Fri. Jun 7, 2019

Room D

International Session | International Session | [ES] E-2 Machine learning

[4D3-E-2] Machine learning: living environment

Chair: Junichiro Mori (The University of Tokyo)
2:00 PM - 3:40 PM Room D (301B Medium meeting room)

[4D3-E-2-01] Privacy-Preserving Resident Monitoring System with Ultra Low-Resolution Imaging and the Examination of Its Ease of Installation
〇Takumi Kimura[1], Shogo Murakami[1], Ikuko Egushi Yairi[1] （1. Sophia University）
2:00 PM - 2:20 PM

[4D3-E-2-02] Trees Detection on Google Street View Images Using Deep Learning and City Open Data
〇Lieu-Hen Chen[1], Hao-Ming Hung[1], Cheng-Yu Sun[1], Eric Hsiao-Kuang Wu[2], Toru Yamaguchi[3], Eri Sato-Shimokawara[3], Hao Chen[1] （1. Nantional Chi Nan University, 2. National Central University, 3. Tokyo Metropolitan University）
2:20 PM - 2:40 PM

[4D3-E-2-03] Scoring and Classifying Regions via Multimodal Transportation Networks
〇Aaron Bramson[1,2,3,4], Megumi Hori[1], Zha Bingran[1], Hirohisa Inamoto[1] （1. GA Technologies, 2. RIKEN Center for Biosystems Dynamics Research, 3. Ghent Univeristy, 4. UNC - Charlotte）
2:40 PM - 3:00 PM

[4D3-E-2-04] Evaluating Road Surface Condition by using Wheelchair Driving Data and Positional Information based Weakly Supervision
〇Takumi Watanabe[1], Hiroki Takahashi[1], Yusuke Iwasawa[2], Yutaka Matsuo[2], Ikuko Eguchi Yairi[1] （1. Sophia Univ., 2. Univ. of Tokyo）
3:00 PM - 3:20 PM

[4D3-E-2-05] Prediction of the Onset of Lifestyle-related Diseases Using Regular Health Checkup Data
〇Mitsuru Tsunekawa[1], Natsuki Oka[1], Masahiro Araki[1], Motoshi Shintani[2], Masataka Yoshikawa[3], Takashi Tanigawa[4] （1. Kyoto Institute of Technology, 2. SG Holdings Group Health Insurance Association, 3. Japan System Techniques Co.,Ltd., 4. Juntendo University）
3:20 PM - 3:40 PM

Room H

International Session | International Session | [ES] E-5 Human interface, education aid

[4H2-E-5] Human interface, education aid: human evaluation

Chair: Naohiro Matsumura (Osaka University)
12:00 PM - 1:40 PM Room H (303+304 Small meeting rooms)

[4H2-E-5-01] Recognition of Kuzushi-ji with Deep Learning Method
〇Xiaoran Hu[1], Mariko Inamoto[2], Akihiko Konagaya[1] （1. Tokyo Institute of Technology, 2. Keisen University）
12:00 PM - 12:20 PM

[4H2-E-5-02] Computerized Adaptive Testing Method using Integer Programming to Minimize Item Exposure
〇Yoshimitsu MIYAZAWA[1], Maomi UENO[2] （1. The National Center for University Entrance Examinations, 2. The University of Electro-Communications）
12:20 PM - 12:40 PM

[4H2-E-5-03] Maximizing accuracy of group peer assessment using item response theory and integer programming
〇Masaki Uto[1], Duc-Thien Nguyen[1], Maomi Ueno[1] （1. University of Electro-Communications）
12:40 PM - 1:00 PM

[4H2-E-5-04] Waveform Processing of Electrocardiogram with Neural Network and Non-contact Measurement using Kinect for Driver Evaluation
〇HAO ZHANG[1], Takashi IMAMURA[1] （1. Niigata University）
1:00 PM - 1:20 PM

[4H2-E-5-05] Probability based scaffolding system using Deep Learning
〇Ryo Kinoshita[1], Maomi Ueno[1] （1. The University of Electro-Communications.）
1:20 PM - 1:40 PM

©The Japanese Society for Artificial Intelligence

International Session | International Session | [ES] E-2 Machine learning

[4D3-E-2] Machine learning: living environment

Chair: Junichiro Mori (The University of Tokyo)

Fri. Jun 7, 2019 2:00 PM - 3:40 PM Room D (301B Medium meeting room)

[4D3-E-2-01] Privacy-Preserving Resident Monitoring System with Ultra Low-Resolution Imaging and the Examination of Its Ease of Installation

○Takumi Kimura[1], Shogo Murakami[1], Ikuko Egushi Yairi[1] （1. Sophia University）

2:00 PM - 2:20 PM

[4D3-E-2-02] Trees Detection on Google Street View Images Using Deep Learning and City Open Data

○Lieu-Hen Chen[1], Hao-Ming Hung[1], Cheng-Yu Sun[1], Eric Hsiao-Kuang Wu[2], Toru Yamaguchi[3], Eri Sato-Shimokawara[3], Hao Chen[1] （1. Nantional Chi Nan University, 2. National Central University, 3. Tokyo Metropolitan University）

2:20 PM - 2:40 PM

[4D3-E-2-03] Scoring and Classifying Regions via Multimodal Transportation Networks

○Aaron Bramson[1,2,3,4], Megumi Hori[1], Zha Bingran[1], Hirohisa Inamoto[1] （1. GA Technologies, 2. RIKEN Center for Biosystems Dynamics Research, 3. Ghent Univeristy, 4. UNC - Charlotte）

2:40 PM - 3:00 PM

[4D3-E-2-04] Evaluating Road Surface Condition by using Wheelchair Driving Data and Positional Information based Weakly Supervision

○Takumi Watanabe[1], Hiroki Takahashi[1], Yusuke Iwasawa[2], Yutaka Matsuo[2], Ikuko Eguchi Yairi[1] （1. Sophia Univ., 2. Univ. of Tokyo）

3:00 PM - 3:20 PM

[4D3-E-2-05] Prediction of the Onset of Lifestyle-related Diseases Using Regular Health Checkup Data

○Mitsuru Tsunekawa[1], Natsuki Oka[1], Masahiro Araki[1], Motoshi Shintani[2], Masataka Yoshikawa[3], Takashi Tanigawa[4] （1. Kyoto Institute of Technology, 2. SG Holdings Group Health Insurance Association, 3. Japan System Techniques Co.,Ltd., 4. Juntendo University）

3:20 PM - 3:40 PM

©The Japanese Society for Artificial Intelligence

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Privacy-Preserving Resident Monitoring System with Ultra Low-Resolution Imaging and the Examination of Its Ease of Installation

Takumi Kimura [*1] Shogo Murakami [*1] Ikuko Egushi Yairi [*1]

[*1] Graduate School of Science and Technology, Sophia University

Monitoring systems using infrared array sensors allow monitoring of residents while protecting their privacy. However, since such a sensor is vulnerable to subtle movements, accuracy of posture classification is low, and limits the locations and methods available for installation. This study proposes a posture classification method with higher accuracy. Over 93% accuracy was achieved in posture classification by RGB conversion of infrared array sensor images and successfully decreased loss due to displacement by DCNN. Additionally, this research considers methods to create artificially simulated data for postural-behavioral study. To check the validity of this method, postures of 3 subjects were examined using a classifier with studied simulation data. Finally, simulation environments with different sensor altitudes and angles were created to examine the ease of installation for the proposed method. As a result, the experiments showed that accuracy was highest at approximately 90% when the sensor was located 50cm below the height of the target and when the tilt angle was within ±2°.

1. Introduction

Resident monitoring systems are useful in detecting abnormal conditions of residents. However, as every move is under inspection, privacy issues arise. As a solution, usage of infrared array sensors has been proposed to preserve privacy as well as to avoid physical burden on the target. [Okada 13][高木 16][楠亀 17]. Such sensors can be placed in various places as they solely rely on temperature data obtained from the infrared sensor to detect the target. Spatial information and light measurements received from the sensors are used to identify the posture and location of the target and to observe their changes.

It is known that sensor installation angle is a factor for decreased classification accuracy, but the analysis on its effects are yet insufficient. Acquisition of learning data for machine learning is key in improving classification accuracy. To solve the aforementioned tasks, this paper assesses the classification accuracy by the single 8x8 infrared array sensor, proposes the methods of artificially creating simulated data to study postures for machine learning, and analyzes the effects caused by the angle of the installed infrared array sensor.

2. A resident monitoring system using ultra low-resolution infrared array sensor imaging

2.1 Posture classification system

To examine posture classification accuracy, a data collection device including an infrared array sensor was developed. A diagram of the device is shown in Figure 2. This device composes of a Raspberry Pi3 Model B mounted with a Grid-EYE (AMG8833) sensor. The Grid-EYE will output an 8x8 pixel image data of the surface temperature for objects detected in the observation space. Temperatures between 0°C -80°C can be detected with a step increment of 0.25°C.

Deep Convolution Neural Network (DCNN), is an effective method for high accuracy image recognition. It is used in this study to examine the obtained infrared image data for use in the production of a posture classifier. Figure 3 shows the structure of the DCNN used.

2.2 Posture classification experiment of a subject using DCNN

A posture classification experiment was conducted on 3 subjects to check the operation of the posture classification system and to evaluate the performance of the DCNN. The first two experiments were conducted in a 9.5m^2 Japanese-style room at roughly 13°C room temperature. Here, the data of a male subject of age 24 and height of 170cm, and a female subject of age 20 and a height of 160cm was obtained. The experiment for the third subject was done in a 20m^2 room with a room temperature of roughly 11°C. The subject was male, of age 22, and was 170cm in height. For all three scenarios, the sensor was placed 140cm from ground level such that the entire body of each subject could be observed.

The subjects were stationed 1-3m away from the sensor and were told to stand, sit, or lie down within the view of the sensor. A total of 14983 frames worth of data were obtained. 10% of the above data were randomly selected and studied by the classifier, then were used to classify the remaining 90%. The result of posture classification, evaluation of accuracy, and recall ratio are shown in Table 1.

Considering practical use, a resident monitoring system needs to be able to detect instantaneous dangerous incidents such as slips and therefore it is desirable for the F-measure to be above 90%. Experimental results showed an F-measure of roughly 87%. It is probable that directly inputting 8x8 temperature data to the DCNN is insufficient for image feature extraction.

However, since DCNN is known to have high accuracy object detection for colored images, focus was placed on converting data

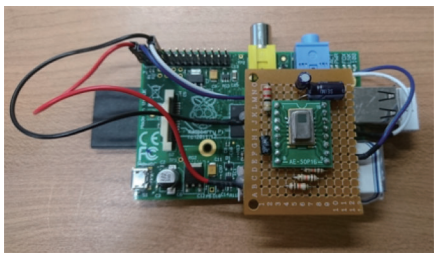


Fig. 1 A posture recognition device equipped with an 8×8 infrared array sensor.

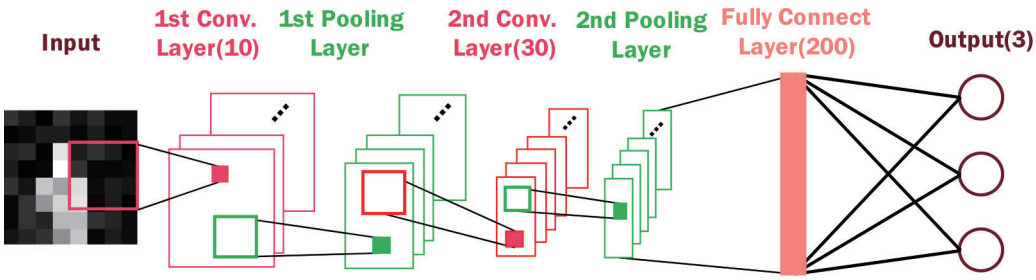


Fig. 2 Illustration of DCNN structure used in 2.2 experiment.

from the infrared array sensor into RGB and inputting them to the DCNN to improve the accuracy of the classifier. [Simonyan 15] This was done by mapping temperature data into a color space and using that as a reference to convert images to RGB. Data were then divided into R, G, and B, and were separately inputted to the DCNN. The classification results after RGB conversion, evaluation of accuracy, and recall ratio are shown in Table 2. In comparison to Table 1, the number of misclassifications decreased, and F-measure was above 90% for all postures. From such results, it could be concluded that RGB conversion of data led to less number of misclassifications. RGB converted data were used for DCNN input hereafter.

3. Learning data generator construction and evaluation of classification accuracy

3.1 Learning data generation procedures

Using the DCNN posture classifying method previously evaluated in 2.2, methods to generate simulation learning data without subject-based experiments were made. The infrared images handled in this research are for a room temperature distribution represented by 8x8 image pixels. Therefore, it can be theorized that placing a human model in the view of the sensor would require less effort while outputting results like those of the subject based experiments. For such a reason, a Unity program including physics engines and functions was used to simulate this environment and was used to produce learning data. This learning data generator was made to arbitrarily set the height and physique of the human model, the tilt and altitude of the sensor, and the size of the room. As for physique, the human model was composed of several body parts including the head, arms, and torso, with each part having its own temperature distribution that could also be arbitrarily changed. For this research, the radiant heat distribution obtained from a real environment experiment was used to set parameters for each body part. As for posture, the human model had 3 types of postures namely, “stand”, “sit” or "lie down”.

After running the learning data generator, ray tracing was performed from the sensor. Whenever the ray hit the human model, the radiant heat information of the human model on the incident spot was recorded. When the ray did not hit the model, the preset radiant heat or temperature of the background was recorded. Simulation by the learning data generator is depicted in Figure 3.

3.2 Learning method evaluation using simulated learning data

DCNN with studied simulation data were used to classify real environment data. The real data used were the same data as those used in 2.2. Simulation data collection was done by placing a human model randomly 1m-3m away from the sensor with three postures— either “stand”, “sit” or "lie down”. Room temperatures

Table 1 Posture classification results of 8x8 infrared array sensor images by DCNN.

Posture classification results

Recognition	Answer		
	Stand	Sit	Lie down
Stand	3582	266	15
Sit	496	5243	435
Lie down	185	590	4171

Posture classification evaluations

	Stand	Sit	Lie down
Accuracy	0.92726	0.84921	0.84331
Recall ratio	0.84025	0.85965	0.90262
F-measure	0.88161	0.85440	0.87196

Table 2 Action classification result of 8x8 RGB converted images of infrared array sensor.

Posture classification results

Recognition	Answer		
	Stand	Sit	Lie down
Stand	3929	137	20
Sit	217	5620	113
Lie down	121	336	4490

Posture classification evaluations

	Stand	Sit	Lie down
Accuracy	0.96158	0.94454	0.90762
Recall ratio	0.92079	0.92237	0.97123
F-measure	0.94074	0.93332	0.93835

and sizes were set according to real data. The human model also corresponded with subjects from the experiment, and the learning data generator was run for 30 minutes outputting 50000 sample data for each simulated subject. Finally, the posture classifier was used to analyze the obtained data. Table 3 shows the posture classification and evaluation results. Results show that average F-measure was relatively low— generally under 80%. From observing misclassified examples, it was hypothesized that such classification accuracy loss occurred due to the existence of high-temperature pixels in the background. Therefore, background elimination was conducted by analyzing the temperature difference in each data, estimating and eliminating background parts, and leaving only the human model in the data. The background elimination process is shown in Figure 4. This elimination method was applied to both simulated learning data and real data used for classification and evaluation. Then, classification was repeated for the second time. Table 4 shows the classification and evaluation results. Results showed that all classification average F-measures increased from under 80% to over 90% accuracy. From such results, it was safe to say that background elimination method was effective when using simulated learning data to classify real data.

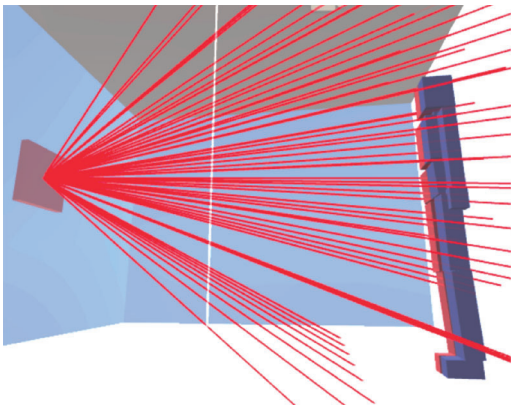


Fig. 3 The learning data generator simulating 8x8 infrared array sensing of a human in a room.

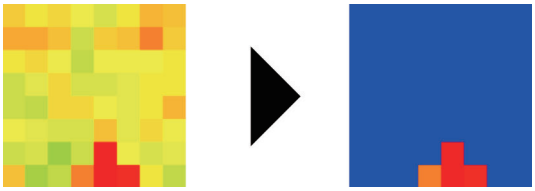


Fig. 4 Process of background elimination algorithm.

4. Evaluation on classification accuracy effects due to the system installation condition

So far, this research had fixed the sensor at a height of 140cm to successfully classify the postures with over 90% accuracy. However, it was never tested to confirm the range of heights this high accuracy rate can be sustained. In implementation, it is highly likely that the sensor will be slightly displaced or tilted from external factors. Therefore, the effect and degree of these factors against classification accuracy were tested, and system installation conditions were considered.

4.1 Evaluation on installation height

Using the learning data generator, the sensor height was changed from 50-170cm with a step increment of 10cm, and the human model height was set at either 170cm, 160cm, or 150cm. Then, for each condition, an evaluation of classification accuracy was performed. The results are shown in Figure 6. For all models, the accuracy peaked when the sensor height was 50cm below the model height.

4.2 Evaluation on sensor tilt

Next, the degree of effect on accuracy by the sensor angle were inspected. The human model height was set at 170cm, and sensor installation height was set at 120cm. Tilt range was set from -5° to 5° with increments of 1°. As for learning data, the preceding data without any tilt was used. Results are summarized in Figure 7. Results show that when the sensor is tilted upwards to 2°, the F-measure is over 90%, but when the tilt reaches 5°, the F-measure is decreased to roughly 86%. When the sensor is tilted down, F-measure is sustained at a high value until -1°. However, a tilt of over -2° drastically decreases the F-measure until under 80% at -

Table 3 Action classification results of the real data by learning with the simulated data.

Posture classification results

Recognition	Answer		
	Stand	Sit	Lie down
Stand	3306	537	216
Sit	959	5129	1274
Lie down	5	427	3130

Posture classification evaluations

	Stand	Sit	Lie down
Accuracy	0.81449	0.69669	0.87872
Recall ratio	0.77424	0.84179	0.67749
F-measure	0.79385	0.76239	0.76509

Table 4 Effects of background image removal on action classification of the real data by learning with the simulated data.

Posture classification results

Recognition	Answer		
	Stand	Sit	Lie down
Stand	4130	0	0
Sit	201	5665	585
Lie down	0	217	4185

Posture classification evaluations

	Stand	Sit	Lie down
Accuracy	1.00000	0.87816	0.95070
Recall ratio	0.95359	0.96311	0.87736
F-measure	0.97624	0.91867	0.91256

5°. Since this research takes privacy preservation as a serious consideration, the number of image pixels used are very low. Therefore, the posture classification relied heavily on high-temperature distributions per row, and a one-row difference gave an extensive impact on classification results.

The high-temperature distribution change caused by the sensor tilt was assumed to be the major cause of the decrease in classification accuracy.

5. Evaluation on classification accuracy effects due to the system installation conditions

In chapter 4, the influence on classification by installation height and sensor tilt were investigated to consider potential external effects upon real implementation. As a result, it was found that installation height yielded highest accuracy at height 50cm below the height of the target, and that sensor tilt largely impacted the classification accuracy.

With these results in mind, real implementation is further considered. Assuming the wall were to be perpendicular, the sensing device could be installed 50cm below the target height after the height of the target is measured. On the other hand, if a wall installation is difficult, then there may be a need to call a specialist to install the device.

A solution to this installation problem could be to create a personalized classifier for the target by inputting the target's room information data into the learning model. Although this method requires meticulous interview on the house conditions, by utilizing the learning data generator one can simulate and obtain data corresponding to the target room and apply the resident monitoring system

6. Conclusions

This paper justified that an infrared array sensor resident monitoring system using an infrared array sensor image with 8x8 pixels would output over 90% accuracy for posture action classification of the target. Noise analysis was performed on a tilt and it was concluded that approximately 90% accuracy was sustained for tilt angle within ±2° by extending the classifier. In addition, posture pattern learning data simulation was taken into consideration, and through comparison against real data, a high accuracy classifier construction was achieved. As for further research, an improvement in simulated learning data, learning data generator, and learning algorithm will be continued for application in a real environment.

Acknowledgements

We would like to show our gratitude to Yuki Kato, Motoharu Sakurai, and all the supporters for their participated and assistance. This research was conducted under the support of 23rd and 24th year of Heisei period research grant from the Support Center for Advanced Telecommunications Technology Research, Grant-in-Aid for Scientific Research B(17H01946), and Grant-in-Aid for challenging Exploratory Research (16K12537).

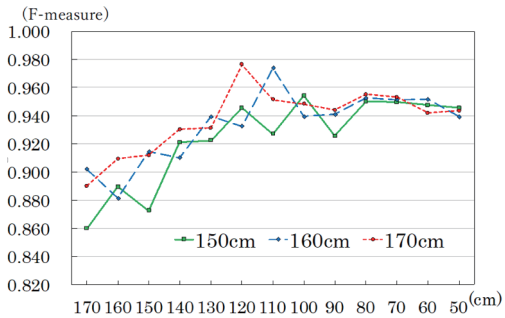


Fig. 5 Accuracy for various sensor altitudes.

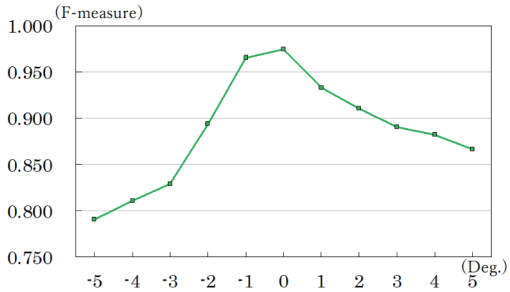


Fig. 6 Accuracy for vertical sensor tilt.

References

[Okada 13] R. Okada and I. Yairi, "An indoor human behavior gathering system toward future support for visually impaired people," Proc. of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, no. 36, Washington, Oct. 2013.

[高木 16] 高木, 高橋, 大塚, "温度センサを用いた高齢者の見守り," 第8回データ工学と情報マネジメントに関するフォーラム (DEIM 2016), P6-1, 福岡, 2016.

[楠亀 17] 楠亀, 米田, 式井, シラワン, 野坂, 久保, "サーモカメラによる非接触温冷感センシング, "Panasonic Technical Journal, pp.118-122, vol.63, no.2, Nov. 2017.

[Simonyan 15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference on Learning Representations (ICLR), May 2015.

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Trees Detection on Google Street View Images Using Deep Learning and City Open Data

Lieu-Hen Chen[*1] Hao-Ming Hung[*1] Cheng-Yu Sun[*1] Eric Hsiao-Kuang Wu[*2] Toru Yamaguchi[*3] Eri Sato-Shimokawara[*3] Hao Chen[*1]

*1 National Chi Nan University, Taiwan *2 National Central University, Taiwan *3 Tokyo Metropolitan University, Japan

For almost every cities and towns, street trees play an important role in representing seasonal change of the street view. Nowadays, lots of countries start promoting open data. Among these data, very useful information related to street trees are well documented with free access by many city governments. At the same time, Google Street View provides the view of a certain surrounding by composing stitched images which are shot by specialized vehicles moving along streets and alleys. However, few research reports have been published on utilizing city open data for trees detection on Google Street View. Therefore, in this study, we aim to perform trees detection on Google Street View Images by utilizing Deep Learning technologies and city open data.

1. Motivation and Research Background

Street Trees are an indispensable component of great neighborhoods. Street trees play an important role for:

1. representing the style of a city, trees provide beauty and aesthetic appeal for the urban landscape.
2. improving environmental quality, especially for reducing air pollution.
3. strengthening urban amenity because trees provide spaces for rest and relaxation.

Therefore, street trees management is always an important issue for city and county governments. It is especially true for cities such as Tokyo and Washington DC where have beautiful cherry blossom seasons.

Nowadays, many countries start promoting open data. The concept of open data is that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. Among the open data provided by some city governments, very useful information related to street trees are well documented with free access. For example, many well documented information of street trees can be accessed form the open database of Taipei and New York city, as shown in figure 1 and figure 2. These useful information includes: tree species, tree height, diameter at breast height, growth status, and position.

Figure 1. Tree's data in Taipei

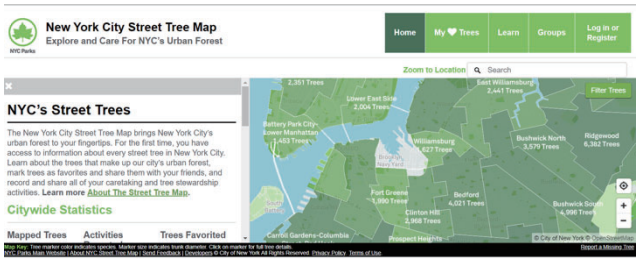


Figure 2. Tree's data in New York

At the same time, Google Map has become the most common tool used for exploring street maps. There are many services available on Google Map. And Google Street View is one of the feature derived from Google Map for improving users' perception of real world. It shows the view of a certain surrounding composed by stitched photographs shot by specialized vehicles moving along streets and alleys. However, these street images are static, no matter which season is now. Moreover, as shown in figure 3, it is still difficult to accurately locate trees on street images by using Google API because of the

Contact:
Lieu-Hen Chen, National Chi Nan University, No. 303, Daxue Rd., Puli Township, Nantou County 545, Taiwan, +886-49-2910-960#4861, +886-49-2915-226, lhchen@csie.ncnu.edu.tw
Hao-Ming Hung, National Chi Nan University, No. 303, Daxue Rd., Puli Township, Nantou County 545, Taiwan, +886-49-2910-960#4861, +886-49-2915-226, d3764291@gmail.com
Cheng-Yu Sun, National Chi Nan University, No. 303, Daxue Rd., Puli Township, Nantou County 545, Taiwan, +886-49-2910-960#4861, +886-49-2915-226, s106321526@mail1.ncnu.edu.tw
Eric Hsiao-Kuang Wu, National Central University, No. 300, Zhongda Rd., Zhongli Dist., Taoyuan City 320, Taiwan, +886-3-4227151, +886-3-4226062, hsiao@csie.ncu.edu.tw
Toru Yamaguchi, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo, Japan 191-0065, +81 42 585 8606, yamachan@tmu.ac.jp
Eri Sato-Shimokawara, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo, Japan 191-0065, +81 42 585 8606, eri@tmu.ac.jp
Hao Chen, National Chi Nan University, No. 303, Daxue Rd., Puli Township, Nantou County 545, Taiwan, +886-49-2910-960#4861, +886-49-2915-226, s107321517@mail1.ncnu.edu.tw

GPS position bias of moving vehicles and the image stitching algorithm.

Figure 3. The shifting problem of tree positions on street picture due to GPS errors

Therefore, in this study, we aim to develop a new approach which detects trees position on Google Street View Images by utilizing Deep Learning technologies and city open data. The remainder of this paper is organized as follows: in section 2, the related works are introduced; in section 3, our prototype system are explained; in section 4, the current experiment results are shown; and finally in section 5, the conclusion are discussed.

2. Related Works

2.1 Trees Detection

In comparison with the number of research works developed for detecting car and people, there are fewer papers published on solving street trees detection problem. In 2006, Wajid Ali proposed a tree detecting method by using the color and texture of trees as the basis for detection in a forest environment [1]. Later, Y. Lu and C. Rasmussen proposed a tree trunk detection algorithm by using contrast templates in 2011 [2]. These studies have good results with limitations of vertical trees on grass fields. In addition, Harri Karrtinen et al. evaluated the quality, accuracy, and feasibility of automatic tree extraction methods, mainly based on laser scanner, which requires high precision and expensive instruments to collect and detect data from trees [3]. It is still very difficult to detect trees images because trees have very complex and various textures and shapes.

2.2 Machine Learning Technologies

Machine learning technologies have been widely applied in solving many problems such as object recognition and route planning, and have achieved many good results in recent years [4]. For example, in1995, C. Cortes and V. Vapnik proposed the Support Vector Machine algorithm which separates the attribute space with a hyperplane to maximize the margin between different classes [5]. Later, in 1997, M. Dorigo, L. Gambardella used the Ant Colony Optimization approach to solve the Traveling Salesman Problem successfully [6].

3. System Implementation

Our system can be divided into the following three stages:

3.1 Adopting Deep Learning approaches for trees image segmentation.

In our prototype system, we trained U-Net and Seg-Net for Deep Learning as shown in figure x. Both U-Net and SegNet are kinds of Convolution Neural Networks. The U-Net was developed for biomedical image segmentation in 2015 [7]. SegNet was developed for autonomous driving applications to enable vehicles understand road scenes in 2016 [8]. For the research convenience, we try to reduce the network size and complexity by pre-categorizing trees images into several groups based on the tree species and seasons. Then each CNN is trained separately for different groups.

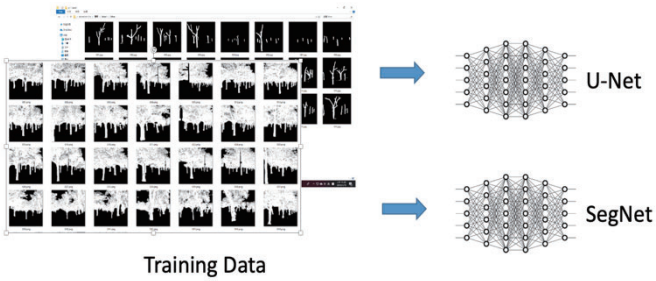


Figure 4. System Flow 1

3.2 Selecting appropriate Deep Learning models for tree segmentation.

By cross referencing the city open data, and the location/compass/time information of Google Street View images, appropriate Deep Learning models are then used for image segmentation of tress. For example, as shown in figure y, our system first determines that there should be a tree contained in the image, according to its GPS and compass information. Then we select the U-Net and Seg-Net which are trained by using deciduous trees images in Spring for image segmentation.

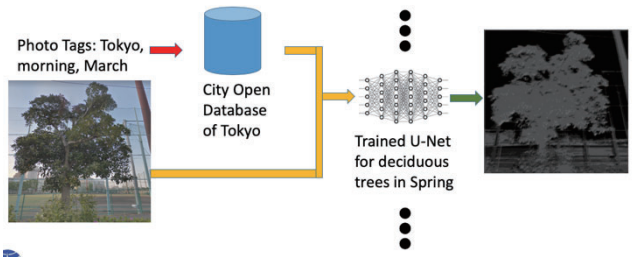


Figure 5. System Flow 2

3.3 With the segmentation results as the guidance, adopting conventional image processing approaches to detect and extract useful information of trees on street pictures.

Currently, we adopted the flood fill algorithm with the segmentation results as the guidance for edge extraction. One of the results is shown in figure 6.

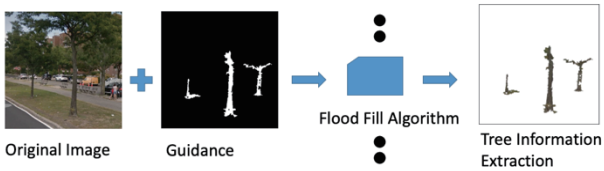


Figure 6. System Flow 3

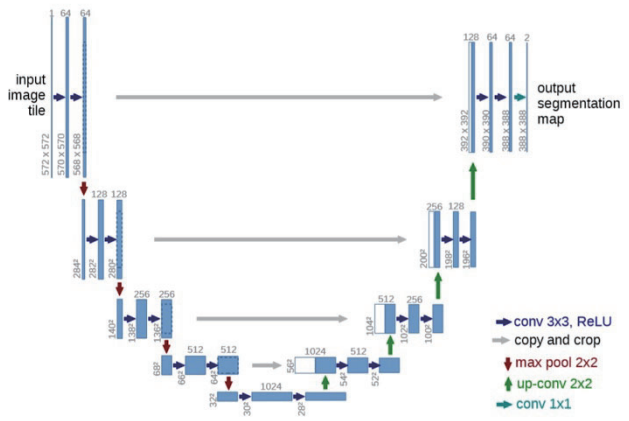


Figure 7. U-Net

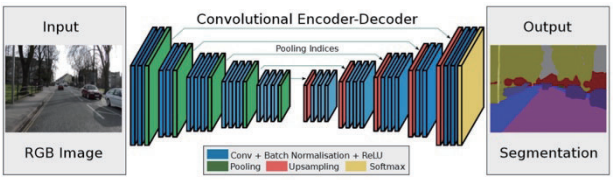


Figure 8. Seg-Net

4. Current Experiment Results

Up to now, we use 100 street view pictures for experiment, Only 40 labeled positive training data are used currently, 30 pictures for testing data, And 30 pictures for negative training data are not adopted yet.

Figure 9. Street Picture in Puli

Figure 10. Segmentation Result use Seg-Net

Figure 10 illustrates an image segmentation result of Seg-Net for a Google Street View picture of Puli, Taiwan in figure 9. The blue regions are for sky, red regions for buildings, white regions for cars, black regions for road, and green regions for plants which are our target for further processing.

Figure 11 is another example of the original street images. And figure 12 illustrates it segmentation results of U-Net trained by deciduous trees images in summer respectively.

Figure 11. A street image which contains multiple trees.

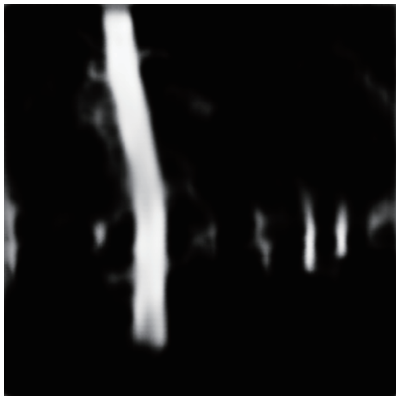


Figure 12. Segmentation Result using U-Net which is trained by trees images in summer.

5. Conclusion

In this research, by utilizing Deep Learning technologies and city open data, we developed a street trees detection method. And we are currently collecting and labeling more images of the 10 most common street trees in 4 seasons for training the Deep Learning models and fine tuning them. There are many parts remained at the conceptual-level, especially at the stage 3, in our prototype system. In addition, more research efforts are required for extracting and separating the shapes of overlapping tree crowns. However, the experiment result still shows that our system has the potentials to:

1. Provide an effective approach for automatically monitoring and managing street trees in smart cities.
2. Severe as a useful tool for users to explore and share the great values of trees in cities and towns.

Acknowledgment

This project is partially supported by No. NSC- 107-2218-E-002 - 048 -.

References

[1] Wajid Ali. “Tree Detection using Color, and Texture Cues for Autonomous Navigation in Forest Environment”. Umea University, 2006.

[2] Y. Lu and C. Rasmussen, "Tree trunk detection using contrast templates," 2011 18th IEEE International Conference on Image Processing, Brussels, 2011, pp. 1253-1256.

[3] Batty M. “Smart cities, big data”. Environment and Planning B: Planning and Design, 39: 191-193, 2012.

[4]ThomasLiebig, NicoPiatkowski, ChristianBockermann, KatharinaMorik, "Dynamic route planning with real-time traffic predic- tions", 2015

[5] CORINNACORTES,VLADIMIRVAPNIK,"Support-VectorNetworks ", 1995

[6] Marco Dorigo, Luca Maria Gambardella, "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem", 1997

[7] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation ", 2015

[8] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", 2016

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Scoring and Classifying Regions via Multimodal Transportation Networks

Aaron Bramson[*1,2,3,4] Megumi Hori[*1] Zha Bingran[*1] Hirohisa Inamoto[*1]

[*1] GA Technologies Inc.
[*2] RIKEN Center for Biosystems Dynamics Research
[*3] Ghent University
[*4] University of North Carolina at Charlotte

In order to better understand the role of transportation convenience in location preferences, as well as to uncover transportation system patterns that span multiple modes of transportation, we score geographic regions according to properties of their multimodal transportation networks. The various scores are then used to classify regions by their dominant mode of transportation, and rank/cluster regions by their transportation features. Specifically, we analyze the train, bus, and road networks of major cities and neighborhoods of Japan to classify them as being train-centric, bus-centric, or car-centric. We also generate scores based on various transportation features to rank cities by their access to public transportation and to categorize/cluster neighborhoods of major cities by their transportation and accessibility properties. We find that business hubs (having low populations) are conveniently reachable via public transportation but vary greatly in their automobile accessibility. Suburban regions have lower connectivity overall but are typically strongly connected to at least one business area. As increasingly rural areas rely more strongly on the road and bus networks, but the network features do not correlate highly with population density.

1. Introduction

Transportation networks can be considered multi-graphs or multilayer networks insofar as there are links of different types connecting nodes representing locations. However, they are also fundamentally geographically embedded which constrains the network structure and requires the inclusion of continuous distance and time weights in discrete network measures. This fusion of network and geographic metrics offers the opportunity to augment network similarity measures as well as fill crucial data gaps about transportation efficiency, accessibility, connectivity, and policies.

2. Data

The geographic foundation of our analysis is a 54,127m^2 (125m inner radius) hexagonal grid covering all of Japan. This is used to define locations as the centers of each hex using GoogleMap's coordinates of Tokyo Station (139.7649361E, 35.6812405N) as a fixed reference point. In order to compare cities and regions within cities, we define a region as all hexes with centroids within 20 km of a selected point. We chose a variety of points across the Tokyo, Kyoto, and Osaka Metropolitan areas to capture a diversity of situations (city centers, suburban bed towns, rural areas, etc.).

2.1 Network Data

We utilize four interwoven networks representing distinct modes of transportation: train/subway, bus/streetcar, road, and walking. The train/subway network represents stations as nodes and train routes as links. In this way, express trains that skip stations are captured by links directly connecting the stations used by that route. The bus network is similarly constructed among bus stops. Our road network is constructed via OpenStreetMaps in which the nodes are intersections and links are road segments; both restricted to roads tagged as *tertiary* or above.

In addition to these networks we include a "walking network". This walking network connects each node of the train network to (1) the closest location of our hex grid as well as (2) any location within 500m of each station. It also connects each bus station to the (1) closest location (2) any location within 200m, and (3) any train station within 200m (when both train and bus networks are included). The third type of link represents a transfer from train to bus. The walking network also connects the nodes of the road network to each location of the hex grid. Finally, we create walking links to convert the location hex grid into a regular k=6 lattice network to allow (slow) transit on foot where no other mode of transportation is available. This walking network is included in all analyses because it is necessary to connect each of the transportation networks to the geographic foundation.

For each link we include a weight equaling the traversal time. For the train and bus data this is set from respective schedules using the average traversal time for that link for that type of train/bus (e.g., local, express). For the road network we calculate the traversal time based on the length of the road segment and the official speed limit (i.e., not considering traffic congestion or actual speeds). For the walking network we assume an average speed of 4kph (15 minutes per km). This slower-than-average speed is used to account for congestion as well indirect walking routes.

In addition to the travel times, we also incorporate a transfer time where appropriate to account for both moving from one platform to another as well as the waiting time for the next train/bus/taxi/etc. Specifically, we add 5 minutes when switching between trains of different lines or types at the same station, and 3 minutes for switching modes: (train ↔ bus, train ↔ road, or bus ↔ road (walking time is already included in the walking link connecting stations, bus stops and intersections).

Contact: Aaron Bramson, GA Technologies, Roppongi Grand Tower 40F, Roppongi 3-2-1, Minato-ku, Tokyo, 106-6290, JAPAN, 03-6230-9180, bramson@ga-tech.co.jp

Although only a rough approximation of the interstitial time gap, it sufficiently summarizes the variance across locations, times of day, walking speeds, congestion conditions, etc. without adding unnecessary complication to the network model.

2.2 Demographic Data

In order to assess practical (versus potential) accessibility we incorporate data regarding the population distribution into our analysis. We take 250m^2 square grid population data obtained from [eStat2018] using grid coordinates from [geoSpacial2018]. Then we resample it to our hex grid using overlap proportions to interpolate the hex populations.

3. Methods

To compare neighborhoods within a city we collect the locations within a 5km radius of multiple secondary and tertiary city centers (these regions overlap). We isolate the transportation networks to within the region of analysis and apply scoring methods to the individual and combined transportation networks. Our most basic evaluation utilizes standard network measures such as diameter, eccentricity profiles, and betweenness profiles along with their time/distance weighted versions. Additionally, we will include specifically geographic and transportation-focused measures such as the profile of times to travel to each regional location, a profile of the number of people reachable within 5, 10, 15, 20 minutes, and the population weighted load on the transportation network to reach the region center.

3.1 Network Measures

To start we calculate several standard network measures (mean degree, mean betweenness, mean eigenvector centrality, mean eccentricity, diameter, clustering coefficient, alpha and beta indices, etc.) on of the following transportation networks: train+walk, bus+walk, road+walk, train+bus+walk, and train+bus+road+walk. We do this for each of several focal areas within Tokyo, Kyoto, and Osaka. This battery of tests allows us to examine both differences in transportation networks for each area and differences among areas for each transportation network.

For each transportation network we calculate the travel times using Dijkstra's algorithm: breadth-first summation of traversed edges' time weights. The core algorithm is augmented to handle transfer times at appropriate junctures. Isochrones are sets of locations binned by travel time, although most of our measures can and do utilize the real-valued traversal times.

3.2 Geotemporal Measures

As a basic measure of accessibility, we compute time-weighted number of hexes reachable form each hex: $\sum_j 1 / t_{ij}$ in which t_{ij} is the shortest time from hex i to hex j. Collecting the population data allows us to determine the *sociability score* of each location; that is, the number of people who can reach each location weighted by the time it takes to reach it. We simplify and generalize the measure from [Biazzo2018] to handle continuous travel time values and averaged edge traversal times. For each hex grid space i we calculate $\sum_j P_j / t_{ij}$ in which P_j is the population of grid space j and t_{ij} is again the shortest time from hex i to hex j. We furthermore include geotemporal versions of certain network measures, such as the time-weighted eccentricity

Figure 1. Isochrone map of Tokyo centered on Tokyo Station using the train+walk network. Darker colors indicate shorter travel times.

(longest, shortest-time path from the center to the periphery) and time-weighted betweenness.

3.3 Machine Learning Techniques

In addition to providing a profile of the multifaceted transportation system, the network and geotemporal measures above are also fuel for clustering and discriminant analysis. We apply an ensemble of available measures of whole-network similarity [Soundarajan2014] (*NetSimile* [Berlingerio2012], N*ormalized LBD* [Richards2010], *Graphlets* [Pržulj2004]) as a basis for distance calculations in addition to standard vector-based methods. Using those distance measures we then apply an ensemble of available unsupervised learning techniques (K-means, spectral clustering, affinity propagation, agglomerative clustering, Gaussian mixture) on the regional profiles to score, cluster, and classify them.

4. Results and Conclusions

This is still a work in progress, but preliminary results reveal cities clustered into those which have a dense rail system (e.g. Tokyo), dense regions that instead rely on buses for public transportation (e.g. Kyoto), and regions with weak public transportation that require automobiles (e.g. small cities and suburbs). Although we expected these features to correlate well with population density; we instead find that other factors heavily influence the type and convenience of a transportation network; factors such as average income, percent of commercial properties, and age demographics.

Within cities we see a familiar pattern of easily accessible central regions with low populations and regions of higher population density further out, with populations again tapering down even further out. These suburban regions often have convenient public transport to the city centers, but locally require buses and/or cars for daily transportation. An analysis of

demographics on the presence of children and elderly within the household should also correlate well with a high score on car-centric transportation. These and other results create a multi-faceted scoring of properties by their transportation and demographic features. Our current efforts aim to summary and visualize these results in an intuitive and interactive way that will lead to greater insights and deeper questions.

While most applications of machine learning to transportation networks aim at traffic prediction, flow efficiency, and rerouting, we are particularly interested in identifying cities with underdeveloped public transportation systems and regions within cities with poor accessibility. Related to the latter point, we will uncover differences in regional accessibility by mode of transportation (e.g., areas that are only convenient if one has access to a car). Identifying under- and over-serviced areas can help in policy decisions including infrastructure planning and housing development. Finally, the fusion of geographic and network measures to score areas by the convenience of, and their reliance on, varying modes of transportation can inform decisions for location services (such as apartment hunting, ride sharing, and new store positioning).

4.1 Future Work

We will extend this analysis by including additional demographic and geographic data in the analysis. Our primary purpose here is scoring and clustering areas by transportation accessibility. Future work will examine the relationship between accessibility and socio-economic factors such as unemployment, income, home-ownership, household structure, age profile, crime, etc. We are also interested in identifying network community structure differences [Bohlin2014] among the transportation modes; that is, which geographic regions are considered to be parts of which neighborhoods when considering different networks. Finally, we wish to pursue question of robustness and efficiency via knockout and detour analyses. This can address response to accidents/failures, and further to identify required structural and throughput changes required to adapt to short-term passenger changes (e.g. the Olympics) and long-term demographic changes (e.g., aging population).

Finally, we are strongly interested in the impact of bicycle ride-sharing programs on transportation flow. Although these programs have long been popular in Europe and China, and bicycles usage is high across Japan, there is very little data or analysis on bicycle usage and its interaction with other transportation modes. The recent growing popularity of bicycle-sharing programs will provide additional data to foster more advanced impact studies.

References

[Berlingerio2012] Michele Berlingerio, Danai Koutra, Tina Eliassi-Rad, Christos Faloutsos. “NetSimile: A scalable approach to size-independent network similarity” *CoRR, Abs*/1209.2684, 2012.

[Biazzo2018] Indaco Biazzo, Bernardo Monechi, and Vittorio Loreto, Universal scores for accessibility and inequalities in urban areas, *arXiv*:1810.03017v1, 2018.

[Bohlin2014] Ludvig Bohlin, Daniel Edler, Andrea Lancichinetti, and Martin Rosvall. "Community detection and visualization of networks with the map equation framework." In *Measuring Scholarly Impact*. Springer, 2014, pp. 3-34.

[eStat2018] Official Statistics of Japan. www.e-stat.go.jp “統計データ/ 国勢調査/ 2015 年/ 5 次メッシュ（250m メッシュ）/ その1人口等基本集計に関する事項” (accessed December 12, 2018).

[geoSpacial2018] Association for Promotion of Infrastructure Geospatial Information Distribution. www.geospatial.jp “都道府県別 250m メッシュ” (accessed December 12, 2018).

[Richards2010] Whitman Richards and Owen Macindoe. “Decomposing Social Networks.” *2010 IEEE Second International Conference on Social Computing* 2010, pp 114-119.

[Soundarajan2014] Sucheta Soundarajan, Tina Eliassi-Rad, and Brian Gallagher. "A guide to selecting a network similarity method." In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 1037-1045. Society for Industrial and Applied Mathematics, 2014.

[Pržulj2004] Nataša Pržulj, Derek G. Corneil, and Igor Jurisica. "Modeling interactome: scale-free or geometric?" *Bioinformatics* 20, no. 18 2004, 3508-3515.

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Evaluating Road Surface Condition by using Wheelchair Driving Data and Positional Information based Weakly Supervision

Takumi Watanabe[*1] Hiroki Takahashi[*1] Yusuke Iwasawa[*2] Yutaka Matsuo[*2] Ikuko E. Yairi[*1]

[*1] Graduate School of Science and Engineering, Sophia University
[*2] Graduate School of Technology Management for Innovation, The University of Tokyo

Providing accessibility information on sidewalks for mobility impaired people is an important social issue. Until now, the authors have evaluated the accessibility of sidewalks by estimating the road surface condition by supervised learning on the accelerometer data mounted on wheelchairs. Video recording and data labeling to accelerometer data based on the video for teacher data require enormous costs and become problematic. This paper proposed and evaluated a novel weakly supervised road surface condition evaluation system of using positional information automatically acquired at driving as a label. The evaluation result showed that weakly supervised learning method using locational label captured detailed features of road surfaces, and classified moving on slopes, curb climbing, moving on tactile indicators, and others with a mean F-score of 0.57 and accuracy of 0.71 close to those of supervised learning method.

1. Introduction

Providing accessibility information on sidewalks for mobility impaired people, such as elderly people and wheelchair users, is one of the important social issues. The conventional methods for gathering accessibility information are as follows: a system that experts evaluate images of sidewalks for each case [Ponsard 06], a crowdsourcing method to recruit volunteers to take pictures of sidewalks and evaluate them [Hara 14, Cardonha 13]. All these methods are based on human power and thus gathering large-scale accessibility information is difficult. Because of the recent expansion of intelligent gadgets, such as smartphones and wristwatch-shaped vital sensors, there is a growing movement of sensing human activities [Swan 13, Nagamine 15]. The authors have been proposing a system which evaluates road surface condition by machine learning using accelerometer data. This system focuses on the fact that the observed values of the accelerometer mounted on a wheelchair is influenced by the condition of the road surfaces. In machine learning, however, video recording and video-based labeling to acceleration data for teacher data require a huge cost and become a serious problem. In various machine learning fields, weakly supervised learning [Zhou 18] methods that do not require conventional detailed teacher labels have been proposed [Oquab 15, Gidaris 18, You 18]. In this paper, the authors propose and evaluate the road surface condition evaluation system by weakly supervised learning which uses positional information labels which can be automatically acquired at the time of driving and thus does not require conventional detailed labels. Our contributions are as follows: we propose a novel method of weakly supervised learning of extracting feature representations of the road surface condition from accelerometer data without conventional detailed labels; we verify the effectiveness of our method using actual data.

Contact: Takumi Watanabe, Graduate School of Science and Engineering, Sophia University, Tokyo, 102-8554 Japan, 03-3238-3280, watanabe@yairilab.net

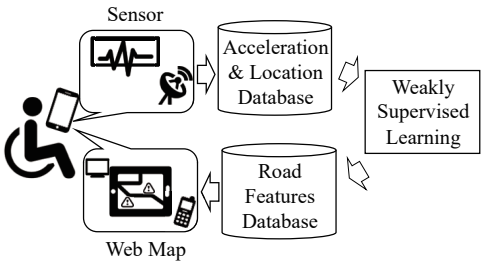


Figure 1 Outline drawing of road surface condition evaluation system by wheelchair sensing and weak supervision.

2. Road surface condition evaluation system

Figure 1 shows an outline of the proposed system. Vibration waveforms of wheelchair movement are collected by an accelerometer mounted on a wheelchair. Extracting road surface information from vibration waveforms using machine learning, the extracted data is accumulated and visualized on a web map. Extracting influence of the road surface condition from the raw accelerometer data is not easy [Lara 12, Liu 17]. Therefore, it is important to convert observed accelerometer data to indexes which represent the condition of the road surface. Some methods for expressing the road surface condition in several discrete classes by creating acceleration data classifiers using machine learning have been proposed [Iwasawa 12, Iwasawa15, Iwasawa 16], and a method for acquiring more detailed road surface features than applied several discrete labels by using feature values extracted from pre-trained DCNN is proposed [Takahashi 18]. However, these methods depend on detailed road surface condition labels and require enormous costs.

3. Road surface condition evaluation by weakly supervised learning

Dataset

The total of nine wheelchair users, including six manual wheelchair users(M1~M6) and three Powered wheelchair users(P1~P3), participated in the experiment. Their actions while

driving about 1.4 km of a specified route around Yotsuya station in Tokyo were measured by an accelerometer (iPod touch) mounted on the lower part of the wheelchair seat, and positioning data of Quasi-Zenith Satellite System (QZSS) was measured at the same time. In order to confirm the situation where the acceleration data sample was acquired, the video of the participant's driving state and the driving road surface condition were taken during the experiment. Acceleration values in the x, y, and z axes of the accelerometer were sampled at 50 Hz, and the total of 1,341,142 samples (about 8 hours) was obtained.

Weakly supervised label

For the training of DCNN, positional information was used as a weakly supervised label as a method of weakly supervised learning. For the positional information in this paper, we checked the location where the accelerometer data was measured by visual observation of the recorded video and used the GPS data (latitude, longitude) acquired on Google Map website as positional information. In assigning labels to the acceleration data, all the sidewalks traveled at the time of the experiment were divided into meshes of an uniform width, and a grid belonging at the time of the measurement of the acceleration data was assigned as a weakly supervised label (as shown in Figure 2). Labels were generated under the conditions of a grid width of 3 m, 4 m, and 5 m.

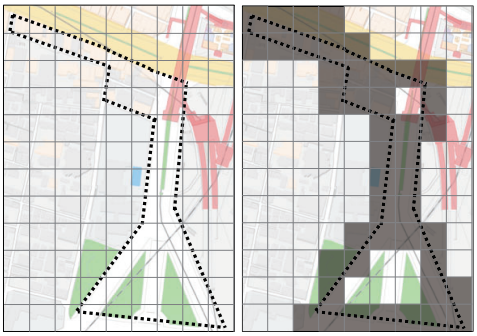


Figure 2 Outline of grids to be weakly supervised label. The left shows how all the sidewalks were divided into meshes, and the right shows how each grid was assigned as a label.

Weakly supervised DCNN

The three axes of acceleration data were segmented into 28502 and 6692 pieces by a sliding window method with a window size of 400 (about 8 seconds) and 100 (about 2 seconds) respectively and overlapping rate of 0.5. As shown in Figure 3, the DCNN used for weakly supervised learning is composed of 7 layers of an input layer, 4 convolutional layers, one fully connected layer, and an output layer. By using the hierarchical structured network and training functions in layers from input to output, feature extractor h and the classifier f those are effective for classification are trained simultaneously.

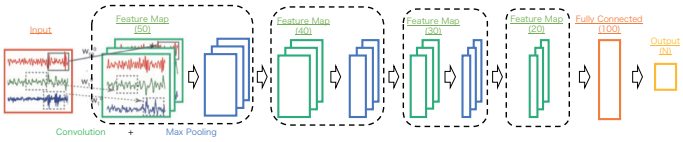


Figure 3 Illustration of the DCNN structure.

Acquisition and Clustering of feature representations

The procedure of acquiring road surface feature representations from weakly trained DCNN and clustering the similar condition road surfaces based on the extracted feature representations is described in order from Step 1 to Step 5.

Step 1: Acquisition of an output pattern of all data

For the DCNN model trained with eight participants data sets as a training data, the remaining one participant data set was input to the DCNN and 100 units output pattern in the fully connected layer was extracted as feature representations of each segmented data.

Step 2: Clustering of feature representations

After compressing the acquired 100-dimensional feature values to a dimension whose cumulative contribution rate exceeds 80% by principal component analysis, clustering was performed on the compressed feature values using the k-means method.

Step 3: Visualization on a map

Clusters generated in Step 2 were color-coded and each point of each cluster was visualized on a map.

Step 4: Analysis of clustering results

Visually comparing the plot result obtained in Step 3 and the recorded video during driving, the road surface condition belonging to each cluster was analyzed.

Step 5: Optimum grid width, window size, and number of clusters

Based on Step 3 and Step 4, the optimum grid width and window size were selected, then a number of clusters that captures the most detailed features of the road surface conditions were selected under best grid width and window size.

4. Qualitative evaluation by clustering

a) Selection of optimum grid width

Plot results with a grid width of 3 m, 4 m, and 5 m with a cluster number of 5 and a window size of 400 were compared.

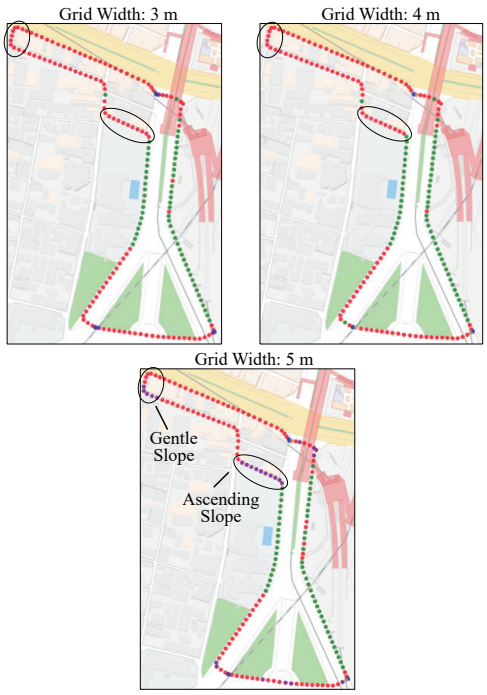


Figure 4 Comparison of clustering results under each condition of a grid width of 3 m, 4 m, and 5 m.

As shown in Figure 4, at the grid width of 5 m, the DCNN captured the features of the ascending slope the most. From this result it is considered that the larger the grid width is, the larger the range of the road surface learned as one label in the DCNN, and DCNN captured an ascending slope where features are easier to read in the larger range.

b) Selection of optimum window size

Plot results with a window size of 100 and 400 with a cluster number of 5 and a grid width of 5 m were compared. As Shown in Figure 5, at the window size of 400, DCNN captured the features of the ascending slope the most. From this result, it is considered that the larger window size is, the larger each segmented training sample in DCNN, and DCNN captured an ascending slope where features are easier to read in the larger range.

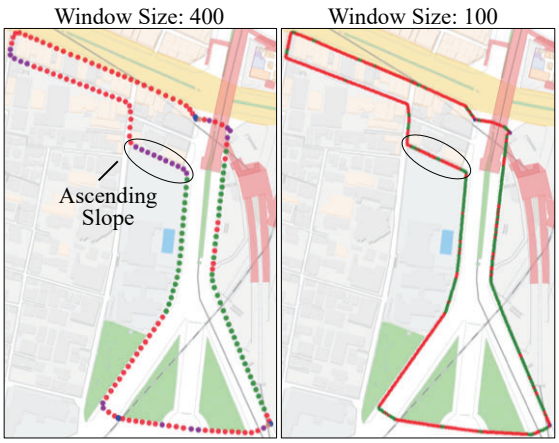


Figure 5 Comparison of clustering results under each condition of a window size of 400 and 100.

c) Selection of the optimum number of clusters

Plot results with the number of clusters 5 to 10 with a grid width of 5 m window size of 400 were compared. As shown in Figure 6, the ascending slope and descending slope were classified into one cluster respectively, and curbs were classified into a specific cluster. Table 1 shows the number of clusters that classified the most detailed road surface condition for each user.

d) Comparison with conventional labeling method

As a result of comparing Figure 5 and clustering result of feature values extracted from detailed labeled trained DCNN, it was shown that weakly supervised method captured more detailed road surface features than the conventional DCNN.

5. Quantitative evaluation of features acquired from weakly supervised DCNN

Evaluation method

Using feature values extracted from weakly trained DCNN as an input, a new classifier was trained as a classification task of four types of labeled road surfaces: slope, curb, braille block, and others. These four types represent typical features of road surfaces, so this method evaluates feature values extracted from weakly supervised DCNN whether they are useful as an indicator of the condition of the road surface.

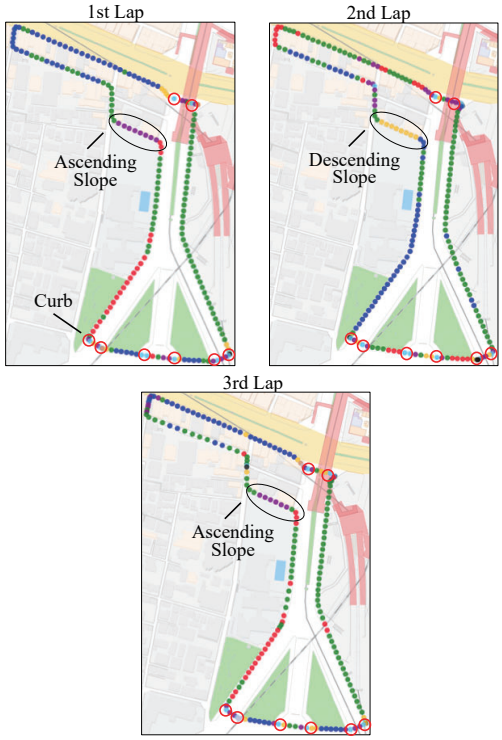


Figure 6 Clustering result with the number of clusters 9 in grid width 5 m and window size 400.
The 1st and 3rd lap are clockwise, so the slope is ascending.
The 2nd lap is counterclockwise, so the slope is descending.

Table 1 Optimum number of clusters for each participant.

Participant	M1	M2	M3	M4	M5	M6	P1	P2	P3
number of clusters	10	8	9	8	9	8	9	9	9

Table 2 Performance comparison between supervised DCNN and weakly supervised DCNN method.

Method	Supervised DCNN	SVM	LR
Mean F-Score	0.58	0.57	0.54
Accuracy	0.81	0.71	0.74

Comparison with Supervised DCNN

Table 2 is a comparison of the classification score. The mean F-score and the accuracy of each class were used as evaluation indexes. Supervised DCNN is the conventional method of training the dataset with DCNN labeled four types of road surfaces. SVM
uses feature values extracted from weakly supervised DCNN as an input and uses Support Vector Machine as a classifier. LR uses feature values extracted from weakly supervised DCNN as an input and uses Logistic Regression as a classifier. As a result, in LR, the mean F-Score was 0.04 points lower than the supervised DCNN, and the accuracy was 0.07 points lower than the supervised DCNN. From this result, it is considered that weakly supervised method misclassified others which occupy more than 70 % of the four labels into a slope, curb, or braille

block, and classified the three types of the road surface to the same degree as Supervised DCNN.

6. Conclusion

In this paper, we proposed a novel method to evaluate road surface condition by weakly supervised learning using positional information as a label and accelerometer data. As a result, it was shown that feature representations acquired by Weakly Supervised DCNN can capture more detailed features of road surfaces than feature values by conventional supervised DCNN, and quantitatively estimate road surface condition. As a future work, we will propose a method to acquire higher-precision feature representations based on new weak label generation method and conduct a detailed analysis of what kind of road surface conditions DCNN with the position information extracts.

Acknowledgments.

We are deeply grateful to all participants of the experiment. This study was supported by Tateishi Science and Technology Foundation in FY2011-2012, the research grant by Chiyoda-ku(CHIYODAGAKU) in FY 2014-2016, and JSPS KAKENHI Grant Number 17H01946.

References

[Ponsard 06] Ponsard, C. and Snoeck, V.: Objective accessibility assessment of public infrastructures, in *10th Int. Conf. on Computers Helping People with Special Needs,* pp. 314–321, Linz, Austria (2006)

[Hara 14] Hara, K.: Scalable methods to collect and visualize sidewalk accessibility data for people with mobility impairments, in *Proc. Adjunct Publication of the 27th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–4, Honolulu, HI, USA (2014)

[Cardonha 13] Cardonha, C., Gallo, D., Avegliano, P., Herrmann, R., Koch, F., and Borger, S.: A crowdsourcing platform for the construction of accessibility maps, in *Proc. 10th Int. Cross-Disciplinary Conf. on Web Accessibility*, p. 26 Rio de Janeiro, Brazil (2013)

[Swan 13] Swan, M.: The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery, *Big Data*, Vol. 1, No. 2, pp. 85–99 (2013)

[Nagamine 15] Nagamine, K., Iwasawa, Y., Matsuo, Y., and Yairi, I. E.: An Estimation of Wheelchair User's Muscle Fatigue by Accelerometers on Smart Devices, in *Adjunct Proc. 2015 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing and Proc. 2015 ACM Int. Symposium on Wearable Computers*, UbiComp/ISWC'15 Adjunct, pp. 57–60, New York, NY, USA (2015)

[Zhou 18] Zhou, Z. H.: A brief introduction to weakly supervised learning, in *National Science Review*, Vol. 5, No. 1, pp. 44-53, (2018)

[Oquab 15] Oquab, M., Bottou, L., Laptev, I., and Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 685-694, Boston, MA, USA (2015)

[Gidaris 18] Gidaris, S., Singh, P., and Komodakis, N.: Unsupervised Representation Learning by Predicting Image Rotations, in *6th Int. Conf. on Learning Representations*, Vancouver, BC, Canada, (2018)

[You 18] You, Z., Raich, R., Fern, X. Z., and Kim, J.: Weakly Supervised Dictionary Learning, in *IEEE Trans. Signal Proc.*, Vol. 66, No. 10, pp. 2527-2541, (2018)

[Lara 12] Lara, O. D. and Labrador, M. A.: A survey on human activity recognition using wearable sensors, in *IEEE Commun. Surveys Tutorials*, Vol. 15, No. 3, pp. 1192–1209, (2012)

[Liu 17] Liu, H., Taniguchi, T., Tanaka, K., Takenaka, Y., and T. Bando: Visualization of driving behavior based on hidden feature extraction by using deep learning, in *IEEE Trans. Intelligent Transportation Systems*, Vol. 18, No. 9, pp. 2477–2489, (2017)

[Iwasawa 12] Iwasawa, Y. and Yairi, I. E.: Life-logging of wheelchair driving on web maps for visualizing potential accidents and incidents, in *PRICAI 2012: Trends in Artificial Intelligence*, pp. 157–169, Springer (2012)

[Iwasawa 15] Iwasawa, Y., Nagamine, K., Matsuo, Y., and Eguchi Yairi, I.: Road Sensing: Personal Sensing and Machine Learning for Development of Large Scale Accessibility Map, in *Proc. 17th Int. ACM SIGACCESS Conf. on Computers & Accessibility,* pp. 335–336ACM (2015)

[Iwasawa 16] Iwasawa, Y., Yairi, I. E., and Matsuo, Y.: Combining Human Action Sensing of Wheelchair Users and Machine Learning for Autonomous Accessibility Data Collection, in *IEICE Trans. on Information and Systems*, Vol. E99-D, No. 4, pp. 115–124 (2016)

[Takahashi 18] Takahashi, H., Nagamine, K., Iwasawa, Y., Matsuo, Y., and Yairi, I. E.: Quantification of Road Condition from Wheelchair Sensing Data Using Deep Convolutional Neural Network, in *IEICE Trans. Information and Systems*, Vol. J101-D, No. 6, pp. 1009-1021 (2018) (In Japanese)

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Prediction of the Onset of Lifestyle-related Diseases Using Regular Health Checkup Data

Mitsuru Tsunekawa*1 Natsuki Oka*1 Masahiro Araki*1 Motoshi Shintani*2
Masataka Yoshikawa*3 Takeshi Tanigawa*4

*1 Kyoto Institute of Techonology *2 SG Holdings Group Health Insurance Association
*3 Japan System Techniques Co.,Ltd. *4 Juntendo University

This study proposes a method for predicting the onset of lifestyle-related diseases using periodical health checkup data. We carefully examined insurance claims data to identify the onsets of the diseases and used them as correct answers for supervised learning. We adopted the undersampling and bagging approach to address the class imbalance problem. We aimed to predict whether lifestyle-related diseases, other than cancer, will develop within one year. The precision and recall of the proposed method were 0.32 and 0.89, respectively. Compared with a baseline that sets thresholds for each examination item and considers their logical sum, it was found that much higher precision could be obtained while maintaining recall, which is meaningful as it allows for the suppression of the number of targets for health guidance, without increasing the negligence of those that are likely to become severely ill.

1. Introduction

Many people have recently begun using Internet mail-order sales, greatly increasing the number of deliveries. Consequently, the social interest in the work environment and health management of courier drivers is increasing. If appropriate health guidance can reduce the occurrence of drivers' lifestyle-related diseases and prevent severe idiopathic illnesses during driving, medical expenses and traffic accidents can be decreased. We, therefore, aimed to use drivers' regular health checkup data to predict the onset of lifestyle-related diseases as accurately as and ensure provision of appropriate health guidance.

Many studies have used machine learning and data mining techniques to predict disease onset from medical data. For example, [Weng 17] highlighted the superiority of machine learning techniques to predict cardiovascular risk from routine clinical data; [Yatsuya 16] predicted the occurrence probability of myocardial infarction or cerebral infarction using health examination results. Morevoer, [Uematsu 17] proposed a model that predicts pneumonia hospitalization using the Lasso logistic regression of regular health checkup data, which is similar to our research model in that it tries to predict using periodical health checkup data of healthy people. Our study objective was to predict whether lifestyle-related diseases, other than cancer, will develop within one year, using regular medical examination data of Courier drivers.

2. Target Data

2.1 Data overview

We used insurance claims and regular health checkup data of employees from the SG Holdings Group Health Insurance Association. Health insurance claims data is created when a person is injured or ill and visits a medical institution, whereas health checkup data is taken regularly (typically once a year). The two datasets were anonymized and linked with a hash code for uniquely identifying patients. In this study, disease onsets were extracted from the insurance claims data and used as correct answers for onset prediction; the health checkup data is used as input for onset prediction.

Health insurance claims data includes disease name codes and medical examination dates etc. Details of the health checkup data are provided later.

In this study, we analyzed the insurance claims data from 1996 to 2017 and the health checkup data from 2006 to 2018, of individuals aged 15-74 years; the total health checkup data was 961,906 sheets for 156,145 people, and the total insurance claims data was 1,617,078 sheets for 108,581 people.

2.2 Disease names as prediction targets

An individual's diagnosed diseases names were obtained by searching through the disease name codes included in the health insurance claims data and corresponded with those in the ICD-10. Table 1 presents the ICD-10 codes and the corresponding disease names to be predicted (hereinafter, referred to as "severe disease names").

2.3 Feature values used for prediction

The following examination items of health checkup data were used as feature values for prediction: Health examination data included not only the numerical data of inspection results, but also the results of a questionnaire on lifestyle habits, and the judgment results of six levels derived from the examination data by medical institutions. Items of abdominal girth and visual acuity judgment, heart rate, visual acuity judgment, fundus judgment, and metabolic judgment were removed, since the ratio of missing values was

Contact: Mitsuru Tsunekawa, Kyoto Institute of Technology, and E-mail: m-tsune@ii.is.kit.ac.jp

Table 1: ICD-10 codes of severe disease names.

ICD-10	disease name
E10	Insulin dependent diabetes mellitus
E11	Non-insulin dependent diabetes mellitus
E14	Diabetes mellitus other than the above
I20	Angina pectoris
I21, I22	Acute myocardial infarction
I42	Cardiomyopathy
I44 I49	Arrhythmia, Conduction defects
I60, I690	Subarachnoid hemorrhage
I61, I691	Intracerebral hemorrhage
I63, I693	Cerebral infarction

50% or more. The health examination data also included findings freely described by doctors; however, we excluded them because natural language understanding is necessary for use of free description.

Health examination data items used as input features are as follows:
Sex; Age; Height; Weight; Body fat percentage; Systolic blood pressure; Diastolic blood pressure; Number of red blood cells; Hemoglobin; Hematocrit; Platelet count; GOT; GPT; γ-GTP; Total cholesterol; HDL cholesterol; LDL cholesterol; Neutral fat; Uric acid; Creatinine; eGFR; HbA1c; Questions about medicine to lower blood pressure, insulin injection, or medicine to lower blood sugar, medicine to ameliorate dyslipidemia, stroke, chronic renal failure, anemia, smoking habits, weight change from the age of 20 years, exercise habits, walking habits, walking speed, weight change over the past year, eating speed, meal just before going to bed, after dinner snacks, skipping breakfast, drinking habits, drinking alcohol amount, sleeping time, willingness to improve lifestyle habits, and willingness to receive health guidance; Judgments on urinary protein and urine sugar; Representative judgment; Judgments on physical measurements, hearing ability, blood pressures, anemia, liver function, renal function, uric acid and gout, blood sugar, sugar metabolism, and urinalysis; Examination judgment.

2.4 Characteristics of data

The data typically have two characteristics. First, considerable imbalance: For example, in 2017, the proportion of people diagnosed with severe diseases was only 4.5%. The learning of such unbalanced data may be greatly affected by the properties of a large number of negative examples, persons who are not diagnosed with severe diseases. Therefore, a method that can successfully learn this imbalanced data must be adopted.

Second, classifying data as positive or negative is not easy. In this study, we aimed to predict whether a person who is healthy at the time of a regular health checkup will be diagnosed with one of the severe diseases within a year of the checkup. Consequently, data was positive if the person will fall sick within a year, and negative if not. Therefore, it was necessary to accurately judge the presence or absence of illness at the time of a health checkup.

The point at which the target disease name first appeared in an employee's insurance claims data was not necessarily the point when he/she first developed the disease. It is not uncommon for individuals with previously diagnosed diseases to join a health insurance association in an industry with large personnel flow. However, because the data used in this study belonged to a health insurance association, it was only available for the period of joining the association; the insurance claims data before entering the association could not be confirmed. The next section describes how we addressed this problem.

3. Data selection and machine learning method

3.1 Data selection

We addressed the classification problem of predicting whether individuals will suffer severe illness within one year, by using medical examination data. The data selection method for positive and negative data was as follows.

First, to address the previously mentioned data availability problem, we used the following method and determined whether the insurance claims data of an individual's first-time diagnosis of a severe disease was actually first. We first calculated the hospital visit interval for the same disease after receiving the diagnosis of a disease. If the hospital visit interval was shorter than the interval between the day of joining the health insurance association and the day of first-time diagnosis of a severe disease, the diagnosis was judged to be actually first; alternatively, if the visit interval was greater, it was judged not to be first. The visit interval was calculated using three interval data, which we regarded as sufficient. The specific procedure was as follows (see also Figure 1):
1) Extract the oldest data (*) with a severe disease name from an individual's insurance claims data. 2) Select three consecutive data with the same disease name that are newer than the extracted data and calculate the hospital visit intervals. 3) Retrieve the individual's oldest insurance claims data (**) and calculate the interval between data (**) and data(*). 4) If the maximum of the three values calculated in "2)" is smaller than that calculated in "3)," regard data (*) as the first-diagnosis data of the disease.

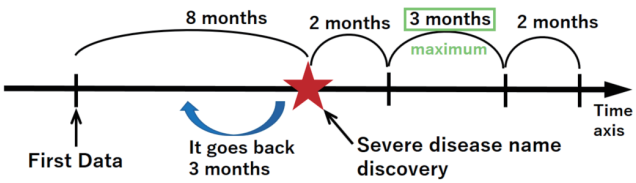


Figure 1: Method of judging the genuineness of the first diagnosis.

Next, the positive health checkup data was chosen from the data included within the range of one year or less before the first appearance of the severe disease. When there were multiple data in the range, we adopted the oldest one. We also considered the amount of changes in the health checkup

data, calculated the differences between the chosen data and the previous data and between the chosen data and the two previous data, and added them to the feature set (Figure 2). Since some items in the health checkup data would have changed with the development of the disease, we considered that the discrimination accuracy improved by explicitly adding the change amount to the feature set.

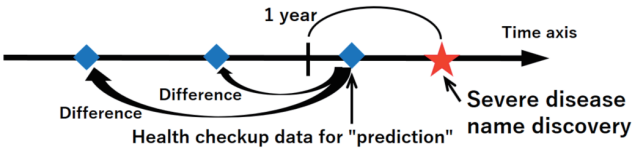


Figure 2: Selection of positive data.

For negative data, however, we excluded data of individuals who had been diagnosed with a severe disease even once and used only the remaining data. Furthermore, if there is no insurance claims data after more than one year from when the health checkup data was extracted, the possibility is that the individual may have been diagnosed with a severe disease within one year of data extraction. This is possible if the person leaves the job. Therefore, to eliminate this possibility, we excluded the data of individuals with no insurance claims data after one year or more from the extracted health examination data. As with the positive data, we calculated the differences using three consecutive health examination data and added them to the feature set (Figure 3).

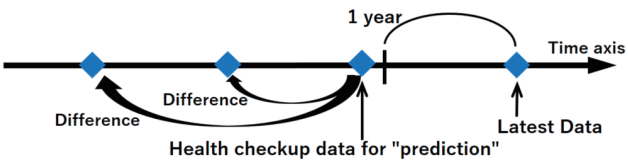


Figure 3: Selection of negative data.

There were cases in which an individual had multiple medical examination data that fit the selection criteria. This was common in both positive and negative data. However, we used only one data per individual to prevent data imbalance. If we did not select the oldest data for positive data, we used data after being diagnosed with one of the severe diseases; however,for negative example, we can select data at any point in time.

Thus, we obtained 1255 positive data and 37664 negative data, with 133 features. The missing values were filled with the median values.

3.2 Machine learning method

In this study, undersampling and bagging [Wallace 11] was adopted as an effective learning method for imbalanced data. Bagging is a method of improving classification accuracy by combining classifiers which are called weak learners. We used decision trees without pruning as classifiers because they were unstable and resulted in higher performance. Undersampling created balanced data.

4. Results and discussion

There were 500 weak learners. Even if the number of weak learners was changed to range between 100 and 500, there was little change in recall and precision; however, if the number was less than 100, the precision decreased. Since the decision tree used for the weak learners was an algorithm not affected by the scale, data scaling was not performed. We used 70% of the dataset for learning and 30% for evaluation. Table 2 presents the confusion matrix of the proposed method. The positive precision and recall were 0.32 and 0.89, respectively.

Table 2: Confusion matrix of the proposed method.

		Predicted class	
		Positive	Negative
Actual class	Positive	334	43
	Negative	700	10600

We used a judgment category table*1 that was officially released by the Japan Society of Ningen Dock as the baseline method. A threshold value for each item was set and discrimination was carried out by logical OR operation on each item. We only used the items that were common to the input features of this study. The total number of items used was 13. The table classified health checkup data into four categories: No abnormality, Mild abnormality, Follow-up required, and Medical treatment required.

The precision-recall curves of the proposed method and the baseline method are shown in Figure 4. The thresholds of three categories, excluding "No abnormality," were used to plot the precision-recall curve of the baseline method. For the proposed method, we changed the ratio of the positive and negative examples in undersampling as follows: 1:0.25, 1:0.5, 1:1, 1:2, 1:4, 1:8, and 1:16. We added another precision-recall curve of the proposed method in which the number of features was reduced to 13 in order to compare with the baseline method using the same features.

Since the graph of the proposed method lies clearly above that of the baseline method, the proposed method can be considered superior to the baseline method. The results demonstrated that much higher precision could be obtained by the proposed method when the recall was about the same degree as the baseline method; increasing precision while maintaining recall is meaningful as it allows for the suppression of the number of targets for health guidance, without increasing the negligence of those who are likely to become severely ill.

We identified the features that were important for classification. The top 3 were HbA1c, metabolism judgement, and the question about taking insulin injection or a drug that lowers blood glucose. Metabolic judgement refers to the judgment of the danger of metabolic syndrome in six levels from the health checkup data. HbA1c is one of the indicators used to judge diabetes, prescribing insulin injection and medication to lower blood glucose is a treatment

*1 https://www.ningen-dock.jp/wp/wp-content/uploads/2013/09/Dock-Hantei2018-20181214.pdf

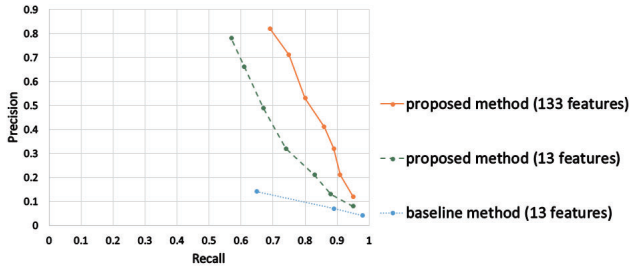


Figure 4: Precision-recall curves of the proposed method and the baseline method.

related to diabetes. Thus, diabetes can be easily distinguished using health checkup data. Among the positive data, the number of diabetes mellitus was as high as 74%; therefore, if identify diabetes, the result would have a high overall accuracy. To confirm this assumption, we focused solely on diabetes and its prediction. Positive cases were defined as persons diagnosed with diabetes mellitus; negative cases comprised patients with severe diseases other than diabetes and healthy people. The created dataset comprised 921 positive and 37998 negative cases. As a result of classification, the positive precision and recall were 0.34 and 0.92, respectively. The confusion matrix is shown in Table 3.

Table 3: Confusion matrix for diabetes prediction using the proposed method.

		Predicted class	
		Positive	Negative
Actual class	Positive	254	23
	Negative	494	10906

Next, to compare with diabetes, we tried to predict the second most frequent angina using the proposed method. As before, positive cases were defined as persons diagnosed with angina pectoris, and negative cases consisted of patients with severe diseases other than angina pectoris and healthy people.

The created dataset comprised 229 positive and 38690 negative cases. As a result of prediction, the positive precision and recall were 0.03 and 0.90, respectively. Table 4 shows the confusion matrix. As evident, the precision reduced and angina pectoris was difficult to discriminate.

Table 4: Confusion matrix for angina pectoris prediction uising the proposed method.

		Predicted class	
		Positive	Negative
Actual class	Positive	62	7
	Negative	1755	9852

5. Conclusion

5.1 Summary

In this study, we proposed a method to predict the onset of lifestyle-related diseases other than cancer, using periodical health checkup data, and a method to select learning data based on the insurance claims data. When all target disease names were identified as positive cases, we obtained positive precision and recall values, 0.32 and 0.89, respectively. Compared to the judgment category table, which the Japan Society of Ningen Dock used as a baseline, it was found that much higher precision could be obtained when the recall was about the same degree.

5.2 Future tasks

In this study, doctors' findings from the health examination data were excluded; however, applying natural language processing to this part will allow the data to be used. As another method of coping with imbalanced data, we plan to use an anomaly detection method that constructs a model using data of healthy people as normal data and detecting data that does not fit the model. We also plan to predict the onset of lifestyle-related diseases after more than one year of a regular medical examination.

A major limitation of this study was that although insulin injection and drugs that lower blood glucose should not be prescribed before receiving a diagnosis of diabetes, the item "Do you take insulin injection or a drug that lowers blood glucose" was used as one of the main items to predict the onset of diabetes. This means that the selection process of positive and negative data need to be reconsidered.

References

[Weng 17] Weng, F. S., Reps, J.,Kai, J., Garibaldi, M. J., and Qureshi, N.: Can Machine-learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?, PLoS One, 12(4), doi:10.1371/journal.pone.0174944 (2017).

[Yatsuya 16] Yatsuya, H., Iso, H., Li, Y., Yamagishi, K., Kokubo, Y., Saito, I., Sawada, N., Inoue, M., and Tsugane, S.: Development of a Risk Equation for the Incidence of Coronary Artery Disease and Ischemic Stroke for Middle-aged Japanese Japan Public Health Center-Based Prospective Study. Circulation Journal, 80(60), 1386-1395 (2016).

[Uematsu 17] Uematsu, H., Yamashita, K., Kunisawa, S., Otsubo, T., and Imanaka, Y.:Prediction of Pneumonia Hospitalization in Adults Using Health Checkup Data, PLoS One, 12(6), doi:10.1371/journal.pone.0180159 (2017).

[Wallace 11] Wallace, C. B.,Small, K., Brodley, E. C., and Trikalinos, A. T.: Class Imbalance, Redux, IEEE 11th International Conference on Data Mining, IEEE Xplore, doi:10.1109/ICDM.2011.33 (2011).

International Session | International Session | [ES] E-5 Human interface, education aid

[4H2-E-5] Human interface, education aid: human evaluation

Chair: Naohiro Matsumura (Osaka University)

Fri. Jun 7, 2019 12:00 PM - 1:40 PM Room H (303+304 Small meeting rooms)

[4H2-E-5-01] Recognition of Kuzushi-ji with Deep Learning Method

〇Xiaoran Hu[1], Mariko Inamoto[2], Akihiko Konagaya[1] （1. Tokyo Institute of Technology, 2. Keisen University）

12:00 PM - 12:20 PM

[4H2-E-5-02] Computerized Adaptive Testing Method using Integer Programming to Minimize Item Exposure

〇Yoshimitsu MIYAZAWA[1], Maomi UENO[2] （1. The National Center for University Entrance Examinations, 2. The University of Electro-Communications）

12:20 PM - 12:40 PM

[4H2-E-5-03] Maximizing accuracy of group peer assessment using item response theory and integer programming

〇Masaki Uto[1], Duc-Thien Nguyen[1], Maomi Ueno[1] （1. University of Electro-Communications）

12:40 PM - 1:00 PM

[4H2-E-5-04] Waveform Processing of Electrocardiogram with Neural Network and Non-contact Measurement using Kinect for Driver Evaluation

〇HAO ZHANG[1], Takashi IMAMURA[1] （1. Niigata University）

1:00 PM - 1:20 PM

[4H2-E-5-05] Probability based scaffolding system using Deep Learning

〇Ryo Kinoshita[1], Maomi Ueno[1] （1. The University of Electro-Communications.）

1:20 PM - 1:40 PM

©The Japanese Society for Artificial Intelligence

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Recognition of Kuzushi-ji with Deep Learning Method: A Case Study of Kiritsubo Chapter in the Tale of Genji

Xiaoran Hu[*1] Mariko Inamoto[*2] Akihiko Konagaya[*1]

[*1] Tokyo Institute of Technology [*2] Keisen University

Reading ancient documents is one of fundamental works on the study of national literature. However, due to the use of kuzushi-ji (classical cursive handwriting characters) in the ancient documents, it requires a lot of knowledge and labor to read. This paper uses an End-to-End method with attention mechanism to recognize the continuous kuzushi-ji in phrases. Compared with the traditional recognition model with Connectionist Temporal Classification (CTC), our approach can get higher accuracy of recognition. The method can recognize phrases written by kana (47 different characters) with the accuracy 78.92% and recognize phrases containing both kanji (63 different characters) and kana with accuracy 59.80%.

1. Introduction

Kuzushi-ji (classical cursive handwriting characters) is general name for the outdated hiragana and kanji that are not used in school education of Japan since 1900. However, the works of Japanese classical literature, especially those before the Edo period, were almost written by kuzushi-ji. As a result, the classical literature can only be read by experts and most literary works are buried without being digitized. In order to deal with this problem, it is necessary to develop the method that aims to reprinting old ancient documents automatically. Recently, some methods were proposed to recognize kuzushi-ji, especially three-character string. The model that combines neural network with connectionist temporal classification won the challenge of recognizing kuzushi-ji (21th PRMU Algorithm Context) [1] in 2017. However, the performance of those models with various length continuous kuzushi-ji is not so good. This paper uses an End-to-End method that can recognize continuous kuzushi-ji with any length in phrases. The method uses the convolutional layers of VGG [2] and BLSTM [3] as encoder, using LSTM with attention mechanism [4] as decoder. To explain the result of two model, this paper use Grad-CAM [5] to visualize the network.

2. Method

2.1 Dataset

The training dataset of kuzushi-ji is from the Center for Open Data in the Humanities [6]. We select 47 kana and 63 kanji characters from 15 books as training dataset. The original training dataset is a set of images with single kana as shown in Fig1. In order to build arbitrary length character dataset, the original dataset does image banalization and combine single character images to form phrases. The model sets the input size of training images as width 32, height 300 pixels. To ensure kana on images are not deformed, the padding place of image is white.

The test dataset of various length continuous kuzushi-ji phrases is from a chapter of ancient roles, Kiritsubo, in the tale of Genji, which is written by unknown writer in Edo period. Separating images into many phrase images have been finished by hands. There are 137 images containing only kana and 159 images containing both kana and kanji.

The book of training dataset and test dataset are written by different persons. So the handwriting style is different between two datasets, which can decrease the accuracy of prediction. So we uses the phrases of the tale of Genji as labels of training dataset to reduce the differences between training and testing of our neural network models.

Fig1. Screenshot of original Training dataset「あ」
http://codh.rois.ac.jp/char-shape/unicode/U+3042/

2.2 Encoder: CNN + BLSTM

An original method for extracting the sequential features form images is to use convolutional neural network, however the native approach does not make use of the spatial dependencies between the features. So, we use the encoder that combines CNN and BLSTM to get the feature vectors from input images.

As shown in Fig2(a), the encoder first uses the convolutional layers to process the images to get robust and high-level features of images. In this process, two dimensional image converts to one dimensional feature map. A two-layer Bidirectional Long-short term memory (BLSTM) network [3] is applied after convolutional neural network to enlarge the range the feature sequences of input images.

2.3 Decoder: LSTM with Attention Mechanism

The decoder of model is LSTM with Attention mechanism. Connectionist temporal classification (CTC) uses a scoring function proposed in 2006[8], which deals with sequence problems where the timing is variable. However, CTC is limited in character recognition in phrases because it does not consider the dependencies between labels. Different from CTC method, attention based LSTM model can predict current label relying on

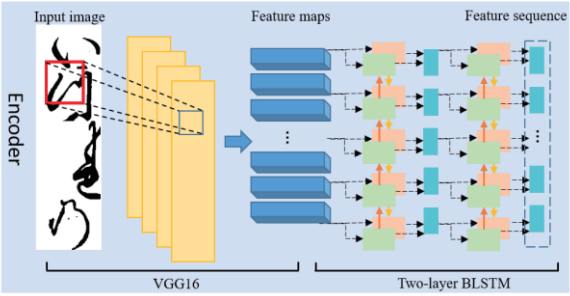
Contact: Akihiko Konagaya, The school of computing, Tokyo institute of Technology, kona@c.titech.ac.jp

the results of previous labels [7]. So, we adopted attention based model as decoder. In order to compare the performance of CTC and attention models, our decoder has the two models in its structure as shown in Fig2(b).

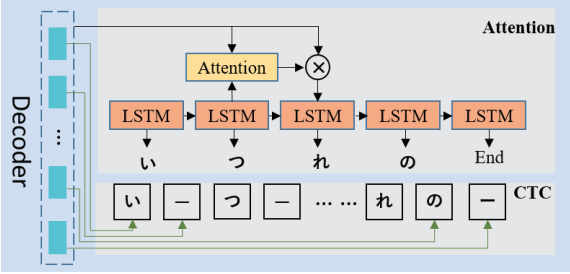
3. Experiments

3.1 Implementation

The structure of CNN in encoder is the convolutional layers of VGG16, which has 5 blocks. The memory units of each layer of BLSTM is 256. CTC based model and attention based model use the same encoder. To fit the input size of CTC model, there is a fully connected layer between BLSTM and CTC. The max step of attention model is set as 11 so that kuzushi-ji's recognition model can recognize up to 10 characters in a phrase. The parameters of CNN use pre-trained weights of VGG16, which are open on GitHub [8].



(a) Encoder structure



(b) Decoder structure of Attention and CTC

Fig2. Encoder and decoder structures of continuous kuzushi-ji recognition model

3.2 Prediction Accuracy on Test Dataset

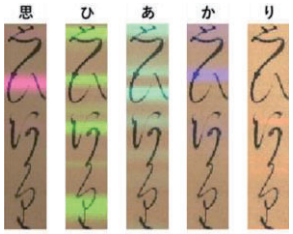
We perform two sets of experiments: one on kana dataset and the other on kana-kanji dataset. Table 1 shows the prediction results of different models. It indicates that the performance of attention model is much better than the one of CTC model. In the both models, the performance on kanji-kana dataset is not so good as kana dataset mainly due to the lack of sufficient number of kanji images in the original dataset.

Table1. The accuracy of prediction

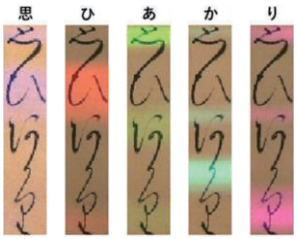
Model	Kana dataset	Kanji-kana dataset
CTC	60.14%	48.16%
Attention	78.92%	59.80%

3.3 Comparison of CTC and Attention Model Results

Fig3 shows the difference of focusing points between two models when recognizing the phrases image "思ひあかり." Both model can give correct prediction, however, compared with CTC based model, attention based model can figure out more precise character positions on the image.



(a) Visualization result of CTC based model



(b) Visualization result of Attention based model

Fig3. Visualization result between two models: the highlight colors mean the positions where neural network focuses on when predicting a target character.

4. Conclusion

This paper proposes a deep learning method to recognize continuous kuzushi-ji phrases using the images of the tale of Genji. Compared with previous model, the proposed model with attention mechanism gets high accuracy of prediction.

Our future task is to improve the performance of recognizing kanji-kana mixed images by enhancing the decoding capability.

Acknowledgements

I would like to express my gratitude to the members of the Konagaya laboratory and Genji Pictorial DB Study Group for valuable comments and discussions in promoting this research.

References

[1] https://sites.google.com/view/alcon2017prmu/コンテスト結果
[2] Simonyan, Karen, and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science, 2014.
[3] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. ICASSP, 2013.
[4] Xu, Kelvin, et al. Show, attend and tell: Neural image caption generation with visual attention. International conference on machine learning. 2015.
[5] Selvaraju, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. ICCV, 2017.
[6] http://codh.rois.ac.jp/index.html.en
[7] Graves, Alex, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd international conference on Machine learning. ACM, 2006.
[8] Dabbish, Laura, et al. Social coding in GitHub: transparency and collaboration in an open software repository. ACM, 2012.

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Computerized Adaptive Testing Method using Integer Programming to Minimize Item Exposure

Yoshimitsu MIYAZAWA*1 Maomi UENO*2

*1The National Center for University Entrance Examinations, Tokyo, Japan,
*2The University of Electro-Communications, Tokyo, Japan

Computerized adaptive testing (CAT) estimates an examinee's ability sequentially and selects test items that have the highest accuracy for estimating the ability. However, conventional CAT selects the same items for examinees who have equivalent ability. Therefore, tests cannot be used practically under circumstances in which the same examinee can take a test multiple times. As described herein, we propose CAT that minimizes item exposure and which adaptively selects different items for examinees of equal ability, while retaining accuracy. This paper presents the method's effectiveness through a simulation experiment and with item pools used by actual test providers. Results confirmed that 1) the proposed method yielded the shortest test length. 2) The proposed method controls exposure and selects different items for different examinees. The non-uniformity of estimation was low. 3) The average item exposure of the proposed method was the lowest.

1. Introduction

Computerized adaptive testing (CAT) estimates an examinee's ability after every answer and selects an item with the highest accuracy of estimating the ability[Linden 10][Ueno 10][Ueno 13].

This item selection presents an important benefit: the number of items to be selected and the testing time can be decreased for an examinee without reducing the accuracy. Furthermore, the abilities of all examinees are measured using the same degree of accuracy. Nevertheless, conventional CAT has the following two shortcomings.

1. Conventional CAT is inapplicable for conditions in which the same examinee takes a test multiple times because the same group of items tends to be selected when the same examinee takes a test multiple times.

2. Conventional CAT provides the same items to examinees of equivalent ability. Therefore, not all items in an item pool can be used effectively. Excessive item exposure engenders disclosure of the item contents to examinees, and might therefore degrade reliability[Way 98].

Items with estimated difficulty and discrimination parameters must be prepared in advance to conduct CAT. Huge amounts of time and other costs thereby arise for test item preparation, especially for high-stakes tests. Accordingly, it is desirable to use all items in an item pool for the implementation of a test. We propose a framework of new CAT that resolves these difficulties in this study.

Constrained CAT (CCAT) is proposed to resolve difficulties posed by excessive exposure [Linden 10, Linden 98]. The method reported by van der Linden et al.[Linden 98, Linden 10], a constrained CAT method, constitutes an item set in which the number of exposed items and necessary answer times, and selects an item with the most information from the item pool at every item selection. This method can control exposure and can select a different item for each examinee. However, deviation of accuracy in the selected items produces great differences among examinees in terms of the test length and accuracy.

We propose a CAT to resolve this difficulty. It can select a different item while maintaining equal accuracy, even for an examinee with the same ability. Specifically, we propose a CAT method that can select items 1) with uniform test length, and 2) with uniform accuracy among tests, but 3) the items are not identical.

This study proposes CAT item selection using integer programming. The approach of the proposed method is described as presented below. 1) A group that satisfies test constraints as Fisher information is assembled using integer programming for minimizing item exposure. 2) An item that has the highest information is selected from the group.

The proposed method uses integer programming for minimizing item exposure and satisfying upper and lower bounds of information. Accordingly, a different item set can be selected, even for an examinee with equal ability. Moreover, the improved diversity of item selection is expected to encourage thorough use of items in an item pool and to mitigate deviation in exposure.

This paper demonstrates the effectiveness of the proposed method through simulation experiments and experiments using actual data.

2. Item Response Theory

In CAT, ability is estimated based on Item Response Theory (IRT)[Lord 80, Lord 68] to select items with the highest estimation accuracy[Baker 04, Linden 16a, Linden 16b]. Item response theory is a recent test theory based on mathematical models for which practical use is in progress lately

Contact: Yoshimitsu MIYAZAWA, The National Center for University Entrance Examinations, 2-19-23, Komaba, Meguro-ku, Tokyo, Japan, 03-5478-1275, miyazawa@rd.dnc.ac.jp

in widely diverse areas related to computer testing.

The two-parameter logistic model (2PLM) has been used for many years as an item response model that is broadly applicable to such binary data. This study also adopts 2PLM, for which the probability of a correct answer given to item i by an examinee j with ability $\theta \in (-\infty, \infty)$ is assumed as

$$p(u_i = 1|\theta) = \frac{1}{1 + \exp[-1.7a_i(\theta - b_i)]}. \tag{1}$$

The standard error of ability estimation based on the item response theory is known to approach the reciprocal of Fisher information asymptotically [Lord 80]. Accordingly, item response theory usually employs Fisher information as an index representing the accuracy.

In 2PLM, the Fisher information is defined when item i provided to an examinee with ability θ using the following equations[Birnbaum 68].

$$I_i(\theta) = \frac{[p'(u_i = 1|\theta)]^2}{p(u_i = 1|\theta)[1 - p(u_i = 1|\theta)]} \tag{2}$$

where

$$p'(u_i = 1|\theta) = \frac{\partial}{\partial\theta} p(u_i = 1|\theta). \tag{3}$$

That result implies that the examinee ability can be ascertained near ability θ using an item with much Fisher information $I_i(\theta)$. Accordingly, it is expected that ability estimation can be implemented by selecting items with much Fisher information at a given ability for each examinee. Based on this concept, an item selection method of computerized based testing, CAT, presents items with much Fisher information. The total of Fisher information of an item set contained in a test presented to an examinee is called test information, which represents the test estimation accuracy.

3. Computerized Adaptive Testing

In CAT, items are selected from a given item set with known item parameters, using the following procedures.

1. The examinee ability is initialized.
2. An item that maximizes Fisher information for given ability is selected from the item pool and is presented to an examinee.
3. The estimated ability of the examinee is updated from the correct/wrong answer data to the item.
4. Procedures 2 and 3 are repeated until the update difference of the estimated ability of the examinee reaches a constant value ϵ or less.

Consequently, for a small number of items to be selected compared with a fixed test, repeating item selection based on maximizing information and estimation of an examinee ability engenders high ability estimation accuracy.

Unfortunately, in CAT, it is highly likely that the same set of items will be selected for examinees who have equal ability. Conventional CAT cannot be used practically under situations in which the same examinee can take a test multiple times. In addition, to follow the normal distribution for ability items with high information, the average value $\theta = 0$ is frequently selected. Therefore, some items in an item pool might not be used effectively. Excessive exposure of items leads to disclosure of the item contents to examinees, and might therefore degrade the test reliability[Way 98].

To resolve this difficulty, we propose a CAT method that minimizes item exposure. It can select a different item while maintaining the same accuracy, even for an examinee with equal ability.

4. Proposed Method

The concept of the proposed method is to assemble a group with test constraints as Fisher information, but with different items using integer programming, to select an item with higher information, and to minimize item exposure from the group. Details of the proposed method are described below.

1. The estimated ability of an examinee is initialized.
2. The group maximizing Fischer information is assembled using the integer programming presented below.

$$\text{Minimize } y = \sum_{i=1}^{I} e_i x_i \tag{4}$$

subject to

$$\sum_{i=1}^{I} I(\theta_k) x_i \geq r_k, k = 1, 2, \cdots, K \tag{5}$$

$$\sum_{i=1}^{I} I(\theta_k) x_i \leq s_k, k = 1, 2, \cdots, K \tag{6}$$

$$\sum_{i=1}^{I} x_i = n(\text{Test length}) \tag{7}$$

To expand the method proposed by Adema (1989) [Adema 92] in which test forms are assembled using integer programming, we propose an optimization problem in which we embed an objective function minimizing item exposure. The lower boundaries and the upper boundaries of test information function at a set of the examinee's ability, $\Theta = \{\theta_1, \ldots, \theta_K\}$, is r_k and s_k. Also, $I(\theta_k)$ denotes the test information function at the examinee's ability θ_k. The exposure count of item i is e_i. If item i is selected into the group, then $x_i = 1$; otherwise $x_i = 0$.

3. An item is selected from a group. Then response data are generated with the given true ability and item parameters.
4. Ability $\hat{\theta}$ is estimated by expected a posteriori (EAP) [Baker 04].

表 1: Results obtained using simulation data

Item pool size	methods	Avg. test length		non-uniformity of estimation	Max. No. exposure item	Avg. exposure item	
500	CAT	19.99	(2.26)	0.03	1000	39.98	(120.93)
	CCAT	23.99	(6.2)	0.21	60	47.98	(23.15)
	Proposal	9.35	(2.15)	0.09	65	18.7	(17.1)
1000	CAT	21.33	(2.15)	0.03	1000	21.33	(82.22)
	CCAT	28.37	(8.98)	0.86	30	28.37	(6.75)
	Proposal	9.73	(2.46)	0.09	42	9.73	(8.74)
2000	CAT	22.48	(2.1)	0.02	1000	11.24	(57.47)
	CCAT	28.56	(11.35)	2.33	15	14.28	(3.19)
	Proposal	9.61	(2.66)	0.08	19	4.8	(4.02)

表 2: Results obtained using actual data

Item pool size	methods	Avg. test length		non-uniformity of estimation	Max. No. exposure item	Avg. exposure item	
978	CAT	14.82	(3.41)	0.05	1000	15.15	(78.5)
	CCAT	26.56	(6.36)	0.19	30	27.15	(8.42)
	Proposal	11.99	(3.15)	0.07	75	12.26	(10.31)

5. Procedures 2–4 are repeated until the update difference of the estimated ability decreases to ϵ or less.

The proposed method selects an item from a different group for each examinee. Accordingly, a different item is expected to be selected, even to an examinee of equivalent ability. Furthermore, improved diversity of item selection is anticipated to encourage thorough use of items in an item pool and to mitigate deviation in exposure.

5. Simulation Experiment

The simulation experiment procedure is the following.

1. An item pool comprising 500, 1000, or 2000 items is generated. The true values of parameters of each item are set randomly from $a_i \sim U(0,1)$, $b_i \sim N(0,1)$.
2. The true ability of an examinee is sampled from $\theta \sim N(0,1)$.
3. The estimated ability of an examinee is initialized to $\hat{\theta} = 0$.
4. An item is selected from an item pool using each method. Then response data are generated with the given true ability and item parameters.
5. The ability $\hat{\theta}$ is estimated using EAP.
6. Procedures 4 and 5 are repeated until the update difference of the estimated ability decreases to ϵ or less. Also, ϵ is set to 0.05, which is used conventionally for actual CAT[Linden 10].
7. Procedures 2–6 are repeated 1000 times to obtain statistical values for the following indices using a delivery pattern and answer data obtained: a) the length of a test, b) the non-uniformity of ability estimation accuracy, and c) the exposure of each item.

Table 1 presents the results. Results confirmed that the proposed method yielded the shortest test length under all conditions. Selecting an item with less information for the initial value of θ rather than selecting an item with more information is known to achieve faster convergence of ability estimation when the initial value of θ is distant from the true ability of an examinee. The proposed method constrains the number of items with uniformity conditions. Therefore, it has a property of not selecting items with an extremely large amount of information only to a certain estimated value. This property shortens the test length, thereby reducing the exposure of items. The proposed method also indicates the smallest standard deviation of test length under all conditions. The proposed method selects items from an item set of uniform information. Therefore, it can render the number of items to take for convergence of the estimated ability uniform.

The non-uniformity of estimation using conventional CAT was lowest. It repeatedly selects some item sets with much information. However, the proposed method controls exposure and selects different items for each examinee. The non-uniformity of estimation was low because of the deviation of measurement accuracy in the selected items.

The maximum exposure was the least when CCAT was used. Results demonstrate that CCAT can constrain maximum exposure directly in the same manner as the fraction of different items. This result is interpreted as attributable to the exposure setting to use as many items in an item pool as possible in this experiment. The averages of exposure obtained using the proposed method were the lowest. Actually, CCAT constrained only the maximum of expo-

sure directly, so that the deviation of exposure could not be controlled.

6. Simulation Using Real Data

This chapter explains evaluation of the effectiveness of the proposed method using real data. An experiment was conducted using an item pool of real data. The item pool contained 978 items. Table 2 presents the experimentally obtained results. Table 2 suggests the following characteristics: 1) The test length produced using the proposed method is the shortest. The standard deviation of test length produced using the proposed method is the smallest, which suggests that the test length selected to examinees has little dispersion. 2) Non-uniformity of estimation by the proposed method is the second highest to CAT, so the same accuracy was maintained using the proposed method. 3) The maximum exposure was smallest by CCAT. The average tended to be small when obtained using the proposed methods.

7. Conclusions

This study has examined a proposed CAT implementation in which different items can be selected, even by an examinee with equivalent ability, while maintaining equal accuracy. Specifically, we proposed a method by which a group is assembled using integer programming: an item is selected from the group. Using simulation experiments and experiments using real data, some points have been verified as benefits of the proposed method.

参考文献

[Adema 92] Adema, J. J.: Implementations of the branch-and-bound method for test construction problems, *Methodika*, Vol. 6, No. 2, pp. 99–117 (1992)

[Baker 04] Baker, F. B. and Kim, S.-H. eds.: *Item Response Theory: Parameter Estimation Techniques*, CRC Press (2004)

[Birnbaum 68] Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability, in Lord, F. M. and Novick, M. R. eds., *Statistical theories of mental test scores*, pp. 397–479, Addison-Wesley (1968)

[Linden 98] Linden, van der W. J. and Reese, L. M.: A model for optimal constrained adaptive testing, *Applied Psychological Measurement*, Vol. 22, No. 3, pp. 259–270 (1998)

[Linden 10] Linden, van der W. J. and Glas, C. A. W. eds.: *Elements of Adaptive Testing*, Springer (2010)

[Linden 16a] Linden, van der W. J. ed.: *Handbook of Item Response Theory, Volume One: Models*, Chapman and Hall/CRC (2016)

[Linden 16b] Linden, van der W. J. ed.: *Handbook of Item Response Theory, Volume Two: Statistical Tools*, Chapman and Hall/CRC (2016)

[Lord 68] Lord, F. and Novick, M. R.: *Statistical Theories of Mental Test Scores*, Addison-Wesley (1968)

[Lord 80] Lord, F. M.: *Applications of Item Response Theory To Practical Testing Problems*, Lawrence Erlbaum Associates (1980)

[Ueno 10] Ueno, M. and Songmuang, P.: Computerized adaptive testing based on decision tree, in *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*, pp. 191–193 (2010)

[Ueno 13] Ueno, M.: Adaptive testing based on bayesian decision theory, in *International Conference on Artificial Intelligence in Education*, pp. 712–716 (2013)

[Way 98] Way, W. D.: Protecting the integrity of computerized testing item pools, *Educational Measurement: Issues and Practice*, Vol. 17, pp. 17–27 (1998)

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Maximizing accuracy of group peer assessment using item response theory and integer programming

Masaki Uto*1 Duc-Thien Nguyen*1 Maomi Ueno*1

*1 University of Electro-Communications

With the wide spread of large-scale e-learning environments, peer assessment has been widely used to measure learner ability. When the number of learners increases, peer assessment is often conducted by dividing learners into multiple groups. However, in such cases, the peer assessment accuracy depends on the method of forming groups. To resolve that difficulty, this study proposes a group formation method to maximize peer assessment accuracy using item response theory and integer programming. Experimental results, however, have demonstrated that the method does not present sufficiently higher accuracy than a random group formation method does. Therefore, this study further proposes an external rater assignment method that assigns a few outside-group raters to each learner after groups are formed using the proposed group formation method. Through results of simulation and actual data experiments, this study demonstrates that the method can substantially improve peer assessment accuracy.

1. Introduction

As an assessment method based on a social constructivist approach, peer assessment, which is mutual assessment among learners, has become popular in recent years. One common use of peer assessment is for summative assessment. The importance of this usage has been increasing concomitantly with the wider use of large-scale e-learning environments [Suen 14, Shah 14]. Peer assessment, however, entails the difficulty that the assessment accuracy of learner ability depends on rater characteristics such as severity and consistency. To resolve that difficulty, item response theory (IRT) models incorporating rater parameters have been proposed [Eckes 11, Uto 18]. The IRT models are known to provide more accurate ability assessment than average or total scores do because they can estimate the ability considering rater characteristics [Uto 16, Uto 18].

In learning contexts, peer assessment has often been adopted for group learning situations such as collaborative learning and active learning [Staubitz 16, Suen 14, Nguyen 15]. Specifically, learners are divided into multiple groups in which they work together, and peer assessment is conducted within the groups. However, in such cases, the ability assessment accuracy depends also on a way to form groups. For example, if a group consists of learners who tend to assess others randomly, their abilities are difficult to be estimated accurately. Therefore, group optimization is important to maximize the accuracy of peer assessment. However, no studies have focused on this issue.

For the reason, this study proposes a new group formation method that maximizes peer assessment accuracy based on IRT. Specifically, the method is formulated as an integer programming (IP) problem that maximizes the lower bound of the Fisher information (FI) measure: a widely used index of ability assessment accuracy in IRT. The method is expected to improve the ability assessment accuracy because groups are formed so that the learners in the same group can assess one another accurately. However, experimental results demonstrated that the method did not present sufficiently higher accuracy than that of a random group formation method. The result suggests that it is generally difficult to assign raters with high FI to all learners when peer assessment is conducted only within groups.

To alleviate that shortcoming, this study further proposes an external rater assignment method that assigns a few optimal outside-group raters to each learner after forming groups using the method presented above. We formulate the method as an IP problem that maximizes the lower bound of the FI for each learner given by assigned outside-group raters. Simulations and actual data experiments demonstrate that assigning a few optimal external raters using the proposed method can improve the peer assessment accuracy considerably.

2. Peer assessment data

This study assumes that peer assessment data $\boldsymbol{U}$ consists of rating categories $k \in \mathcal{K} = \{1, \cdots, K\}$ given by each peer-rater $r \in \mathcal{J} = \{1, \cdots, J\}$ to each learning outcome of learner $j \in \mathcal{J}$ for each task $t \in \mathcal{T} = \{1, \cdots, T\}$. Letting u_{tjr} be a response of rater r to learner j's outcome for task t, the data $\boldsymbol{U}$ are described as $\boldsymbol{U} = \{u_{tjr} \mid u_{tjr} \in \mathcal{K} \cup \{\text{-1}\}, t \in \mathcal{T}, j \in \mathcal{J}, r \in \mathcal{J}\}$, where $u_{tjr} = -1$ denotes missing data.

Furthermore, this study assumes that peer assessment is conducted by dividing learners into multiple groups for each task $t \in \mathcal{T}$. Here, let x_{tgjr} be a dummy variable that takes 1 if learner j and peer r are included in the same group $g \in \mathcal{G} = \{1, \cdots, G\}$ for task t, and which takes 0 otherwise. Then peer assessment groups for task t can be described as $\boldsymbol{X}_t = \{x_{tgjr} \mid x_{tgjr} \in \{0, 1\}, g \in \mathcal{G}, j \in \mathcal{J}, r \in \mathcal{J}\}$. Consequently, when peer assessment is conducted among group members, the rating data u_{tjr} become missing data if learners j and r are not in the same group ($\sum_{g=1}^{G} x_{tgjr} = 0$).

The purpose of this study is to estimate the learner ability accurately using IRT for peer assessment [Uto 16] from the data $\boldsymbol{U}$ by optimizing the groups $\boldsymbol{X} = \{\boldsymbol{X}_t \mid t \in \mathcal{T}\}$.

Contact: Masaki Uto, 1-5-1, Choufugaoka, Choufu-shi, Tokyo, Japan, 042-443-5627, uto@ai.lab.uec.ac.jp

3. IRT for peer assessment

The IRT for peer assessment [Uto 16] has been formulated as a graded response model that incorporates rater parameters. The model defines the probability that rater r responds in category k to learner j's outcome for task t as

$$P_{tjrk} = P^*_{tjrk-1} - P^*_{tjrk}, \tag{1}$$
$$P^*_{tjrk} = [1 + \exp(-\alpha_t \gamma_r (\theta_j - \beta_{tk} - \varepsilon_r))]^{-1}$$

Here, θ_j denotes the ability of learner j; γ_r reflects the consistency of rater r; ε_r represents the severity of rater r; α_t is a discrimination parameter of task t; and β_{tk} denotes the difficulty in obtaining category k for task t ($\beta_{t1} < \cdots < \beta_{tK-1}$); $P^*_{tjr0} = 1$, and $P^*_{tjrK} = 0$.

In IRT, the standard error estimate of ability assessment is defined as the inverse square root of the FI. More information implies less error of the assessment. Therefore, FI can be regarded as an index of the ability assessment accuracy. For the above model, FI of rater r in task t for a learner with ability θ_j is calculable as

$$I_{tr}(\theta_j) = \alpha_t^2 \gamma_r^2 \sum_{k=1}^{K} \frac{\left(P^*_{tjrk-1} Q^*_{tjrk-1} - P^*_{tjrk} Q^*_{tjrk}\right)^2}{P^*_{tjrk-1} - P^*_{tjrk}}, \tag{2}$$

where $Q^*_{tjrk} = 1 - P^*_{tjrk}$.

The FI of multiple raters for learner j in task t is definable by the sum of the information of each rater. Therefore, when peer assessment is conducted within group members, the FI for learner j in task t is calculable as shown below.

$$I_t(\theta_j) = \sum_{\substack{r=1 \\ r \neq j}}^{J} \sum_{g=1}^{G} I_{tr}(\theta_j) x_{tgjr} \tag{3}$$

A high value of FI $I_t(\theta_j)$ signifies that the group members can assess learner j accurately. Therefore, if we form groups to provide great amounts of FI for each learner, then the ability assessment accuracy can be maximized.

4. Group formation method

Based on this idea presented above, we formulate the group formation optimization method (designated as *PropG*) as an IP problem that maximizes the lower bound of FI for each learner. Specifically, *PropG* for task t is formulated as the following IP problem.

$$\text{maximize} \quad y_t \tag{4}$$
$$\text{subject to} \quad \sum_{\substack{r=1 \\ r \neq j}}^{J} \sum_{g=1}^{G} I_{tr}(\theta_j) x_{tgjr} \geq y_t, \quad \forall j, \tag{5}$$
$$\sum_{g=1}^{G} x_{tgjj} = 1, \quad \forall j, \tag{6}$$
$$n_l \leq \sum_{j=1}^{J} x_{tgjj} \leq n_u, \quad \forall g, \tag{7}$$
$$x_{tgjr} = x_{tgrj}, \quad \forall g, j, r \tag{8}$$

The first constraint requires that FI for each learner j be larger than a lower bound y_t. The second constraint restricts each learner as belonging to one group. The third constraint controls the number of learners in a group. Here, n_l and n_u represent the lower and upper bounds of the number of learners in group g. In this study, $n_l = \lfloor J/G \rfloor$ and $n_u = \lceil J/G \rceil$ are used so that the numbers of learners in respective groups become as equal as possible. This IP maximizes the lower bound of FI for learners. Therefore, by solving the problem, one can obtain groups that provide as much FI as possible to each learner.

4.1 Evaluation of group formation methods

To evaluate the effectiveness of *PropG*, we conducted the following simulation experiment. 1) For $J = 30$ and $T = 5$, the true IRT model parameters were generated randomly. 2) For the first task $t = 1$, learners were divided into $G \in \{3, 4, 5\}$ groups using *PropG* and a random group formation method (designated as *RndG*). For *PropG*, the FI values were calculated using the true parameter values. 3) Given the created groups and the true model parameters, peer assessment data were sampled randomly for the current task t based on the IRT model. 4) Given the true rater and task parameters, the learner ability was estimated from the data generated to date. 5) RMSE between the estimated ability and the true ability were calculated. 6) Procedures 2) – 5) were repeated for the remaining tasks. 7) After 10 repetitions of the procedures described above, the average values of RMSE were calculated.

Fig. 1 presents the results. Results demonstrate that RMSE decreases with the decreasing number of groups G or with increasing numbers of tasks or learners because the number of data for each learner increases. Generally, the increase of data per learner is known to engender improvement of the ability assessment accuracy [Uto 16]. Comparing the group formation methods, however, *PropG* does not decrease RMSE sufficiently. The results indicate that it is difficult to form groups to sufficiently increase the peer assessment accuracy. To overcome this shortcoming, we further propose the assignment of outside-group raters to each learner, given the groups created using *PropG*.

5. External rater assignment

The proposed external rater assignment method (designated as *PropE*) is formulated as an IP problem that maximizes the lower bound of information for learners given by the assigned outside-group raters. Specifically, given a group formation $\boldsymbol{X}_t$, *PropE* for task t is defined as folllows.

$$\text{maximize :} \quad y'_t \tag{9}$$
$$\text{subject to :} \quad \sum_{r \in \boldsymbol{C}_{tj}} I_{tr}(\theta_j) z_{tjr} \geq y'_t, \quad \forall j \tag{10}$$
$$\sum_{r \in \boldsymbol{C}_{tj}} z_{tjr} = n^e, \quad \forall j \tag{11}$$
$$\sum_{j=1}^{J} z_{tjr} \leq n^J, \quad \forall r \tag{12}$$
$$z_{tjj} = 0, \quad \forall j \tag{13}$$

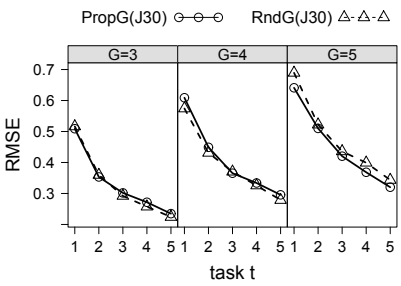


Figure 1: RMSE values of group formation methods in the simulation experiment.

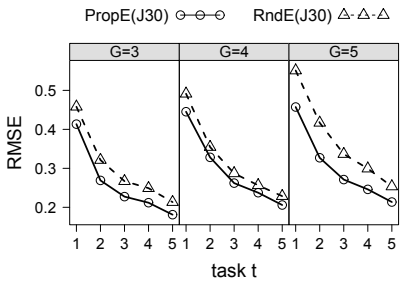


Figure 2: RMSE values of external rater assignment methods for each G and t in the simulation experiment.

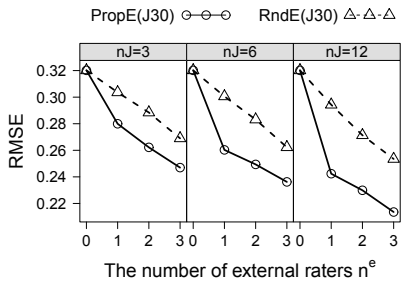


Figure 3: RMSE values of external rater assignment methods for each n^J and n^e in the simulation experiment.

Here, $\boldsymbol{C}_{tj} = \{r \mid \sum_{g=1}^{G} x_{tgjr} = 0\}$ is the set of outside-group raters for learner j in task t given a group formation $\boldsymbol{X}_t$. In addition, z_{tjr} is a variable that takes 1 if external rater r is assigned to learner j in task t; it takes 0 otherwise. Furthermore, n^e denotes the number of external raters assigned to each learner; n^J is the upper limit number of outside-group learners assignable to each rater. Here, n^e and n^J must satisfy $n^J \geq n^e$. The increase of n^J makes it easier to assign optimal raters to each learner, although differences in the workload among the learners increases.

The first constraint in the IP restricts that the FI for each learner given by the assigned outside-group raters must exceed a lower bound y'_t. The second constraint requires that n^e number of outside-group raters must be assigned to each learner. The third constraint restricts that each learner can assess at most n^J number of outside-group learners. The objective function is defined as the maximization of the lower bound of the FI for learners given by assigned external raters. Therefore, by solving the proposed method, an external rater assignment z_{tjr} is obtainable so that n^e outside-group raters with high FI are assigned to each learner.

5.1 Evaluation of external rater assignment

To evaluate the performance of the proposed method, we conducted the following simulation experiment, which is similar to that conducted in 4.1. 1) For $J = 30$ and $T = 5$, the true model parameters were generated randomly. 2) For the first task $t = 1$, learners were divided into $G \in \{3, 4, 5\}$ groups using *PropG*. Then, given the created groups, $n^e \in \{1, 2, 3\}$ outside-group raters were assigned to each learner using *PropE* and a random assignment method (designated as *RndE*). Here, we changed the value of n^J for $\{3, 6, 12\}$ to evaluate its effects. In *PropG* and *PropE*, FI was calculated using the true parameter values. 3) Peer assessment data were sampled randomly for current task t following the IRT model, given the true model parameters, the formed groups and the rater assignment. 4) The following procedures were identical to procedures 4) – 7) of the previous experiment.

Fig. 2 shows the RMSE for each t and G when $n^J = 12$ and $n^e = 3$, and Fig. 3 shows the RMSE for each n^e and n^J when $G = 5$ and $t = 5$. In Fig. 3, the results for $n^e = 0$ correspond to *PropG*. Results show that the accuracy of the external rater assignment methods tends to increase concomitantly with decreasing number of groups and increasing number of tasks and assigned external raters n^e because the number of rating data for each learner increases. Furthermore, Fig. 3 shows that both external rater assignment methods reveal the lower RMSE than *PropG* in all cases, which suggests that the addition of the external raters is effective to improve the ability assessment accuracy. Comparison of the external rater assignment methods reveals that *PropE* presented higher accuracy than *RndE* in all cases. Furthermore, the RMSE difference between *PropE* and *RndE* tends to increase with increasing n^J value because the increase of n^J makes it easier to assign optimal raters to each learner.

From those results, we infer that the proposed method can improve the peer assessment accuracy efficiently when a large value of n^J and a small value of n^e are given.

6. Usage in actual e-learning situations

PropG and *PropE* require IRT parameter estimates to calculate FI. Although the experiments described above used the true parameter values, they are practically unknown. Therefore, this section presents a description of how to use *PropG* and *PropE* when the IRT parameters are unknown in actual e-learning situations. We consider the following two assumptions for using *PropG* and *PropE* in an e-learning course. 1) More than one task is offered in the course. 2) All tasks were used in past e-learning courses at least once, and past learners' peer assessment data corresponding to the tasks were collected. Although the second assumption might not necessarily be satisfied in practice, it is necessary to estimate the task parameters.

Under the second assumption, we can estimate the task parameters. Given task parameter estimates, we can use *PropG* and *PropE* through the following procedures under the first assumption. 1) For the first task, peer assessment is conducted using randomly formed groups. 2) The rater parameters and learner ability are estimated from the obtained peer assessment data. 3) For the next task, group formation and external rater assignment are conducted using *PropG* and *PropE* given the parameter estimates. 4) Repeat procedures 2) and 3) for remaining tasks.

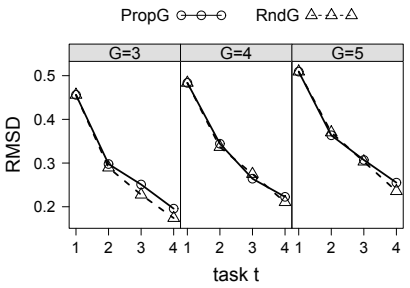


Figure 4: RMSDs of group formation methods in the actual data experiment.

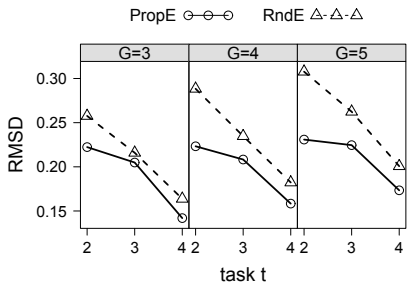


Figure 5: RMSDs of external rater assignment methods for each G and t in the actual data experiment.

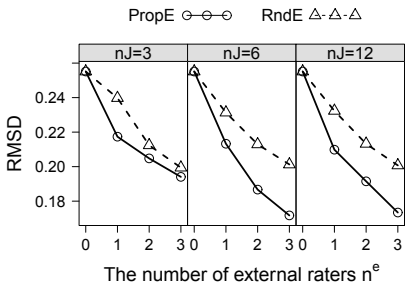


Figure 6: RMSDs of external rater assignment methods for each n^J and n^e in the actual data experiment.

7. Actual data experiment

This section evaluates the effectiveness of *PropG* and *PropE* using actual peer assessment data based on the above usage. We gathered actual data using the following procedures. 1) As subjects for this study, 34 university students were recruited. 2) They were asked to complete four essay writing tasks offered in NAEP. 3) After the participants completed all tasks, they were asked to evaluate the essays of all other participants for all four tasks using a rubric with five rating categories. Furthermore, we collected additional rating data (designated as *five raters' data*) for task parameter estimation. The data consist of ratings assigned by 5 graduate school students to the essays gathered in the experiment above.

Using the actual data, we conducted the following experiments. 1) The task parameters in the IRT model were estimated using the five raters' data. 2) Given the task parameter estimates, the rater parameters and learner ability were estimated using the full peer assessment data. 3) For the first task, $G \in \{3, 4, 5\}$ groups were created randomly. 4) The peer assessment data without peer-rater assignment were changed to missing data. 5) From the peer assessment data up to the current task, the rater parameters and learner ability were estimated given the task parameters estimated in Procedure 1). 6) RMSD between the ability estimates and that estimated from the complete data in Procedure 2) was calculated. 7) For the next task, $G \in \{3, 4, 5\}$ groups were formed by *PropG* and *RndG*. Then, given the groups formed by *PropG*, $n^e \in \{1, 2, 3\}$ external raters were assigned to learners by *PropE* and *RndE* under $n^J \in \{3, 6, 12\}$. Here, *PropG* and *PropE* used the task parameters obtained in Procedure 1) and the current estimates of ability and rater parameters to calculate FI. 8) For the remaining tasks, procedures 4) – 7) were repeated. 9) After repeating the procedures described above 10 times, the average values of the RMSD were calculated.

Fig. 4 presents results of each group formation method. Figs. 5 and 6 show those of the external rater assignment methods. Fig. 5 presents results for each $t \geq 2$ and $G \in \{3, 4, 5\}$ when $n^J = 12$ and $n^e = 3$. Fig. 6 shows those for each n^e and n^J when $G = 5$ and $t = 4$. Results show similar tendencies to those obtained from the simulation experiments. Specifically, comparing the group formation methods, *PropG* does not improve the accuracy much, while the assessment accuracy is improved drastically by introducing external raters. Furthermore, the proposed external rater assignment method realizes the higher accuracy than the random assignment method when n^J is large and n^e is small.

8. Conclusion

This study proposed the group formation method and external rater assignment method to improve peer assessment accuracy using IRT and IP. The experimentally obtained results showed that the external rater assignment method, which assigns a few optimal outside-group raters to each learner, improved the accuracy dynamically, although the proposed group formation method did not improve the accuracy sufficiently.

References

[Eckes 11] Eckes, T.: *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*, Peter Lang (2011)

[Nguyen 15] Nguyen, T., Uto, M., Abe, Y., and Ueno, M.: Reliable Peer Assessment for Team-project-based Learning using Item Response Theory, pp. 144–153 (2015)

[Shah 14] Shah, N. B., Bradley, J., Balakrishnan, S., Parekh, A., Ramchandran, K., and Wainwright, M. J.: Some Scaling Laws for MOOC Assessments, in *KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014)* (2014)

[Staubitz 16] Staubitz, T., Petrick, D., Bauer, M., Renz, J., and Meinel, C.: Improving the peer assessment experience on MOOC platforms, in *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pp. 389–398ACM (2016)

[Suen 14] Suen, H. K.: Peer assessment for massive open online courses (MOOCs), *The International Review of Research in Open and Distributed Learning*, Vol. 15, No. 3, pp. 312–327 (2014)

[Uto 16] Uto, M. and Ueno, M.: Item Response Theory for Peer Assessment, Vol. 9, No. 2, pp. 157–170 (2016)

[Uto 18] Uto, M. and Ueno, M.: Empirical comparison of item response theory models with rater's parameters, *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32 (2018)

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Waveform Processing of Electrocardiogram with Neural Network and Non-contact Measurement using Kinect for Driver Evaluation

Hao ZHANG [*1]

Takashi IMAMURA [*2]

[*1]Graduat School of Science and Technology, Niigata University

[*2] Faculty of Engineering, Niigata University

In recent years, the measurement method of biological information using non-contact sensor have become popular. And, in-vehicle measurement systems for a driver evaluation are also needed in order to decrease the serious traffic accidents. In this research, improvement method of heart rate measurement with high accuracy by using the non-contact sensor “Kinect” is examined. As the reliable biological information, the Holter monitor of electrocardiograph is combination used for measurement experiment. Moreover, to extract high level features from electrocardiogram, a neural network based autoencoder has also constructed. In this paper, about the constructed neural network, the effectiveness of abnormal R wave detection through learning for waveform characteristics of electrocardiogram and the possibility of application for reconstruction of the heart rate signal from Kinect are discussed.

1. Introduction

According to a highway survey of 1000 drivers, the driver who experienced sleepiness during the operation was 78% of the total. In addition, "Front Carelessness" occupies about half of all accidents and about 40% in the death accident. It is thought that the factor which exists in this "front carelessness" is different by the driver, and there is insufficient attention and concentration as one of the big factors. Therefore, it is necessary to improve the attention to the operation, to operation of the quick handle, and the quick brake, the existence of the obstacle on the road, the sudden change of the curvature of the road, so maintenance and control of tension are necessary.

Recently, the driver's load is reduced by the various driving assistance, there is a possibility that attention is insufficient by lowering the tension. On the other hand, if the tension is too high, there may be a possibility that the physical strength and the mental state are exhausted early, and sleepiness may occur. It is thought that the extreme state of the tension seems to lead to the lowering of the accident avoidance capacity. It is said that the tension of the driver can be evaluated using electroencephalogram, blood pressure, electrocardiogram (ECG) [Deguchi 2006]. The purpose of this study is to develop a system to evaluate the feeling of tension by measuring the R-R interval variation of drivers during driving.

In this research, two kinds of technique to measure R-R interval variation are used. First one is acquisition for ECG with an electrocardiograph by using multifunctional wireless Holter monitor and recorder "CarPod" (Medilink inc.). It is an typical method of contact measurement, and it is also used in medical test for sleep apnea by its high accuracy. It can be used in daily life environment without disturbance for several behaviors or works. However, it is not suitable for the actual situation of vehicle driver monitor. Because, it is required to put the 3 or 5 electrodes on human skin directory.

Second one is sensing method using non-contact sensing such as Infrared sensors or cameras. "Kinect" (Microsoft Corp.) is typical example of such kinds of sensor which has already been put into practical use as a product of home game interface. And several researchers were tried to acquire the heart rate waveform by the changes of the luminance value (RGB) from the human face [Nakamura 2015].

On the other hand, although it is easy to acquire the electrocardiogram itself, there are still many unresolved parts about the analysis method and application method of information obtained from the detected waveform. Fundamental analysis is based on frequency analysis. For example, Yokoyama et al. Proposes a method to evaluate heart rate variability using the power spectrum which is frequency information of electrocardiogram and the amplitude spectrum extracted from the amplitude information [Kiyoko 1999]. In addition, recent developments in deep learning have been studied and applied to electrocardiograms. Takahashi et al. Has proposed the feature extraction from the electrocardiogram using stained convolutional extraction autoencoder (SCDAE) [Takahashi 2017]. Their research showed that robust feature extraction is possible for waveform changes due to observation objects and heart rate variability by using autoencoder.

In this paper, we propose a neural network (NN) - based autoencoder, which extracts high - level features from normal waveforms, and detects a location with abnormal electrocardiograms compared to required examination data. And, it is also possible to compare the electrocardiogram feature extracted by the autoencoder with the heartbeat waveform obtained from Kinect and to analyze the heartbeat characteristics from the RGB information.

Contact: ZHANG Hao, Niigata University,8050 Ikarashi-2nocho Nishiku-Niigata-JAPAN,950-2181,025-262-7475, ima@eng.niigata-u.ac.jp

2. The Proposed Method

2.1 Summary of Techniques

The fixed length data divided by the sliding window method in the electrocardiogram waveform does not consider labels for heartbeat interval or arrhythmia, so it is easy to use the acquired data [Noriyasu 2018]. Autoencoder (Ae) is a dimension compression algorithm using neural network. the weight of the AE is trained through neural network layer used to reconstruct the input data, which is of high dimensionality. As a result of the training, we can obtain a higher-level representation of the input data [Takahashi 2017].

The proposed method consists of three steps: Generate Partial Time-series, Autoencoder Pre-learning, Abnormal Inspection and R Wave Detection.

- Generate partial time-series of fixed length data from normal and abnormal data
- Autoencoder pre-learning and extracting normal data features
- Inspecting data abnormally with the trained autoencoder
- Constructing the R wave detector using discrete wavelet transform (MODWT)

(1) Generate Partial Time-series

ECG data X divided into fixed length D by slide window method

$$X=(X_1,\dots,X_R)$$

Here, $X_r$is the r-th section, and R is the total number of spaces. The fixed length D and shift of the sizes of stride S are given as parameters.

(2) Autoencoder Pre-learning

Input $X=(X_1,\dots,X_R)$,We use NN based deep autoencoder in MATLAB 2017A. As shown in Fig. 2, it is possible to obtain a low dimensional feature which holds information representing data.

The encoder section of the autoencoder is represented by encoder (・), and the decoder section is represented by decoder(・). The feature of the r-th section is expressed by

$$Z_r=\mathrm{encoder}(X_r)$$

As a result of applying the decoder to the encoder result Z_r which is the low dimensional feature quantity,

$$Y_r=\mathrm{decoder}(Z_r)$$

is similar to the original signal X_r.

The loss function for calculating how much the neural network matches the original data is the mean square error (mase sparse) between the original signal X_r and the decoded result Y_r,

$$E=\frac{1}{2R}\sum_{r=1}^{R}(X_r-Y_r)^2$$

Which is an indicator of poor performance of the neural network.

(3) Abnormal Inspection

An autoencoder has the function of bringing an abnormal waveform closer to a normal waveform. Using this function, it is possible to check waveform abnormality.

(4) R Wave Detector

The R-wave detector was constructed using the discrete wavelet transform (MODWT) to emphasize the R peak of the ECG waveform.

'Sym4' wavelet is similar to 'QRS' complex, so it is suitable for 'QRS' detection. For comparison, the results of extracted 'QRS' complex and the results with 'sym4' wavelet are shown in Fig.1.

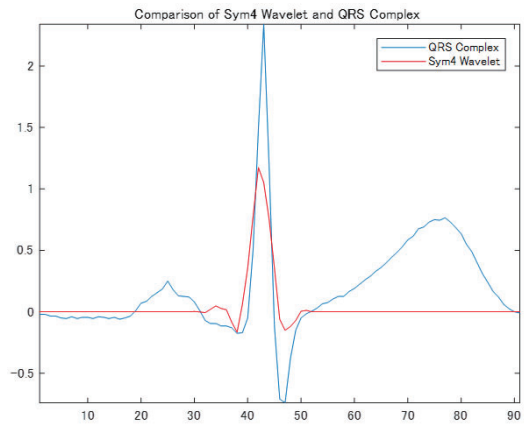


Fig.1 Comparison of 'QRS' and 'sym4' wavelets

3. Experiments

3.1 Measurement Experiment

Measurement experiments were made of electrocardiogram and Kinect. The resting state was maintained for more than 1 hour before starting the measurement. In the measurement at rest, the subject was chosen to be the most comfortable sitting position. The measurement time of the electrocardiograph and Kinect is half hour and one minute. The subjects were healthy 25 years old men without arrhythmia, heart disease, and autonomic neuropathy.

3.2 Numerical Experiment

In this experiment, we rely on MATLAB 2017A, fixed length D = 100 Samples,S =1 Samples in the sliding window system, hidden layer set 160, and an autoencoder was constructed as shown in Fig. 2.

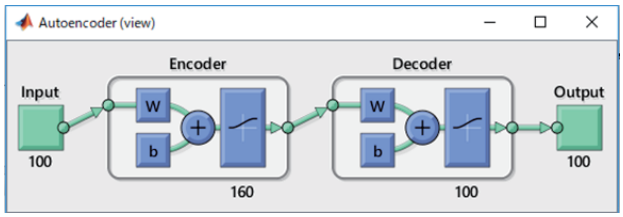


Fig. 2 Block diagram of Autoencoder

The normal part of the measured ECG waveform was chosen, and 10000 samples were trained as training data and 5000 samples were tested.

Fig. 3 shows the learning curve of Ae. The mean square error (mase sparse) between the decoded result Y_r and the original signal X_r is only 3.5195e-05 when the number of calculations is 3000, which shows that high accuracy learning has been performed.

This is because the amount of information can be greatly lost by reducing the dimension of the autoencoder. Fig.4 shows the following. We calculate the square error (MSE) of the restored signal Y_r and the original data X_r by selecting a random one waveform feature. Error division of individual data is shown. It seems to have succeeded in maintaining the waveform characteristic and restoring it.

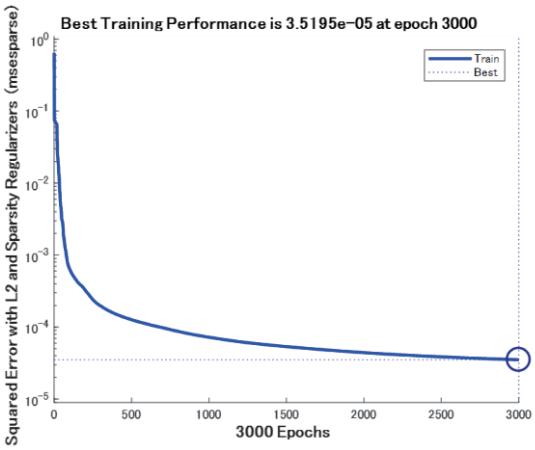


Fig. 3 Learning Curve of Ae

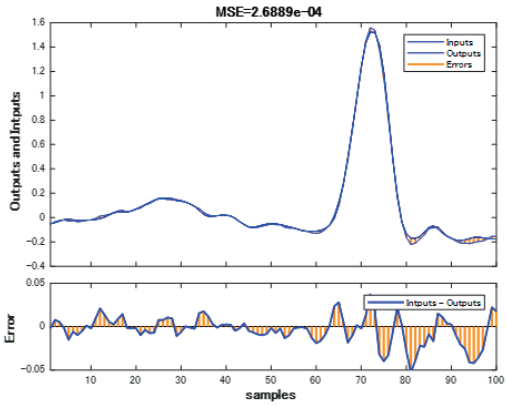


Fig. 4 Waveform Feature

And, the threshold of the error is set to 0.55 (the minimum value exceeding the maximum value of the training data error), and through the test data experiment, the error becomes 0 in the normal case of the electrocardiogram. In addition, two R-wave anomalies were found by the pre-learning Autoencoder from 5-minute data.

Finally, we measured the total heart rate 330 and average heart rate 66 for 5 minutes using MODWT.

4. Heart Rate Measurement Using Kinect in Driving Situation

In order to examine the characteristics of heart rate signal in driving situation, experimental setup using several measurement method as shown in previous section and driving simulator is constructed. In this section, several measurement problem not only signal processing issue but also installation of measurement devices are described through the experimental setups.

4.1 Driving Heart Rate Measurement Using Kinect

As the human heart contracts, the volume of the blood vessel changes. The amount of reflection of light also changes according to the changes. It is able to obtain the reflected light from the capillary of the face with Kinect. Then, it becomes possible to acquire a heartbeat waveform.

However, environmental light conditions will be always change in actual driving situation. And the driver also has driving behavior such as maneuvering the steering wheel on the curve or at the intersection. According to them, drastically changes of the light condition and driving behavior have possibility of making disturbance for visual sensing. Therefore, position of installation for Kinect is also important to consider in order to acquire the driver's face in stable.

4.2 Fixture Design with DS-6000

In order to measure the driver to a better angle with Kinect, the fixture points and its device is designed considering with several structures and dimensions of cockpit of the typical vehicle. In this research, driving simulator "DS-6000" (Mitsubishi Precision Inc.) is used for driving environment. It is simulated for the driving cockpit of typical type of car and their component such as driver's sheet, steering wheel, shift lever and pedals. Especially, driving instruments such as speed and gas meter are also set in the back of steering wheel.

For in-vehicle measurement, it is also important to keep enough fields of view for driver. On the other hand, to make better acquisition of driver's face, it is expected to put the non-contact sensors in front of the face. To satisfy these constraints, the upper side of the front windows is selected and the fixture of Kinect sensor was designed as shown in Fig. 5.

Fig. 5 Experimental setup of Kinect on Driving simulator

4.3 Measurement Result

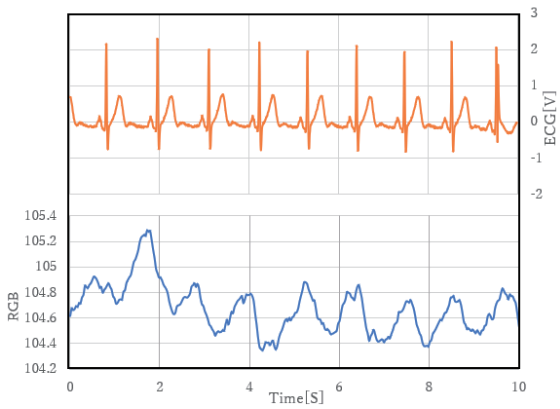


Fig. 6 Comparison of RGB heartrate waveforms and ECG waveforms

The experiment in driving situation with a subject driver has been conducted using constructed experimental setup as shown in previously. The measurement results in this experiment as shown in Fig.6. In this figure, red line shows heart rate waveform obtained by the electrocardiograph, and blue one shows the heart beat waveform measured by Kinect.

Through the simple comparison, heart beat have weakness of overall waveform for typical heart rate signal while heart rate signal has stability. However, it has possibility that the peak points of heart beat can be indicate the R wave characteristic based on the waveform reconstruction techniques.

5. Conclusion

In this paper, we propose a new method of analysis of electrocardiogram waveform data using an NN based autoencoder. At first, the waveform feature of a high level was extracted after the longtime waveform of the electrocardiogram was cut into the fixed length by the slide window system, and the high accuracy automatic encoder was prepared by adjusting various parameters. In addition, abnormal electrocardiogram was passed through a trained autoencoder, and abnormal R wave was accurately detected. Finally, more accurate heart rate information was obtained using MODWT.

As a future problem, the heart rate data of the RGB acquired from the face by Kinect has some correspondence with the electrocardiogram waveform, but it is not easy to detect heartrate. Therefore, it is expected that Kinect's heartbeat data will be processed with an autoencoder that extracts high level waveform features, and concrete connection with electrocardiogram will be examined. With that continuation, we will experiment a plurality of subjects and measure the R-R interval to achieve the final purpose of evaluating the driver's tension.

References

[Deguchi 2006] Mitsuo Deguchi, Junichi Wakasugi, Tatsuya Ikegami, Shinji Nanba, Masaki Yamaguchi, Evaluation of Driver Stress Using Motor-vehicle Driving Simulator, IEEJ Trans, SM,Vol.126,No.8, 2006.

[Nakamura 2015] Koru Nakamura, Tsukasa Sugiura, Takata Tomohiro, Tomoaki Ueda, KINECT for Windows SDK Programming, syuuwa System, 2015.

[Kiyoko 1999] Kiyoko Yokoyama, Junichirou Ushida, Takayoshi Yoshioka, Osamu Ozeki, Yosaku Watanabe, Kazuyuki Takata, Proposal of the Evaluation Method of the Heart Rate Variability using the Frequency Information of the Electrocardiogram, T.IEE Japan, Vol.199-C, 1999.

[Takahashi 2017] Thu Takahashi, Keiichi Ochiai, Yusuke Fukazawa, Feature extraction and arrhythmia classification from 2-lead ECG using stacked convolutional denoising autoencoders, 2017 Information Processing Society of Japan, IPSJ SIG Technical Report, 2017.

[Noriyasu 2018] Noriyasu Omata, Yoshihiro Nakanura, Susumu Shirayama, Extracting Long-term Patterns from Electrocardiogram Waveform using Deep Learning, The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Probability based scaffolding system using Deep Learning

Ryo Kinoshita Maomi Ueno

Graduate school of Informatics and Engineering, The University of Electro-Communications.

Recently, a great deal of interest in the learning science field has arisen in the use of software to scaffold students in complex tasks. However, most of those software tools have been unable to adapt to individuals To solve the problem, IRT-based approaches to predict student's performance have been proposed. These studies show predicting students' correct answer probability with high accuracy is of critical importance. However, IRT-based approach doesn't predict student's performance accurately when the test data are sparse or imbalanced. To achieve high accuracy in those situations, we proposed a novel scaffolding system based on deep learning. We show proposed method can predict student's performance more precisely than traditional IRT method.

1. Introduction

The leading metaphor of human learning has recently been transferred from instructionism to social constructivism. In the context, the Zone of Proximal Development(ZPD) was introduced by [Vygotsky 62, Vygotsky 78]. [Bruner 96] also emphasized the social nature of learning and reported importance of "scaffolding". He defined scaffolding as steps taken to reduce the degrees of freedom in carrying out some task so that children can concentrate on difficult task.

To carry out effective scaffolding, teachers need to estimate student's ability on ZPD and predict their performances after scaffolding. Therefore, [Collins 89] worked on a new assessment method called "dynamic assessment," which provided a cascadig sequence of hints (called graded hints) to understand how much supports students needed to complete tasks.

Recently, a great deal of interest in the learning science field has arisen in the use of software to scaffold students in complex tasks. However, most of those software tools have been unable to adapt to individuals

[Ueno 15, Ueno 18] proposed a novel Item Response Theory(IRT) to represent the individual student's development as the increase of the latent ability variable and then to provide optimal help by predicting the performance given several hints. They assumed that the optimal student's correct answer probability exists for a student's successful performance. Their results revealed that the adaptive hint function is the most effective for learning when they determined 0.5 to be the correct answer probability. Thus, to predict students' correct answer probability with high accuracy is of critical importance.

However, IRT-based approach has some disadvantages. 1) IRT is not robust when students' data are sparse or imbalanced. 2) General IRT models assumes that students' ability has only a single dimensionality

Meanwhile, there are many studies which utilize deep neural networks(deep learning) for educational data mining[Piech 15, Le 18]. They reported high accuracies in various tasks, because deep neural networks can be trained to achieve high accuracy for training data. Furthermore, deep neural networks can includes several abilities in the hidden layers and be relatively robust.

Therefore, we propose a novel scaffolding system for adaptive learning based on deep learning, Deep Response Model(DRM), to predict students' performances with high accuracy. It consists of two independent layers, Student Layer and Item Layer, combined their outputs to predict students' performances

The experiment shows that DRM can predict students' performances more precisely than traditional IRT model in time series data.

2. Dynamic Assessment

The scaffolding process requires dynamic assessment to predict learner performance when a teacher ' s help is presented to them, as explained previously. Brown and her team compared the performance of children's responses to IQ test items under two conditions. The first was "static assessment," which involved children trying to solve problems under conventional test conditions, for which they received no help or guidance. The same children were also tested on the same items under dynamic conditions of providing a series of graded hints. The results demonstrated that dynamic assessment provided a stronger basis for predicting learning outcomes than static measures did. The most important result was that the greatest learning gain tended to be achieved by children who only needed the minimum number of hints. The magnitude of the 'gap' between assisted and unassisted performance indicated by the amount of help needed was therefore prognostic of individual differences in learning outcomes. Assessing how much help a learner needed to succeed provided more decisive information about readiness for learning than determining how often they failed when doing the same, untutored tasks. Consequently, dynamic assessment integrated the assessment of learners' prior knowledge with the task of helping them to learn.

An important difficulty associated with previous studies is that the number of hints needed was not a reliable mea-

連絡先: Ryo Kinoshita,Graduate school of Informatics and Engineering,The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, 182-8585, Japan,kinoshita@ai.is.uec.ac.jp

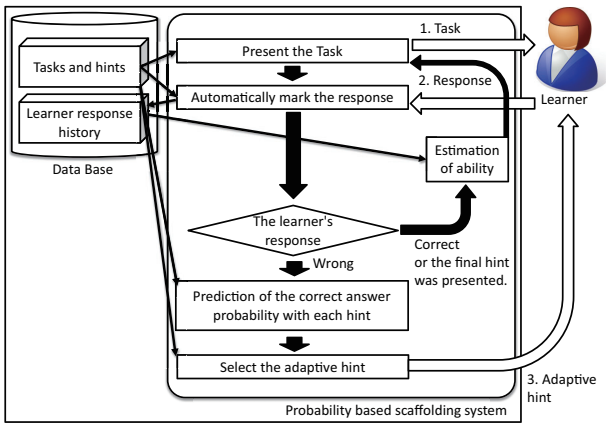


Fig. 1: Outline of the probability-based scaffolding system

sure of dynamic assessment because it depended strongly on the task difficulty. In addition, earlier studies were unable to predict how much support a learner needed for solving a task that had not been presented to the learner.

3. Item Response Theory

This section briefly introduces IRT, a recent test theory based on mathematical models, which is being used widely in various areas and adaptive learning. It has three main benefits:

1)It can assess ability while minimizing the effects of heterogeneous or aberrant items, which have low estimation accuracy.

2)The students' responses to different items can be assessed on the same scale.

3)It predicts the individual probability of correct response to a problem based on past response data.

This section introduces the two-parameter logistic model(2PLM), which is an extremely popular IRT model. For the 2PLM, u_j denotes the response of a student to item j. In this model, $u_j = 0$ shows the student answered item j incorrectly and $u_j = 1$ shows the student answered item j correctly.

In the 2PLM, the probability of correct answer given to item j by a learner with ability $\theta \in (-\infty, \infty)$ is assumed as

$$P_j(u_j = 1|\theta) = \frac{1}{1 + exp(-1.7a_j(\theta - b_j))} \tag{1}$$

where $a_j \in [0, \infty)$ is the jth item's discrimination parameter expressing the discriminatory power for students' abilities of item j, and $b_j \in (-\infty, \infty)$ is the jth item's difficulty parameter expressing the degree of difficulty of item j.

4. Proposed Deep Learning Model

We developed a scaffolding system to solve the programming trace problem. Fig. 1 depicts an outline of the system framework.

Especially, this paper introduce proposed model for estimating students' ability and predicting the correct answer probability. Fig. 2 shows a process of the proposed method. It consists of two neural networks, Student Layer and Item Layer, to separate students' ability from item traits. Then, it combines two outputs to predict students' response.

Proposed method predicts a response $u_{ij} \in \{0, 1\}$ given by student $i \in \{1, ..., I\}$ for item $j \in \{1, ..., J\}$ as follows.

The input of Student Layer, $\boldsymbol{s}_i \in \mathbb{R}^I$ is a one-hot vector, where only the ith element is 1 and the others are 0. Then, it calculates two-layered feed forward networks.

$$\boldsymbol{\theta}_1^{(i)} = tanh\left(\boldsymbol{W}^{(\theta_1)}\boldsymbol{s}_i + \boldsymbol{b}^{(\theta_1)}\right) \tag{2}$$

$$\boldsymbol{\theta}_2^{(i)} = tanh\left(\boldsymbol{W}^{(\theta_2)}\boldsymbol{\theta}_1^{(i)} + \boldsymbol{b}^{(\theta_2)}\right) \tag{3}$$

$$\theta_3^{(i)} = \boldsymbol{W}^{(\theta_3)}\boldsymbol{\theta}_2^{(i)} + b^{(\theta_3)} \tag{4}$$

We use hyperbolic tangent as an activate function:

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{5}$$

$W^{(\theta_1)}$,$W^{(\theta_2)}$are weight matrices:

$$\boldsymbol{W}^{(\theta_1)} = \begin{pmatrix} w_{11}^{(\theta_1)} & w_{12}^{(\theta_1)} & \dots & w_{1I}^{(\theta_1)} \\ w_{21}^{(\theta_1)} & w_{22}^{(\theta_1)} & \dots & w_{2I}^{(\theta_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\theta}_1|1}^{(\theta_1)} & w_{|\boldsymbol{\theta}_1|2}^{(\theta_1)} & \dots & w_{|\boldsymbol{\theta}_1|I}^{(\theta_1)} \end{pmatrix}$$

$$\boldsymbol{W}^{(\theta_2)} = \begin{pmatrix} w_{11}^{(\theta_2)} & w_{12}^{(\theta_2)} & \dots & w_{1|\boldsymbol{\theta}_1|}^{(\theta_2)} \\ w_{21}^{(\theta_2)} & w_{22}^{(\theta_2)} & \dots & w_{2|\boldsymbol{\theta}_1|}^{(\theta_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\theta}_2|1}^{(\theta_2)} & w_{|\boldsymbol{\theta}_2|2}^{(\theta_2)} & \dots & w_{|\boldsymbol{\theta}_2||\boldsymbol{\theta}_1|}^{(\theta_2)} \end{pmatrix}$$

$W^{(\theta_3)}$ is a weight vector.

$$\boldsymbol{W}^{(\theta_3)} = \begin{pmatrix} w_1^{(\theta_3)} & w_2^{(\theta_3)} & \dots & w_{|\boldsymbol{\theta}_2|}^{(\theta_3)} \end{pmatrix}$$

$$\boldsymbol{b}^{(\theta_1)} = \left(b_1^{(\theta_1)}, b_2^{(\theta_1)} ... b_{|\boldsymbol{\theta}_1|}^{(\theta_1)}\right),\ \boldsymbol{b}^{(\theta_2)} = \left(b_1^{(\theta_2)}, b_2^{(\theta_2)} ... b_{|\boldsymbol{\theta}_2|}^{(\theta_2)}\right)$$

are bias parameter vectors, and $b^{(\theta_3)}$ is a bias parameter.

This model considers the last value of Student Layer,$\theta_3^{(i)}$, as a student parameter of i

In the same way, the input of Item Layer, $\boldsymbol{q}_j \in \mathbb{R}^J$ is a one-hot vector, where only the jth element is 1 and the others are 0. Then, it calculates two-layered feed forward networks.

$$\boldsymbol{\phi}_1^{(j)} = tanh\left(\boldsymbol{W}^{(\phi_1)}\boldsymbol{q}_j + \boldsymbol{b}^{(\phi_1)}\right) \tag{6}$$

$$\boldsymbol{\phi}_2^{(j)} = tanh\left(\boldsymbol{W}^{(\phi_2)}\boldsymbol{\phi}_1^{(j)} + \boldsymbol{b}^{(\phi_2)}\right) \tag{7}$$

$$\phi_3^{(j)} = \boldsymbol{W}^{(\phi_3)}\boldsymbol{\phi}_2 + b^{(\phi_3)} \tag{8}$$

$\boldsymbol{W}^{\phi_1}$,$\boldsymbol{W}^{(\phi_2)}$ are weight matrices:

$$\boldsymbol{W}^{(\phi_1)} = \begin{pmatrix} w_{11}^{(\phi_1)} & w_{12}^{(\phi_1)} & \dots & w_{1J}^{(\phi_1)} \\ w_{21}^{(\phi_1)} & w_{22}^{(\phi_1)} & \dots & w_{2J}^{(\phi_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\phi}_1|1}^{(\phi_1)} & w_{|\boldsymbol{\phi}_1|2}^{(\phi_1)} & \dots & w_{|\boldsymbol{\phi}_1|J}^{(\phi_1)} \end{pmatrix}$$

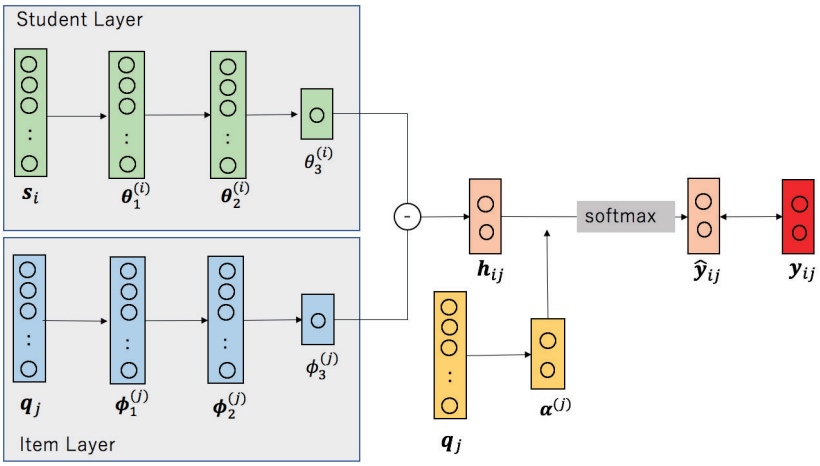


Fig. 2: Proposed Deep Learning model

$$\boldsymbol{W}^{(\phi_2)} = \begin{pmatrix} w_{11}^{(\phi_2)} & w_{12}^{(\phi_2)} & \dots & w_{1|\phi_1|}^{(\phi_2)} \\ w_{21}^{(\phi_2)} & w_{22}^{(\phi_2)} & \dots & w_{2|\phi_1|}^{(\phi_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\phi_2|1}^{(\phi_2)} & w_{|\phi_2|2}^{(\phi_2)} & \dots & w_{|\phi_2||\phi_1|}^{(\phi_2)} \end{pmatrix}$$

$\boldsymbol{W}^{(\phi_3)}$ is a weight vector:

$$\boldsymbol{W}^{(\phi_3)} = \begin{pmatrix} w_1^{(\phi_3)} & w_2^{(\phi_3)} & \dots & w_{|\phi_2|}^{(\phi_3)} \end{pmatrix}$$

$$\boldsymbol{b}^{(\phi_1)} = \left(b_1^{(\phi_1)}, b_2^{(\phi_1)} ... b_{|\phi_1|}^{(\phi_1)}\right), \boldsymbol{b}^{(\phi_2)} = \left(b_1^{(\phi_2)}, b_2^{(\phi_2)} ... b_{|\phi_2|}^{(\phi_2)}\right)$$

are bias parameter vectors, and,$b^{(\phi_3)}$is a bias parameter.

This model considers the last value of Item Layer,$\phi_3^{(j)}$, as an item parameter of j.

Then, a student parameter and an item parameter are combined to predict the student response.

Specifically, proposed model outputs response probabilities $\hat{\boldsymbol{y}}_{\boldsymbol{i,j}} = (\hat{y}_{i,j}^{(0)}, \hat{y}_{i,j}^{(1)})$ using hidden layer $\boldsymbol{h}^{(i,j)} = (h_0^{(i,j)}, h_1^{(i,j)})$ and softmax function.

$$\boldsymbol{h}^{(i,j)} = (W^{(y)})^T(\theta_3^{(i)} - \phi_3^{(j)}) + \boldsymbol{b}^{(y)} \tag{9}$$

$$\begin{aligned} \hat{y}_{i,j}^{(c)} &= softmax(\boldsymbol{\alpha}^{(j)} \circ \boldsymbol{h}^{(i,j)}) \\ &= \frac{exp(\alpha_c^{(j)} * h_c^{(i,j)})}{\sum_c exp(\alpha_c^{(j)} * h_c^{(i,j)})} \end{aligned} \tag{10}$$

$\boldsymbol{W}^{(y)} = (w_1^{(y)}, w_2^{(y)})$, $b^{(y)} = (b_1^{(y)}, b_2^{(y)})$ are weight vector and bias vector. $\circ$ in equation(10) indicates Hadamard product. In the process, we implement attention function $\boldsymbol{\alpha}$, which is calculated based on the question vector $\boldsymbol{q}_j$ as follows:

$$\boldsymbol{\alpha}^{(j)} = softmax(\boldsymbol{W}^{(\alpha)}\boldsymbol{q}_j + \boldsymbol{b}^{(\alpha)}) \tag{11}$$

$\boldsymbol{W}^{(\alpha)}$ is a weight matrix:

$$\boldsymbol{W}^{(\alpha)} = \begin{pmatrix} w_{11}^{(\alpha)} & w_{12}^{(\alpha)} & \dots & w_{1J}^{(\alpha)} \\ w_{21}^{(\alpha)} & w_{22}^{(\alpha)} & \dots & w_{2J}^{(\alpha)} \end{pmatrix}$$

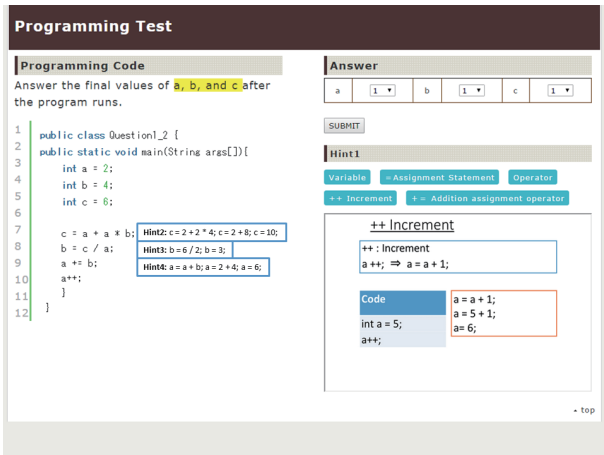


Fig. 3: Dynamic Assessment System

$b^{(\alpha)} = (b_1^{(\alpha)}, b_2^{(\alpha)})$ is a bias vector.

With attention, the model can consider other traits of items like a discrimination or label balance.

This model are trained to reduce the following loss function.

$$Loss = -\sum_i \sum_j \sum_{c \in (0,1)} \boldsymbol{y}_{i,j}^{(c)} \log \hat{\boldsymbol{y}}_{i,j}^{(c)} \tag{12}$$

Loss shows a classification error between response probability $\hat{\boldsymbol{y}}$ and true response $\boldsymbol{y}$, where $\boldsymbol{y}_{i,j}^{(c)} = 1$ when student i answered item j correctly, and otherwise, $\boldsymbol{y}_{i,j}^{(c)} = 0$.

5. Actual Data Experiment

This section presents a description of evaluation of the effectiveness of the proposed model using actual time-series data. Actual data were gathered through the following procedures using the dynamic assessment system we developed. The system is shown in Fig. 3.

(1)82 university students who are not familiar with programming were enrolled.

(2) They tried to solve the 19 items in order with hints about grammar of programming.

Table. 1: Prediction accuracy of learner's performance

Item	2	3	4	5	6	7	8	9	10	
DRM	**52.50%**	**72.15%**	**62.82%**	**58.44%**	75.00%	**61.04%**	**62.67%**	67.11%	**63.16%**	
DRM w/o attention	**52.50%**	**72.15%**	58.97%	57.14%	**76.32%**	**61.04%**	58.67%	65.79%	60.53%	
IRT	51.22%	69.51%	56.10%	53.66%	69.51%	57.32%	58.54%	**68.29%**	57.32%	
Item	11	12	13	14	15	16	17	18	19	Average
DRM	**70.42%**	**67.57%**	**80.26%**	**79.73%**	**87.84%**	65.76%	**97.37%**	**92.00%**	73.68%	**71.64%**
DRM w/o attention	69.01%	**67.57%**	71.05%	**79.73%**	83.78%	**67.57%**	96.05%	**92.00%**	**76.32%**	70.34%
IRT	59.76%	62.20%	76.83%	74.39%	81.71%	60.98%	86.59%	84.15%	73.17%	66.73%

In this experiment, u_{ij} are defined as follows:

$$u_{ij} = \begin{cases} 1 & \text{(student } i \text{ answered item } j \text{ correctly} \\ & \text{without any hints)} \\ 0 & (otherwise) \end{cases}$$

The data includes much missing, which account for over 7%.

We employ Chainer*1, one of frameworks on deep learning, for implementation and optimization of proposed model. The parameters are updated iteratively for 1500 times using batch learning. We use adaptive moment estimation(Adam)[Kingma 14] as an optimizer.

On the other hand, we adopt 2 PLM as a comparative method. The parameters were estimated using Markov chain Monte Carlo algorithm with following prior distributions.

$$\theta \sim N(0,1), \quad \log a \sim N(0,1), \quad b \sim N(1,0.4) \tag{13}$$

$N(\mu, \sigma)$ denotes normal distribution with mean μ and standard deviation σ.

Using the actual data, we calculate the accuracy as following:
1) All models are trained using full data and hold only their item parameters, Item Layer weights and bias.
2) Student parameters are estimated using actual data between item 1 to item $j(j = 1, ..., J-1)$.
3) Student responses for item $j+1$ are predicted.
4) Each agreement rate between their predictions and true responses are calculated

Table 1 presents the results. As shown in table 1, the proposed model predict students' response the most precisely for the almost all of items. In addition, result shows the average of agreement rates of the proposed model over all items are much higher than traditional IRT model.

Furthermore, proposed method with attention achieve higher accuracy than without attention. Thus, the effectiveness of attention was shown.

Here, Bonferroni's multiple comparison test using p-values of Wilcoxon signed-rank test shows there are significant differences between three models at 0.05 significance level. It showed proposed model outperformed the conventional model in terms of response predictions.

*1 https://chainer.org/

6. Conclusion

This study proposed a novel scaffolding system using deep learning method. Experiments conducted with actual data demonstrated that the proposed model can predict students' responses more precisely than a traditional model.

Although this study specifically addressed only response prediction, it can be useful for response prediction with hints and has potential to apply for adaptive learning. These are left as subjects for future work.

Reference

[Bruner 96] Bruner, J.: The Culture of Education, Harvard University Press (1996)

[Collins 89] Collins, A., Seely Brown, J., and E Newmann, S.: Cognitive Apprenticeship: Teaching the Craft of Reading, Writing, and Mathematics, Vol. 18, pp. 453–494 (1989)

[Kingma 14] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv:1412.6980 (2014)

[Le 18] Le, C. V., Pardos, Z. A., Meyer, S. D., and Thorp, R.: Communication at Scale in a MOOC Using Predictive Engagement Analytics, in Artificial Intelligence in Education - 19th International Conference, AIED 2018, London, UK, June 27-30, 2018, Proceedings, Part I, pp. 239–252 (2018)

[Piech 15] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., and Sohl-Dickstein, J.: Deep Knowledge Tracing, in Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. eds., Advances in Neural Information Processing Systems 28, pp. 505–513, Curran Associates, Inc. (2015)

[Ueno 15] Ueno, M. and Miyasawa, Y.: Probability Based Scaffolding System with Fading, in Conati, C., Heffernan, N., Mitrovic, A., and Verdejo, M. F. eds., Artificial Intelligence in Education, pp. 492–503, Cham (2015), Springer International Publishing

[Ueno 18] Ueno, M. and Miyazawa, Y.: IRT-Based Adaptive Hints to Scaffold Learning in Programming, IEEE Transactions on Learning Technologies, Vol. 11, No. 4, pp. 415–428 (2018)

[Vygotsky 62] Vygotsky, L.: Thought and Language, Harvard University Press (1962)

[Vygotsky 78] Vygotsky, L.: Mind in Society, Harvard University Press (1978)