VSR: A Unified Framework for Document Layout Analysis combining Vision, Semantics and Relations

Peng Zhang¹
Can Li¹
Liang Qiao¹
Zh anzhan Cheng²¹
Shiliang Pu¹
Yi Niu¹
Fei Wu²

1. Hikvision Research Institute, Hangzhou, China
2. Zhejiang University, Hangzhou, China
Background

◆ Document Layout Analysis

- Vision
- Semantics
- Relations

Image
Table
Text
Title
Background

- Multimodal document layout analysis frameworks
  - NLP-based framework

Components generation: (e.g. PDF parsing)

Feature Extraction: vision + semantics

Relation Module: Sequence labeling

Limitations:
  - Insufficient capabilities in layout modeling
Multimodal document layout analysis frameworks

- CV-based framework

**Limitations:**
- Limited semantics
- Simple and heuristic modality fusion strategy
- Lack of relation modeling between components
Background

- Multimodal document layout analysis frameworks
  - VSR (Vision, Semantics and Relations)

Advantages:
- Semantics at multiple granularities (Character & Sentence)
- Two-stream network and adaptive aggregation module to exploit vision and semantics effectively
- A GNN-based relation module to support relation modeling in both NLP- and CV-based methods
Method

- Two-stream ConvNets
- Multi-scale Adaptive Aggregation
- Relation Module
Method

- Two-stream ConvNets

**input** (document image): \( V_0 = x \in \mathbb{R}^{H \times W \times 3} \)

**output** (multi-scale visual features):
\[
\{V_2, V_3, V_4\} \quad V_i \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times c_i^V}
\]

**input** (text embedding maps):
\[
S_0 = \text{LayerNorm}(\text{Chargrid} + \text{Sentgrid}) \in \mathbb{R}^{H \times W \times c_0^S}
\]

- character granularity
- sentence granularity

**output** (multi-scale semantic features):
\[
\{S_2, S_3, S_4\} \quad S_i \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times c_i^S}
\]
Method

- Multi-scale Adaptive Aggregation

\[
\{V_2, V_3, V_4\} \quad V_i \in \mathbb{R}^{H_i \times W_i \times C_i}
\]

\[
\{S_2, S_3, S_4\} \quad S_i \in \mathbb{R}^{H_i \times W_i \times C_i}
\]

\[
AM_i = h\left(g\left([V_i, S_i]\right)\right)
\]

\[
FM_i = AM_i \odot V_i + (1 - AM_i) \odot S_i
\]

\[
\{FM_2, FM_3, FM_4\}
\]

[·]: concatenation

\(g\): convolutional layer
\(h\): activation function
Method

- Relation Module

Nodes: $Z = \{z_1, \cdots, z_N\}$

self-attention

Updated Nodes: $Z' = \{z'_1, \cdots, z'_N\}$

node features: $z_j = \text{LayerNorm}(f_j + e_j^{pos}(b_j))$

visual features: $f_j = \text{RoIAlign}(FM, b_j)$

position embeddings: $e_j^{pos}(b_j)$

probabilities: $\hat{p}_j = \text{Softmax}(\text{Linear}_{cls}(z'_j))$

regression coordinates: $\tilde{b}_j = \text{Linear}_{reg}(z'_j)$
Outline

- Background
- Method
- Experiments
## Experiment

- **Datasets**

| Dataset         | Num of Samples | Metric          | Classes                                                                 | Support tasks                             |
|-----------------|----------------|-----------------|-------------------------------------------------------------------------|-------------------------------------------|
| Article Regions | 822            | mAP             | Title, Authors, Abstract, Body, Figure, Figure Caption, Table, Table Caption, References | CV-based method                           |
| PubLayNet       | 360K           | AP@IOU 0.5-0.95 | Text, Title, List, Figure, Table                                         |                                           |
| DocBank         | 500K           | F1-score mAP    | Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table, Title | CV-based method + NLP-based method         |
Experiment

- SOTA results
  - Article Regions

**Table 1. Performance comparisons on Article Regions dataset**

| Method                              | Title | Author | Abstract | Body | Figure Caption | Table Caption | Reference | mAP  |
|-------------------------------------|-------|--------|----------|------|----------------|---------------|-----------|------|
| Faster RCNN [31]                    | -     | 1.22   | -        | 87.49| -              | -             | -         | 46.38|
| Faster RCNN *w/ context* [31]      | -     | 10.34  | -        | 93.58| -              | -             | -         | 70.3 |
| Faster RCNN *reimplement*           | 100.0 | 51.1   | 94.8     | 98.9 | 94.2           | 91.8          | 97.3      | 67.1 | 90.8 | 87.3 |
| Faster RCNN *w/ context reimplement*| 100.0 | 60.5   | 90.8     | 98.5 | **96.2**       | 91.5          | **97.5**  | 64.2 | 91.2 | 87.8 |
| VSR                                 | **100.0** | **94** | **95** | **99.1** | **95.3** | **94.5** | **96.1** | **84.6** | **92.3** | **94.5** |

Note: missing entries are because those results are not reported in their original papers.
## Experiment

- **SOTA results**
  - PubLayNet

### Table 2. Performance comparisons on PubLayNet dataset.

| Method          | Dataset | Text  | Title | List  | Table | Figure | AP   |
|-----------------|---------|-------|-------|-------|-------|--------|------|
| Faster R-CNN [43] | val     | 91.6  | 84    | 88.6  | 96    | 94.9   | 91   |
| Mask R-CNN [43]  |         |       |       |       |       |        |      |
| VSR             |         | 96.7  | 93.1  | 94.7  | 97.4  | 96.4   | 95.7 |
| Faster R-CNN [43] | test    | 91.3  | 81.2  | 88.5  | 94.3  | 94.5   | 90   |
| Mask R-CNN [43]  |         |       |       |       |       |        |      |
| DocInsightAI     |         | 94.51 | 88.31 | 94.84 | 95.77 | 97.52  | 94.19|
| SCUT            |         | 94.3  | 89.72 | 94.25 | 96.62 | 97.68  | 94.51|
| SRK             |         | 94.65 | 89.98 | 95.14 | 97.16 | 97.95  | 94.98|
| SiliconMinds    |         | 96.2  | 89.75 | 94.6  | 96.98 | 97.6   | 95.03|
| VSR             |         | 96.69 | 92.27 | 94.55 | 97.03 | 97.90  | 95.69|
Experiment

- SOTA results
- DocBank

### Table 3. Performance comparisons on DocBank dataset in F1 Score.

| Method             | Abstract | Author | Caption | Equation | Figure | Footer | List | Paragraph | Reference | Section | Table | Title | Macro Average |
|--------------------|----------|--------|---------|----------|--------|--------|------|-----------|-----------|---------|-------|-------|---------------|
| BERT\_base         | 92.94    | 84.84  | 86.29   | 81.52    | 100.0  | 78.05  | 71.33| 96.19     | 93.10     | 90.81   | 82.96 | 94.42 | 87.70         |
| RoBERT\_base       | 92.88    | 86.18  | 89.44   | 82.48    | 100.0  | 80.14  | 73.53| 96.46     | 93.41     | 93.37   | 83.89 | 95.11 | 88.91         |
| LayoutLM\_base     | 98.16    | 85.95  | 95.97   | 89.47    | 100.0  | 89.57  | 89.48| 97.88     | 93.38     | 95.98   | 86.33 | 95.79 | 93.16         |
| BERT\_large        | 92.86    | 85.77  | 86.50   | 81.77    | 100.0  | 78.14  | 69.60| 96.19     | 92.84     | 90.65   | 83.20 | 94.30 | 87.65         |
| RoBERT\_large      | 94.79    | 87.24  | 90.81   | 83.70    | 100.0  | 83.92  | 74.51| 96.65     | 93.34     | 94.07   | 84.94 | 94.61 | 89.88         |
| LayoutLM\_large    | 97.84    | 87.83  | 95.56   | 89.74    | 100.0  | 91.46  | 90.04| 97.90     | 93.32     | 95.96   | 86.79 | 95.52 | 93.50         |
| X101                | 97.17    | 82.27  | 94.35   | 89.38    | 88.12  | 90.29  | 90.51| 96.82     | 87.98     | 94.12   | 83.53 | 91.58 | 90.51         |
| X101+LayoutLM\_base| 98.15    | 89.07  | 96.69   | 94.30    | 99.90  | 92.92  | 93.00| 98.43     | 94.37     | 96.64   | 88.18 | 95.75 | 94.78         |
| X101+LayoutLM\_large| 98.02   | 89.64  | 96.66   | 94.40    | 99.94  | 93.52  | 92.93| 98.44     | 94.30     | 96.70   | 88.75 | 95.31 | 94.88         |
| VSR                | **98.29**| **91.19**| **96.32**| **95.84**| **99.96**| **95.11**| **94.66**| **98.66**| **95.05**| **97.11**| **89.24**| **95.63**| **95.59**     |

### Table 4. Performance comparisons on DocBank dataset in mAP.

| Models  | Abstract | Author | Caption | Equation | Figure | Footer | List | Paragraph | Reference | Section | Table | Title | mAP |
|---------|----------|--------|---------|----------|--------|--------|------|-----------|-----------|---------|-------|-------|-----|
| Faster RCNN | 96.2    | 88.9 | 93.9 | **78.1** | 85.4 | 93.4 | 86.1 | 67.8 | 89.9 | 76.7 | 77.2 | **95.3** | 86.3 |
| VSR      | **96.3** | **89.2** | **94.6** | 77.3 | **97.8** | 93.2 | **86.2** | **69.0** | **90.3** | **79.2** | **77.5** | **94.9** | **87.6** |
**Ablation results**
- Effects of multi-granularity semantic features

---

**Table 5. Effects of semantic features at different granularities.**

| Vision | Semantics | Title | Author | Abstract | Body | Figure | Figure | Table | Table | Reference | mAP |
|--------|-----------|-------|--------|----------|------|--------|--------|-------|-------|-----------|-----|
| ✓      | Char      | 100.0 | 51.1   | 94.8     | 98.9 | 94.2   | 91.8   | 97.3  | 67.1  | 90.8      | 87.3|
| ✓      | ✓         | 100.0 | 71.4   | 96.5     | 98.9 | 95.6   | 93.6   | 96.9  | 68.6  | 89.9      | 90.2|
| ✓      | ✓         | 100.0 | 60.2   | 95.5     | 99.0 | 97.8   | 93.2   | 98.9  | 73.0  | 91.2      | 89.8|
| ✓      | ✓         | 100.0 | 84.3   | 96.1     | 98.7 | 95.7   | 92.5   | 99.4  | 71.4  | 92.4      | **92.3**|
Experiment

- Ablation results
  - Effects of two-stream network with adaptive aggregation

Table 6. Effects of two-stream network with adaptive aggregation.

| Method                  | Title | Author | Abstract | Body | Figure Caption | Table | Table Caption | Reference | mAP  | FPS  |
|-------------------------|-------|--------|----------|------|----------------|-------|---------------|-----------|------|------|
| Single-stream at input level | R101  | 94.7   | 58.7     | 82.7 | 98.1           | 97.9  | 96.3          | 91.8      | 63.7 | 91.5 |
|                         | R152  | 100.0  | 50.5     | 85.3 | 97.9           | 98.0  | 94.4          | 93.3      | 62.6 | 90.5 |
| Single-stream at decision level | R101  | 99.5   | 67.6     | 95.1 | 98.8           | 95.0  | 93.2          | 96.6      | 70.7 | 91.3 |
|                         | R152  | 100.0  | 80.2     | 91.0 | 99.4           | 96.0  | 92.4          | 98.3      | 73.8 | 91.7 |
| VSR                     | R101  | 100.0  | 84.3     | 96.1 | 98.7           | 95.7  | 92.5          | 99.4      | 71.4 | 92.4 |
|                         |       |        |          |      |                |       |               |           |      |      |
Experiment

- Ablation results
  - Effects of relation module

### Table 7. Effects of relation module.

| Method   | Title | Author | Abstract | Body  | Figure | Figure caption | Table | Table caption | Reference | mAP |
|----------|-------|--------|----------|-------|--------|---------------|-------|---------------|----------|-----|
| Faster RCNN w/o RM | 1     | 51.1   | 94.8     | 98.9  | **94.2** | 91.8          | 97.3  | 67.1          | 90.8     | **87.3** |
| Faster RCNN w/ RM  | 1     | **88.4** | **99.1** | **99.1** | 85.4   | **92.6**      | **98.0** | 79.2          | **91.6** | **92.6** |
| VSR w/o RM        | 1     | 84.3   | 96.1     | 98.7  | **95.7** | 92.5          | 99.4  | 71.4          | 92.4     | **92.3** |
| VSR w/ RM         | 1     | **94**  | 95       | **99.1** | 95.3   | **94.5**      | 96.1  | **84.6**      | 92.3     | **94.5** |

**Fig. 4.** Qualitative comparison between VSR w/wo RM. Introducing RM effectively removes duplicate predictions and provides more accurate detection results (both labels and coordinates). The colors of semantic labels are: Figure, Body, Figure Caption.
https://davar-lab.github.io/index.html
See Far, Go Further