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

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

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



Method

Experiments

♦ Document Layout Analysis



The average of total serum IgE levels was highest in 8-17

3.4. Blood Cell Essissiphii Court. The blood cell contings count was also compared between groups. The essimple cell proportion was 4.2 \pm 0.4% in patients sensitized only spring polkens, while it was significantly higher (3.7 \pm 0.4% in patients cannitated so both personal altergens and sprin polens (P=0.1046, Mann-Whitney U test) (Figure 20.1164; 2). The blood cell essimplied court showed the on

3.5. Allegic Semitiation in Arbina. Fifty-nise patients (Audit, 13 Children) had been previously diagnosed (et aluth), 13 Children) had been previously diagnosed (et aluth), 13 Children) had been been diagnose with authora. Sensitiation to any allegon was detected in 58% of patients with arbina (34/98). Twenty-sai (48/98) that of 59 patients with arbina (34/98) to print pollom (Table 3) Approximately half of the authora patients (58/9), 30/98 were unmitted to perennial allegons. Seven percent or



pollen, while 50% (94/598) in patients without athms were essentiated exclusively to these allergers. Thirty-seven percent of pollents with a persons achina diagnosis (2259) were esentiated to both spring and personal allergers, which was guarkizantly lapler than that observed in particuts without softma (20% (27/598) (P = 0.0017, dis-separe test).

Mean total serum [ql] levels in patients with asthean see $NT = 80 \, \mathrm{UMm}$, while those in patients without astim were $2.3 \pm 2.7 \, \mathrm{U/m}$. U = 0.0001 compared to patient without astim sealman, Mann-Withray U Levol. Blood evision-phil or proportion in patients with asthman was $3.4 \pm 0.0\%$. Distincts without arthman, the proportion was $3.9 \pm 0.29 \, \mathrm{Mes}$ decision-phil cell proportion in a patient with asthman was $3.4 \pm 0.0\%$. Blood existing level of the proportion of the $3.9 \pm 0.29 \, \mathrm{Mes}$ desirable phil cell proportion in patients with asthman axis or application; by larger than those in patients without patients.

4. Discussio

Among exemination, an language of by the server allegacy specific, light level, do not at these correspond with the printient symptoms. We found that approximately twice, a most printies was essentialed to shell garge plates and support of the server and the se

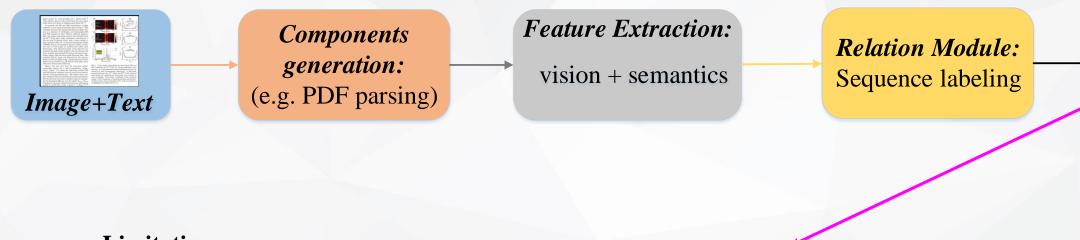


Vision Semantics

Relations

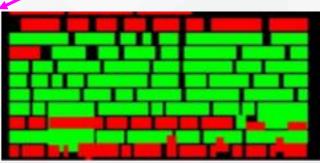


- **♦** Multimodal document layout analysis frameworks
 - > NLP-based framework



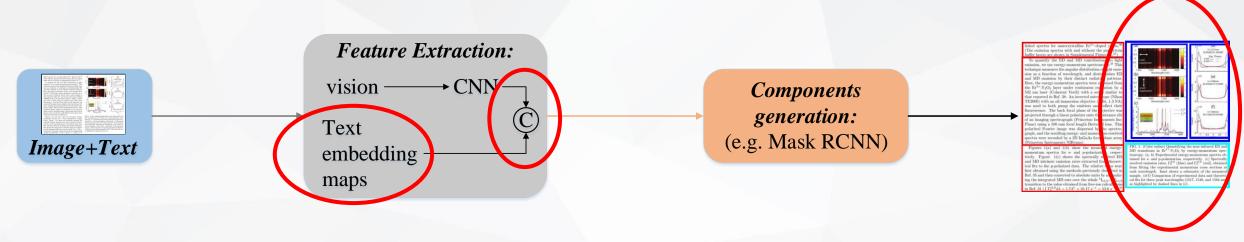
Limitations:

• Insufficient capabilities in layout modeling



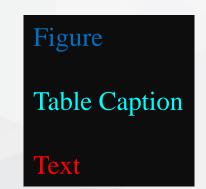


- **♦** Multimodal document layout analysis frameworks
 - > CV-based framework



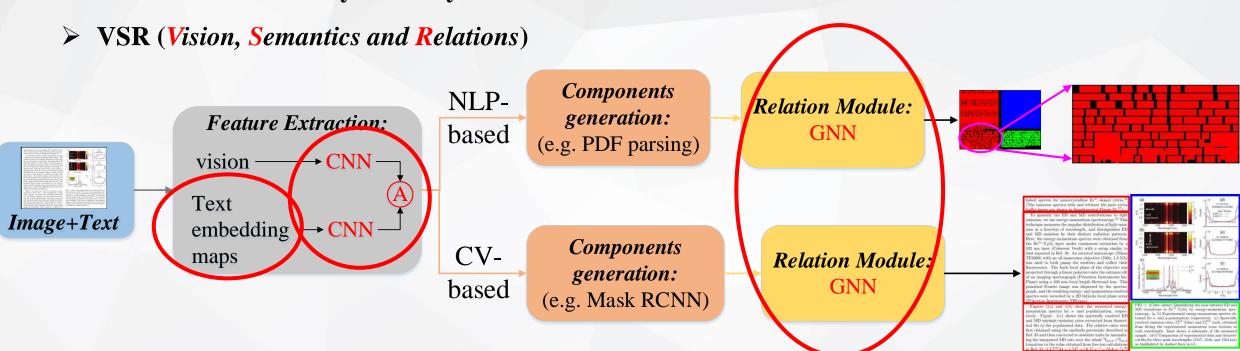
Limitations:

- Limited semantics
- Simple and heuristic modality fusion strategy
- Lack of relation modeling between components





♦ Multimodal document layout analysis frameworks



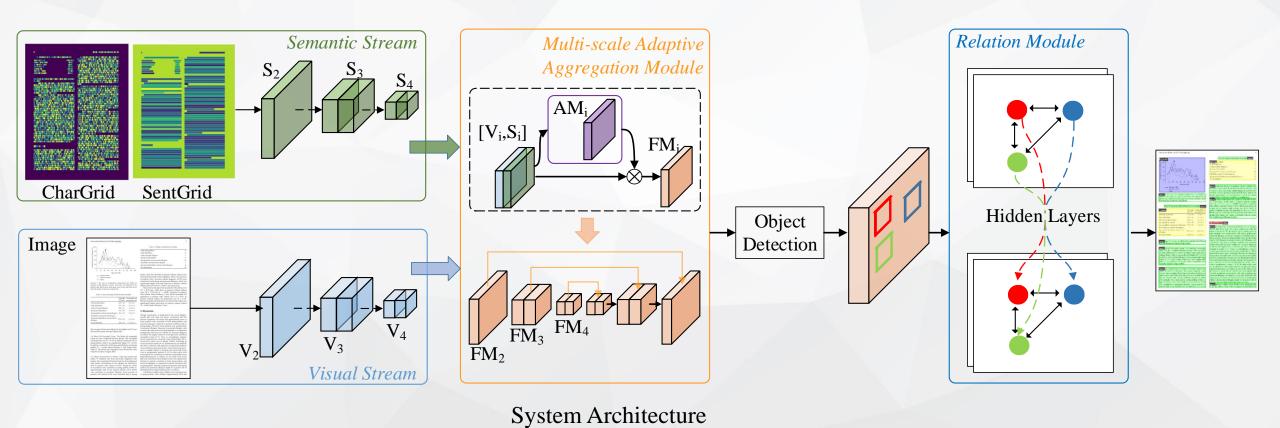
Advantages:

- 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

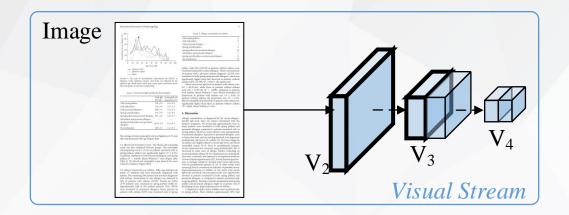
Experiments

- > Two-stream ConvNets
- ➤ Multi-scale Adaptive Aggregation
- > Relation Module



Method

> Two-stream ConvNets

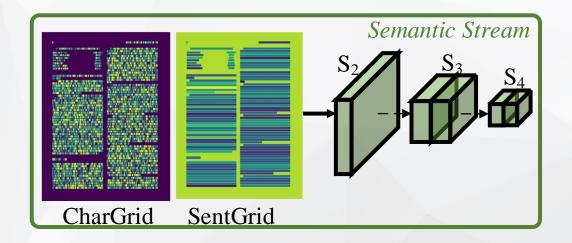


input (document image):

$$V_0 = x \in R^{H \times W \times 3}$$

output (multi-scale visual features):

$$\{V_2, V_3, V_4\} \quad V_i \in R^{\frac{H}{2^i} \times \frac{W}{2^i} \times C_i^V}$$



input (text embedding maps):

$$S_0 = LayerNorm(Chargrid + Sentgrid) \in R^{H \times W \times C_0^S}$$

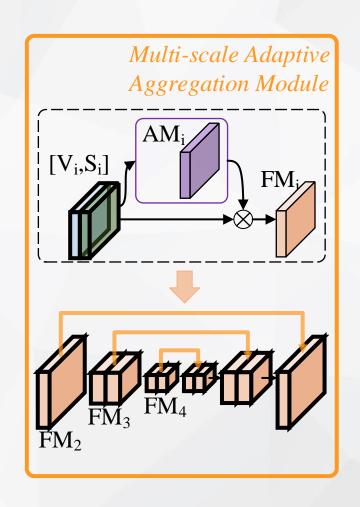
character granularity sentence

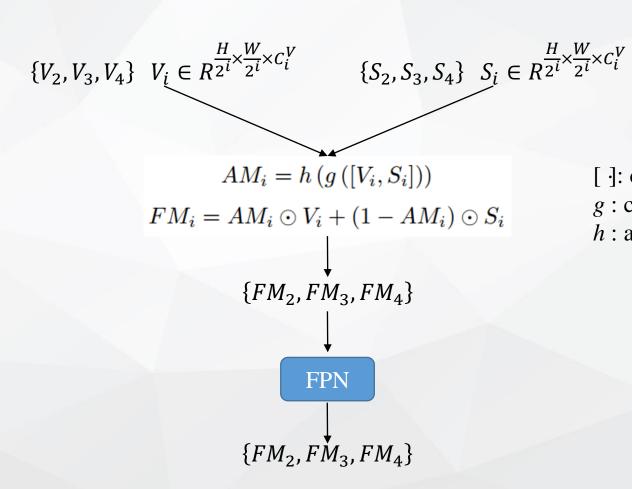
granularity

output (multi-scale *semantic* features):

$$\{S_2, S_3, S_4\}$$
 $S_i \in R^{\frac{H}{2^i} \times \frac{W}{2^i} \times C_i^V}$

➤ Multi-scale Adaptive Aggregation





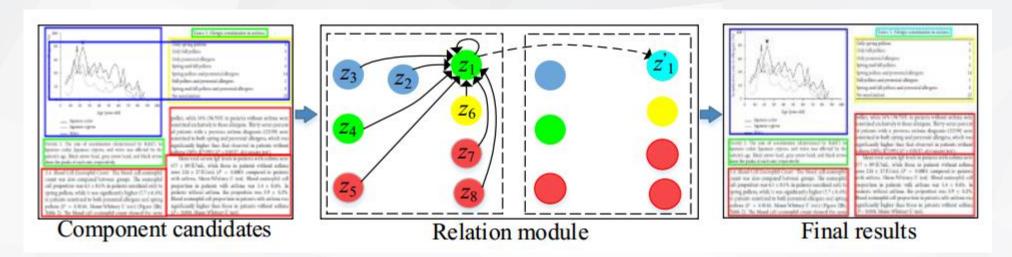
[]: concatenation

g: convolutional layer

h : activation function

Method

> Relation Module



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

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

visual features: $f_j = RoIAlign(FM, b_j)$
position embeddings: $e_j^{pos}(b_j)$

probabilities:
$$\widetilde{p_{j}^{c}} = Softmax(Linear_{cls}(z_{j}^{\prime}))$$

regression
$$\widetilde{b_{j}} = Linear_{reg}(z_{j}^{\prime})$$

coordinates:

Method

Experiments

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

> SOTA results

> Article Regions

Table 1. Performance comparisons on Article Regions dataset

Method	Title	Author	Abstract	Body	Figure	Figure Caption	Table	Table Caption	Reference	mAP
Faster RCNN [31]	-	1.22	-	87.49	-	-	-	-	-	46.38
Faster RCNN w/ context [31]	_	10.34	-	93.58	-	-	-	30.8	-	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 [31]	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.

- > SOTA results
 - PubLayNet

Table 2. Performance comparisons on PubLayNet dataset.

N	Method	Dataset	Text	Title	List	Table	Figure	AP
Faster	RCNN [43]		91	82.6	88.3	95.4	93.7	90.2
Mask	RCNN [43]	val	91.6	84	88.6	96	94.9	91
	VSR		96.7	93.1	94.7	97.4	96.4	95.7
Faster	RCNN [43]		91.3	81.2	88.5	94.3	94.5	90
Mask	RCNN [43]		91.7	82.8	88.7	94.7	95.5	90.7
Doc	InsightAI		94.51	88.31	94.84	95.77	97.52	94.19
	SCUT	test	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
Silic	conMinds		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

NLP-based:

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
$RoBERTa_{base}$	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
$RoBERTa_{large}$	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

CV-based:

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		nantics Sentence	Title	Author	Abstract	Body	Figure	Figure Caption	Table	Table Caption	Reference	mAP
			100.0	51.1	94.8	98.9	94.2	91.8	97.3	67.1	90.8	87.3
\checkmark	\checkmark		100.0	71.4	96.5	98.9	95.6	93.6	96.9	68.6	89.9	90.2
\checkmark		\checkmark	100.0	60.2	95.5	99.0	97.8	93.2	98.9	73.0	91.2	89.8
$\sqrt{}$	\checkmark	\checkmark	100.0	84.3	96.1	98.7	95.7	92.5	99.4	71.4	92.4	92.3

> 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	Figure Caption	Table	Table Caption	Reference	mAP	FPS
Single-stream	R101	94.7	58.7	82.7	98.1	97.9	96.3	91.8	63.7	91.5	86.2	19.07
at input level	R152	100.0	50.5	85.3	97.9	98.0	94.4	93.3	62.6	90.5	85.8	18.15
Single-stream	R101	99.5	67.6	95.1	98.8	95.0	93.2	96.6	70.7	91.3	89.8	19.79
at decision level	R152	100.0	80.2	91.0	99.4	96.0	92.4	98.3	73.8	91.7	91.4	16.43
VSR	R101	100.0	84.3	96.1	98.7	95.7	92.5	99.4	71.4	92.4	92.3	13.94

Ablation results

> Effects of relation module

Table 7. Effects of relation module.

		Title	Author	Abstract	Body	Figure	Figure caption	Table	Table caption		
Faster RCNN	w/o RM	1	51.1	94.8	98.9	94.2	91.8	97.3	67.1	90.8	87.3
	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
	w/ RM	1	94	95	99.1	95.3	94.5	96.1	84.6	92.3	94.5

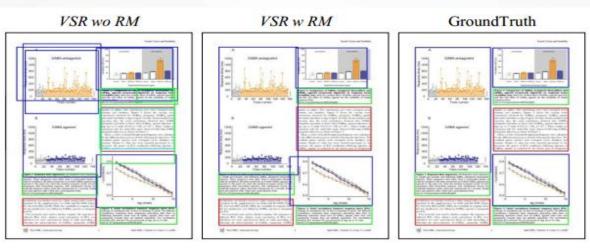


Fig. 4. Qualitative comparison between VSR w/wo RM. Introducing RM effectively removes duplicate predictions and provides more accurate detection results (both labels and coordinates). The colors of semantic labels are: Figure, Body, Figure Caption.

https://davar-lab.github.io/index.html



Publications

Datasets

Competitions Activities About Us

News



DavarOCR release 2021/07/23

We have published the general OCR toolbox and benchmark DavarOCR!



We have 2 papers published on ICPR 2020!



2021/05/03

We won the 1st place in both two tasks in ICDAR 2021 SLP Competition!



3 papers at ICDAR 2021 2021/04/28

We have 3 papers accepted by ICDAR 2021!



2 papers at AAAI 2021

2021/02/07

We have 2 papers accepted by AAAI 2021!



1 paper at TIP 2020 2020/12/04

We have 1 paper published on IEEE Trans. on Image Processing (TIP)!



1 paper at MM 2020 2020/10/12

We have 1 paper accepted by ACMMM 2020!



1 paper at AAAI 2020 2020/02/07

We have 1 paper accepted by AAAI 2020 (Oral)!







See Far, Go Further



