Experimental Application of Semantic Segmentation Models Fine-Tuned with Synthesized Document Images to Text Line Segmentation in a Handwritten Japanese Historical Document

Sayaka Mori and Tetsuya Suzuki[a]

Department of Electronic Information Systems, College of Systems Engineering and Science,
Shibaura Institute of Technology, Saitama, Japan

Keywords: Text Line Segmentation, Historical Document, Deep Learning, Semantic Segmentation, Data Synthesis.

Abstract: Because it is difficult even for Japanese to read handwritten Japanese historical documents, computer-assisted transcription of such documents is helpful. We plan to apply semantic segmentation to text line segmentation for handwritten Japanese historical documents. We use both synthesized document images resembling a Japanese historical document and annotations for them because it is time-consuming to manually annotate a large set of document images for training data. The purpose of this research is to evaluate the effect of fine-tuning semantic segmentation models with synthesized Japanese historical document images in text line segmentation. The experimental results show that the segmentation results produced by our method are generally satisfactory for test data consisting of synthesized document images and are also satisfactory for Japanese historical document images with straightforward formats.

1 INTRODUCTION

Transcription of Japanese historical documents is not only a fundamental task in historiography and Japanese literature, but it has also gained importance in recent years with efforts to transcribe earthquake historical documents and use them for disaster prevention (Rekihaku, National Museum of Japanese History et al.,).

It is difficult even for Japanese to read handwritten Kana, which is a type of Japanese characters, used in historical documents because they are quite different from those currently used.

For this reason, text line segmentation, which is one of the elemental technologies in transcribing historical Japanese documents, is helpful.

We plan to use semantic segmentation for text line segmentation for handwritten Japanese historical documents. Because constructing a large set of manually annotated document images for machine learning training data is time-consuming, we automatically synthesize a lot of document images resembling Japanese historical document and their annotations, which are center line images of text lines, using a modified version of our system (Inuzuka and Suzuki, 2021).

The purpose of this research is to evaluate the effect of fine-tuning semantic segmentation models with synthesized Japanese historical document images in text line segmentation.

The organization of this paper is as follows: In Section 2, we explain the characteristics of the target handwritten Japanese historical document. We then summarize related work in Section 3. We propose our method in Section 4. Section 5 describes an experiment to select the best semantic segmentation model among seven models. Section 6 describes an experiment to apply the best semantic model to the target document. Finally, we provide concluding remarks in Section 7.

2 THE TARGET DOCUMENT

We selected ”The Tales of Ise” (Reizei, 1994) as a target handwritten Japanese historical document because the format is relatively simple. As a result, it is easy to generate document images resembling those of the document.

Fig.1 shows the characteristics of its page layout, which make it difficult to segment text lines.

[a] https://orcid.org/0000-0002-9957-8229

Mori, S. and Suzuki, T.
Experimental Application of Semantic Segmentation Models Fine-Tuned with Synthesized Document Images to Text Line Segmentation in a Handwritten Japanese Historical Document.
DOI: 10.5220/0012433100003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 826-832
ISBN: 978-989-758-684-2; ISSN: 2184-4313

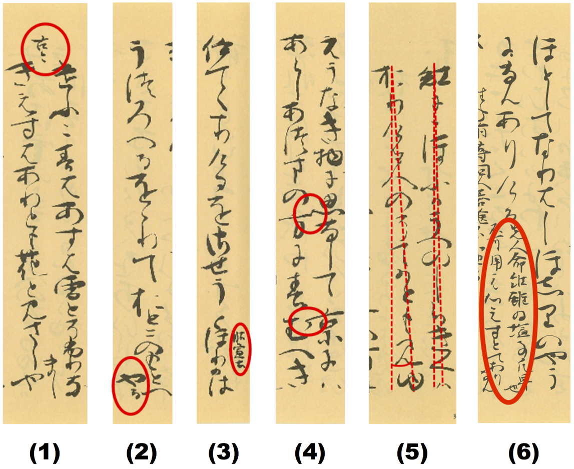


Figure 1: Characteristics of the target document's layout adopted from (INUZUKA and SUZUKI, 2020) (Circles and lines are added in document images scanned from (Reizei, 1994)).

- Fig.1 (1) shows an example where poems in the document are accompanied by notes representing the sources of the poems above them.
- Fig.1 (2) shows an example where some text lines are folded at the bottom of pages.
- Fig. 1 (3) shows an example where some text lines have readings or additional explanations next to them.
- Fig.1 (4) shows an example where some characters in adjoining text lines touch each other.
- Fig.1 (5) shows an example where the center lines of some text lines are leaning.
- Fig.1 (6) shows an example where there exist paragraphs with additional paragraphs next to them.
- There are characters written in a cursive style.

3 RELATED WORK

3.1 KuroNet

KuroNet (CLANUWAT et al., 2019) is a deep neural network that detects the locations of characters and recognizes the characters in a given image of a historical Japanese document. It, however, does not recognize text lines in the document images.

If we integrate a service such as KuroNet with text line segmentation, we will be able to determine the reading order of characters in historical Japanese documents.

3.2 Text Line Segmentation by a Fully Convolutional Network and post-processing

Barakat et al. proposed a method for extracting text lines from handwritten Arabic documents where some characters touch each other and the orientations of the text lines are not regular (Barakat et al., 2018). This method consists of two steps. The first step is a semantic segmentation by a fully convolutional network, which detects blobs of text lines where one text line may be divided and two text lines may touch. The second step is a post-processing that connects disconnected components to refine over-segmented situations. The authors used a publicly available challenging handwritten dataset (Ha et al., 1995).

3.3 Text Line Segmentation by the YOLOv3 Object Detection Algorithm

Our research group applied the YOLOv3 object detection algorithm based on a deep neural network to text line segmentation in Japanese historical handwritten documents (Inuzuka and Suzuki, 2021).

One of the problems of using deep neural networks for text line segmentation is the creation of a training data set because it is time-consuming to manually annotate a large set of document images.

Therefore, we implemented a document image synthesis system in Python and used it to generate both synthesized document images resembling target documents and bounding boxes surrounding text lines as annotations.

Fig.2 shows the document image synthesis process. The system consists of four command line interfaces: `fonts`, `format`, `typeset`, and `print`.

- The `fonts` command randomly selects at most given number of fonts for each Japanese Kana from Kuzushiji-49(Clanuwat et al., 2018) which is a data set of deformed Kana. It finally records the resulting font data in a font data file.
- The `format` command outputs a format resembling the format of a target document based on a document model to a format file. It randomly generates various paragraphs and text lines.
- The `typeset` command reads both a font data file and a format file, and then records the result of the typesetting in a file.
- The `print` command reads both a typesetting file and a font data file, and then generates both document images and bounding boxes surrounding text

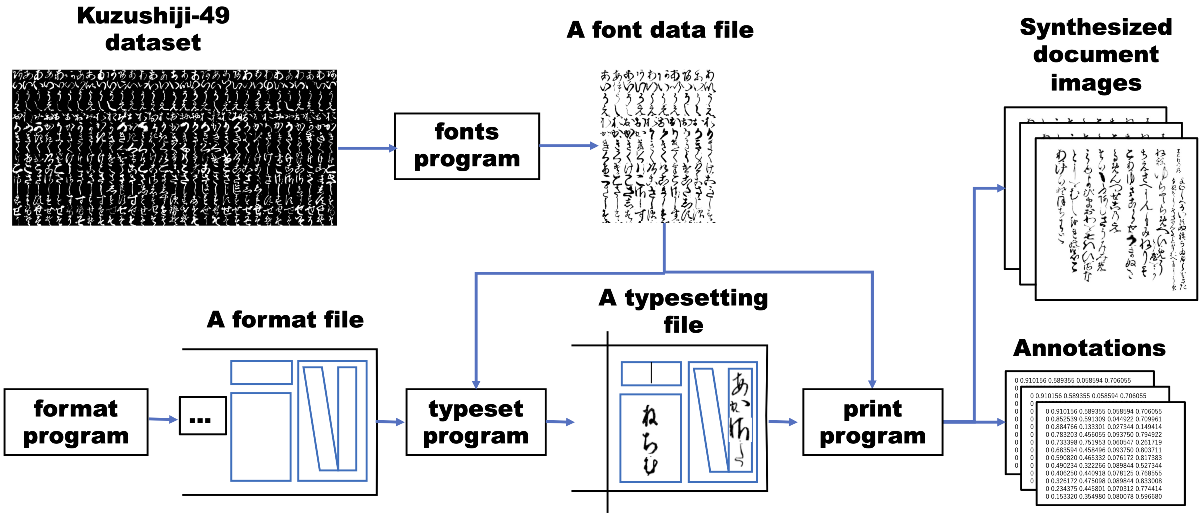


Figure 2: The document image synthesis process amended from (INUZUKA and SUZUKI, 2020).

lines as annotations according to the typesetting file.

Each text line in synthesized document images is a meaningless character sequence because it consists of randomly selected characters from the font data file.

The experiment results in (Inuzuka and Suzuki, 2021) show that a YOLOv3 model trained on a set of synthesized document images is competitive with one trained on a set of manually annotated document images.

However, the segmentation results by the YOLOv3 model trained on a set of synthesized document images include under- and over-segmentation. Fig.3 and Fig.4 show a page image of the target document described in Section 2 overlaid with the manually annotated ground truth and the segmentation results of the YOLOv3 model trained on a set of synthesized document images respectively.

4 OUR TEXT LINE SEGMENTATION METHOD

We propose a text line segmentation method using semantic segmentation for handwritten Japanese historical documents as follows.

We generate a set of synthesized document images resembling a target document and their annotations. The annotations are images with center lines of the synthesized documents. To generate them, we use a modified version of the document image synthesis system described in Section 3.3, enabling it to output center line images of text lines as annotations.

Subsequently, we train a CNN model for semantic segmentation using the set of synthesized document images and their annotations.

When provided with a document image, we apply the trained CNN model to obtain a segmentation result.

5 EXPERIMENT 1

We experimented with model selection.

5.1 Method

We used the following semantic segmentation models and the encoder provided by the Web site ”Segmentation models pytorch” (Iakubovskii, 2019).

Architecture. Unet++, MAnet, Linknet, FPN, PSPNet, PAN and DeepLabV3+

Encoder. EfficientNet

We used the modified system to generate 10,000 images resembling the target document described in Section 2, each containing around 10 lines, and increased the total number of images to 100,000 using data augmentation. The size of each image is 256 × 256. Fig.5(1) shows a page of a synthesized document image overlaid with center lines.

We used the dataset for both training and testing, and trained each model using the optimizer Adam for 40 epochs.

5.2 Results

Table 1 shows the average IoU values of the models for test data. The highest average IoU value among

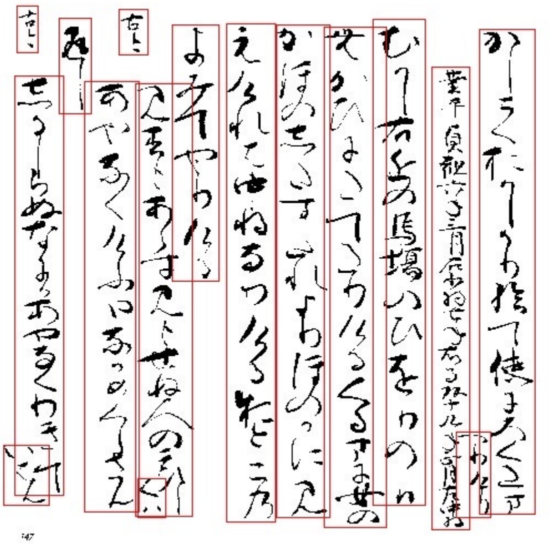


Figure 3: A page image of "The Tails of Ise" (Reizei, 1994) overlaid with the ground truth cited from (INUZUKA and SUZUKI, 2020).

Table 1: Average IoU values for test data.

Model	Average IoU values for test data
Unet++	0.917
Linknet	0.908
MAnet	0.906
FPN	0.883
DeepLabV3+	0.865
PAN	0.865
PSPNet	0.662

them is 0.917, corresponding to the Unet++ model, and the lowest IoU value among them is 0.662, corresponding to the PSPNet model.

Fig.5 and Fig.6 show segmentation two examples of the segmentation results for test data produced by the Unet++ model. Fig.5 shows a successful segmentation result and Fig.6 shows an under-segmentation result where neighboring two center lines are touching.

5.3 Evaluation

The Unet++ model outperforms the other six tested models in terms of IoU. The segmentation results produced by the Unet++ model are generally satisfactory.

The Unet++ model, however, produced under-segmentation where neighboring two center lines are touching as shown in Fig.6. To solve this problem, we may need to adjust the thickness of center lines as annotations.

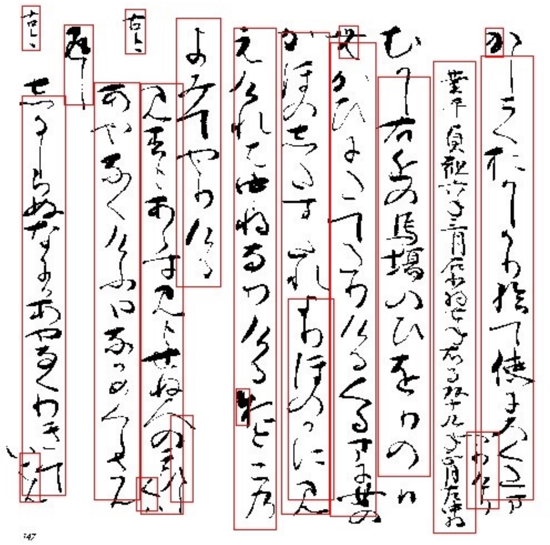


Figure 4: A page image of "The Tails of Ise" (Reizei, 1994) overlaid with segmentation results by YOLOv3 cited from (INUZUKA and SUZUKI, 2020).

6 EXPERIMENT 2

We applied the Unet++ model fine-tuned with synthesized document images to the handwritten Japanese historical document described in Section 2 and evaluated the segmentation results.

6.1 Method

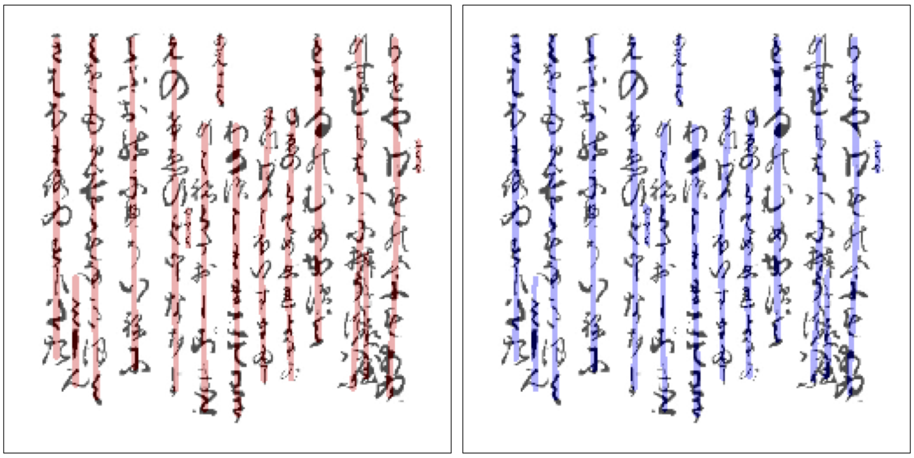
The differences from the method in Experiment 1 are as follows.

- Only the Unet++ model was tested.
- The size of each synthesized image is not 256 × 256 but 512 × 512.
- We applied the trained Unet++ model to the scanned and binarized document image of "The Tales of Ise" consisting of 166 pages (Reizei, 1994).

6.2 Results

Fig.7, Fig.8, Fig.9, and Fig.10 show examples of the segmentation results produced by the Unet++ model.

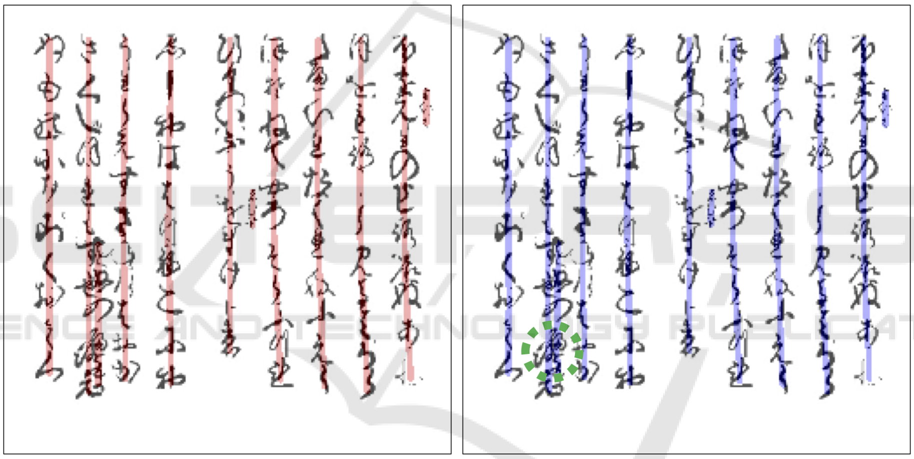
- Fig.7 shows a successful result.
- Fig.8 also shows under-segmentation, where the segmentation of the source of a poem and the segmentation of the poem are connected.
- Fig.9 also shows under-segmentation, where the center points of characters within the dotted-line circle are not included in any segmentation.



(1) A synthesized document image overlaid with center lines

(2) The same synthesized document image overlaid with segmentation results

Figure 5: A successful segmentation result for test data by Unet++.



(1) A synthesized document image overlaid with center lines

(2) The same synthesized document image overlaid with segmentation results

Figure 6: A segmentation result with under-segmentation for test data by Unet++. Neighboring two center lines are touching in the dotted-line circle.

- Fig.10 shows over-segmentation, where the left part and the right part of a character are in two distinct segmentations.

6.3 Evaluation

As shown in Fig. 7, the Unet++ model produced satisfactory segmentation results for document images with straightforward formats.

The under-segmentation shown in Fig.8 will be resolved by incorporating additional classes, such as the ”source of poem” class in segmentation though we employed only the ”text line” class in the experiment.

The under-segmentation shown in Fig.9 will be resolved by increasing the number of short text lines in the training data set.

An idea to resolve the over-segmentation shown in Fig.10 is to include more characters with separate parts in synthesized document images used for training data.

7 CONCLUSION

We evaluated the effect of fine-tuning semantic segmentation models with synthesized Japanese historical document images in text line segmentation. The annotations in the experiments are the center line im-

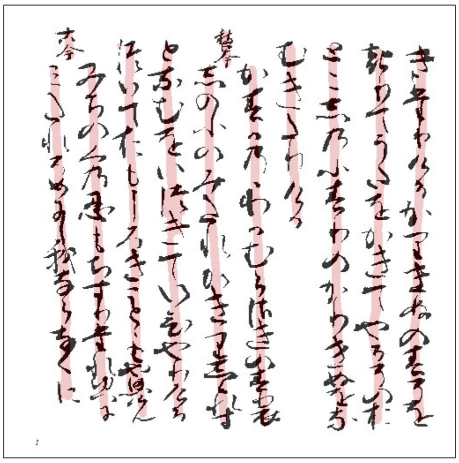


Figure 7: A page image of ”The Tails of Ise” (Reizei, 1994) overlaid with a successful text line segmentation result produced by Unet++.

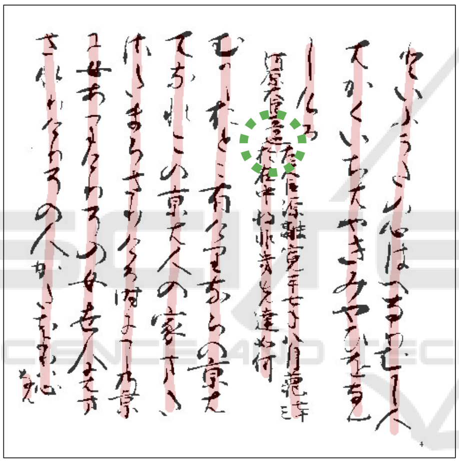


Figure 8: A page image of ”The Tails of Ise” (Reizei, 1994) overlaid with an under-segmentation result produced by Unet++. The center line of the source of a poem and the second center line of the poem are connected in the dotted-line circle.

ages of text lines. We synthesized document images resembling a target handwritten Japanese historical document by putting randomly generated texts. We selected the Unet++ model as the best among the seven semantic segmentation models in terms of IoU. Segmentation results produced by the Unet++ model for test data consisting of synthesized document images are generally satisfactory. Segmentation results produced by the Unet++ model for the target document are as follows. If the format of the document is straightforward, the results are satisfactory. We, however, often observed under- and over-segmentation. The problems may be resolved by either increasing the number of classes in semantic segmentation or adjusting the biases in the training dataset.

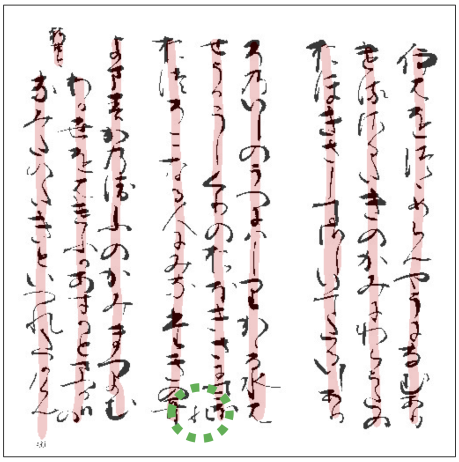


Figure 9: A page image of ”The Tails of Ise” (Reizei, 1994) overlaid with an under-segmentation result produced by Unet++. A character in the dotted-line circle is not in any segmentation.

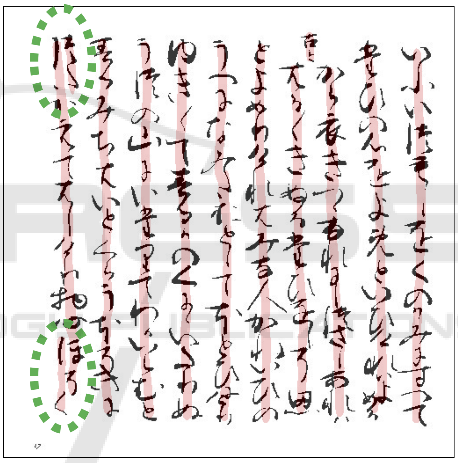


Figure 10: A page image of ”The Tails of Ise” (Reizei, 1994) overlaid with over-segmentation results produced by Unet++. The separate parts of some characters are in different segments in the dotted-line circles.

Future work would include the following. One of our future works is to improve the performance by modifying the method for construction of both synthesized document images and annotations. In addition, we need to apply our method to other historical documents to confirm the applicability of our method.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP22K12736.

REFERENCES

Barakat, B. K., Droby, A., Kassis, M., and El-Sana, J. (2018). Text line segmentation for challenging handwritten document images using fully convolutional network. In *16th International Conference on Frontiers in Handwriting Recognition, ICFHR 2018, Niagara Falls, NY, USA, August 5-8, 2018*, pages 374–379. IEEE Computer Society.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. (2018). Deep learning for classical japanese literature. *CoRR*, abs/1812.01718.

CLANUWAT, T., LAMB, A., and KITAMOTO, A. (2019). Kuronet: Pre-modern japanese kuzushiji character recognition with deep learning. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 607–614. (in English).

Ha, J., Haralick, R., and Phillips, I. (1995). Document page decomposition by the bounding-box project. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 1119–1122 vol.2.

Iakubovskii, P. (2019). Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch.

INUZUKA, N. and SUZUKI, T. (2020). Text line segmentation for japanese historical document images using deep learning and data synthesis. *SIG Technical Reports (CH)*, 2020-CH-122(4):1–6.

Inuzuka, N. and Suzuki, T. (2021). Experimental application of a japanese historical document image synthesis method to text line segmentation. In *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*, pages 628–634. INSTICC, SciTePress.

Reizei, T. (1994). *Tales of Ise (photocopy)*. Kasama Shoin.

Rekihaku, National Museum of Japanese History, Earthquake Research institute, the University of Tokyo, and Research Group for Historical Earthquakes, Kyoto University. MINNA DE HONKOKU. https://honkoku.org/index_en.html. Accessed: 2023-11-05.