and LAM than highly proficient and mixed proficiency pairs (Villamor and Rodrigo 2017a). To extend this previous finding, ANOVA was performed on the CRQA results based on the pairs' proficiency levels. Results showed that successful pairs have significantly lower CRQA values than unsuccessful pairs in all proficiency levels, except for LMAX and TT.

The CRPs, fixation maps, and scan patterns were also examined. Most of the CRPs and fixation maps of the unsuccessful cases exhibit more heavily clustered recurrence points and their scan pattern similarities are more pronounced compared to the successful cases in all proficiency levels. In summary, among the CRQA metrics, LMAX and TT are the only metrics that are not significant across categories based on program category and proficiency level of the pairs. The next section presents the collaboration patterns as inferred from the CRQA results of the successful and unsuccessful pairs.

Collaboration patterns of successful pairs

In the discussion that follows, RR, DET, and L are combined as these three metrics are more interconnected than the others. As the recurrence points increase (increase in RR), the chances of forming diagonal lines (DET) are also higher. When the density of recurrence points falls within the diagonal line, the more the pairs are considered closely coupled. L measures the average of these diagonal line lengths. LMAX and TT are not included for the reason discussed earlier.

- Successful pairs have more “low RR/DET/ L .”

The successful pairs may prefer more individual work. It is possible that the individuals in the successful pairs search for errors independently of each other even when working on the same program together. Our evidence shows that successful pairs had more instances where they worked together on the same program (Villamor and Rodrigo 2018b). However, working together may not necessarily mean that they are looking at the same locations in the program and follow the same scanpaths.

To verify this, the loci and sequence similarities were computed and compared when pairs worked together and did not work together. The difference in loci similarity is not significant, but successful pairs working together have significantly lower sequence similarity compared to when they did not work together ($M_{\text{WorkedTogether}} = 10.68$, $SD_{\text{WorkedTogether}} = 3.84$; $M_{\text{NotTogether}} = 13.26$, $SD_{\text{NotTogether}} = 4.25$; $t = 3.448$, $p = 0.001$). This proves that when successful pairs work together, it does not guarantee that their scanpaths would be identical.

The lower turnout of these CRQA metrics may also be an indication that successful pairs chat more frequently. We also have evidence that shows that they indeed chatted more than the unsuccessful pairs (Villamor and Rodrigo 2018b). How does chatting more frequently affect or reduce these CRQA metrics? When pairs chat, sudden gaze transitions are more common. These sudden gaze transitions occur when the current point-of-regard is suddenly shifted to the chat window or moves away from the chat window towards a particular location on the stimulus. These sudden fixation shifts characterized by longer saccades (i.e., eye movements occurring between fixations) may reduce the possibility of having recurrent fixations because the chat windows of the pairs may not be positioned in the same locations on the screen or could be positioned far away from the source code text. Findings revealed that successful pairs have significantly longer average saccade lengths than unsuccessful pairs ($M_{\text{Successful}} = 130$ px, $SD_{\text{Successful}} = 27$ px; $M_{\text{Unsuccessful}} = 117$ px, $SD_{\text{Unsuccessful}} = 23$ px; $t = -3.863$, $p = 0.000$).

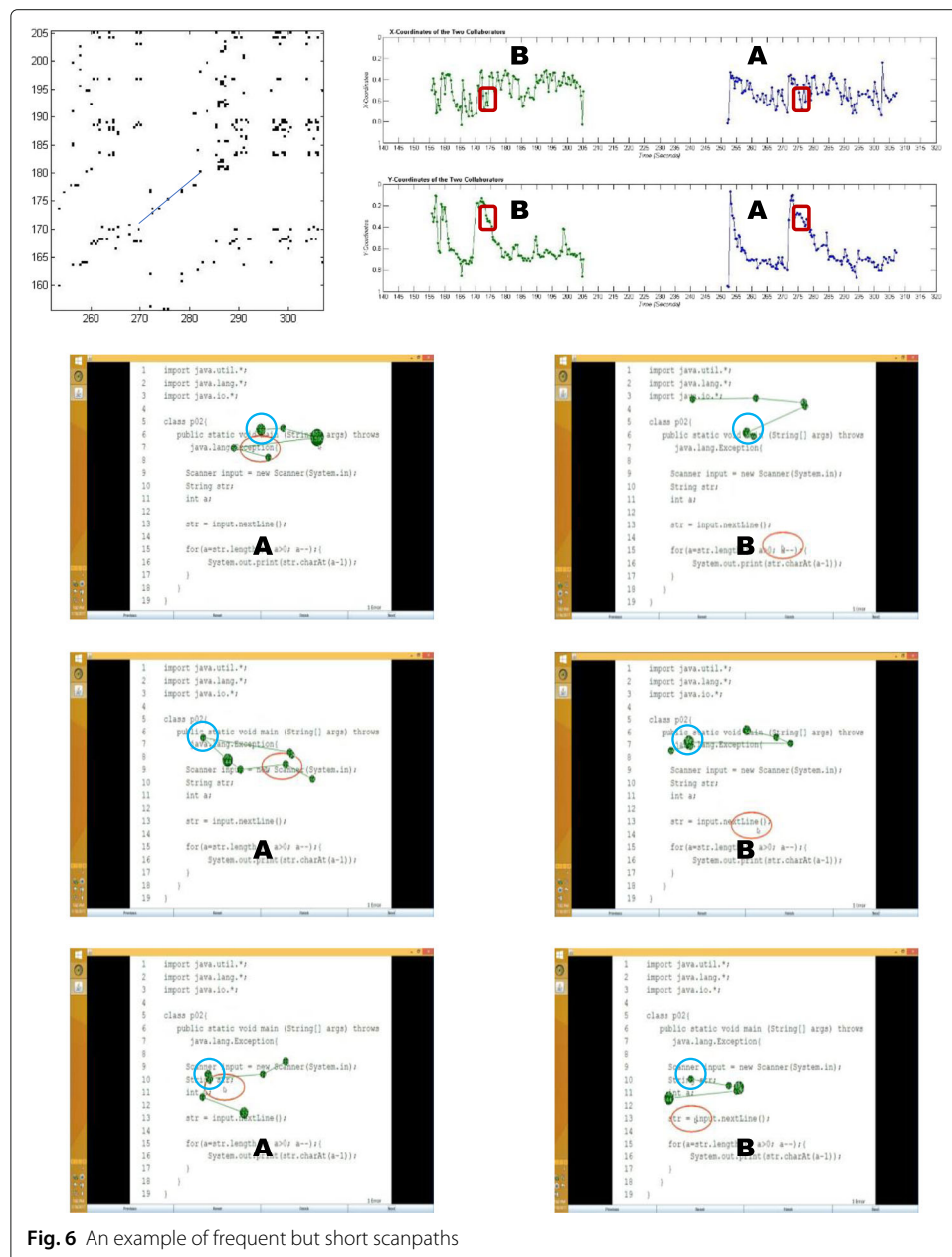
The successful pairs' preference for more individual work and chatting more frequently may be a sign that they are more confident. It was found that the successful pairs mostly chat with partners to confirm the errors that they had already identified. Hence, they only chatted as needed, which did not necessarily require gaze coordination because their chat windows may be positioned at different locations on the screen. A combination of these two: working independently and chatting frequently as needed, increases accuracy and confidences of decisions.

To further verify that successful pairs are more confident, the participants' self-efficacy ratings, were investigated where all participants personally rated themselves how confident they are in doing programming tasks as identified in the Computer Programming and Self-Efficacy Scale of Ramalingam and Wiedenbeck (1998). The rating scales are as follows: 7—absolutely confident, 6—mostly confident, 5—fairly confident, 4—50/50, 3—slightly confident, 2—mostly not confident, and 1—not at all confident. The rating results of the individuals in the pairs were averaged to get the pair rating. Results showed that successful pairs are indeed more confident than the unsuccessful pairs given their average rating of 5.03 (SD = 0.53) which is *fairly confident*, compared to the unsuccessful pairs average rating of 4.49 (SD = 0.71) which is borderline 50/50. The difference is significant at $t = -6.720$ and $p = 0.000$.

- Successful pairs have more “low ENTR.”

This is an indication that successful pairs frequently share similar but shorter scanpaths. An example of frequent but shorter scanpaths is shown in Fig. 6. The CRP on the left shows recurrence points that are aligned diagonally but at certain intervals. These recurrence points are found somewhere in the enclosed portions of the scan patterns on the right side. The series of video snapshots below show the corresponding locations of these diagonally aligned recurrence points on the stimulus. This means that the low ENTR of the successful pairs is due to having more incidences similar to this illustration.

Frequent but shorter similar scanpaths may be a sign that successful pairs are more strategic and more precise when searching for errors because they need to look at only fewer elements on the screen, possibly those locations where errors are more likely to occur. Because majority of the successful pairs are highly proficient, then these pairs may



use their knowledge from experience to look at certain locations in the program where they think the errors are most likely to be found. This confirms the findings of Sharma et al. (2018) suggesting that having a low ENTR is the result of having a small number of elements looked over a fixed period of time where they coined the term “*focused gaze*” to refer to this scenario.

- Successful pairs have more “low LAM.”

This connotes that successful pairs transition faster once they find the errors in a program; hence, they finish sooner and find more errors. This is an indication that they may have less comprehension problems. Their shorter total fixation time on-target compared to unsuccessful pairs ($M_{\text{Successful}} = 18.7$ s, $SD_{\text{Successful}} = 15.97$ s; $M_{\text{Unsuccessful}} = 27.4$ s, $SD_{\text{Unsuccessful}} = 31.2$ s; $t = 2.764$, $p = 0.006$) proves that successful pairs indeed find bugs faster and do not dwell on them once found.

This may also mean that successful pairs are more time-conscious and, hence, more particular about finishing on time. They make sure that they can proceed immediately to the next programs so that they can find more bugs within the allotted time limit. Upon examination of the chat logs and using the phrases such as “let’s go,” “let’s move on,” “moving on,” and “let’s proceed,” it was found that 30% of the time the successful pairs explicitly used these phrases to signal their partners to proceed to the next program. The unsuccessful pairs used these phrases 26% of the time.

Collaboration patterns of unsuccessful pairs

As per successful pairs, RR, DET, and L are collectively discussed, and LMAX and TT are not included.

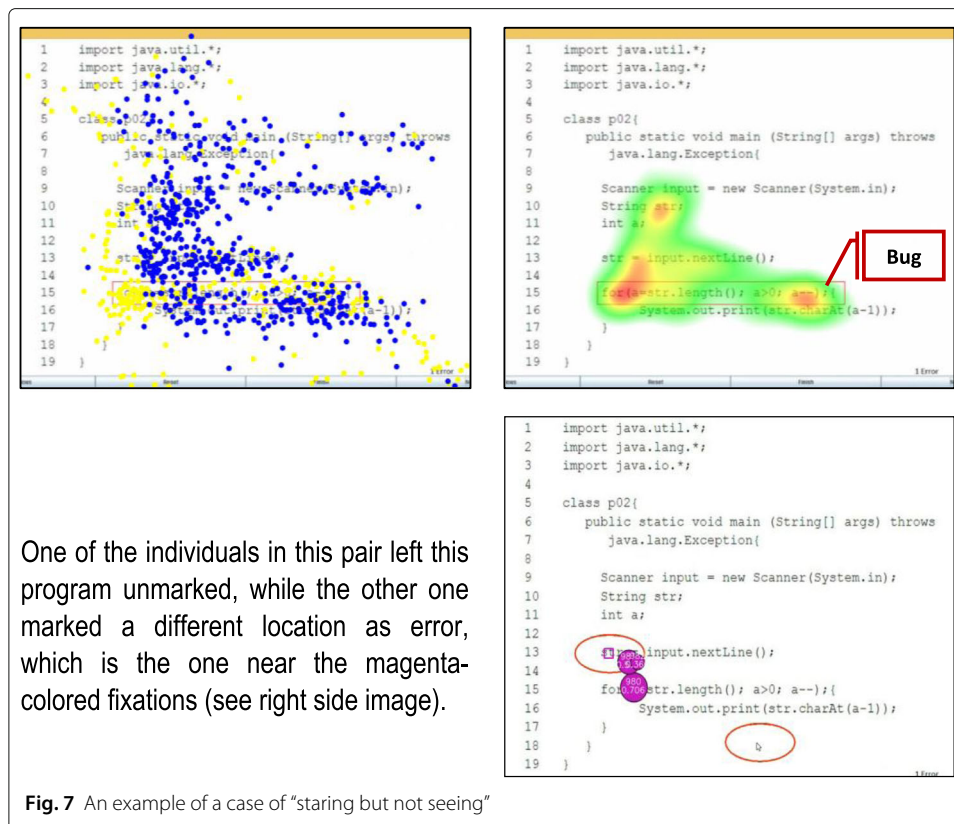
- Unsuccessful pairs have more “high RR/DET/ L .”

This is an indication that unsuccessful pairs may have more incidences of “staring but not really seeing,” that is, fixating but not actively encoding information about the program. Their heavily clustered fixations resulting to a possible increase in RR are often an index of greater uncertainty in recognizing a target item (Jacob and Karn 2003). They may be looking at the same area repetitively because they do not know where else to look, and this may have resulted to more similar scanpaths. The unsuccessful pairs’ higher loci and sequence similarities prove that they had more instances where they had looked at the same locations on the screen and had shared more similar scanpaths brought about by looking repetitively at the same locations back and forth.

Figure 7 shows a fixation map and heatmap of one of the unsuccessful pairs showing fixation clusters on the error and increased attention on the line that contains the error but failed to mark the error. The error in this program is the extra semi-colon right after the for loop. This is an example of a case of “staring but not seeing.”

- Unsuccessful pairs have more “high ENTR.”

The unsuccessful pairs tend to have more complicated scanpath relationships, which means that the length of their scanpath similarities tend to vary a lot. This is a sign of



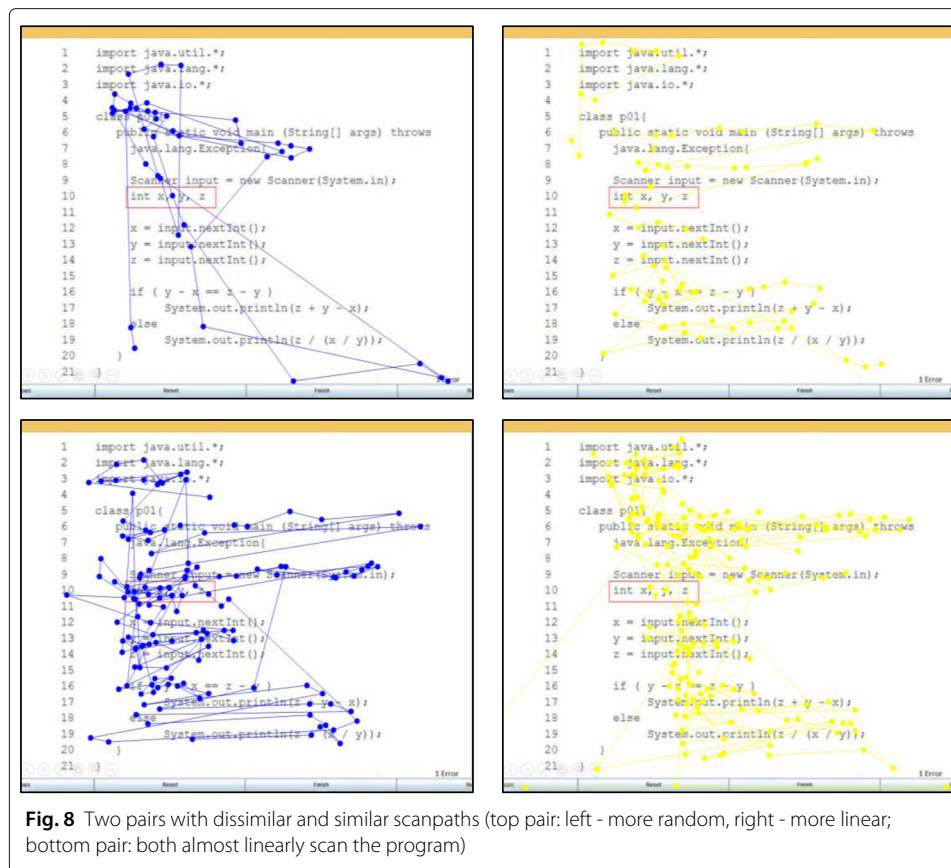
varying levels of uncertainties when looking for errors. Some pairs may have very different scanpaths while others follow almost the same scanpaths when looking for bugs. Figure 8 shows two unsuccessful pairs with different and similar scanpaths, respectively. The first pair (top) with different scanpaths is less certain where the error is. In the second pair (bottom), even if both individuals have almost alike scanpaths, only one is more certain where the error is.

A high ENTR indicates that both individuals look at more elements on the screen at a given time window, which is called “*unfocused gaze*” (Sharma et al. 2018). This implies that unsuccessful pairs tend to extend their search to other parts of the program or may repetitively look at the same areas because they do not know where else to look for the errors or they may have already seen them but not realizing that those are already the errors.

- Unsuccessful pairs have more “high LAM.”

This implies that unsuccessful pairs need lengthy consideration of the program and they are more likely to get stuck in the program; hence, they finish slower and find fewer bugs. This is also an indication that they might be having more comprehension problems. The unsuccessful pairs’ significantly longer total fixation time on-target proves that in most cases they were having difficulty deciding whether the ones they were looking at were actually the errors or they simply did not understand what the program was all about.

This may also suggest that unsuccessful pairs have the habit of skipping a problem because they have a hard time understanding or answering it, then get back to it several minutes after, and still being faced with the same comprehension problem. This pattern



may have caused the visually recurring patterns or uniformly sized laminar structures that increase the value of LAM.

To verify this, the chat logs of the pairs were examined to see manifestations of programs being skipped and answered at a later time. Results showed that unsuccessful pairs had more incidences where they skipped a problem and got back to it compared to the successful pairs. The result, however, is not significant, but this pattern of skipping problems, returning to it and tracing it in a similar manner as before combined with the unsuccessful pairs' higher fixation counts may have resulted to the visually recurring patterns and larger clusters of recurrence points found on their CRPs.

Figure 9 shows an example where one of the unsuccessful pairs decided to skip a program timestamped at 3:33:17 PM then got back to it at 3:59:49 PM and still encountered the same comprehension problem. Underneath it is the corresponding CRP of this pair with the recurring patterns.

Implications on computer science education

If we think about collaboration in the classroom in CS education, one of the strategies that comes to mind is to engage students in pair programming sessions. Prior research show that pair programming has been beneficial to students' learning and self-esteem. However, despite these benefits studies have also shown that pair programming may not be for everybody and it could actually do more harm than good as in the case of the more struggling students in programming, which typically happens when pair programming is

have higher fixation counts and have more occurrences of high CRQA metrics. Since the CRQA results are significant, the textures found on the CRPs of the successful and unsuccessful pairs were examined. It was observed that the CRPs of the successful pairs contain traces of more single and isolated points, bands of white spaces and empty regions, and few rectangular segments of recurrence points. On the other hand, the CRPs of the unsuccessful pairs manifest heavily clustered recurrence points or larger laminar states and visually recurring patterns.

The collaboration patterns of the successful and unsuccessful pairs were then characterized as interpreted from their respective CRQA results and CRP textures. The successful pairs are found to have more preference for individual work at specific times, they may have frequently shared similar but shorter scanpaths, have more frequent scanpath transitions, and transition faster so they find errors more quickly. The unsuccessful pairs, on one hand, may need lengthy consideration of the program, may have shared similar scanpaths that are longer, follow a certain pattern in searching for errors, may look at the same region repeatedly because they are clueless as to where else to look, use trial-and-error in debugging, and usually exhibit program comprehension problems.

The results of this study is inconsistent with prior literature, i.e., higher degree of gaze coupling is tantamount to better collaboration. However, the issues raised by Nüssli (2011) regarding gaze cross-recurrence and collaboration were confirmed in this study. The higher degree of gaze coupling by the unsuccessful pairs were due to the following: (1) they were looking at the same place together due to chance, (2) the collaborators' gazes were directed to specific regions on the screen at different moments in time, which is not a result of a conversational process, and (3) they used the same manner of reading source codes, which was similar to reading ordinary text, resulting to more or less looking at the same area at the same time. It does not necessarily follow, therefore, that a pair with a higher degree of gaze coupling is more coupled or have better collaboration.

Utilizing CRPs is good when we are to analyze the temporal evolution of joint attention of two individuals in a collaborative task. It is easy to tell from the CRPs if a pair has low visual synchronization or they have good visual coordination. However, the downside of using CRPs is that it cannot give us information as to "where" the pairs coupled their gazes during the interaction. Other methods are used to supplement the CRPs in order to get this kind of information. Hence, to obtain more substantial results, it is recommended in our future work to consider examining other streams of data such as the pairs' discourse data and the fixation recordings among other things.

Abbreviations

CRP: Cross-recurrence plot; CRQA: Cross-recurrence quantification analysis; DET: Determinism; ENTR: Entropy; *L*: Average diagonal length; LAM: Laminarity; LMAX: Longest diagonal length; RR: Recurrence rate; TT: Trapping time

Acknowledgments

The authors would like to thank Ateneo de Davao University, Ateneo de Manila University, Ateneo de Naga University, University of Cordillera, University of San Carlos, and University of Southeastern Philippines for allowing us to conduct the eye tracking experiment.

Authors' contributions

MV carried out the study and drafted the manuscript. MMR contributed to the review of the manuscript. Both authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The data cannot be shared because we informed the participants that their data will be kept confidential when we sought for ethical clearance for this research. It was explicitly stated in the ethical clearance that only the collaborators in the multi-institution team will have access to the data.

Competing interests

The authors declare that they have no competing interests.

Author details

¹College of Information and Computing University of Southeastern Philippines, Bo. Obrero, Davao City, Philippines.

²Department of Information Systems and Computer Science Ateneo de Manila University, Loyola Heights, Quezon City, Philippines.

Received: 1 February 2019 Accepted: 14 November 2019

Published online: 23 December 2019

References

- Andrist, S., Collier, W., Gleicher, M., Mutlu, B., Shaffer, D. (2015). Analyzing gaze coordination with epistemic network analysis. *Frontiers in Psychology*, 6, 1016.
- Cherubini, M., Nüssli, M.A., Dillenbourg, P. (2010). This is it!: Indicating and looking in collaborative work at distance. *Journal of Eye Movement Research*, 3(5), 1–20.
- Hailikari, T., Katajauvuri, N., Lindblom-Ylänne, S. (2008). The relevance of prior knowledge in learning and instructional design. *American Journal of Pharmaceutical Education*, 72(5), 113.
- Hannay, J.E., Dybæ, T., Arisholm, E., Sjøberg, D.I. (2009). The effectiveness of pair programming: A meta-analysis. *Information and Software Technology*, 51(7), 1110–1122.
- Iwanski, J.S., & Bradley, E. (1998). Recurrence plots of experimental data: To embed or not to embed? *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 8(4), 861–871.
- Jacob, R.J., & Karn, K.S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The Mind's Eye*. <https://doi.org/10.1016/b978-044451020-4/50031-1> (pp. 573–605).
- Jermann, P., Mullins, D., Nüssli, M.A., 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.
- Kuriyama, N., Terai, A., Yasuhara, M., Tokunaga, T., Yamagishi, K., Kusumi, T. (2011). Gaze matching of referring expressions in collaborative problem solving. In *International Workshop on Dual Eye Tracking in CSCW (DUET 2011)*.
- Levenshtein, V.I. (2002). Bounds for deletion/insertion correcting codes. In *Proceedings IEEE International Symposium on Information Theory*. <https://doi.org/10.1109/isit.2002.1023642> (p. 370): IEEE.
- Marwan, N., Romano, M.C., Thiel, M., Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6), 237–329.
- Murphy, L., Fitzgerald, S., Hanks, B., McCauley, R. (2010). Pair debugging: a transactive discourse analysis. In *Sixth international workshop on Computing education research*. <https://doi.org/10.1145/1839594.1839604> (pp. 51–58): ACM.
- Nüssli, M.A. (2011). *Dual eye-tracking methods for the study of remote collaborative problem solving*. PhD thesis: École Polytechnique Fédérale de Lausanne.
- Olsen, J.K., Ringenberg, M., Alevén, V., Rummel, N. (2015). Dual eye tracking as a tool to assess collaboration. In *ISLG 2015 fourth workshop on intelligent support for learning in groups* (pp. 25–30).
- Pietinen, S., Bednarik, R., Glotoya, T., Tenhunen, V., Tukiainen, M. (2008). A method to study visual attention aspects of collaboration: eye-tracking pair programmers simultaneously. In *Proceedings of the 2008 symposium on Eye tracking research and applications*. <https://doi.org/10.1145/1344471.1344480> (pp. 39–42): ACM.
- Ramalingam, V., & Wiedenbeck, S. (1998). Development and validation of scores on a computer programming self-efficacy scale and group analyses of novice programmer self-efficacy. *Journal of Educational Computing*, 19(4), 367–381.
- Richardson, D.C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science*, 29(6), 1045–1060.
- Schilbach, L. (2015). Eye to eye, face to face and brain to brain: novel approaches to study the behavioral dynamics and neural mechanisms of social interactions. *Current Opinion in Behavioral Sciences*, 3, 130–135.
- Schinkel, S., Dimigen, O., Marwan, N. (2008). Selection of recurrence threshold for signal detection. *The European Physical Journal-Special Topics*, 164(1), 45–53.
- Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-supported collaborative learning*, 8(4), 375–39.
- Sharma, K., Olsen, J.K., Alevén, V., Rummel, N. (2018). Exploring causality within collaborative problem solving using eye-tracking. In *Lifelong Technology-Enhanced Learning Lecture Notes in Computer Science* (pp. 412–426). Cham: Springer.
- Villamor, M., & Rodrigo, M.M. (2017a). Characterizing collaboration based on prior knowledge in a pair program tracing and debugging eye-tracking experiment. In *15th National Conference on Information Technology Education (NCITE 2017)*.
- Villamor, M., & Rodrigo, M.M. (2017b). Exploring lag times in a pair tracing and debugging eye-tracking experiment. In *25th International Conference on Computers in Education* (pp. 234–236).
- Villamor, M., & Rodrigo, M.M. (2017c). Impact of prior knowledge and acquaintanceship on collaboration and performance: a pair program tracing and debugging eye-tracking experiment. In *25th International Conference on Computers in Education* (pp. 186–191).
- Villamor, M., & Rodrigo, M.M. (2018a). Do friends collaborate and perform better?: A pair program tracing and debugging eye-tracking experiment. In *18th Philippine Computing Science Congress* (pp. 9–16).
- Villamor, M., & Rodrigo, M.M. (2018b). Impact of pair programming dynamics and profiles to pair success. In *26th International Conference on Computers in Education* (pp. 123–132).

- Voßkühler, A., Nordmeier, V., Kuchinke, L., Jacobs, A.M. (2008). Ogama (open gaze and mouse analyzer): Open-source software designed to analyze eye and mouse movements in slideshow study designs. *Behavior research methods*, 40(4), 1150–1162.
- Webber Jr., C.L., & Zbilut, J.P. (2005). Recurrence quantification analysis of nonlinear dynamical systems, In *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 26–94).
- Zbilut, J.P., Giuliani, A., Webber Jr, C.L. (1998). Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters A*, 246(1-2), 122–128.
- Zheng, B., Hajari, N., Atkins, M.S. (2016). Revealing team cognition from dual eye-tracking in the surgical setting, In *Ninth Biennial ACM Symposium on Eye Tracking Research and Applications*. <https://doi.org/10.1145/2857491.2884062> (pp. 321–322): ACM.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
