Journal of Japan Association for Earthquake Engineering, Vol. 25, No. 8, 2025

AUTOMATIC DIGITIZATION OF JMA STRONG-MOTION SEISMOGRAMS RECORDED ON SMOKED PAPER USING DEEP LEARNING —A CASE FOR THE 1940 KAMUI-MISAKI-OKI EARTHQUAKE—

Mitsuko FURUMURA[1] and Ritsuko S. MATSU'URA[2]

[1] Member, Dr. Sci., Director, Earthquake Research Center, Association for the Development of Earthquake Prediction,
Tokyo, Japan, furumura@erc.adep.or.jp
[2] Ph D., Senior Research Fellow, Earthquake Research Center, Association for the Development of Earthquake Prediction,
Tokyo, Japan, matsuura@adep.or.jp

ABSTRACT: Furumura et al. (2023) trained convolutional neural network (CNN) models to automatically digitize the waveforms scanned from the Japan Meteorological Agency (JMA) analog strong-motion seismographs recorded on smoked paper. We validated the CNN models using automatically digitized seismograms of the 1940 Kamui-Misaki-Oki earthquake by applying CNN models to the scanned images of the seismograms that were not used for CNN training. We compared the resulting data with manually traced data. The automatically digitized data agreed well with the manually traced data in most cases, although some data required correction. Using the CNN models substantially reduced the effort required to digitize analog records.

Keywords: *JMA smoked-paper seismograms, Automatic digitization, Convolutional neural network model, 1940 Kamui-Misaki-Oki earthquake*

1. INTRODUCTION

Furumura et al.[1)] built convolutional neural network (CNN) models (patent pending[1]) using deep learning to automatically digitize the waveforms in scanned images of strong-motion seismograms that were originally recorded on smoked paper using images from the Headquarters for Earthquake Research Promotion (HERP) data retrieval system of analog Japan Meteorological Agency (JMA) seismograms[2)] as well as waveform data manually traced from the images. They used the algorithm developed by Long et al.[3)] in the neural network for deep learning. The waveform position at each time point on the X coordinate was regressed from the image, where the X direction of the seismogram image was the time axis, and the Y direction was the amplitude axis; that is, a single amplitude value was estimated. The

[1] Application Number : 2021-110400.

large-amplitude analog recordings were located in an arc around the arm's pivot and had multiple amplitude values at any given time. Multiple *y* values were also generated for a given *x* value when the baseline shifted away from the pivot position. Therefore, the images and manual data were morphed to obtain a single amplitude value *y* for a given time *x*, and then training was performed. In addition, this method does not require data preprocessing, such as eliminating waveforms other than the target waveform in the image, eliminating baselines at other time periods, or connecting time marks, unlike conventional automatic waveform digitizing methods (e.g., Teves-Costa et al.[4], Wang et al.[5], Bartlett et al.[6], and Ishii and Ishii[7]). This considerably reduces the time and effort required to digitize analog waveforms.

In this study, we automatically digitized the strong-motion seismograms from the 1940 Kamui-Misaki-Oki earthquake of Japan using the CNN models trained by Furumura et al.[1]. The waveform and frequency characteristics of the results were compared with those of the manually traced data, and the effectiveness of the CNN models was verified.

2. DIGITIZING RECORDS

The JMA magnitude of the 1940 Kamui-Misaki-Oki earthquake, which occurred at 00:08 JST on August 2, 1940, was 7.5 at a depth of 10 km, according to Utsu[8), 9)]. The earthquake triggered a tsunami that hit

Table 1 Strong-motion smoked-paper records that were automatically digitized

Station	Strong-Motion Seismograph	Component	Start of Record (JST)				End of Record (JST)			
			Year	Month	Day	Time	Year	Month	Day	Time
Mori	CMO type	UD EW NS	1940	8	1	06:12	1940	8	2	−
Aomori	CMO type	UD EW NS	1940	8	1	08:37	1940	8	2	−
Akita	Omori type	UD EW NS	1940	8	1	−	1940	8	2	−
Yamagata	Imamura type	EW NS	1940	8	1	21:04	1940	8	2	08:10
Maebashi	CMO type	UD EW NS	1940	8	1	18:00	1940	8	2	−
Niigata	Nakamura type	EW NS	1940	8	1	20:54	1940	8	2	−
Wajima	CMO type	*UD** EW NS	1940	8	1	05:15	1940	8	2	05:05

* Neither manual tracing nor automatic digitizing was performed because the amplitude of the UD component recorded at the Wajima Station was small. CMO, Central Meteorological Observatory; −, unknown.

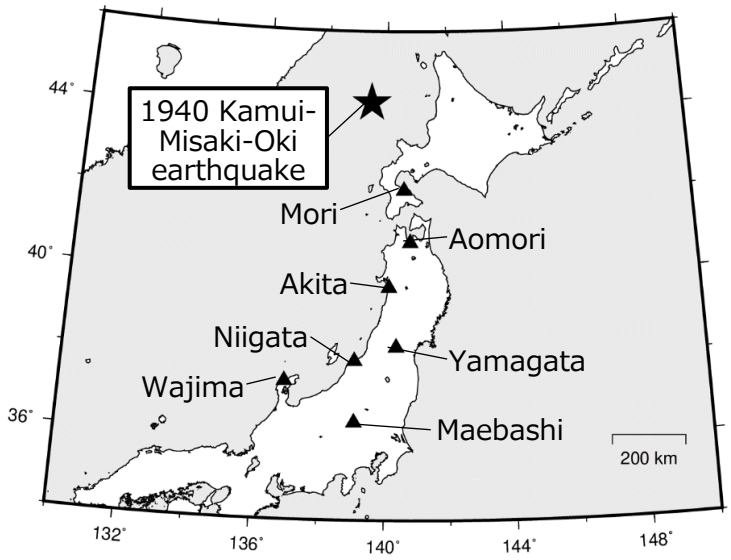


Fig. 1 Locations of earthquake epicenter according to Utsu[8), 9)] (★) and stations for which automatic digitization was performed (▲)

Teshio, Haboro, Tomamae, and other areas along the Sea of Japan coast, killing 10 people and wiping out many boats[10]. In contrast, the maximum seismic intensity was 4 on the JMA scale, which was felt along the coast of the Sea of Japan, where the ground shaking caused minimal damage[11]. Kagami[12] examined articles published in local newspapers about the earthquake and found that buildings were damaged in a wide range of areas, such as Otaru, Yoichi, and Iwanai. The earthquake was observed at meteorological stations throughout Japan and in areas under Japanese occupation at that time[10]. Scanned images of the strong-motion seismograms recorded at 22 of these stations are available in the HERP data-retrieval system of the JMA analog seismograms[2]. We automatically digitized the strong-motion seismograms from seven of these stations (Table 1 and Fig. 1) using original TIFF images that were scanned at 400 dpi with a color scanner (2806; Array Corporation, Japan). The original strong-motion seismogram paper was approximately 30 cm wide and 70–80 cm long, and ground shaking was recorded with a needle attached to an arm that scratched into a thin layer of soot adhered to the paper.

3. OVERVIEW OF THE CNN MODEL

Furumura et al.[1] built a CNN model with deep learning using scanned images of strong-motion seismograms and manually traced waveform data from these images. The model was initially trained using data from damage-causing earthquakes that mainly consisted of large-amplitude waveforms (Model A). We collected manual tracings of waveforms with smaller amplitudes in parallel with the training, and the model was retrained, including these waveforms (Model B). The results of the validation of two CNN models showed that the trained models were suitable for automatic waveform digitization. We also found that the models did not produce satisfactory results when the amplitude of the target waveform was large or when the waveform contained many high frequencies. An investigation of several cases revealed the following: (1) the accuracy of automatic digitization can be increased for large amplitudes via reducing the image height in advance, and (2) the accuracy of the onscreen determinations of arm length and pivot position shift decreases when high frequencies are included because of the almost-vertical movement of the drawing needle. Fine-tuning these parameters (arm length and pivot position shift) increases automatic digitization accuracy.

4. AUTOMATIC DIGITIZATION OF WAVEFORMS

The waveforms were automatically digitized as follows:
(1) An image of the target waveform for each component was cropped.
(2) The effects of the finite arm length and pivot position shift were removed.
(3) Automatic digitization was performed using CNN models.
(4) Data were corrected (if necessary).
The series of operations (1)–(4) were executed with a single program using a graphical user interface (GUI, referred to as a ‘digitizing program’). The details of this process are described below.

4.1 Cropping image of target waveform of each component of interest

The original TIFF image was opened using the digitizing program, and the rectangle surrounding the waveform of one component of interest as well as the left and right edges of the baseline containing the waveform were specified on the image. This operation rotated and cropped the image so that the baseline was horizontally positioned at the vertical center of the cropped image. The resulting cropped image was a grayscale PNG image.

4.2 Removing effects of finite arm length and pivot position shift

The following transformations were applied to the cropped image to remove the effects of the finite arm length and pivot position shift:

$$x = x_a - L_p(1 - \cos\theta) \tag{1}$$

$$y = y_a - c_p \tag{2}$$

$$L_p = L\frac{R}{25.4},\ c_p = c\frac{R}{25.4},\ \theta = \sin^{-1}\left(\left(y_a - c_p\right)\Big/ L_p\right) \tag{3}$$

where (x_a, y_a) and (x, y) are the coordinates (in pixels) of each pixel in the cropped image before and after removing the effects of the finite arm length and pivot position shift in the cropped image, respectively (Fig. 2; see Furumura et al.[1)] for details). O_T represents the origin of the coordinates after removing the effects of the finite arm length and the pivot shift position. O_A represents the origin before removing the effects. X values are positive on the X axis to the right of the coordinate origin. Y values are positive on the Y axis downward from the coordinate. L (mm) is the arm length, and c (mm) is the shift from the pivot position where the arm is fixed. L_p (px) and c_p (px) are converted to pixels using the resolution R (dpi) of the scanned image and used for coordinate transformation. L (mm) and c_p (px) are the input parameters of the digitizing program. We determined the appropriate L (mm) and c_p (px) values by checking the transformed image because the transformed image was immediately displayed on the screen when these parameters were set using slider bars.

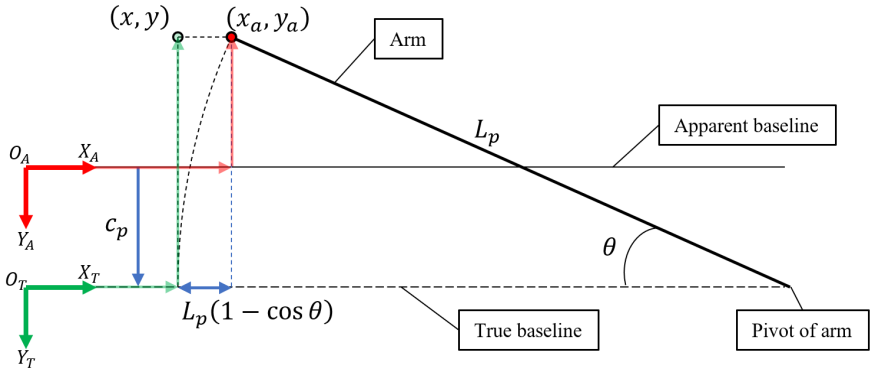


Fig. 2 The relationship between the coordinates (x_a, y_a) before removing the effect of the finite arm length as well as the baseline shift and the coordinates (x, y) after their removal, where O_A and O_T are the origin of the coordinates before and after removal, respectively.

4.3 Automatic digitization using CNN models

The CNN model (Model A) trained by Furumura et al.[1)] was used for automatically digitizing the images. The digitizing program displays the results of automatic digitizing superimposed on the image, allowing the user to visually judge the accuracy of the digitization. The user can retry automatic digitization after reducing the image height or readjusting L and c by looking at the waveform features if the automatically digitized result does not accurately trace the waveform.

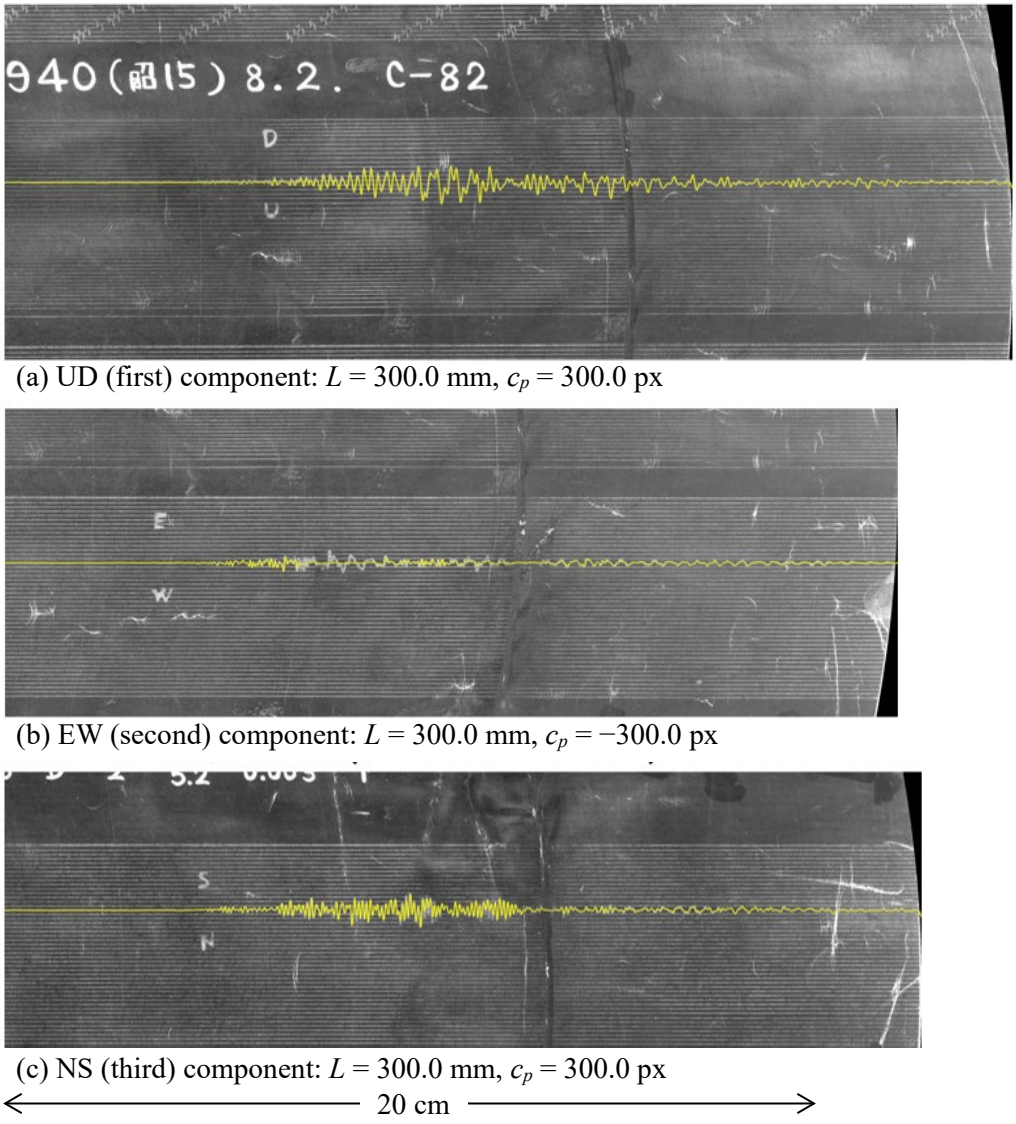
4.4 Data correction

The data can be manually corrected if reducing the image height or readjusting L and c does not substantially increase the accuracy of the digitization. The digitizing program provides menus for specifying multiple correct points (the program redigitizes the image so that the specified multiple correct points are used), moving points (moving the specified multiple points together), adding points, deleting points, and stretching the amplitude of an arbitrary range (adjusting only the Y coordinate according to the amplitude). The data can be corrected using these options, if necessary.

The automatic digitization results are provided as the coordinate values (x, y) at which the effects of the finite arm length and pivot position shift in the cropped image are removed. Thus, the result can be transformed back to the coordinates (x_a, y_a) in the cropped image using Eqs. (1)–(3), then transformed back to the coordinates in the original image using the positional relationship with the original image. The transformation in Eqs. (1)–(3) is used for automatic digitization, as described by Furumura et al.[1]. A complex transformation is required to accurately remove the effect of a finite arm length using the coordinate values (x_a, y_a), as described by Sasaki et al.[13] and Čadek[14].

5. COMPARISON OF AUTOMATIC AND MANUAL DIGITIZING RESULTS

The results of automatic digitization obtained using the methods described in Sections 4.1 to 4.3 (without correcting the data as described in Section 4.4) were compared with the seismogram images and manually digitized data to confirm the validity of the proposed method. The manually digitized data were obtained after tracing the waveform of the original TIFF image using Adobe Illustrator and removing the effects of the finite arm length and pivot position shift using the same arm length L and pivot position shift c_p values as in the automatic digitizing.



(a) UD (first) component: L = 300.0 mm, c_p = 300.0 px

(b) EW (second) component: L = 300.0 mm, c_p = −300.0 px

(c) NS (third) component: L = 300.0 mm, c_p = 300.0 px

Fig. 3 Images of strong-motion seismograms from the Akita Station and results of automatic digitization (yellow line)

Figure 3 shows the results of the automatic digitization of a strong-motion seismogram recorded at the Akita Station using the proposed method, which are superimposed on the images with yellow lines. The arm length L and pivot position shift c_p shown in Fig. 3 were used for the transformation in Eqs. (1)–(3). The automatic digitization adequately traced almost all of the waveforms of the three components, although the written characters and baselines of the other time periods were not erased. The UD component did not noticeably differ within the entire period.

Figure 4 compares the waveforms and Fourier spectra of the automatically and manually digitized data. The horizontal (X-axis) coordinates of the waveform were converted into time units using the paper feed rate (30 mm/min[2]) of the mechanical strong-motion seismographs because the coordinates of the digitized points were provided in pixels. The automatically and manually digitized waveforms are displayed such that the beginning of the waveform occurs at 0 s. The baseline shift in the UD-component waveform that occurred with manual digitization was not reproduced with the automatic digitization, and the EW- and NS-component waveforms showed slight difference. The automatically and manually digitized waveforms were almost the same, except for some points. The Fourier spectra of the digitized data overlapped between 0.08 Hz (period of 12.5 s) and 0.6 Hz (period of 1.67 s)[3]. The results for the UD component (red) were similar. The automatic digitization adequately captured the period characteristics of the strong-motion records, as mechanical strong-motion seismographs are narrowband seismographs with natural periods of approximately 5–6 s.

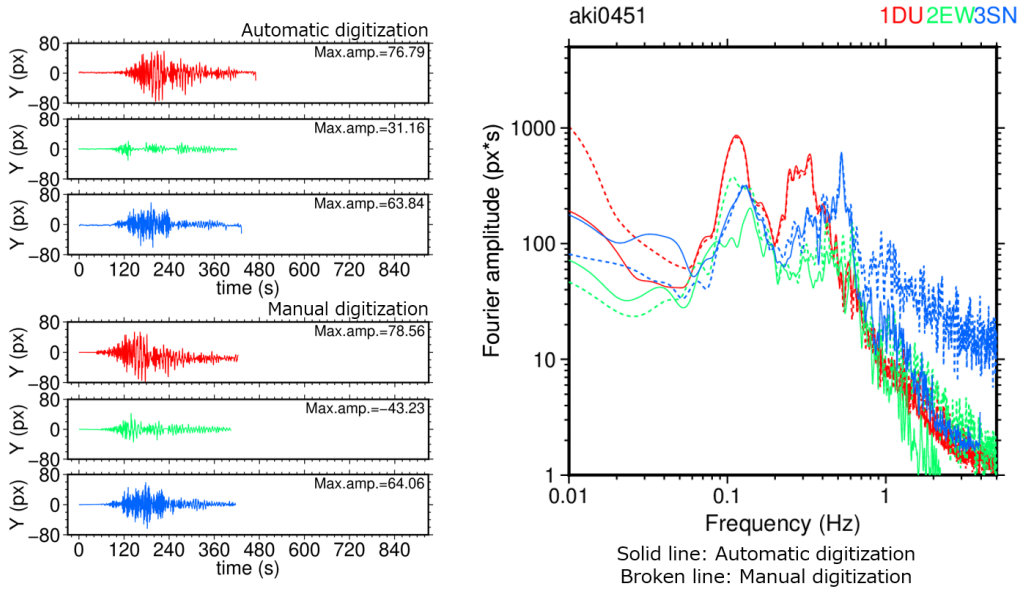


Fig. 4 Comparison of waveforms and Fourier spectra obtained via automatic and manual digitization at the Akita Station. The waveforms were converted from x to seconds and are displayed so that the beginning of the waveform occurs at 0 s. Red, green, and blue represent the UD, EW, and NS components, respectively. The 80 px on the vertical axis of the waveform corresponds to 5.08 mm (= 80.0 px / 400 dpi × 25.4 mm/in).

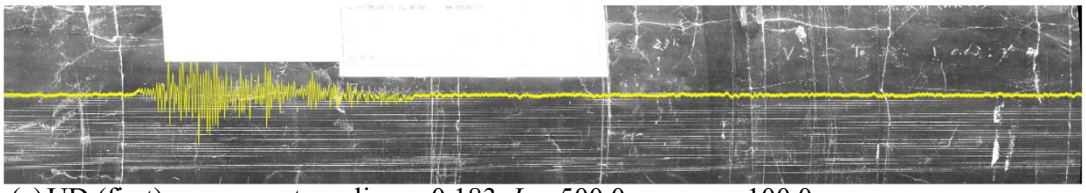
Another example of the automatic digitization results from the Mori Station is shown in Fig. 5. The observed records had many high frequencies, and their amplitudes were large. The large-amplitude peak portion decreased and considerably deviated from the observed waveform when the image was first automatically digitized without image height reduction. Reducing the image height in the digitizing program, such that the peak amplitude was within a certain range (reduction ratios of 0.183 and 0.100

[2] The target seismogram recorded the time mark of each minute. The paper feed speed estimated from the minute marks was approximately 25 mm/min.

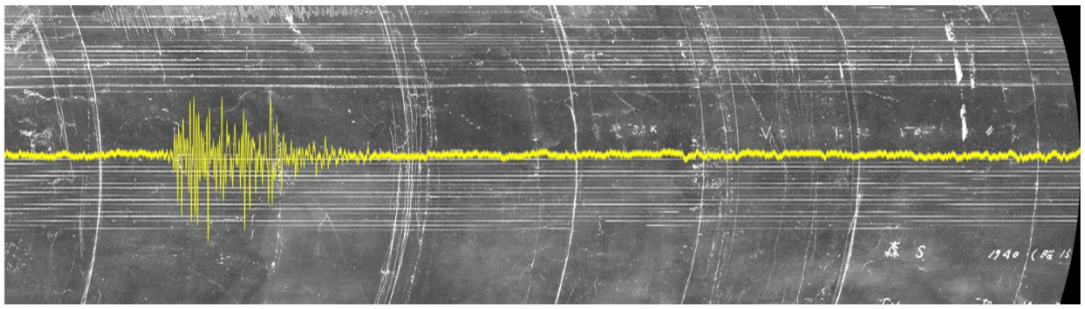
[3] The range was 0.067 Hz (period of 15 s) to 0.5 Hz (period of 2 s) assuming a paper feed speed of 25 mm/min.

for the UD and NS components, respectively) before automatic digitization, resulted in a substantial increase in the digitization accuracy of the large-amplitude waveform. The yellow lines appear thicker in Fig. 5 in the automatic digitization results for the two components with adjusted image heights because the image heights were restored to their original size. The UD and NS components were approximately six and ten times thicker than the EW component (the image height was not adjusted), respectively. The resolution of the small-amplitude portion decreased as the image height reduced, indicating that image height should only be reduced in selected cases. The EW component was partially broken with a constant value at a certain amplitude. The automatic digitization results of this part of the image were almost in line with the baseline, indicating that a saturated waveform also requires attention during automatic digitization. The observed waveform was sufficiently accurately traced during the unsaturated period.

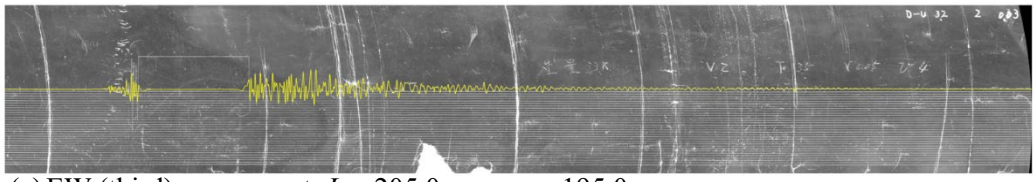
Figure 6 compares the waveforms and Fourier spectra of the automatic digitization results (the UD and NS components were the result of reautomatic digitization) with those of manually digitized data. The x-coordinates were converted to time units, setting the paper feed rate to 30 mm/min[4], as in Fig. 4. The Fourier spectra were calculated for the time periods before (Fig. 6, (1)) and after (Fig. 6, (2)) the saturated part because the EW component was partially saturated. The automatically digitized waveforms differed from the manually digitized waveforms in the baseline distortion and saturated portions; however, no other differences were noted. The automatically digitized Fourier spectra data agreed with the manually digitized data between approximately 0.1 and 0.9 Hz (periods of 10 and 1.25 s, respectively), indicating that this method is sufficiently effective for digitizing strong-motion seismograms, as was the case for seismograms from the Akita Station.



(a) UD (first) component: scaling = 0.183, L = 500.0 mm, c_p = 100.0 px



(b) NS (second) component: scaling = 0.100, L = 195.0 mm, c_p = 80.0 px



(c) EW (third) component: L = 205.0 mm, c_p = 195.0 px

40 cm

Fig. 5 Images of strong-motion seismograms from the Mori Station and their automatic digitization results (yellow line); the UD and NS components were automatically digitized with height-reduced images and then restored to the original image height.

[4] The target seismogram contained no clear time mark. The paper feed speed of the CMO strong-motion seismograph was approximately 30 mm/min according to the Seismology and Volcanology Research Department, the Meteorological Research Institute, Japan[15].

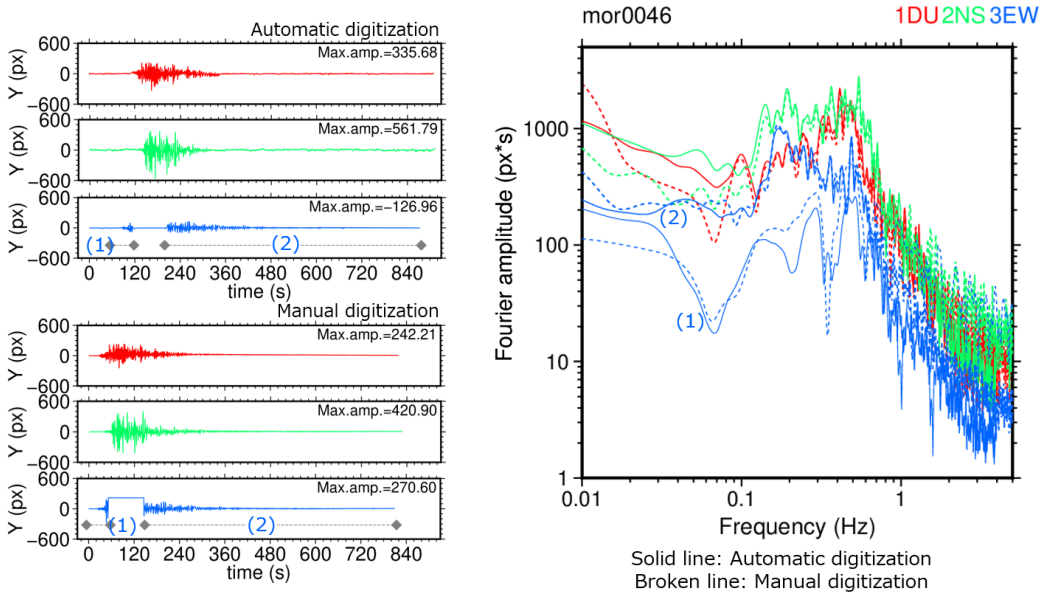


Fig. 6 Comparison of waveforms and Fourier spectra obtained via automatic and manual digitizing at the Mori Station. The waveforms were converted from *x* to seconds and are displayed such that the beginning of the waveform is 0 s. Red, green, and blue represent the UD, NS, and EW components, respectively. The Fourier spectra for the EW component waveform were separately calculated for the time periods before (1) and after (2) the waveform was saturated. The 600 px on the vertical axis of the waveform corresponds to 38.1 mm.

The waveforms from the other stations were also automatically digitized, and the Fourier spectra were calculated at a paper feed rate of 30 mm/min. The horizontal components at the Aomori and Niigata Stations were automatically digitized after adjusting the image height, and the Fourier spectra were calculated from the results. The baselines of the two components at the Yamagata Station were slightly distorted, and the waveforms were not accurately traced in the later small-amplitude phases. Otherwise, no significant difference in the waveforms was observed, as at the Akita and Mori Stations. In addition, we confirmed that the shape and level of the Fourier spectra strongly agreed with the manual digitization results in the frequency range in which the mechanical strong-motion seismograph operated properly, particularly at frequencies where the spectrum contained peaks.

6. CONCLUSIONS

The CNN models that were constructed by Furumura et al.[1)] using scanned images and manually traced waveform data were used to automatically digitize the strong-motion waveforms of the 1940 Kamui-Misaki-Oki earthquake. Because the algorithm developed by Long et al.[3)] was used to construct the CNN models, the following approach was applied to use these models: first, one component was extracted from the original image; second, the effects of the arm length and pivot position shifts in the analog waveform were removed; and, finally, automatic digitization was performed. The waveforms and Fourier spectra of the automatic and manual digitization results were compared. The waveforms were successfully digitized almost automatically, except for some parts, and the period characteristics were consistent with those of the observed records within the operating range of the seismograph.

Automatic digitization was reperformed after adjusting the image height for the two components at the Mori Station to reduce the digitizing errors at large amplitudes. The image height also needed to be adjusted for the two horizontal components at the Aomori and Niigata Stations. The resolution of the

small-amplitude portion decreased with decreasing image height. We recommend combining the results with and without adjusting the height if the large-amplitude portion is long. We recommend attempting automatic digitization without height adjustment and then correcting the deviation from the waveform in the image if the large-amplitude portion is short.

A seismogram generally contains other components of the target waveform as well as baselines and records from other time periods. Text or symbols may be written on the paper, and soot peeling or scratches may have occurred before the varnishing process or during storage. These features interfere with the automatic digitization because they appear white or near-white in the scanned image; previous automatic digitization studies have implemented various measures to eliminate the noise and avoid the reading of these non-target areas. Progress in the automatic digitization of analog records may have been slow because these preprocessing steps were necessary in previous automatic digitization efforts. The proposed method does not require time-consuming preparation; therefore, the digitization process is more efficient than previously proposed methods.

The original scanned TIFF images were used in this study. These images can be provided upon request because the file size of the TIFF images was too large (more than 200 MB) for online viewing. Instead, the HERP data retrieval system of the JMA analog seismograms[2)] allows users to freely view and download images in jpeg2000 format, compressing the file to 1/200 of its original size. Inverting the black and white tones facilitates the observation of the waveforms in smoked-paper seismograms, and we created black-and-white inverted jpeg2000 images, which are also available for viewing and downloading. Please contact us with the image number and other information so that we may provide you with an original high-resolution TIFF image if you wish to use the image from the record in the HERP data-retrieval system of JMA analog seismograms[2)]. See http://www.susu.adep.or.jp/ for usage details.

Although this method enables automatic digitization, visual inspection remains necessary. We hope many researchers will use this method to quantify analog waveforms, promoting further studies based on valuable past records.

ACKNOWLEDGMENT

This study was conducted as part of the Supporting Project for the Headquarters for Earthquake Research Promotion (HERP), sponsored by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan. The figures were created using GMT.

REFERENCES

1) Furumura, M., Ogawa, Y., Sakamoto, K. and Matsu'ura, R. S.: Automatic Digitization of JMA Strong-Motion Seismograms Recorded on Smoked Paper—An Attempt Using Deep Learning, *Seismological Research Letters*, Vol. 94, No. 6, pp. 2712–2724, 2023.
2) Furumura, M., Iwasa, K., Suzuki, Y., Demachi, T., Ishibe, T. and Matsu'ura, R. S.: Data Retrieval System of JMA Analog Seismograms in the Headquarters for Earthquake Research Promotion of the Japanese Government, *Seismological Research Letters*, Vol. 91, No. 3, pp. 1403–1412, 2020.
3) Long, J., Shelhamer, E. and Darrell, T.: Fully Convolutional Networks for Semantic Segmentation, *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015. (Open Access version is available at https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html) (last accessed on March 21, 2024)
4) Teves-Costa, P., Borges, J. F., Rio, I., Ribeiro, R. and Marreiros, C.: Source Parameters of Old Earthquakes: Semi-Automatic Digitization of Analog Records and Seismic Moment Assessment, *Natural Hazards*, Vol. 19, No. 2–3, pp. 205–220, 1999.
5) Wang, M., Jiang, Q., Liu, Q. and Huang, M.: A New Program on Digitizing Analog Seismograms, *Computers and Geosciences*, Vol. 93, pp. 70–76, 2016.

6) Bartlett, A. H., Lichtner, B. A., Nita, M., Yashar, B. and Bartlett, L E.: SKATE: A Web-Based Seismogram Digitization Tool, *Seismological Research Letters*, Vol. 89, No. 5, pp. 1886–1893, 2018.
7) Ishii, M. and Ishii, H.: DigitSeis: Software to Extract Time Series from Analogue Seismograms, *Progress in Earth and Planetary Science*, Vol. 9, No. 50, pp. 1–16, 2022.
8) Utsu, T.: Catalog of Large Earthquakes in the Region of Japan from 1885 through 1980, *Bulletin of the Earthquake Research Institute, University of Tokyo*, Vol. 57, pp. 401–463, 1982 (in Japanese with English abstract).
9) Utsu, T.: Catalog of Large Earthquakes in the Region of Japan from 1885 through 1980 (Correction and Supplement), *Bulletin of the Earthquake Research Institute, University of Tokyo*, Vol. 60, pp. 639–642, 1985 (in Japanese with English abstract).
10) Central Meteorological Observatory: *Geophysical Review*, No. 492, August 1940, 156 pp., 1940 (in Japanese).
11) Usami, T., Ishii, H., Imamura, T., Takemura, M. and Matsu'ura, R. S.: *Materials for Comprehensive List of Destructive Earthquakes in Japan, 599–2012*, University of Tokyo Press, 724 pp., 2013 (in Japanese).
12) Kagami, H.: Literature Survey of Damage Due to the off Kamui-Misaki Earthquake of August 2, 1940, *Architectural Institute of Japan Journal of Technology and Design*, Vol. 24, pp. 457–460, 2006 (in Japanese with English abstract).
13) Sasaki, Y., Tamura, K. and Aizawa, K.: Analysis of Long-Period Ground Motions Based on Displacement Seismograph Records by JMA (5)—Analysis of the Miyagi-Ken-Oki Earthquake of 1978—, *Technical Note of Public Works Research Institute*, Vol. 2664, 207 pp., 1988 (in Japanese).
14) Čadek, O.: Studying Earthquake Ground Motion in Prague from Wiechert Seismograph Records, *Beiträge zur Geophysik*, Vol. 98, No. 5, pp. 438–447, 1987.
15) Seismology and Volcanology Research Department, Meteorological Research Institute, Japan: Strong-Motion Seismograph Model 83 for the Japan Meteorological Agency Network, *Technical Reports of the Meteorological Research Institute*, Vol. 7, 132 pp., 1983 (in Japanese with English abstract).

(Original Japanese Paper Published: September, 2024)
(English Version Submitted: April 25, 2025)
(English Version Accepted: May 14, 2025)