Abstract

In this work, we look at the problem of structure extraction from document images with a specific focus on forms. Forms as a document class have not received much attention, even though they comprise a significant fraction of documents and enable several applications. Forms possess a rich, complex, hierarchical, and high-density semantic structure that poses several challenges to semantic segmentation methods. We propose a prior based deep CNN-RNN hierarchical network architecture that enables document structure extraction using very high resolution (1800 × 1000) images. We divide the document image into overlapping horizontal strips such that the network segments a strip and uses its prediction mask as prior while predicting the segmentation for the subsequent strip. We perform experiments establishing the effectiveness of our strip based network architecture through ablation methods and comparison with low-resolution variations. We introduce our new rich human-annotated forms dataset, and we show that our method significantly outperforms other segmentation baselines in extracting several hierarchical structures on this dataset. We also outperform other baselines in table detection task on the Marmot dataset. Our method is currently being used in a world-leading customer experience management software suite for automated conversion of paper and PDF forms to modern HTML based forms.

1. Introduction

Semantic structure extraction for documents has been explored in various works [17, 18, 46, 48]. Document structure extraction has been performed for digitising documents to make them reflowable which is useful in web based services [1, 14, 23, 35]. In this work, we look at a complex class of documents, i.e., Forms. Forms are documents that are used to capture data by various organisations in domains such as government services, finance, administration, and healthcare. Such industries which have been using paper forms or flat PDF forms would want to digitize them by converting them into an appropriate digitised version [35] (such as an HTML). Once these forms are made reflowable, they can be used on many devices with different form factors [1, 14]. This availability across devices automatically increases the ease of doing business or provide services since people can interact with them easily. Additionally, form digitization enables other capabilities such as better handling of data filled in digitized version, applying validation checks on data filled in fields, consistent form design control, auto-filling similar fragments etc.

(a) Fragment of a form having large empty spaces (b) Semantic structure induced around these empty spaces

Figure 1: Fragment of a form document

To make paper forms reflowable, we need to extract its semantic structure at multiple levels of hierarchy. Some low level elementary structure such as text can be extracted from the PDFs of forms. However, these PDFs do not contain any metadata about other higher-order structures and therefore, there is a need to extract such constructs. The semantic structure of forms is different from regular documents. Much of previous work looks at normal documents comprising of coarse structural elements that span a large area in the document image e.g., paragraphs, figures, lists, tables [48, 38, 30]. Forms have dense and intricate semantic structures as shown in figure 2 (right). In many forms, the structure is induced due to the presence of large empty ar-
The hierarchical semantic structure in forms consists of composite entities (such as Text Blocks, Text Fields, Choice Fields, Choice Groups, Lists, Tables etc.) that comprise of some basic entities like TextRuns and Widgets. We define a TextRun as a group of words present in a single line and widgets as empty spaces provided to fill information in forms. A TextBlock is a logical block of self-contained text comprising of one or more TextRuns, a Text Field comprises of a group of one or more Widgets and a caption Text Block describing the content to be filled in the field. Choice Fields are Boolean fields used for acquiring optional information. A Choice Group is a collection of such Choice Fields and an optional Choice Group Title which is a Text Block that describes instructions regarding filling the Choice Fields. Figure 2 (left) illustrates different semantic structures present in a form document at different hierarchical levels.

In this work, our goal is to extract such structures from the image of the Form PDF. We propose a method to extract the lower-level elements like TextRuns and Widgets along with higher-order structures like Fields, Choice Groups, Lists and Tables. Document structure extraction has been studied extensively with recent works mostly employing deep learning based fully convolution neural networks [18, 46, 48]. They extract structures through performing semantic segmentation [29, 6, 32] over document image. Such methods perform well at extracting coarser structures in documents, but do not perform well at extracting closely spaced structures in form images (as discussed in the Experiments section). Since they process the entire image in a single forward pass, due to memory constraints, they downscale the original image before providing it as input to their model. Down-scaling of input makes it difficult to disambiguate closely spaced structures, especially in dense regions and leads to merging of different structures.

Also to extract different objects, it is required to obtain precise boundaries between them to avoid merge. To do this, we:

- Propose a new prior and sub-strips based segmentation network pipeline mechanism to train a document segmentation network on a very high resolution ($1800 \times 1000$) image. Further, our network architecture does not require pre-training with Imagenet [9] or other large image datasets.
- Propose a multi-branch CNN-RNN based segmentation network that performs a hierarchical semantic segmentation for forms and documents.
- Introduce a new human-annotated Forms Image data set, which contains bounding boxes of a complete hierarchy of semantic classes and text annotations for all the text present in the images.
- Compare our method with semantic segmentation based baselines outperforming them significantly and show ablations highlighting the importance of strip based high resolution training and our network architecture design.

Our method is currently being used in a world leading customer experience management software suite for automated conversion of paper and pdf forms to modern HTML based forms.

We propose a prior based strip segmentation to mitigate the memory limitation on GPU while training a neural network on high resolution images. The intuition is to divide the image into overlapping horizontal strips and segment each strip individually. Strip based segmentation method can potentially fail to predict continuous semantic structures that span across multiple strips. Hence we introduce a prior, where each image strip’s prediction is cached on the GPU and provided as prior concatenated with the input while predicting the segmentation mask of the subsequent strip. In addition, the multi-branch CNN-RNN network was introduced to train multi-level hierarchy segmentation together...
in a single network, so that it learns to predict consistent
segmentation masks across these hierarchies [4, 28].
Structures that typically span a large area of a form or
document like tables and lists could be processed at a lower
resolution but they significantly benefit from the CNN-RNN
architecture. We demonstrate this in section 4.3.3.

2. Related Work

Document structure analysis started as heuristic-based
methods [16, 10, 15, 44] based on handcrafted features [25]
for extracting paragraphs and graphics. Other approaches,
like connected components, cut based methods, and others,
were also used for extracting text areas[15] and physical
layouts[44]. Most of these approaches can be classified
into either top-down or bottom-up approaches. The top-
down methods focus on extracting text-lines and aggregat-
ing them into paragraphs. The bottom-up approaches go
about detecting layout by subdividing the page into blocks
and columns.

Most of the recent deep-learning based approaches are
based on fully-convolution neural network (FCN)[48, 18,
46] and avoid any heuristic-based approaches. These FCN’s
are trained to generate semantic segmentation [29] for the
rasterized version of the document. FCNs have also been
used to locate and recognize handwritten annotations in old
documents [24]. [47] proposed a joint text detection and
recognition model. They used a region proposal network
which detects the beginnings of text lines, a line following
model predicts a sequence of short bounding boxes along
the text-line, which is then used to extract and recognize
the text. We employ our high resolution segmentation net-
work to extract and disambiguate closely spaced textruns
and textblocks from form images.

Table detection has been the key focus of some works
like [17, 18, 26, 2, 36]. In [17], table region candidates
were chosen based on some loose rules which were later
filtered using a CNN. In [18], an FCN was proposed,
having a multi-scale architecture which had two branches
where one was dedicated to table segmentation while the
other was used to find contours. After that, an additional
CRF(Conditional Random Field) was used to refine the seg-
mentation output further. We propose a multi-branch ar-
britecture to segment hierarchical structures that overlap in
same region in a form. For tables, we compare our model
with [18] on marmot dataset, one of the largest publicly
available table evaluation dataset [12]. While there are
other works [33, 38] that perform table decomposition into
rows and columns (which our model is capable of doing),
we discuss table detection only in the scope of this paper.
Other works like [43] introduced a large dataset of 5.5 mil-
lion document labels focusing on detecting bounding boxes
for figures on this dataset using an Overfeat [40] network,
trained over image embedding generated using ResNet-101.

It is evident that FCN based segmentation approaches
have led to great advancement in document structure ex-
traction. However, a few approaches have also tried other
network architectures and input modalities such as text.
[22, 27] are some of the multi-modal approaches proposed
to extract named entities from invoices. Other network ar-
chitectures such as Graph Neural Networks (GNNs) have
been explored in [37, 34] for detecting tables in invoice
documents and parsing table structure, respectively. In a re-
lated domain of document classification also, CNN based
methods have been explored in recent times. [42] used
them for document verification. Moreover, [49] proposed
HAN to create sentence and document embedding in a hi-
erarchical fashion using a multi-level attention mechanism.

Document classification has also been explored using multi-
modal models [3] by extracting visual and textual features
from MobileNet[19] and FastText [5] respectively. These
features are later on concatenated to learn a better classifi-
cation model.

Multi-modal Semantic segmentation has been proposed
in [48] to extract figures, tables, lists, and several other
types of document structures. A text embedding map for
the entire page of the document image gets concatenated
with the visual feature volume in a spatially coherent man-
er such that there is a pixel to text correspondence. They
also introduce a region consistency loss in addition to the
cross-entropy based segmentation loss. We use their ap-
proach as one of our baselines on forms. Another baseline
against which we compare our model is the DeepLabV3+
[7], which is the current state-of-the-art semantic segmen-
tation network.

Our model is a CNN-RNN based semantic segmenter.
CNN-RNN based networks have been used earlier [45, 41]
for semantic segmentation on natural images however later
research using CNN based architectures [20, 21] achieved
better results. In this work, we share a new CNN-
RNN based hierarchical semantic segmentation network
and show that it is more effective for form documents as
compared to [7, 48]. We share an ablation study showing
that replacing RNNs in our architecture with 1D dilated [50]
convolution degrades performance for form documents.

3. Methodology

In this section, we discuss our proposed model which is
used to extract various structures from the form documents
like widgets, fields, textblocks, choicegroups etc. For doc-
uments, especially in the case of forms, the semantic struc-
ture extends extensively in both vertical and horizontal di-
rections. For many structures such as widgets and fields,
it may even extend to empty spaces in the input image re-
quiring the model to predict objects in parts where there is

\footnote{http://www.icst.pku.edu.cn/cpdp/sjzy/index.htm}
no explicit visual signal. Also, the higher-level structures are composed of lower-level elements and it is necessary to make fine grained predictions at different levels. This leads to our motivation to use a multi-branch CNN-RNN based semantic segmenter to capture long range relationships and predict multiple masks at different hierarchies that are mutually consistent. Finally, to address the issue of dense text documents and forms, we modify the network input mechanism by enabling a tile stitching behavior in our network while performing segmentation to train it at higher resolutions.

3.1. Network Pipeline

Fully convolution segmentation networks [29, 6, 32] takes entire input image and processes it to generate the segmentation mask. Due to this, the network’s memory footprint increases significantly with the input image resolution. Since the GPUs used for training have limited memory, it limits the resolution at which the input can be provided during training. In this section, we discuss our network pipeline, which addresses and mitigates this limitation.

We convert the RGB input image into grayscale and resize the grayscale image having height and width \((I_H \times I_W)\) to \((H \times w)\) such that \(I_W\) scales to \(w\) and \(I_H\) gets scaled by the same ratio, i.e., \(w/I_W\). The resulting image is further cropped or padded with zeros to a size of \(h \times w\). \(h\) is kept larger than \(H\) to accommodate elongated document images. We divide the input image into overlapping horizontal strips. Let \(S_h\) be strip height, \(O_h\) be overlap height between consecutive strips, \(SegNet\) is our segmentation network, and \(SegMsk\) denote segmentation mask. Following this notation, Algorithm 1 describes our method where the network predicts the segmentation mask of different strips in succession. Each strip’s mask prediction uses the predicted mask for the previous strip as prior. The first strip uses a zeroed out mask as prior. It is pertinent to point out that the gradients do not flow from one strip to its previous strip while training. Since \(SegNet\) predicts multiple segmentation masks corresponding to the structures at different levels of hierarchy, instead of giving predictions for all structures as prior, we select as prior only a few channels corresponding to classes like background, border, text, widget, and choice groups. We copy logits corresponding to these classes from multiple segmentation output masks to one prior mask having many channels with each channel dedicated to one class.

As stated earlier, we use a CNN-RNN based network architecture to predict precise and uniform segmentation masks. Our network broadly comprises of three components, Image Encoder (IE), Context Encoder (CE) and Decoders (DE). We concatenate a prior mask to each image strip and feed it into the Image Encoder (IE) to generate features at multiple granular levels. The final features of IE, are then processed through a RNN based Context Encoder (CE), which generates features capturing contextual dependencies. The features extracted from CE at different layers represent separate levels of semantic hierarchy like lower level (text, widget) to higher levels (e.g., fields, choice-fields, and choice groups). All these sets of features from multiple levels of CE and IE is then passed to separate Decoders (DE) to generate segmentation masks for different semantic structure levels. We would now explain each of these modules in greater detail.

3.2. Network Architecture

3.2.1 Image Encoder

The figure 3 depicts the architecture of Image Encoder (IE) that comprises of multiple convolution layers, max-pooling layers, and dilated convolutions [50]. As shown in the figure, the first conv layer has \(3 \times 3\) kernel with a 48 channel output. The parameters of the remaining layers are highlighted using the same notation. Each convolution layer has a stride of 1 unless specified, the third conv layer in Image Encoder has a stride of 2 and is denoted by “1/2” in the figure. Similarly, all the max-pooling
layers have a stride of 2 by default. The dilated convolution block at the end of Image Encoder, consists of four dilated conv layers, that work in parallel on the same feature volume at different dilation rates. The output of these dilated convolutions is concatenated and passed on to the Context Encoder. We extract several intermediate features \((\text{detail}_1, \text{detail}_2, \text{detail}_3, \text{detail}_4)\) from the Image Encoder that act as skip connections \([31]\) and are used by the separate decoders.

### 3.2.2 Context Encoder

The context encoder (CE) is composed of five 2D RNN layers; each consisting of 2 bidirectional RNNs \([39]\) working in both horizontal and vertical directions. Each 2D RNN first processes the feature volume in vertical direction, and subsequently, its outputs are processed in a horizontal direction. We use the intersection RNN cell \([8]\) as the basic unit in our RNN layers. The first 2D RNN has a state-size of 196 in both forward and backward directions, and outputs a 3D volume with 392 channels. Similarly, the rest of the 2D RNNs have a state-size of 144 for forward and backward units, that produce a volume of 288 channels, as shown in Fig 3. Each of the 2D RNN’s output, starting from the second 2D RNN, is fed to a CNN decoder to predict the segmentation mask at one level of the hierarchy.

### 3.2.3 Decoder

The network consists of multiple decoders for generating segmentation maps for different levels of semantic structure. Each decoder operates on a different feature volume from Context Encoder. It up-samples it by passing it through a transposed convolution layer \([32]\). The up-sampled features are subsequently passed through another conv layer. Finally, these features are concatenated with another feature volume \(\text{detail}_4\), obtained from Image Encoder. The decoder branch repeats the sequence of such operations multiple times, as shown in the Fig 3. The first decoder that predicts the lowest level of the semantic structure (TextRun and Widget) takes an extra feature volume \(\text{detail}_1\) from IE as input to predict the segmentation mask. The other decoders have similar architecture but separate set of weights. Each convolution in the decoder branch has a stride of 1, and each transpose conv, depicted as \(\text{convT}\) in the figure 3, has a stride of 2 by default. The different decoders are used to predict segmentation masks for different spatially overlapping classes like widget and fields. Such a network design helps in segregating the classes according to hierarchy since the container groups and their constituent classes are predicted in separate masks.
4. Experiments

4.1. Datasets

**Forms Dataset**: We used our rich Forms Dataset comprising of 52,490 human annotated Form images. The Form images were extracted from the corresponding PDF Form documents. We had these Form images annotated to represent the hierarchy and spatial boxes of different form elements and structures such as Text Runs, Widgets, Text Blocks, Choice Group Titles, Text Fields, Choice Fields, Choice Groups, Tables and Lists. We split the dataset into 48,256 images for training and 3,234 images for validation. We used a set of 1,000 separate test images for the final evaluation of our model with the baselines and to perform ablation studies. We plan to release this test set to help advance research in this field.

**Marmot Dataset**: We evaluate and compare our model trained on Forms Dataset on the Marmot Dataset. This dataset is one of the largest publicly available Table evaluation datasets. It contains a total of 2000 document images corresponding to approximately equal number of English and Chinese documents.

4.2. Implementation Details

We set $w = 1000$, $h = 1800$, $S_h = 600$, $O_h = 200$ for the SegNet model defined in Section 3. We slice the high resolution input image into 4 overlapping horizontal strips. All the convolution and deconvolution layers have ReLU activation. We train our model at a batch size of 32 on 8 Tesla V100 GPUs in parallel. We use AdaDeltaOptimizer [51] to train the parameters of our model with an exponential decaying learning rate using $1 \times 10^{-1}$ as the starting learning rate and a decay factor of 0.1. Please refer to Figure 3 for specific configuration details of different network layers. To enable the network to predict concise masks, we use convex hull [13] to determine segmentation masks.

4.3. Results

4.3.1 Model Evaluation and Ablation Studies

We evaluate our high resolution model comprising of 4 decoders. The decoders predict structures in ascending order of hierarchy. The first decoder predicts elementary structure regions consisting of TextRuns and Widget. The second decoder predicts structures next in the hierarchy – TextBlocks and ChoiceGroupTitles. The third decoder extracts text Fields and ChoiceFields and finally the fourth decoder predicts ChoiceGroups. We add another class - Border, surrounding each structure and make the network predict this class to enable it to disambiguate different objects and generalise better.

We subsequently refer to this network configuration as **HighResNet**. We perform ablations establishing gains from our high resolution segmentation network by comparing it against: 1) LowResNet - a low resolution variation of HighResNet that takes input image at 792 resolution and predicts hierarchical segmentation masks for the entire image in a single forward pass without dividing it into strips; 2) NoPriorNet - A HighResNet variation where we divide the input image into 3 horizontal strips with no overlap between consecutive strips. In this variant, the segmentation mask predicted for a strip is not given as prior for the subsequent strip prediction; 3) NoRNNNet - where the 2d horizontal(vertical) RNNs in our network architecture are replaced with horizontal(vertical) 1d dilated convolution blocks such that each block comprises of four dilated convolution layers each having a kernel size of $1 \times 9$ ($9 \times 1$) and dilation rate of 1, 2, 4, 8 respectively. We use pixel mean Intersection over Union (MIoU) to evaluate different models. We summarise MIoU scores for different ablations in Table 1.

**Compare HighResNet with LowResNet**: It can be seen that by extracting hierarchical structure in high resolution, HighResNet is able to improve the MIoU scores significantly with 2.4%, 5.2%, 0.8%, 11.4%, 6%, 15.1%, 12.4% absolute improvements for TextRuns, Widgets, TextBlocks, ChoiceGroupTitles, Text Fields, ChoiceFields and Choice Groups respectively.

**Compare HighResNet with NoPriorNet**: Adding predicted segmentation mask as prior while making prediction for the subsequent strip in a page improves the MIoU scores marginally for TextRun, TextBlock, ChoiceGroupTitle with notable improvements of 1.6%, 1.4%, 1.9% and 2.1% in absolute MIoU for Widget, Text Field, Choice Field and Choice Group respectively. The most significant improvement is observed for ChoiceGroups. Being a larger hierarchical construct, it is probable for a choice group to span across consecutive strips. Hence, adding prior assists in completing the segmentation mask across strips.

**Compare HighResNet with NoRNNNet**: It can be seen that using RNN to capture context and long distance relationship yields better performance compared to 1d dilated convolution blocks. HighResNet yields 1.2%, 2.2%, 3.4% and 2% absolute improvement in Text Run, Choice Group Title, Choice Field and Choice Group respectively.

We estimate recall and precision on structure extraction task and compare with ablation methods. To perform this, we consider a predicted structure as correct match if the IoU of its predicted mask is above a certain threshold (0.7) with an expected structure mask. We report the results in Table 2 where we obtain similar results with HighResNet performing better than other ablation methods.
4.3.2 Comparison with Baselines

We consider two semantic segmentation baselines - DeepLabV3+ (DLV3+) [7], which is the state of the art for semantic segmentation tasks on natural images and Multimodal Fully Convolutional Network (MFCN) [48] designed for extracting several complex structures in documents. The baseline segmentation models segment the input image into a flattened hierarchy while our forms comprise of hierarchical structures. To address this, we process the output of the penultimate layer of the baseline models through 4 separate FC layers to obtain hierarchical masks using data schema similar to HighResNet.\(^4\) We train the baselines on RGB images at a resolution of 792×792 following an aspect ratio preserving resize. For MFCN, loss for different classes is scaled according to pixel area covered by elements of each class (calculated over the dataset) as described in their work. Table 1 compares the MIoU of our approach with the baselines. As can be seen, our model HighResNet significantly outperforms both the baselines on all form structures.

Figure 4(Left) illustrates segmentation masks predicted by different baseline methods and our model on a sample form image. It can be seen that while the baseline methods merge different elements and hierarchical structures such as TextBlocks and Fields, our model predicts crisp segmentation masks extracting all such structures almost perfectly. For choice group, it can be seen that the baseline methods predict incomplete segmentation mask while our model is able to capture long range dependencies among its constituent elements and predict a complete mask.

4.3.3 Evaluation on Other Higher Order Constructs

In this section, we discuss the performance of our model at extracting other higher order structures like Lists and Tables. These structures are relatively more evident and span large regions in a page reducing the need to disambiguate them in high resolution. Consequently, we train a separate low resolution (792×792) version of our proposed network analogous to LowResNet which we refer to as LowResNet-TL in order to predict these structures. In order to evaluate the performance of this network, we also train a separate network for the two baselines for extracting Tables and Lists. Table 3 compares the MIoU of our method with the baseline models and the Figure 4 (Right) illustrates the network outputs for the task of Table and List segmentation. It can be seen that LowResNet-TL significantly outperforms both the baselines.

\(^4\)We trained separate models of each baseline corresponding to each hierarchical structure but it did not provide any further improvements.

| Method          | TextRun | Widget | TextBlock | ChoiceGroup | Text Field | Choice Field | Choice Group |
|-----------------|---------|--------|-----------|-------------|------------|--------------|--------------|
| DLV3+           | 65.8    | 58.2   | 68.7      | 16.9        | 61         | 24.2         | 28.2         |
| MFCN            | 77.5    | 49.2   | 69.9      | 25.1        | 39.3       | 26.4         | 30.3         |
| LowResNet       | 88      | 75.4   | 87        | 60.9        | 77.2       | 63.4         | 67.3         |
| NoPriorNet      | 89.6    | 79     | 87.7      | 71.6        | 81.8       | 76.6         | 77.6         |
| NoRNNNet        | 89.2    | 80.1   | 87.2      | 70.1        | 82.8       | 75.1         | 77.7         |
| HighResNet      | 90.4    | 80.6   | 87.8      | 72.3        | 83.2       | 78.5         | 79.7         |

Table 1: Mean IoU of different baseline methods and ablation models on elements and several hierarchical form structures

| Model          | TextRun | Widget | TextBlock | ChoiceGroup | Text Field | Choice Field | Choice Group |
|----------------|---------|--------|-----------|-------------|------------|--------------|--------------|
| LowResNet      | P       | R      | P         | R           | P          | R            | P            |
| NoPriorNet     | 76.6    | 55.9   | 59.4      | 59.2        | 60.0       | 57.3         | 65.7         |
| NoRNNNet       | 78.2    | 63.3   | 64.7      | 68.8        | 66.7       | 70.3         | 75.1         |
| HighResNet     | 78.4    | 63.7   | 63.1      | 68.6        | 66.6       | 69.2         | 73.0         |
| DLV3+ (list only) | –       | 54.5   | –         | –           | 62.2       | 73.0         | 66.9         |
| DLV3+ (table only) | 65.1    | 71.3   | 69.5      | 79.7        | 65.5       | 74.5         | 67.3         |
| MFCN           | 49.4    | 44.9   | 34.7      | 50.0        | 37.8       | 41.2         | 26.1         |

Table 2: Precision-Recall numbers for the different hierarchical form structures on the different ablation models computed with an IoU threshold of 0.7

| Method      | Table | List |
|-------------|-------|------|
| DLV3+ [7]   | 66.8  | 56.7 |
| DLV3+ (list only) | –     | 54.5 |
| DLV3+ (table only) | 65.1 | –    |
| MFCN [48]   | 49.4  | 44.9 |
| LowResNet-TL| 78.5  | 68.2 |

Table 3: Comparison of MIoU scores of our method with the baselines for Table and List on Forms Dataset.
Table 4: Comparison of Table precision-recall numbers on the Marmot Dataset

| Method                  | English | Chinese |
|-------------------------|---------|---------|
|                         | P  | R  | P  | R  |
| **Fang [11]**           | 58.0 | 49.0 | 89.0 | 80.0 |
| **MSMT-FCN (0.8 IoU)**  | 75.3 | 70.0 | 77.0 | 76.1 |
| **MSMT-FCN (0.9 IoU)**  | 47.0 | 45.0 | 49.3 | 49.1 |
| **LowResNet-TL (0.8 IoU)** | 72.7 | 76.8 | 70.0 | 79.1 |
| **LowResNet-TL (0.9 IoU)** | 61.2 | 64.6 | 62.3 | 70.5 |

Additionally, we also compare the precision-recall numbers of the Table predictions of LowResNet-TL on the Marmot Dataset with other prior methods – Fang [11] and Multi-Scale Multi-Task FCN (MSMT-FCN) [18] in Table 4. It can be seen that LowResNet-TL performs similar to MSMT-FCN for an IoU threshold of 0.8. However, LowResNet-TL performs significantly better than MSMT-FCN for a higher IoU threshold of 0.9 indicating the architecture is able to predict much crisper predictions.

5. Conclusion

In this paper, we propose a novel neural network architecture and training mechanism to extract document structure on very high resolution form images. We observe that a much higher resolution segmentation is beneficial for extracting structure, particularly on forms since they possess highly dense regions. We propose a prior based strip segmentation approach where we divide the input image into overlapping horizontal strips, perform segmentation on the strip and use the prediction output as prior while segmenting the subsequent strip. We propose a multi-trunk hierarchical CNN-RNN based encoder decoder model as our underlying segmentation network that captures long range contextual dependencies while segmenting different hierarchical constructs. Various ablation studies show the effectiveness of our high resolution segmentation approach and network architecture design. We compare our method with different
semantic segmentation baselines outperforming them significantly on our Forms Dataset for several structures such as TextBlocks, Fields, Choice Groups etc. Additionally, our model trained on Forms Dataset outperforms prior art for table detection on Marmot dataset. We also introduce our Forms Dataset and plan to release the test set of this dataset.

Our method is currently being used in a world leading customer experience management software suite for automated conversion of paper and pdf forms to modern HTML based forms.

References

[1] Hassan Alam and Fuad Rahman. Web document manipulation for small screen devices: A review. In Web Document Analysis Workshop (WDA), 2003.

[2] Saman Arif and Faisal Shafait. Table detection in document images using foreground and background features. In 2018 Digital Image Computing: Techniques and Applications (DICTA), pages 1–8. IEEE, 2018.

[3] Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. Multimodal deep networks for text and image-based document classification. arXiv preprint arXiv:1907.06370, 2019.

[4] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. Machine Learning, Jul 1997.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062, 2014.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018.

[8] Jasmine Collins, Jascha Sohl-Dickstein, and David Sussillo. Capacity and trainability in recurrent neural networks. In ICLR, 2017.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[10] Dimitrios Drivas and Adnan Amin. Page segmentation and classification utilising a bottom-up approach. In Proceedings of 3rd International Conference on Document Analysis and Recognition, volume 2, pages 610–614. IEEE, 1995.

[11] Jing Fang, Liangcai Gao, Kun Bai, Ruiheng Qiu, Xin Tao, and Zhi Tang. A table detection method for multipage pdf documents via visual separators and tabular structures. In 2011 International Conference on Document Analysis and Recognition, pages 779–783. IEEE, 2011.

[12] Jing Fang, Xin Tao, Zhi Tang, Ruiheng Qiu, and Ying Liu. Dataset, ground-truth and performance metrics for table detection evaluation. In 2012 10th IAPR International Workshop on Document Analysis Systems, pages 445–449. IEEE, 2012.

[13] Ronald L Graham and F Frances Yao. Finding the convex hull of a simple polygon. Journal of Algorithms, 4(4):324–331, 1983.

[14] Aditya Gupta, Anuj Kumar, VN Tripathi, S Tapaswi, et al. Mobile web: web manipulation for small displays using multi-level hierarchy page segmentation. In Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology, pages 599–606. ACM, 2007.

[15] Jaekyu Ha, Robert M Haralick, and Ilhsin T Phillips. Document page decomposition by the bounding-box project. In Proceedings of 3rd International Conference on Document Analysis and Recognition, volume 2, pages 1119–1122. IEEE, 1995.

[16] Jaekyu Ha, Robert M Haralick, and Ilhsin T Phillips. Recursive xy cut using bounding boxes of connected components. In Proceedings of 3rd International Conference on Document Analysis and Recognition, volume 2, pages 952–955. IEEE, 1995.

[17] Leipeng Hao, Liangcai Gao, Xiaohan Yi, and Zhi Tang. A table detection method for pdf documents based on convolutional neural networks. In 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pages 287–292. IEEE, 2016.

[18] Dafang He, Scott Cohen, Brian Price, Daniel Kifer, and C Lee Giles. Multi-scale multi-task fcn for semantic page segmentation and table detection. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 1, pages 254–261. IEEE, 2017.

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[20] Simon Jegou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, July 2017.

[21] X. Jin, X. Li, H. Xiao, X. Shen, Z. Lin, J. Yang, Y. Chen, J. Dong, L. Liu, Z. Jie, J. Feng, and S. Yan. Video scene parsing with predictive feature learning. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 5581–5589, Oct 2017.

[22] Anoop Ravendira Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards understanding 2d documents. arXiv preprint arXiv:1809.08799, 2018.

[23] Mohamed Khemakhem, Axel Herold, and Laurent Romary. Enhancing usability for automatically structuring digitised dictionaries. 2018.

[24] Andreas Kölsch, Ashutosh Mishra, Saurabh Varshneya, Muhammad Zeshan Afzal, and Marcus Liwicki. Recogniz-
[49] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pages 1480–1489, 2016. 3

[50] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015. 3, 4

[51] Matthew D Zeiler. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012. 6