2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)

Instance Segmentation of Newspaper Elements Using Mask R-CNN

Abdullah Almutairi
Department of Information Science
Kuwait University
abdullah.almutairi@ku.edu.kw

Meshal Almashan
School of Engineering
The University of Tokyo
almashan@mlab.t.u-tokyo.ac.jp

***Abstract*—Newspaper digitization has gained wide interest around the world. Archives of digitized newspapers contain a wealth of information that spans decades. To extract this abundance of information, optical character recognition (OCR) techniques are used. However, as a first step, the newspaper pages should be logically deconstructed into articles to gain meaningful knowledge. This is difficult due to the complex layout of newspapers and the various styles, shapes, and languages of newspaper articles. Newspaper pages also contain other elements besides articles, such as advertisements that come in multiple shapes and forms, and top headers that contain information about the newspaper's issue and page. Therefore, it is important to detect these elements before information extraction begins.**

In this paper, we present a deep learning solution for the problem of newspaper page semantic segmentation of the main newspaper elements (articles, advertisements, and page headers). We employed the instance segmentation method mask R-CNN [1] to create a language-agnostic model that logically deconstructs a newspaper page raw image into its main elements based only on its visual features. We show the results of experiments that display the accuracy and robustness of our model.

***Index Terms*—newspaper article segmentation, semantic segmentation, deep learning, mask R-CNN**

I. Introduction

Newspaper publishers around the world are digitizing current and previous issues of their newspapers, which are then collected into archives containing a wealth of information spanning many decades. OCR technology has been widely used to extract this information. However, newspaper pages should first be deconstructed into articles to obtain meaningful information. This is difficult because newspapers have a complex layout. Furthermore, each newspaper has a custom layout and style that affects the article's look. Each page contains other elements besides articles, such as a page header, which lists the newspaper's name as well as the date, page section, and page number. These page headers differ from other newspapers in content, style, and layout. A newspaper page can also contain advertisements that are visually unique, some being graphical and some textual.

In this paper, we propose a deep learning solution using the mask R-CNN method to segment and classify newspaper elements. The mask R-CNN uses a convolutional neural network (CNN) to do instance segmentation on objects in images. We will use this method to create a general language-agnostic model to semantically segment newspaper pages raw images into its main elements (articles, advertisements, and a page header) without any preprocessing or post processing steps. To train this model, we have gathered and fully annotated newspaper pages from various countries, languages, and newspapers. Previous solutions to this problem were mostly rule-based and did not generalize the varying styles and layouts of the different newspapers.

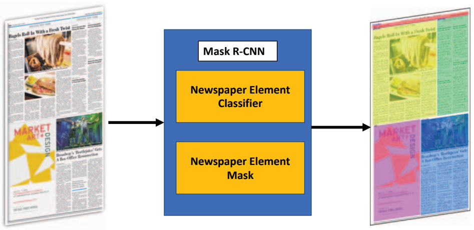


Fig. 1: The newspaper element segmentation and classification model. The input to the model is the raw pixels of the newspaper page, the output is the page with masks and labels on all elements on the page.

To demonstrate the accuracy and robustness of our model, we show experimental results of our model on various newspapers and languages, only some of which the model was trained on.

II. Related Work

Logical decomposition of newspaper pages into articles and other elements has gained a wide interest. The main focus was on segmenting articles, since they are the main element in a newspaper. Most research focused on extracting articles from newspapers with a specific language, such as, English [2] Arabic [3] and French [4]. To the best of our knowledge, there has been no previous work on finding an article segmentation model that showed generality on multiple languages.

The main approach used for article segmentation is a bottom-up approach where low-level components of an article or a page, such as titles, paragraphs, columns, separator lines and pictures were classified, extracted, and then logically clustered together to form an article. For the low-level components extraction step, different methods were used. Run-length smoothing algorithm (RLSA) [5] with a connected component analysis were used in [3]. Palfrey et al [4] used a conditional

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00223

random field (CRF). For the clustering of the components (also called tracking or article identification) step, Bansal et al [2] used the structured labeling method called the fixed point model. While many papers [6] [4] [7] [8] used rule-based heuristics to solve the component clustering.

Machine learning techniques were also used to solve article segmentation. Elanwar et al [9] used an ensemble of support vector machines to classify text and graphic elements of Arabic documents, which outperformed RLSA but did not yield good results on the complex layout of newspapers. Hadjer et al used a neural network with one hidden layer to detect the newspaper's logical components [10]. Deep learning techniques were used scarcely to solve document segmentation. The closest paper to our work is Meier et al [11], who used a fully convolutional network (FCN) to segment newspaper articles. However, their model does not detect the type of element that paper segment belongs to. Also, it does not handle non-rectangular shaped articles. Using FCN in contrast to mask R-CNN for article segmentation could be troublesome, since FCN does not handle instance segmentation and only gave a shared mask to adjacent articles.

III. Instance Segmentation of Newspaper Elements

A. Description of Mask R-CNN

The mask R-CNN classifies objects in an image providing them with a bounding box and a segmentation mask. We will use this method to semantically segment newspaper pages into three main elements (articles, advertisements, and page headers). The mask R-CNN consists of two stages: The first stage is a small neural network called a Region Proposal Network (RPN), which scans different-sized regions of the newspaper page to find the regions that most likely contain elements. These are called candidate regions. The best candidate regions are chosen by the RPN and moved to the second stage. In the second stage, features are extracted from each proposed region to perform classification, bounding box regression, and a pixel-level binary mask for each region in parallel. Having a pixel-level mask is crucial for article segmentation since many articles have a non-rectangular shape and cannot be captured accurately with a rectangular bounding box.

B. Network Architecture

The mask R-CNN uses ResNet [12] with 101 layers and a Feature Pyramid Network (FPN) [13] as a convolutional backbone for the network to extract features of the newspaper elements across the whole page. The network head extends the ResNet-FPN backbone and adds a fully convolutional branch for mask prediction.

C. Dataset for Training the Network

To train our model to achieve generality and language-agnostic abilities, we have chosen to use multiple newspapers in different languages. The languages of the newspapers chosen are English, German, French, and Arabic. Since Arabic is written from right to left while English is written left to right, the layout and order of the newspaper elements in our data will differ based on the language of the newspaper. Having newspapers with various languages and different writing systems in our training dataset will make the model focus on the common features of newspaper elements while ignoring the language-related features. Each page of our training data has been fully annotated, with each region of the page labeled as one of the three main elements (article, advertisement or page header). The small regions left without a label (i.e. white spaces or vertical/horizontal divider lines between the elements) will be implicitly considered as the background class. We have gathered and annotated 750 pages for the training dataset and 99 pages for the validation dataset to be used for the model selection.

In order to produce an accurate model for segmenting newspapers, a large number of training data is required. Since our training data was small, a couple of techniques were used to alleviate this problem. The first technique was using image augmentation methods on the current training data. We have horizontally flipped 50% of the images in the training data to increase the size of the dataset, having flipped text in in the dataset will make the model focus on the layout of the articles instead on the used text. The second technique was using the transfer learning method [14] where the weights from a pre-trained network backbone on the MS COCO dataset [15] were used for our network. Only the weights of the network heads are trained on our dataset.

IV. Implementation Details

For our experiments, we use the mask R-CNN implementation developed by Matterport, Inc. [16], which only has subtle differences between their implementation and the original mask R-CNN paper. The model was trained with 100 epochs. The training was done on an Nvidia Tesla V100 high performance GPU.

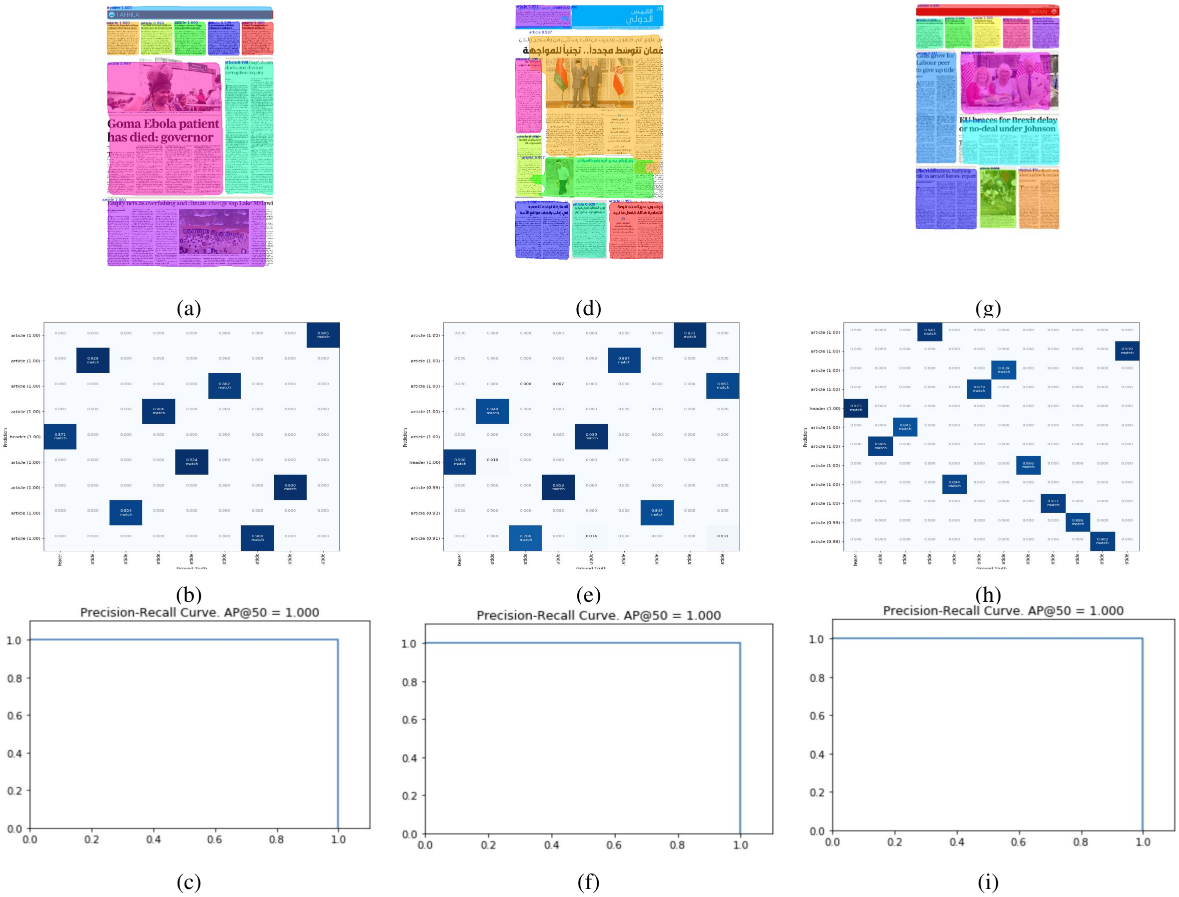
V. Experiments

To show the performance of our newspaper entity segmentation model, we first display the accuracy and inference results of the trained model on the validation data. Then, to show the accuracy of our model on different newspaper layouts and languages, we show the results of running our model inference on the following datasets:

(i) Pages from our validation dataset.
(ii) Pages not used for training the model but from newspapers present in the training dataset.
(iii) Pages from newspapers not present in our training dataset.
(iv) Pages from newspapers and langauges not present in the training dataset (e.g. Russian and Japanese newspapers).

A. Results on the Validation Dataset

We trained the model for 100 epochs with early stopping for choosing the best weight. Table I shows the validation loss on the classification and the bounding boxes for the RPN and mask R-CNN of the model. The table also shows the loss on the mask from the mask R-CNN. The mean average precision


(a)

(d)

(g)

(b)

(e)

(h)

(c)

(f)

(i)

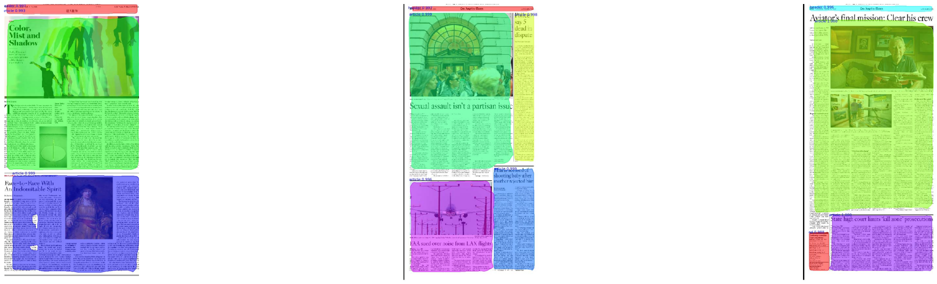
Fig. 2: The model's results on the validation dataset. The first row show a newspaper page with the learned masks and classes of its elements, the second row show the confusion matrix for each page, the third row show the precision-recall curve for each page.

(mAP) with the intersection over union (IoU) threshold of 50% for the validation dataset was 81.6%.

TABLE I: THE TRAINED MODEL VALIDATION LOSS

RPN class loss	0.012
RPN bounding box loss	0.283
Mask R-CNN class loss	0.202
Mask R-CNN bounding box loss	0.12
Mask R-CNN mask loss	0.13
mAP$_{50}$	**81.6**

Figure 2 shows a sample of our model's results on the validation dataset. We show the newspaper pages with the learned masks overlaid on them, each mask having the predicted class label with the prediction confidence score. We also show the confusion matrix for each page; each row in the confusion matrix indicates a predicted element in the page, and each column represents the ground truth elements of that page. The value in each cell in the matrix represents the percentage that each pixel of the predicted mask of that row corresponds with the pixels of the mask from the ground truth column. The darkness of the shade of the cell represents the magnitude of the percentage. If the predicted mask for an element correctly corresponds with the ground truth element mask with an IoU threshold of 50%, the word match will be present in the cell. The last row in the figure show the precision-recall curve for each page, still using the IoU threshold of 50%. The x-axis represents the precision of our model's prediction for the page, while the y-axis represents the recall. The average precision (AP) of our model's prediction for that page is computed from the area under the curve and is displayed above the figure. Each page in the sample is from a different newspaper with figures 2a and 2g showing pages from an English newspaper and figure 2d from an Arabic newspaper. The pages and the corresponding confusion matrices confirm that our model segments and classifies each newspaper element correctly and with a high confidence score. The page headers are correctly classified in each page even though each has a different style and layout. The articles, with their wide variety of layouts, are also correctly classified in each page. One of the interesting results in figure 2d is the ability of our model to segment non-rectangular shaped articles with some success. Another interesting result is correctly classifying and segmenting an



(a)

(b)

(c)

Fig. 3: The model's results on pages not used for training the model but from newspapers present in the training dataset. (a) is from the Wall Street Journal, (b) and (c) are from the Los Angeles Times. Each newspaper element in the pages were correctly classified and segmented.

article next to the page header, even though it is an uncommon newspaper layout. From the precision-recall curve, we see that the average precision (AP) of our model on all pages in the sample are 100%.

B. Results on Pages Not Used for Training the Model but from Newspapers Present in the Training Dataset

We ran our model on pages from newspapers present in the training dataset. Figure 4 shows a sample of the segmented pages results from the Wall Street Journal and the Los Angeles Times. Our model segments and classifies the page headers, articles and accompanying pictures correctly. However, the article segments in figure 3a had gaps in their masks. Moreover, non-rectangular shaped articles were not fully segmented, but the generated segments were also non-rectangular corresponding to the shape of the article.

C. Results on Pages from Newspapers Not Present in our Training Dataset

In this section, we show the results of our model on pages from newspapers that were not used for training the model. We use pages from the Boston Globe (figure 4c) and pages from the Daily Telegraph (figures 4a and 4b). The articles in these samples were correctly segmented and classified. The half-page advertisement in figure 4a and the full page advertisement in figure 4c were correctly segmented and classified. All the page headers were all also correctly segmented and classified.

D. Results on Pages from Languages not Present in the Training Dataset

To explore the performance of our model on untrained languages, we tested it on a Russian newspaper and a Japanese newspaper. Unlike English and other European languages, the Russian language uses Cyrillic alphabets. We ran our model on the Russian newspaper Voenno-Promyshlennyy Kurier, figure 5 shows the results. As in the results for the other trained languages, our model segments and classifies the articles correctly, with the exception of the false positive article in the middle of the page. The page header was also accurately segmented and classified.

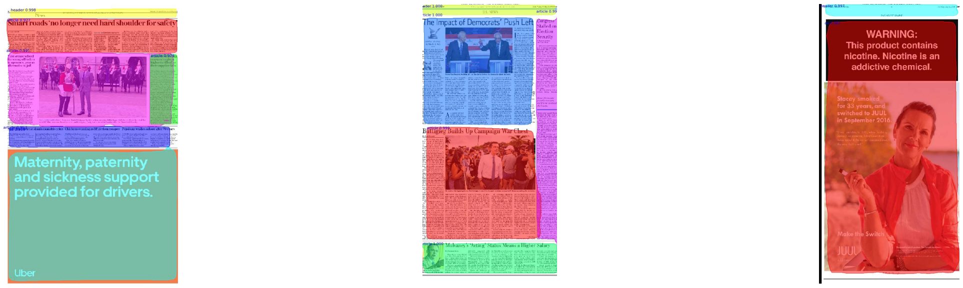
For our final experiment, we ran our model on the Japanese newspaper Mainichi Shimbun. The Japanese language and print layout are very different from its European and Arabic counterparts. The Japanese writing system is different in that sentences are written in vertical lines from right to left in most cases but can be written in horizontal lines from left to right. This unique writing system creates a different layout for articles in Japanese newspapers, but in most cases, articles are rectangular and separated by vertical and horizontal lines.

Figure 5b shows the results of our model on the Japanese newspaper. The model struggles to segment and classify all the articles due to the different layout, and some articles are not segmented and classified or an article is classified as more than one article. However, the model managed to segment and classify most of the advertisements and page headers.

The results of our experiments show the ability of our model to accurately classify and distinguish between the newspaper's main elements for a number of languages with different alphabets and orientations, even though it struggles with languages with a unique layout such as the Japanese language. This is achieved with a small training dataset and may be improved upon by increasing the size of the dataset.

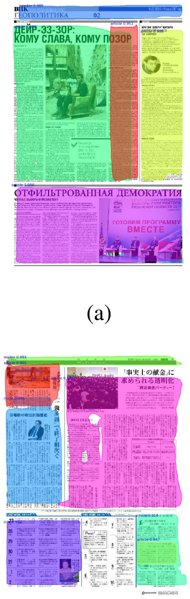
VI. Conclusion

In this paper, we have presented a general language-agnostic model for segmenting and classifying newspaper elements (articles, advertisements, and page headers). These newspaper elements are segmented and classified based only on their visual features. The segmentation and classification will happen on the raw page image without the need for any preprocessing or postprocessing steps. The model was created using the deep learning instance segmentation method called mask R-CNN. We have presented experimental results of our model that show its accuracy and robustness on various newspapers and languages. Segmenting Japanese newspaper articles was limited due to their unique layout caused by their vertical lines of text.



(a) (b) (c)

Fig. 4: The model's results on pages from newspapers not present in our training dataset but from a language in the training dataset. (a) and (b) are from British newspaper the Daily Telegraph and (c) is from the American newspaper the Boston Globe. Each newspaper element in the pages were correctly segmented and classified.



(a)

(b)

Fig. 5: The model's results on newspapers with languages it was not trained on. (a) is a page from a Russian newspaper, (b) is from a Japanese newspaper.

References

[1] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2017.322

[2] A. Bansal, S. Chaudhury, S. D. Roy, and J. Srivastava, "Newspaper article extraction using hierarchical fixed point model," in *2014 11th IAPR International Workshop on Document Analysis Systems*. IEEE, 2014, pp. 257–261.

[3] K. Hadjar and R. Ingold, "Arabic newspaper page segmentation," in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, Aug 2003, pp. 895–899.

[4] T. Palfray, D. Hebert, S. Nicolas, P. Tranouez, and T. Paquet, "Logical segmentation for article extraction in digitized old newspapers," in *Proceedings of the 2012 ACM symposium on Document engineering*. ACM, 2012, pp. 129–132.

[5] K. Y. Wong, R. G. Casey, and F. M. Wahl, "Document analysis system," *IBM journal of research and development*, vol. 26, no. 6, pp. 647–656, 1982.

[6] P. E. Mitchell and H. Yan, "Newspaper layout analysis incorporating connected component separation," *Image and Vision Computing*, vol. 22, no. 4, pp. 307–317, 2004.

[7] K. Chaudhury, A. Jain, S. Thirthala, V. Sahasranaman, S. Saxena, and S. Mahalingam, "Google newspaper search–image processing and analysis pipeline," in *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009, pp. 621–625.

[8] F. Liu, Y. Luo, M. Yoshikawa, and D. Hu, "A new component based algorithm for newspaper layout analysis," in *Proceedings of Sixth International Conference on Document Analysis and Recognition*. IEEE, 2001, pp. 1176–1180.

[9] R. Elanwar, W. Qin, and M. Betke, "Making scanned arabic documents machine accessible using an ensemble of svm classifiers," *Int. J. Doc. Anal. Recognit.*, vol. 21, no. 1-2, pp. 59–75, Jun. 2018. [Online]. Available: https://doi.org/10.1007/s10032-018-0298-x

[10] K. Hadjar and R. Ingold, "Logical labeling of arabic newspapers using artificial neural nets," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. IEEE, 2005, pp. 426–430.

[11] B. Meier, T. Stadelmann, J. Stampfli, M. Arnold, and M. Cieliebak, "Fully convolutional neural networks for newspaper article segmentation," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 414–419.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *CoRR*, vol. abs/1411.1792, 2014. [Online]. Available: http://arxiv.org/abs/1411.1792

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *Lecture Notes in Computer Science*, p. 740–755, 2014. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10602-1_48

[16] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," https://github.com/matterport/Mask_RCNN, 2017.