

Detecting Learners' Weak Points Utilizing Time Intervals of Pen Strokes

Kazuya Kishi ^{*} , Motoki Miura ^{*}

Abstract

We consider that most learners tend to focus on and review only problems they incorrectly answered after performing exercises. To deepen their understanding, it is also necessary for learners to review problems that took time to answer even though they were answered correctly. However, it is difficult to judge which problems took time to answer with an ordinary pen and paper. Therefore, we adopt a digital pen to help learners to recognize parts of a problem that took long to complete. Using the pen-stroke interval data obtained by a digital pen, we can discover the weak points of a learner. In this study, we implemented the method and evaluated it by comparing the weak points extracted by system, learners and evaluators. The results confirm that the method can detect weak points with an accuracy of about 50% to 60%.

Keywords: digital pen, weak point, writing data

1 Introduction

In recent years, with the development of information technology, various digital devices are now used in education. The Japan Ministry of Education, Culture, Sports, Science and Technology (MEXT) aims to expand ICT education by introducing one unit of digital equipment by 2020 as the “information vision of education” [1]. Some schools already have classes that use electronic blackboards and tablets. In addition, there are many people who use e-learning, which is online learning that offers a wide range from business people to students. As information technology is now used in education, MEXT released the final report for the “Roundtable on Computers in Education for the 2020s” project [2]. In the report, two issues are specified. One is that the cost of constructing digital equipment, wireless LANs, and other networks and systems should be considered. The other is that conventional digital devices require user expertise and are, hence, not accessible to teachers and children.

A digital pen is a tool for digitizing a learner's handwriting. A digital pen does not expect expertise because its usage is similar to that of conventional paper and pen. Therefore, a digital pen is straightforward and intuitive for learners [3][4]. With the Anoto-type digital pen, paper that has a special dot pattern embedded within it is used, and the writing

^{*} Kyushu Institute of Technology, Fukuoka, Japan

is digitized by reading this pattern with a camera installed in the pen's tip. Therefore, as in ordinary classes, writing data can be collected by a pen and paper. We believe that we can support learner reviews using those data.

We consider that most learners tend to focus on and review only problems they incorrectly answered after exercises. However, in exams such as the National Center Test for University Admissions and the Synthetic Personality Inventory (SPI)¹, many problems could be solved if there was sufficient time, but the time allowed for these tests is short and limited. In other words, we consider that it is necessary to review problems that take time to answer even though the learners answered them correctly to deepen their understanding. It is difficult to judge which problems took longer to answer with an ordinary pen and paper. Even if the learner could understand which problem he/she was slow to solve, it is difficult to determine which part of a problem took long during the solution process. Therefore, we adopt a digital pen to help the learner to recognize the part of the problem that was difficult to solve. By collecting writing data with a digital pen, the learner can focus on a part of the solution process that takes long.

2 Purpose of the study

Using the conventional pen and paper, a learner can check whether the answer is correct or not. Then, the learner can review the problems with incorrect answers as one set of weak points. However, it is difficult to determine a problem that took long, even if the learner answered it correctly, and it is not easy to find out where the learner spent the most time in the solution process. Therefore, by acquiring writing data using a digital pen, we can discover a learner's weaknesses, not only for the problem but also with respect to parts of the problem. Focusing on a part of the problem that a student was slow to solve will make conducting more effective learning possible. In this paper, we evaluate whether a learner's weak points could be detected by using pen-stroke interval data.

3 Related works

Asai et al. estimated the stumbling state of a learner using the proportion of each writing mode (air stroke, writing stroke, and erasing stroke) as a feature value [5]. The feature value is calculated using the online handwritten data over an arbitrary time window. As a result, it was found that by setting the length of the time window appropriately, stumbling could be detected with precision and a recall rate exceeding 80%. In a study to estimate the state of the learner from the writing data, Asai et al. also examined the relationship between a student's pen activity and his/her state of frustration while solving handwritten problems [6]. The numbers of writing and erasing strokes, the ratio of time during which the pen is moving at more than 5 pixels/sec, the mean values of the pen's pressure factor, writing stroke speed, and air-stroke speed were utilized as features. They could detect a student's frustration with 87% precision and a recall rate of 90%. Yu et al. classified mental workload levels by examining online writing features of curvature and velocity [7]. Their classification accuracy ranged from 43% to 55%. In our research, we use the times of pen strokes as a detection method.

Nakatsuka et al. proposed a method of detecting stagnant problem-solving periods using time intervals of pen strokes [8]. To obtain the necessary criteria, multiple criteria

¹The SPI is a test that measures personality and ability.

were compared using DP matching. In our research, we do not compare with other people but detect within individuals.

Deguchi et al. examined a review-problem extraction method and investigated whether it is important to use answer time in a review support system [9]. Using results for SPI, they confirmed that it is possible to use the answer time to extract review questions because the solution time is correlated to solution confidence and problem difficulty. By contrast, we focus on the process rather than on the whole question by considering the weak points of a learner to be the part of the problem that took long to solve.

4 Method

To detect learners' weaknesses, we use the time between pen strokes during an exercise using a digital pen. The general purpose of the exercise is for the learner to self-check whether he/she can solve problems within an appropriate length of time and to find his/her weak points. There are two types of exercises: pseudo-examination and practice. Pseudo-examination is similar to a real performance test and is conducted within a given time limit. Practice is a kind of rehearsal, and no time limit is given. In this research, we assume that the learner treats the practice as an exercise. After the practice, the learner determines his/her weak points, and strengthens them by repetition.

We explain the definition of pen-stroke intervals used in this research. "Pen-down" is the state when the pen is in contact with a paper, and "Pen-up" is the state when it does not contact any paper. The stroke interval is defined as the duration from the pen-up of the previous stroke to the pen-down of the next stroke. However, we consider that the duration of writing should be included for further estimation. Therefore, we define our stroke interval as the duration from the start time of the pen-down of the previous stroke to the start time of the pen-down of the next stroke.

To detect learner weaknesses, we need to define an appropriate threshold for the obtained time intervals during a learner's exercise. Generally, to solve all the problems of exercise within a given time limit, the learner should master his/her weak-point problems. If the learner could solve the weak-point problems in a shorter time, the total solution time would be reduced. We believe that time-saving is important for better scores and understanding. Therefore, we calculate the expected time saved using the total solution time and the given time limit. We determine a threshold time so as to collect time saved from longer time durations until the total time saved becomes equal to the expected time saved. In our method, we define a weak-point stroke to be one whose interval is greater than the threshold time.

One of the advantages of using the intervals of pen strokes is that we can focus on the details of the solution process. Given the solution time of a whole problem, we can understand that the whole problem is a weak point, but it is difficult to know which part of the problem took a long time to complete. In addition, the concept of time is easy for students to understand, and it was thought that it would be easy for learners to accept it when implemented as a system.

5 Implementation

In this section, we describe the implementation of the method. We have developed a system that loads handwritten note data of digital pens and analyzes them to extract weaknesses.

To enable the learner to recognize the weak points, the system highlights the weak-point strokes in blue, and the strokes written on the same line after a weak-point stroke are colored in red. The related strokes are extracted using the height of strokes written after the weak-point stroke. Furthermore, the thickness of the colored strokes is doubled.

The threshold time is calculated using the method described in Section 4. This value is one of the indices learners can use to recognize the weak points they should fix.

5.1 Additional functionality

Basically, the system highlights strokes using the calculated threshold. However, to improve the effectiveness of the system, we implemented a slider to change the threshold value (Figure 1). When the threshold is small and the number of weak points is large, by gradually decreasing the value of the threshold using the slider, the learner can check all weak points in order from the most time-consuming places. In addition, if the number of weak points is small, the learner can deepen his/her understanding of the problem by finding more weak points by decreasing the value of the slider. The slider value can be changed in units of 1 sec.

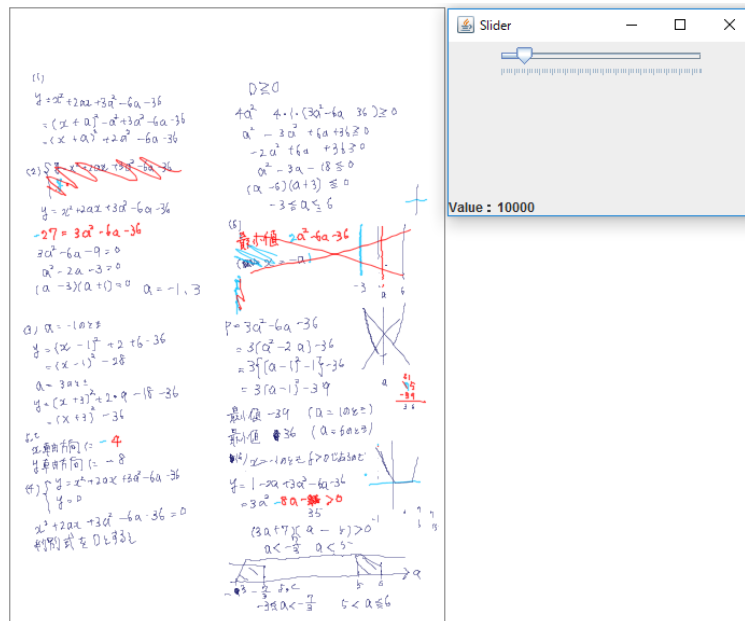


Figure 1: Weak point coloring and slider (blue: strokes detected as weak point, red: lines containing a stroke detected as a weak point)

6 Experiment

We asked nine students of the Kyushu Institute of Technology to solve one major problem in a mathematics test of the National Center Test, and gathered written data during that time. The test participants were denoted by letters A to I. In this experiment, we defined two time limits: one was an expected solution time of 15 min, and the other was a hard limit time of 25 min. In order to evaluate the system, not only the system but learners and

evaluators also extracted weak points. In our research, we focus on comparing the weak points extracted by the system and the evaluators to determine how well the system can find weak points. In addition, we collected learners' self extracted weak points. The weak points were also compared with the weak points extracted by the system for considering how the system supports learners' self extracted weak points.

6.1 Weak points extracted by the system

A threshold value was set for each writing stroke, and a stroke with a time difference larger than the threshold value was considered to be a weak point. We set the expected solution time to 15 min, but we could not set the threshold for participants A and F because they solved the problem in less time. Therefore, we removed their data from the following analysis. Moreover, the major problem consisted of six small questions, but participants D and G could not solve the sixth question. Therefore, the threshold was calculated using the expected solution time of 12 min (Figure 2).

6.2 Weak points extracted by evaluators

The evaluators extracted weak points as a third party, that is, a teacher side. We asked three other students of the Kyushu Institute of Technology to be evaluators. While collecting the writing data, we recorded the learners' videos using two cameras (Figure 3). The evaluators were asked to watch the videos and extract the learners' weak points. The evaluators were familiar with the answer because they had solved the problem before. The evaluators watched a video and recorded both the times a learner seemed to become confused and the time the learner left the confused state. After that, we analyzed the times and used those that at least two evaluators identified as confused states. Strokes written during this time were treated as weak-point strokes extracted by the evaluator. Figure 4 shows an example of the weak-point strokes extracted by the evaluators.

6.3 Weak points extracted by learners

To ascertain whether the system can extract weak points that learners can not find by themselves, we examined how many weak points can be extracted by the learners. After a learner solved the problem, we asked him/her where he/she had weak points. The learners then marked the weak points they had noticed on the paper (Figure 5). In addition, we asked them about their impression of the system using a questionnaire (Figure 6).

6.4 Evaluation

We evaluated the system by comparing the weak points extracted by the system with those extracted by the evaluators. For each of the seven participants' writing data, we calculated precision and recall rates, and obtained F values. The formula for obtaining each value is as follows:

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F = \frac{2*Precision*Recall}{Precision+Recall}$$

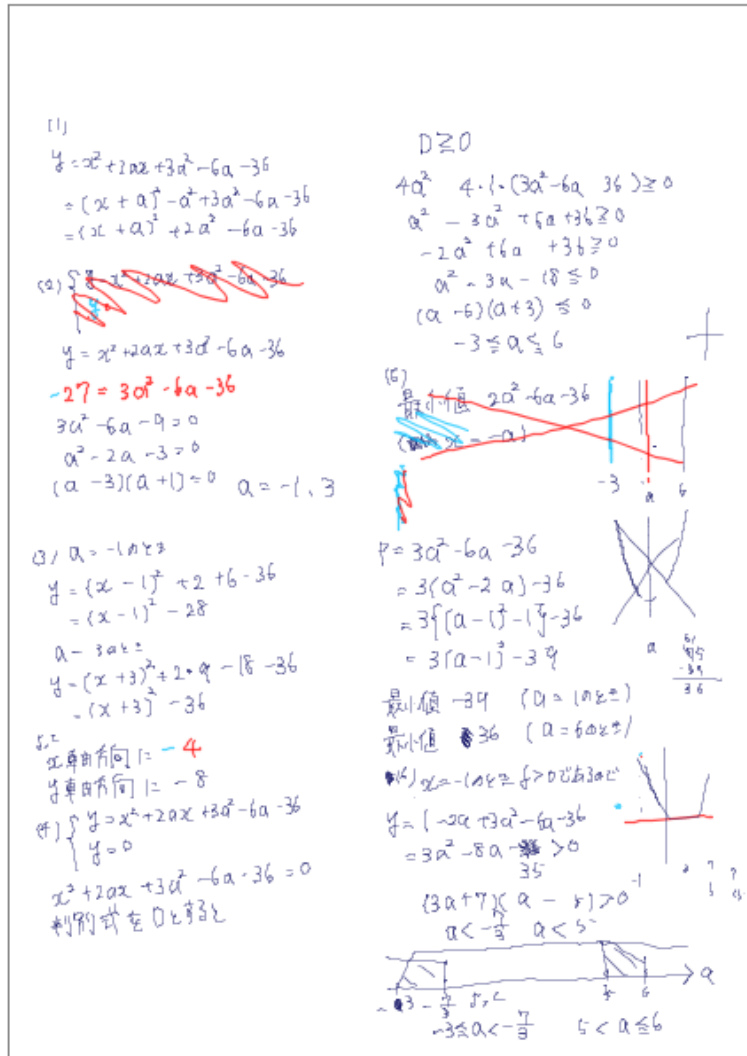


Figure 2: Weak points extracted by the system

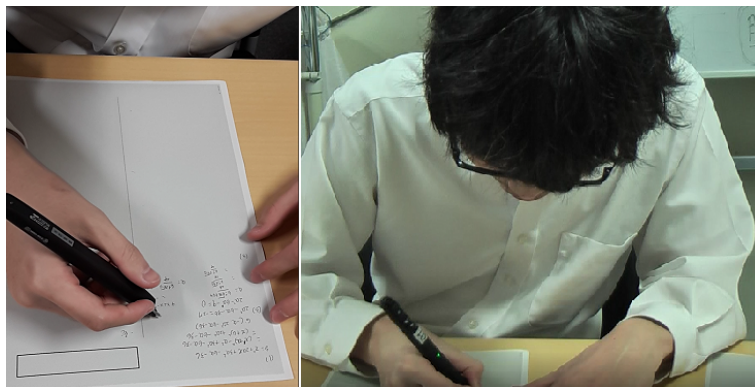


Figure 3: A learner's hand (left) and the learner's state (right)

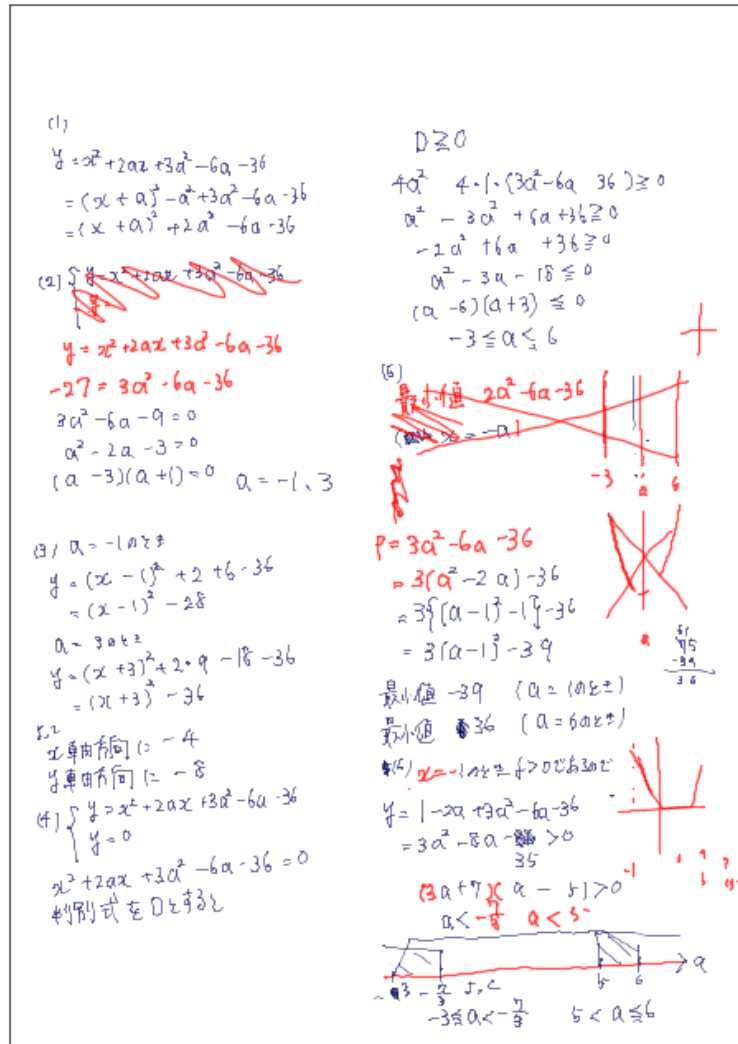


Figure 4: Weak points extracted by evaluators

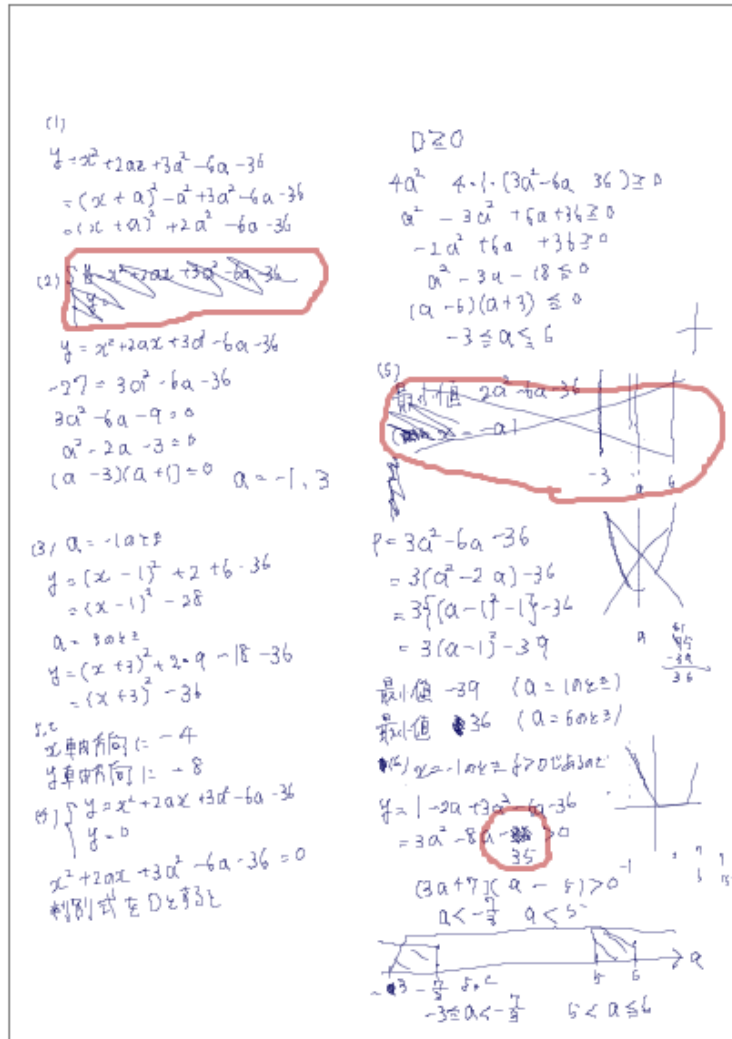


Figure 5: Weak points extracted by the learner

- Questionnaire (For 5 grades, 5 is very good, 1 is very bad.)
- 1, Please tell me the impression of solving the problem. (Free description)
 - 2, How was the comfort of using the digital pen? (1 to 5)
 - 3, Have you found weak points in the system? (1 to 5)
 - 4, Please tell me the threshold value that makes the weak point easier to see or review.
 - 5, Do you really want to use the system? (1 to 5)
 - 6, Please tell me the impression of the system. (Free description)

Figure 6: Content of the questionnaire

Here, TP is the number of parts for which both the system and evaluators extracted weak points, FP is the number of parts that the evaluators did not extract but were extracted by the system, and FN is the number of parts that the evaluators extracted but the system did not extract. Basically, a part here is one line. If two or more strokes were colored blue by the system in one line, we divided the weak-point line into two parts. In addition, with regard to detailed parts such as graphs and small written calculations, we determined the parts by watching the learners' videos.

7 Results

A graph of the pen stroke intervals of collected writing data is shown in Figure 7. The results are compared in the following sections.

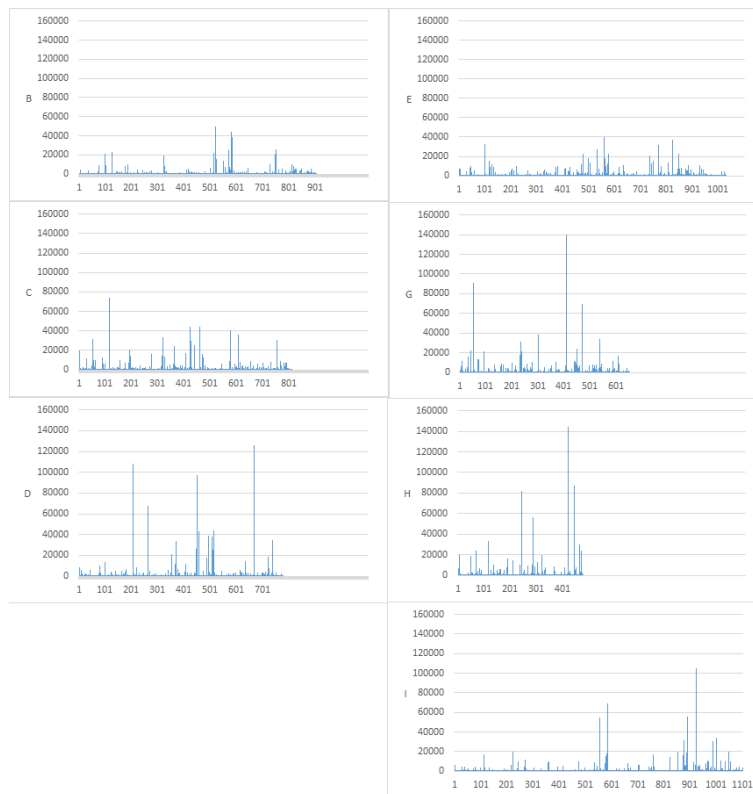


Figure 7: Pen-stroke intervals (The vertical axis shows the interval (msec) of the pen stroke, and the horizontal axis shows the pen stroke number)

7.1 Comparison of the weak points extracted by the system and evaluators

The weak points extracted by the system and by the evaluators are compared in Table 1. In addition, precision, recall, and F value are shown in Figures 8 and 9. Because the relationship between precision and recall is a trade-off, the F value is used for evaluation. Table 1 shows that the system can detect weak points with an accuracy of about 50% to 60%.

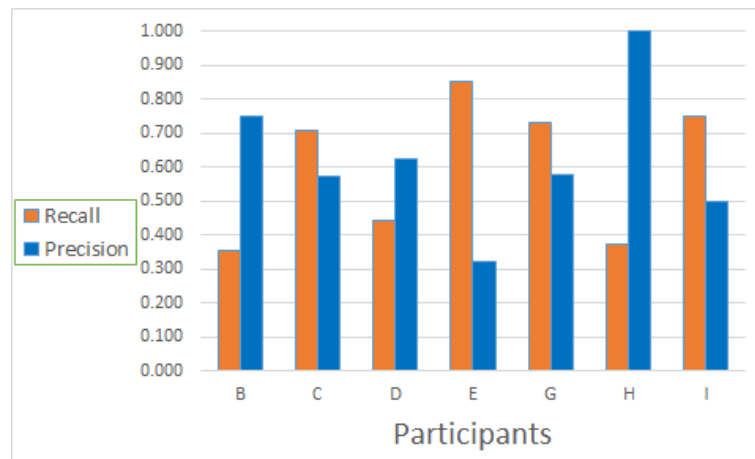


Figure 8: Precision and Recall

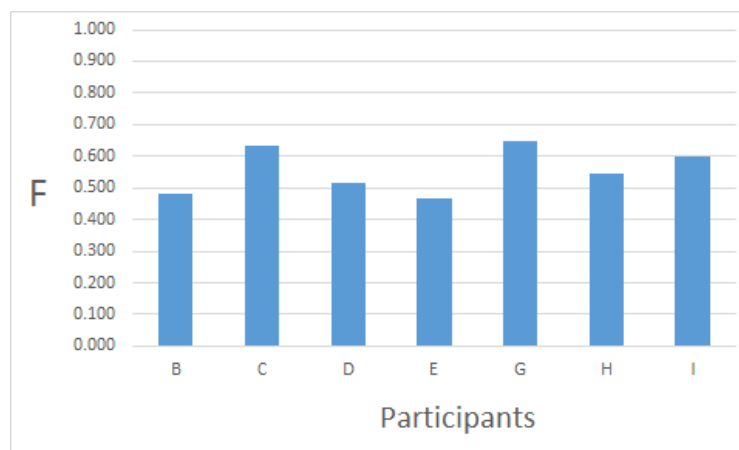


Figure 9: F values

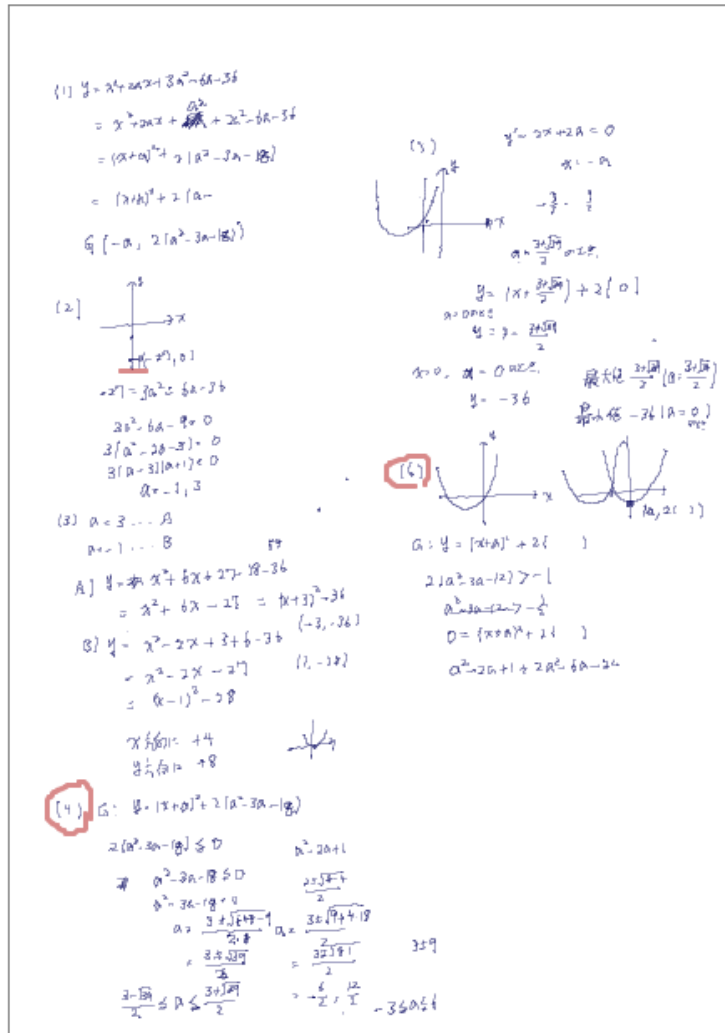


Figure 10: Weak points extracted by the learner (Participant E)

Table 1: Comparison of the weak points extracted by the system and those extracted by the evaluators

Participants	B	C	D	E	G	H	I
TP	6	24	15	29	22	3	15
FP	2	18	9	61	16	0	15
FN	11	10	19	5	8	5	5
Precision	0.75	0.57	0.63	0.32	0.58	1.00	0.50
Recall	0.35	0.71	0.44	0.85	0.73	0.38	0.75
F	0.48	0.63	0.52	0.47	0.65	0.55	0.60
Threshold(sec)	16	6	6	3	6	62	9

7.2 Comparison with learner weak points

One of the weak points extracted by a learner is shown in Figure 10. If the learner marked the question number, we considered all the answer strokes of the question to be weak points. Most learners marked the whole problem, not a part of the problem. In addition, the learner often recognized and marked weak-point strokes when he/she found mistakes during problem solving. Fewer weak points were extracted by the learner than by the other two means (the system and evaluators).

Furthermore, the result of the questionnaire is shown in Table 2.

Table 2: Questionnaire responses

Participants	B	C	D	E	G	H	I
(2)Pen's comfort	4	3	4	5	4	3	3
(3)Weak points	4	5	4	5	5	4	3
(4)Threshold(sec)	20	20	25	12	24	30	6
(5)Evaluation	3	5	4	5	4	4	4

8 Discussion

8.1 Threshold setting

8.1.1 Comparison of weak point detection by proposed method and manual setting

Table 3 shows the results when weak points are detected by the learner's self-determined threshold. The self-determined threshold is selected by the learner with moving the slider after the experiment. We call the self-determined threshold as manual threshold.

Comparing the detection result with the threshold by our proposed method and the detection result with the manual threshold, our proposed method can find weakness accurately for C, D, G. For B and I, the accuracy is slightly falling but we think it is not bad, since it will be set automatically and an index to finish the answer within the time. However, the accuracy differs greatly for E and H.

Figure 11 shows the standard deviation of the time interval of the pen stroke and the expected time saved, which are used to set the threshold. Comparing the results of participants B, E, and G, we found that when the standard deviation was small and the expected

Table 3: Detection with manual threshold

Participants	B	C	D	E	G	H	I
TP	6	10	10	10	7	6	20
FP	0	2	2	7	0	0	23
FN	12	16	21	12	13	1	1
Precision	0.33	0.39	0.32	0.46	0.35	0.86	0.95
Recall	1.00	0.83	0.83	0.59	1.00	1.00	0.47
F (manual setting)	0.50	0.53	0.47	0.51	0.52	0.92	0.63
F (proposed method)	0.48	0.63	0.52	0.47	0.65	0.55	0.60
Manual Threshold(sec)	20	20	25	12	24	30	6

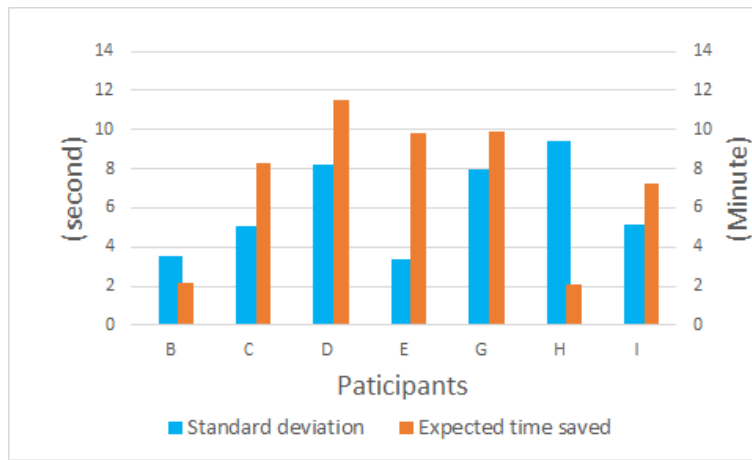


Figure 11: The relationship between standard deviation of pen-stroke intervals and the difference between answer times and target answer times

time saved was long, the threshold decreased. In that case, because the number of strokes extracted as weak points was large, it was difficult for the learner to determine the true weak point. In addition, comparing the results of participants H and B, we found that the threshold increased when the standard deviation was large and the expected time saved was short. In that case, because few strokes were extracted as weak points, the learner could not find his/her true weak point. In these cases, we conclude that setting the threshold value by the proposed method is not suitable.

8.1.2 Alternative threshold setting method

We consider a new method of setting a threshold value. According to the concept of the proposed method so far, if the expected time saved is the same, the threshold becomes higher as the standard deviation is higher, and the threshold becomes higher as the expected time saved is smaller when the standard deviation is the same. Therefore, in order to prevent the threshold from becoming too large or too small, a new threshold setting is derived by the following expression.

$$NewThreshold = Threshold * \frac{ExpectedTimeSaved(s)}{StandardDeviation(s)*60}$$

This formula reduces the threshold value when the standard deviation is large, and decreases the threshold value when the scraping time is small. These data are shown in Table 4. The result of detecting the weak point with the new threshold is shown in Table 5. We conducted a t-test at the effective level of 5% on the F value according to the original proposed method and the F value by the new method, and it was found that there was a significant difference at $p = 0.038$. From this result, we found that weak points can be accurately detected by the new threshold setting method than the original proposed method. In addition, we conducted a t-test at an effective level of 5% for F value by a new method and F value by manual setting, and no significant difference was observed at $p = 0.266$. However, we consider that the new method is more useful than the manual setting since the F value of 6 out of 7 participants is high.

Table 4: Setting new threshold

Participants	B	C	D	E	G	H	I
Expected Time Saved(min)	2.17	8.25	11.5	9.83	9.95	2.12	7.22
Standard Deviation(sec)	3.50	5.06	8.24	3.38	7.94	9.43	5.18
Time Saved / Deviation	0.62	1.63	1.40	2.91	1.25	0.22	1.39
Threshold(sec)	16	6	6	3	6	62	9
New Threshold(sec)	10	10	8	9	8	14	13

Table 5: Detection with new threshold

Participants	B	C	D	E	G	H	I
TP	10	17	13	15	21	7	12
FP	3	7	8	15	12	6	6
FN	9	12	20	10	9	0	8
Precision	0.53	0.59	0.39	0.60	0.70	1.00	0.60
Recall	0.77	0.71	0.62	0.50	0.64	0.54	0.68
F	0.63	0.64	0.48	0.55	0.67	0.70	0.63
New Threshold(sec)	10	10	8	9	8	14	13

8.1.3 When making the threshold constant

We have set thresholds for each learner and detected weak points. The threshold difference between learners decreases according to the new threshold setting method. Therefore we verified the effectiveness of the individual thresholds rather than an average threshold. The average threshold was 10.3sec. Weak point of each learner was detected by the system, and the result is as shown in Table 6. We conducted a t-test at an effective level of 5% on the F value by the new method and the F value by the average threshold, and it was found that there was a significant difference at $p = 0.039$. From this result, we found that the detection performance improves by the individual thresholds for learners.

Table 6: Detection with average threshold(10.3sec)

Participants	B	C	D	E	G	H	I
TP	10	16	12	9	15	7	13
FP	2	6	6	10	9	9	9
FN	9	12	20	14	12	0	8
Precision	0.53	0.57	0.38	0.39	0.56	1.00	0.62
Recall	0.83	0.73	0.67	0.48	0.63	0.44	0.59
F	0.65	0.64	0.48	0.43	0.59	0.61	0.61

8.2 System usefulness

According to the results of the questionnaire, many learners were able to find weak points using the system, and their opinion that they wanted to use the system was strong. In addition, the threshold set by the learner was larger than the one set by our method in many cases. Regarding the comparison of weak points extracted by the learner (Figure 5) with those extracted by the system (Figure 2), because the system extracts many weak points, the system can identify more learner weak points. Furthermore, learners who marked a question as a weak point, because they could find the part of the question that took long to solve, could see the details of their weak points. In the free-description section of the questionnaire, we could obtain the following positive comments:

- “I think the coloring of weak points is good.”
- “The weak point detection based on exact time duration, not personal feelings, is useful.”
- “Threshold time setting is effective for finding weak points.”

In addition, we obtained the following negative feedback from the participants:

- “Time needed to read the questions should be removed from the stroke intervals.”
- “Wrong answers and mistakes that were written quickly could not be found correctly..”
- “Please indicate whether the answer is correct or not.”
- “Mental arithmetic tends to be detected as a weak point. I think this is not fair.”
- “I would have to prepare a PC to use the system.”
- “It is troublesome to move the slider by myself.”

We consider that these negative points could be improved for further development.

8.3 Comparison with related research

We were able to detect weak points with accuracy of about 50% to 60%. The detection target may be slightly different, but the detection rate is equivalent to that of similar researches[5][8]. It is easy for the learner to change the threshold value because detecting

factor is only pen-stroke interval. Furthermore, we found that it is possible to detect weak points by using individual time intervals. This merit is effective when the matching with other's time intervals failed.

9 Conclusion

In this paper, we implemented a method to detect learners' weak points using pen-stroke intervals and evaluated them. Comparing weak points extracted by the system with those extracted by the evaluator, we obtained precision, recall, and F values. As a result, it is possible to detect weak points with an accuracy of about 50% to 60%. The questionnaire results show that learners can find their weak points using the system. Moreover, they evaluated the system positively. In addition, more weak points were extracted by the system than by a learner. Thus, the learner can find many weak points by using the system.

The experimental results show that our system obtained 60% accuracy. However, we believe that this accuracy is not enough for detecting weak points. To further improve the detection accuracy, we should reconsider the threshold-setting method. For example, instead of setting the threshold value using the total solution time, we should set the threshold value for each problem. In addition, we considered the detection of weak points. It may be possible to obtain higher accuracy by adding other factors. For example, the meaning of sentences, which can be obtained from handwritten data such as the relationship of strokes erased by double lines and the physical or psychological status of the learner could be utilized to improve the system. Furthermore, it is necessary to verify what kind of educational effect the learner has by the iterative use of the system.

Acknowledgment

The part of this research was supported by the fund of JSPS KAKENHI Grant-in-Aid for Scientific Research (C): Grant Number 15K00485.

References

- [1] Ministry of Education, Culture, Sports, Science and Technology. The Vision for ICT in Education - Toward the Creation of a Learning System and Schools Suitable for the 21st Century -. http://www.mext.go.jp/b_menu/houdou/23/04/_icsFiles/afieldfile/2012/08/03/1305484_14_1.pdf, April 2011.
- [2] Roundtable on informatization of education for the 2020s. "Roundtable on informatization of education for the 2020s" final summary. http://www.mext.go.jp/b_menu/houdou/28/07/_icsFiles/afieldfile/2016/07/29/1375100_01_1_1.pdf, 2016.(in Japanese).
- [3] Motoki Miura, Taro Sugihara, and Susumu Kunifuji. Improvement of digital pen learning system for daily use in classrooms. In *Educational Technology Research*, Vol. 34, pp. 49-57, 2011.
- [4] Seiichi Uchida, Marcus Liwicki, Masakazu Iwamura, Koichi Kise, and Shinichiro Omachi. Digital pen. In *Journal of the Institute of Image Information and Television Engineers*, Vol.64, No.3, pp. 293-298, 2010.(in Japanese).

- [5] Hiroki Asai, Akari Nozawa, Shogo Sonoda, and Hayato Yamana. Student's stumble detection using online handwriting data. In DEIM Forum A8-4, 2012.(in Japanese).
- [6] Hiroki Asai and Hayato Yamana. Detecting student frustration based on handwriting behavior. UIST' 13.
- [7] Kun Yu, Julien Epps, and Fang Chen. Mental workload classification via online writing features. In 12th International Conference on Document Analysis and Recognition, vol. 00, pp. 1110-1114, 2013.
- [8] Chihiro Nakatsuka, Yoshitaka Morimura, and Atsushi Hashimoto. Detecting answer stagnation point using time intervals of pen strokes. In JSiSE research report 30(7), pp. 71-74, March 2016.(in Japanese).
- [9] Hiroyuki Deguchi, Yuya Onishi, Tetsuya Ohya, Hironori Koyama, and Masashi Kawasumi. Review support system which selects the questions by using answering time. In FIT2010, 2010.(in Japanese).