# Instance Aware Document Image Segmentation using Label Pyramid Networks and Deep Watershed Transformation

Xiao-Hui Li[1,2], Fei Yin[1,2], Tao Xue[3], Long Liu[3], Jean-Marc Ogier[4], Cheng-Lin Liu[1,2,5]
[1]*National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences*
*95 Zhongguancun East Road, Beijing 100190, P.R. China*
[2]*University of Chinese Academy of Sciences, Beijing, P.R. China*
[3]*Tencent Co. Ltd, Beijing, P.R. China*
[4]*L3i Laboratory, University of La Rochelle, 17042 La Rochelle Cedex 1, France*
[5]*CAS Center for Excellence of Brain Science and Intelligence Technology, Beijing, P.R. China*
*Email: {xiaohui.li, fyin, liucl}@nlpr.ia.ac.cn, {emmaxue, dragonliu}@tencent.com, {jean-marc.ogier}@univ-lr.fr*

***Abstract*—Segmentation of complex document images remains a challenge due to the large variability of layout and image degradation. In this paper, we propose a method to segment complex document images based on Label Pyramid Network (LPN) and Deep Watershed Transform (DWT). The method can segment document images into instance aware regions including text lines, text regions, figures, tables, etc. The backbone of LPN can be any type of Fully Convolutional Networks (FCN), and in training, label map pyramids on training images are provided to exploit the hierarchical boundary information of regions efficiently through multi-task learning. The label map pyramid is transformed from region class label map by distance transformation and multi-level thresholding. In segmentation, the outputs of multiple tasks of LPN are summed into one single probability map, on which watershed transformation is carried out to segment the document image into instance aware regions. In experiments on four public databases , our method is demonstrated effective and superior, yielding state of the art performance for text line segmentation, baseline detection and region segmentation.**

***Keywords*-document image segmentation, instance segmentation, label pyramid network, deep watershed transformation**

## I. Introduction

Automatic document image segmentation, including region segmentation, text line segmentation and baseline detection, plays an important role in document image understanding and information extraction.

Existing document image segmentation methods can be grouped into two major categories: traditional methods and [1]–[4] deep learning based methods [5]–[16]. Traditional methods either split the document image into smaller regions progressively (top down) or group small elements (pixels or connected components) into larger regions (bottom up), they depend on sophisticated handcrafted features and heuristic rules which are prone to errors. On the contrary, deep learning based methods use semantic segmentation frameworks [17] to predict each pixel's label, then connected component analysis is used to obtain document regions [13] [15]. Despite the prediction ability from deep learning, it is hard to separate regions which are close or even interfering. Fig. 1 and Fig. 4 show that for regions overlapping each other, the segmentation performance degrades rapidly. To overcome this problem, some works only predict the central areas of regions, e.g., boundary shrunken regions [13] or central text line areas [5]–[8]. This can greatly alleviate the false merging problem, but burdensome postprocessing procedures are still needed to obtain the precise boundaries.

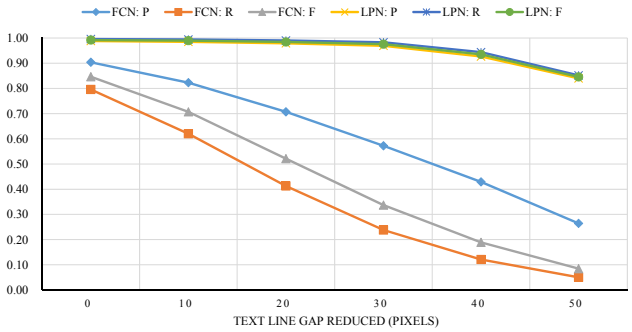


Figure 1: Text line segmentation results of FCN and LPN on synthesized CASIA-HWDB databases with gradually reduced text line gaps (IoU=0.75).

In this paper, we propose a method based on Label Pyramid Networks (LPN) and Deep Watershed Transformation (DWT) [18] to segment documents into instance aware regions. The backbone of our LPN can be Fully Convolutional Networks (FCN) [17] with any structures, e.g., U-Net [19], while the label map pyramid is transformed from the original single label map through distance transformation and multi-level thresholding. In training, label map pyramids on training images are provided to the LPN model for multi-task learning, while in segmentation, the outputs of multiple tasks of LPN are summed into one single probability map, on which watershed transformation is carried out to segment the original image into instance aware regions. Our method utilizes hierarchical boundary information of regions, and can separate extremely near even overlapping regions without using burdensome postprocessing procedures. On four public

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00088

databases (CASIA-HWDB, Bozen, cBAD and Maurdor) and three segmentation tasks (text line segmentation, baseline detection and region segmentation), our method demonstrates superior performance.

The rest of this paper is organized as follows. Section 2 briefly reviews related works. Section 3 gives details of the proposed method. Section 4 presents the experimental results, and Section 5 draws concluding remarks.

## II. RELATED WORKS

Comprehensive surveys of traditional document segmentation methods have been given in [2] and [3]. In this section, we will focus on deep leaning based methods closely related to this work.

**Document semantic segmentation.** Semantic segmentation aims to give pixel-level (or small patch) labeling of image. Chen et al. [12] use simple Convolutional Neural Networks (CNN) to classify image patches into text, decoration and comment. Pastor-Pellicer et al. [5] use CNN to predict main text body pixels in a sliding window way. Other methods use FCN for pixel labeling tasks. For instance, Meier et al. [13], Yang et al. [14], He et al. [15] and Xu et al. [16] use FCN to classify pixels into different categories including text, figure, table and decoration.

**Text line segmentation.** Most existing deep leaning based text line segmentation methods [5]–[8] share a common strategy that they only predict the centerline pixels to avoid false merging of adjacent text lines, but they require burdensome postprocessing procedures to obtain complete text line boundaries, which limit their performance for documents with complex layout or severe image degradation.

**Baseline detection.** Baseline detection [20] can help text line recognition systems, especially those for Latin scripts, and thus, has triggered some research works. Quirós et al. [9] use *A-net* and *M-net* to predict baseline pixels and then use a contour extraction approach to extract baseline curves. Oliveira et al. [10] use U-Net to predict baseline pixels, and then apply hysteresis thresholding to obtain connected components and then baseline polygons. Grüning et al. [11] use U-Net with multi-scale attention to predict baseline and separator superpixels and then use Delaunay neighborhood system to cluster these super pixels into baselines.

**Document region segmentation.** In many cases, document analysis systems need to obtain semantic region segmentation results. To achieve this goal, Meier et al. [13] binarize the pixel-level output of FCN with a threshold of 0.35 and then use background pixels to separate the article regions. He et al. [15] use a threshold of 0.5 to binarize the probability map generated by fusing multi-scale multi-task deep FCN and CRF, and then apply connected components analysis to get locations of table regions. These methods are not satisfactory to handle documents of complex layout and severe degradation, however.

Our method is inspired by the works of Hayder et al. [21] and Bai et al. [18] for instance segmentation where label map pyramids and deep watershed transformation have been used. To the best of our knowledge, our work is the first to handle document image segmentation in instance aware manner and totally based on deep learning framework.

## III. PROPOSED METHOD

The framework of the proposed method is illustrated in Fig. 2. On the training stage, the original label maps are transformed into label map pyramids witch are provided to the backbone FCN model for multi-task learning; while on the test stage, the outputs of all tasks are summed into single probability maps on which watershed transformation is carried out to segment the original images into instance aware regions. In this section, we will give details of our method for instance aware document image segmentation. Firstly, we introduce the label map pyramids we use in our entire work, including its generation and normalization. Then the structures of the backbone U-Net are presented, especially the crucial modifications from previous works. After that, we introduce the loss function we used to train our model. At last, the deep watershed transformation used to segment the probability maps into separate regions is introduced.

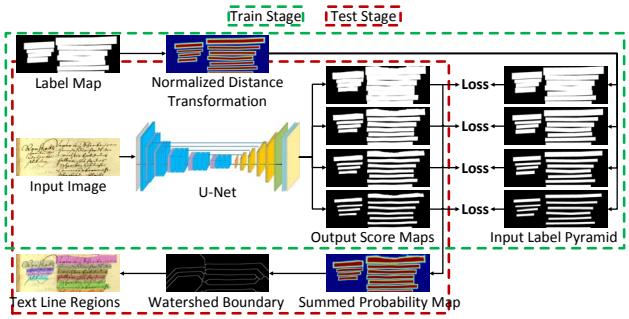


Figure 2: Framework of the proposed LPN.

### A. Label Map Pyramids

To obtain the label map pyramid, we first convert the original label map into a distance map through distance transformation on which each pixel's value is the nearest distance to the boundary it belongs to. Then we truncate the distance map so that the max value of the distance map is or less than 64, which limit the boundaries to a certain value thus can easy the model training. After that, we normalize the distance map so that each region's min value and max value is 0 and 255, respectively. At last, we binarize the distance map with multi-level thresholding to get a label map pyramid containing $N$ (in our work, $N = 4$) binary label maps.As we can seen in Fig. 2, the label map pyramids contains gradually shrunken regions which implicitly integrate information of region boundaries.

It is worth mentioning that for baseline detection we draw baselines with gradually thinner stroke width as label map

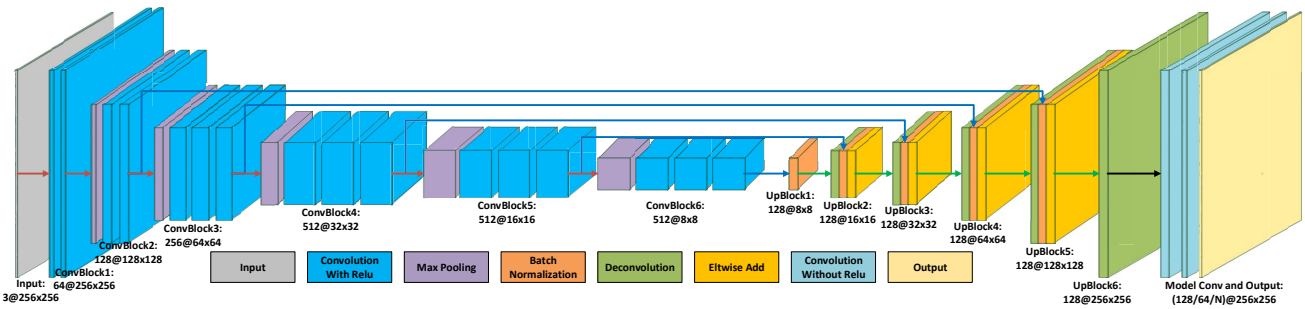


Figure 3: Network architecture of the backbone U-Net.

pyramid. In segmentation, after watershed transformation based region segmentation, we use region center lines as the detected baselines.

The basic idea of transforming single label map into $N$ binary label maps is similar with the precious work [21]. However, different with their one-hot formulation, our $N$ binary maps have gradually shrunken (nested) regions thus pixels can be foreground in multiple label maps. What's more, [21] rely on detected object bounding boxes to distinguish different instances, while in this work, we utilize watershed transformation to segment instance aware regions, which is simpler and more direct.

### B. Network Architecture

Generally speaking, document images are very different from natural scene images in the richness of color and texture, and the background spaces inside document regions maybe much larger than those between adjacent regions. These difficulties propose new challenges to the design of network structures.

The architecture of our backbone FCN is illustrated in Fig. 3, and the parameter configurations are clearly depicted on it. In all convolution layers, we use kernel size of $3 \times 3$ and stride size of $1 \times 1$; while in the max pooling layers and deconvolution layers, we use kernel size of $2 \times 2$ and stride size of $2 \times 2$; what's more, in the batch normalization layers we use kernel size of $1 \times 1$ and stride size of $1 \times 1$. To maintain the sizes of feature maps before and after convolution, we use zero padding accordingly. Our network is originated from the well-known U-Net [19] but with the following crucial modifications to improve its performance:

First, on the base of the original VGG16 network, we add a deeper convolution block *ConvBlock6* with three convolution layers to enhance the presentation capacity and enlarge the receptive field of the network.

Second, in convolution block3, block4, block5 and block6, we use atrous convolution with a fixed atrous rate value of 2 to further enlarge the receptive fields of the network.

Third, in the skip connection layers, we use batch normalization with $1 \times 1$ kernels to reshape the original convolution layers to regular batch norm layers with fixed channel numbers so that they can go through addition operation with the deconvolution layers in an element-wise way.

The last but the most important, the output layer contains $N$ channels and the pixel values of each channel represent the probabilities they belongs to certain classes. In training, these $N$ channels are used for multi-task learning under the supervision of input label map pyramid; while in segmentation, these $N$ channels are summed into one single probability map which has greater values on region centers and smaller values on region boundaries, thus can facilitate the subsequent watershed transformation.

### C. Loss Function

The LPN is trained on images with label map pyramids through multi-task learning. For each task (predicting one label in the label map pyramid, through one-versus-all training), we use weighted cross entropy as our loss function in which the weights are set by inverse class frequency in each batch. To be specific, our loss function is defined as follows:

$$CE(p_t) = -\alpha_t log(p_t) \tag{1}$$

where

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{2}$$

is probability of ground truth label, and

$$\alpha_t = \begin{cases} \frac{background\_pixel\_number}{total\_pixel\_number} & \text{if } y = 1 \\ \frac{foreground\_pixel\_number}{total\_pixel\_number} & \text{otherwise} \end{cases} \tag{3}$$

is the inverse class frequency loss weight. Because we have multi-task to train simultaneously, we sum the loss of each task to get total loss of our model.

### D. Deep Watershed Transformation

After obtaining the multi-task output probability maps, we sum and average them into one single probability map which has greater values on region centers and smaller values on region boundaries. This probability map is quantized to $N + 1$ classes (markers) to reduce noise. Then we perform watershed transformation on this quantized probability map to obtain watershed boundaries which can separate adjacent regions. This watershed boundary map is multiplied with the probability map, then connected component analysis is applied to get final region segmentation results. An example

of deep watershed transformation based region segmentation is shown in Fig. 2.

Watershed transformation on original gray level images is prone to over segmentation due to the great number of local maximum or minimum points. On the contrary, the probability maps that our deep LPN model generates have very smooth values, and the local maximum points inside the same regions usually merge into connected areas, which is very suitable for watershed transformation.

The previous work [18] also involve the idea of using watershed transformation on deep feature maps for instance segmentation but they require semantic segmentation results as input and predict boundary direction maps and then watershed energy maps on which a fixed threshold is used to yield the final predictions, which is different from ours.

## IV. EXPERIMENTS

### A. Databases

We conducted our experiments on four public available databases: CASIA-HWDB [22], Bozen [23], cBAD-TrackB [20] and Maurdor [24].

CASIA-HWDB is a handwritten Chinese database with totally 5090 pages including 4075 pages used for training and 1015 pages used for testing. Text line segmentation on the original CASIA-HWDB database is easy, the reason we choose this database is that we have full control of its ground truth thus we can synthesize more challenging images by gradually narrow the gaps between adjacent text lines.

Bozen database consists of 400 pages written in Early Modern German including 350 pages used for training and 50 pages used for testing. Text line and baseline ground truth information is available in form of PAGE XML. This dataset is quite challenging for text line segmentation and baseline detection because most pages consist many difficulties such as bleed through, touching text lines and marginalia.

cBAD database [20] consists of two sub database namely TrackA[Simple Documents] and TrackB[Complex Documents]. In this paper, we only use TrackB for our experiments. The cBAD-TrackB database consists of 1380 pages, out of which 270 pages with PAGE XML ground truths are used for training and 1010 pages without ground truths are used for testing.

Maurdor is an extremely heterogeneous and challenging database for document region segmentation with multilingual (French, English, Arabic) and both handwritten and printed contents. It consists of overlapping and nesting regions of multiple classes. This dataset is composed of 8129 pages with GEDI XML ground truth, out of which 6129 pages are used for training, 1000 pages are used for validation and the rest 1000 pages are used for testing.

### B. Metrics

For text line segmentation on CASIA-HWDB and Bozen, we report the well known precision (P), recall (R) and their harmonic mean (F) under various intersection-over-union (IoU) levels, ranging from 0.5 to 0.95. For baseline detection on Bozen, we use the evaluation tool provided on the website[1] of ICDAR2017 competition on baseline detection. However, we don't have ground truth of cBAD-TrackB test set, so we upload and test our results on the aforementioned website and also report P, R and F values. For document region segmentation on Maurdor, we use Jaccard index and ZoneMap [25] as our metrics.

### C. Implementation Details

For all the experiments we conducted, we use the same parameter configurations and training approaches as follow: VGG16 model pre-trained on ImageNet is used to initialize our network, then the parameters are optimized by stochastic gradient descent (SGD) with batch size of 32 for 200k iterations on a Titan Xp GPU server of 12G memory. We crop the original document images into small patches with size of $256 \times 256$ pixels due to GPU memory limitation. The initial learning rate is $1.0 \times 10^{-6}$, then we deduce it by a factor of 10 for every 50k iterations. It's worth noting that we apply parallel training on two Titan Xp GPUs for Maurdor because it is a much larger database than the others. Another difference when training Maurdor is that we use *one-vs-all* strategy for multi-class training because the regions in Maurdor can overlap and nest with each other so that one pixel may have multiple labels.

### D. Experimental Results

**Text Line Segmentation.** To study LPN's ability of handling extremely near even overlapping regions, we conduct experiments on a series of synthesized CASIA-HWDB datasets with gradually narrowed gaps between adjacent text lines. Text line segmentation on the original CASIA-HWDB database is easy, however, as we can see from Fig. 1 and Table I, with the gaps between adjacent lines become smaller and smaller, the lines are more and more difficult to separate using the baseline FCN. On the contrary, our LPN can always segment the lines with high performance regardless the gap size between adjacent lines. Examples of segmentation results can be seen in Fig. 4 and Fig. 6. We also show the performance of our LPN and baseline FCN on CASIA-HWDB and Bozen at different IoU levels in Fig. 5. As we can see, our method can achieve impressive results on both two datasets under wide range of IoU levels.

**Baseline Detection.** Baseline detection results on Bozen and cBAD-TrackB are shown in TABLE II. The methods we list here are the wining method of ICDAR2017 competition on baseline detection and three other newly proposed methods with high performance. As we can see, our method outperforms the others with large margin on Bozen, and achieves comparable results with state-of-the-art methods on cBAD-TrackB. Please keep in mind that our

[1]https://scriptnet.iit.demokritos.gr/competitions/5/1/

method doesn't apply any post-processing procedures or data augmentation strategies, which have been proved can further improve the model performance. Interestingly, our method achieves highest recall among all the methods, which is more meaningful for the entire document processing system because the subsequent OCR systems can filter out most of the false detected baselines. Examples of baseline detection results can be seen in Fig. 6.

Table I: Text line segmentation on synthesized CASIA-HWDB data sets with gradually narrowed gaps (IoU=0.75).

| Method | | Gap Narrowed (pixels) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 10 | 20 | 30 | 40 | 50 |
| FCN | P | 0.9036 | 0.8230 | 0.7070 | 0.5723 | 0.4292 | 0.2644 |
| | R | 0.7958 | 0.6198 | 0.4129 | 0.2385 | 0.1215 | 0.0505 |
| | F | 0.8463 | 0.7071 | 0.5213 | 0.3367 | 0.1894 | 0.0848 |
| LPN | P | 0.9879 | 0.9846 | 0.9787 | 0.9695 | 0.9264 | 0.8398 |
| | R | 0.9957 | 0.9940 | 0.9904 | 0.9827 | 0.9436 | 0.8515 |
| | F | 0.9918 | 0.9893 | 0.9845 | 0.9760 | 0.9350 | 0.8456 |

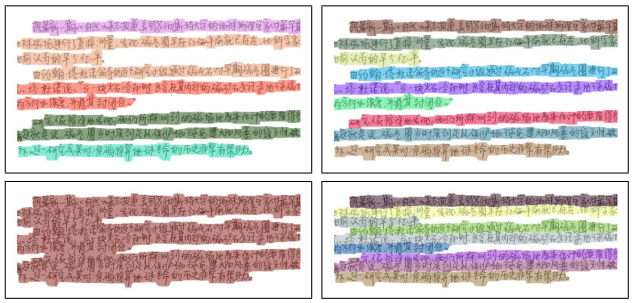

Figure 4: Text line segmentation on CASIA-HWDB. **Top**: FCN and LPN on original images; **Bottom**: FCN and LPN on synthesized images with 50 pixels of gap narrowed.

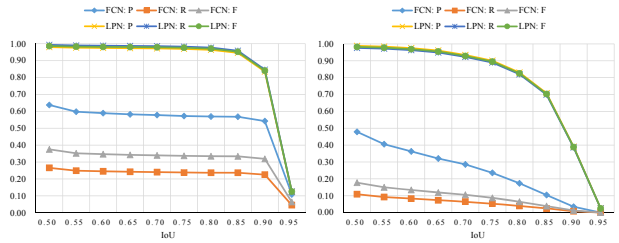


Figure 5: Performance at different IoU levels. **Left**: CASIA-HWDB with 30 pixels of gaps narrowed; **Right**: Bozen.

**Document Region Segmentation.** Document region segmentation results on Maurdor are shown in TABLE III. S1, S2, S3 and S5 are four systems from [25]. Our LPN model outperforms previous works and the baseline FCN with large margin under the metric of ZoneMap. Generally speaking, document region segmentation is a much more challenging task than text line segmentation and baseline detection because the inner-class variation of regions can be extremely dramatic. As we can see from Fig. 7, text regions in different categories of documents can have totally different sizes, shapes and visual appearances. The different configurations of our methods when matching the segmented regions and the reference regions in the process of calculating ZoneMap

Table II: Baseline detection on Bozen and cBAD-TrackB.

| Method | Bozen | | | cBAD-TrackB | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| DMRZ [20] | – | – | – | 0.8540 | 0.8630 | 0.8590 |
| Multi-Task [9] | 0.9580 | 0.9910 | 0.9740 | 0.8480 | 0.8540 | 0.8510 |
| dhSegment [10] | – | – | – | 0.8260 | 0.9240 | 0.8720 |
| ARU-Net [11] | 0.9765 | 0.9734 | 0.9750 | **0.9260** | 0.9180 | **0.9220** |
| Proposed | **0.9948** | **0.9986** | **0.9967** | 0.8864 | **0.9509** | 0.9176 |

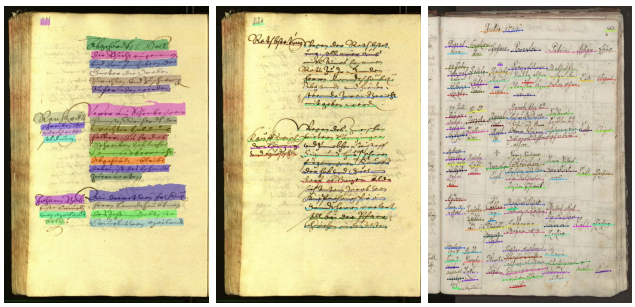

Figure 6: **Left**: Text line segmentation on Bozen; **Middle**: Baseline detection on Bozen; **Right**: Baseline detection on cBAD-TrackB.

scores are shown in TABLE IV. Compared with previous works and the baseline FCN, our LPN model generates most *Match* pairs and fewest *Merge* pairs, which demonstrates once again the great ability of our model to handle extremely near and overlapping regions. However, the *Split* pairs of LPN is more than that of FCN, which means our model tend to make more over-segmentation errors. How to solve this problem is still worthy of further research. Region segmentation results can be seen in Fig. 7.

Table III: Document region segmentation results on Maurdor.

| Method | ZoneMap | | | Jaccard |
|---|---|---|---|---|
| | $\alpha_c = 0.0$ | $\alpha_c = 0.5$ | $\alpha_c = 1.0$ | |
| S1 | 90.0 | 107.1 | 124.1 | 0.150 |
| S2 | 60.1 | 75.9 | 91.8 | 0.315 |
| S3 | 31.2 | 57.3 | 83.4 | 0.190 |
| S5 | 52.2 | 62.4 | 72.7 | 0.287 |
| FCN | 22.90 | 29.61 | 36.32 | **0.8656** |
| LPN | **17.81** | **23.57** | **29.32** | 0.8647 |

Table IV: Different configurations used to calculate ZoneMap.

| Method | Total | Match | Merge | Split | FA | Miss |
|---|---|---|---|---|---|---|
| S1 | 50145 | 7855 | 3226 | 10122 | 21236 | 7706 |
| S2 | 30625 | 8852 | 4710 | 5025 | 2324 | 9714 |
| S3 | 32846 | 13034 | 4784 | 4225 | 6851 | 3552 |
| S5 | 26418 | 8233 | 4534 | 4231 | 2246 | 7174 |
| FCN | 11353 | 7427 | 2928 | **239** | 368 | 391 |
| LPN | 20091 | **16953** | **1619** | 940 | **326** | **253** |

## V. Conclusion

In this paper, we propose a method based on Label Pyramid Network and Deep Watershed Transform for instance aware document region segmentation. Our method can handle extremely near even overlapping regions without using burdensome postprocessing procedures. The impressive performance achieved on four public available databases and three segmentation tasks demonstrate the effectiveness

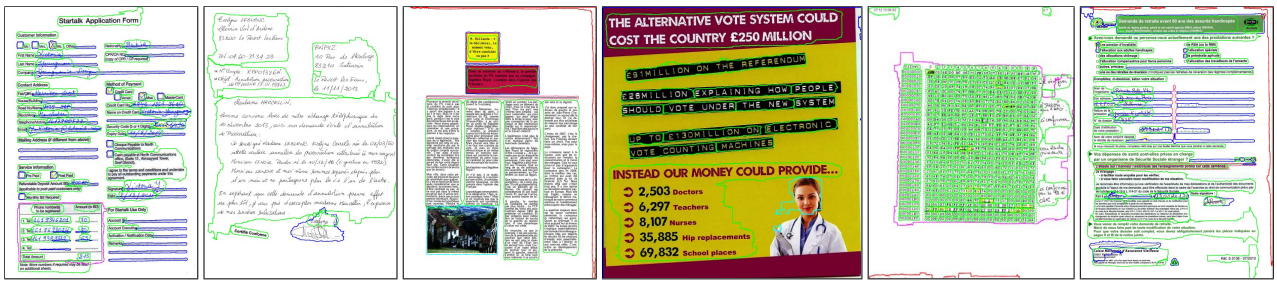


Figure 7: Document region segmentation on Maurdor.(text; photographic image; line drawing; graphics and subtypes; table; separator; noise.)

and superiority of the proposed method. However, there still exist some errors especially the over-segmentation errors, how to merge the over-segmented subregions belonging to the same instance is still an important problem worthy of further research.

In the future, we are planning to combine our method with structured prediction frameworks to predict the relationship of adjacent subregions and try to solve the problem of over-segmentation.

## ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC) Grants 61721004, 61573355, 61733007.

## REFERENCES

[1] A. M. Namboodiri and A. K. Jain, "Document structure and layout analysis," in *Digital Document Processing*, 2007, pp. 29–48.

[2] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal of Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 123–138, 2007.

[3] F. Shafait, D. Keysers, and T. Breuel, "Performance evaluation and benchmarking of six-page segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 941–954, 2008.

[4] A. Corbelli, L. Baraldi, C. Grana, and R. Cucchiara, "Historical document digitization through layout analysis and deep content classification," in *Proceedings of the 23rd International Conference on Pattern Recognition*, 2016.

[5] J. Pastor-Pellicer, M. Z. Afzal, M. Liwicki, and M. J. Castro-Bleda, "Complete system for text line extraction using convolutional neural networks and watershed transform," in *Proceedings of the 12th IAPR Workshop on Document Analysis Systems*, 2016, pp. 30–35.

[6] Q. N. Vo and G. Lee, "Dense prediction for text line segmentation in handwritten document images," in *Proceedings of the 23rd IEEE International Conference on Image Processing*, 2016, pp. 3264–3268.

[7] G. Renton, C. Chatelain, S. Adam, C. Kermorvant, and T. Paquet, "Handwritten text line segmentation using fully convolutional network," in *Proceedings of the 14th International Conference on Document Analysis and Recognition*, vol. 5, 2017, pp. 5–9.

[8] T. M. Breuel, "Robust, simple page segmentation using hybrid convolutional mdlstm networks," in *Proceedings of the 14th International Conference on Document Analysis and Recognition*, vol. 1, 2017, pp. 733–740.

[9] L. Quirós, "Multi-task handwritten document layout analysis," *arXiv preprint arXiv:1806.08852*, 2018.

[10] S. A. Oliveira, B. Seguin, and F. Kaplan, "dhsegment: A generic deep-learning approach for document segmentation," *arXiv preprint arXiv:1804.10371*, 2018.

[11] T. Grüning, G. Leifert, T. Strauß, and R. Labahn, "A two-stage method for text line detection in historical documents," *arXiv preprint arXiv:1802.03345*, 2018.

[12] K. Chen, M. Seuret, J. Hennebert, and R. Ingold, "Convolutional neural networks for page segmentation of historical document images," in *Proceedings of the 14th International Conference on Document Analysis and Recognition*, vol. 1, 2017, pp. 965–970.

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

[14] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. L. Giles, "Learning to extract semantic structure from documents using multimodal fully convolutional neural networks," in *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4342–4351.

[15] D. He, S. Cohen, B. Price, D. Kifer, and C. L. Giles, "Multi-scale multi-task fcn for semantic page segmentation and table detection," in *Proceedings of the 14th International Conference on Document Analysis and Recognition*, vol. 1, 2017, pp. 254–261.

[16] Y. Xu, W. He, F. Yin, and C.-L. Liu, "Page segmentation for historical handwritten documents using fully convolutional networks," in *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, vol. 1, 2017, pp. 541–546.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[18] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2858–2866.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[20] M. Diem, F. Kleber, S. Fiel, T. Grüning, and B. Gatos, "cbad: Icdar2017 competition on baseline detection," in *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, vol. 1, 2017, pp. 1355–1360.

[21] Z. Hayder, X. He, and M. Salzmann, "Boundary-aware instance segmentation," in *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5696–5704.

[22] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Casia online and offline chinese handwriting databases," in *Proceedings of the 11th International Conference on Document Analysis and Recognition*, 2011, pp. 37–41.

[23] J. A. Sánchez, V. Romero, A. H. Toselli, and E. Vidal, "Read dataset bozen," http: //dx.doi.org/10.5281/zenodo.218236.

[24] S. Brunessaux, P. Giroux, B. Grilheres, M. Manta, M. Bodin, K. Choukri, O. Galibert, and J. Kahn, "The maurdor project: improving automatic processing of digital documents," in *Proceedings of the 11th International Workshop on Document Analysis Systems*, 2014, pp. 349–354.

[25] O. Galibert, J. Kahn, and I. Oparin, "The zonemap metric for page segmentation and area classification in scanned documents," in *Proceedings of the 21st IEEE International Conference on Image Processing*, 2014, pp. 2594–2598.