ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Recognizing Multiple Text Sequences from an Image by Pure End-To-End Learning

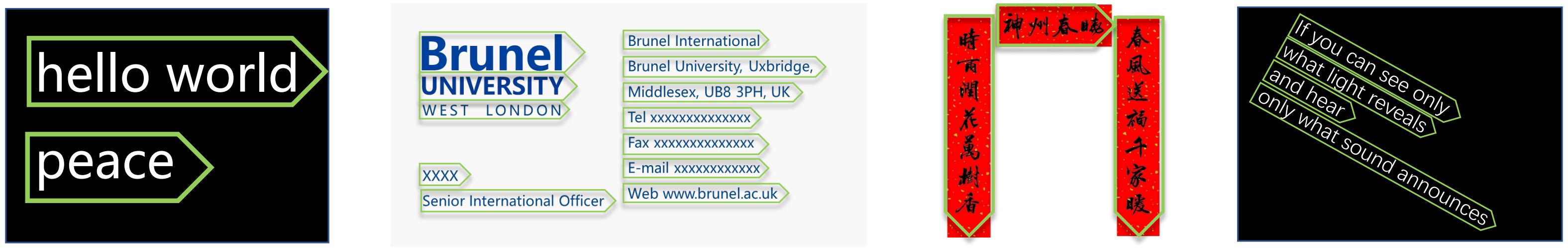
Zhenlong Xu[1] Shuigeng Zhou[1] Fan Bai[1]
Zhanzhan Cheng[2] Yi Niu[2] Shiliang Pu[2]
[1]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai 200433, China
[2]Hikvision Research Institute, China

OUTLINE

- Motivation
- Method
- Experiments
- Conclusion

Motivation

Problem: Recognizing **multiple text sequences** from an image by **pure end-to-end learning**.



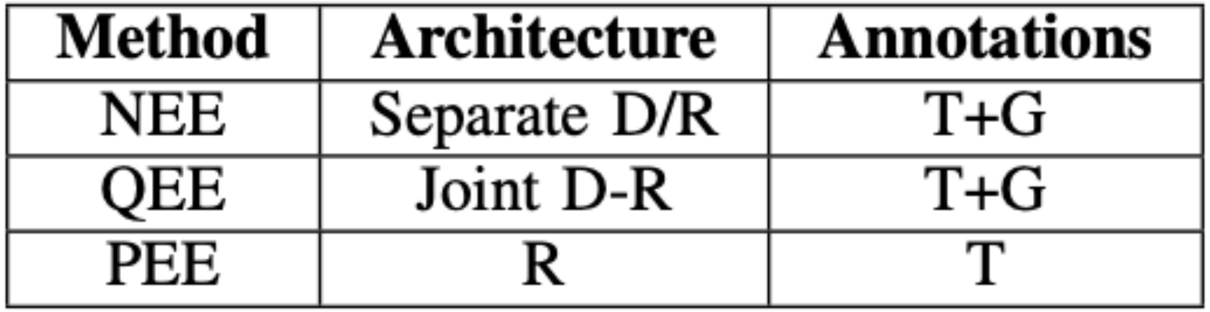
(a) (b) (c) (d)

MSR

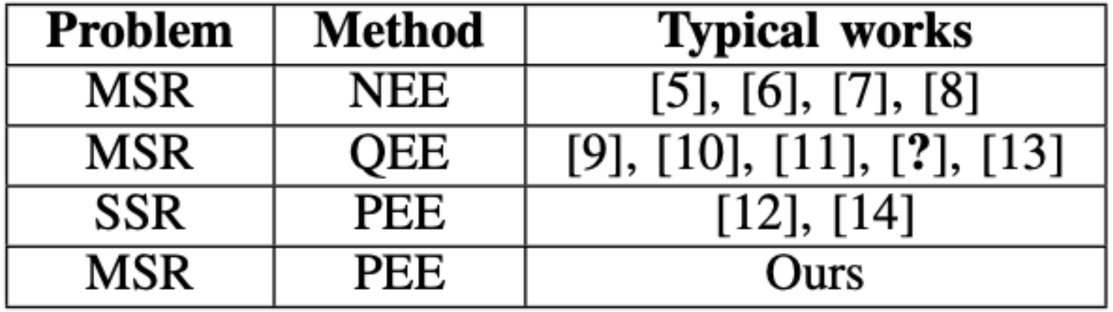
Annotations:

only text; no location;

PEE



Method	Architecture	Annotations
NEE	Separate D/R	T+G
QEE	Joint D-R	T+G
PEE	R	T



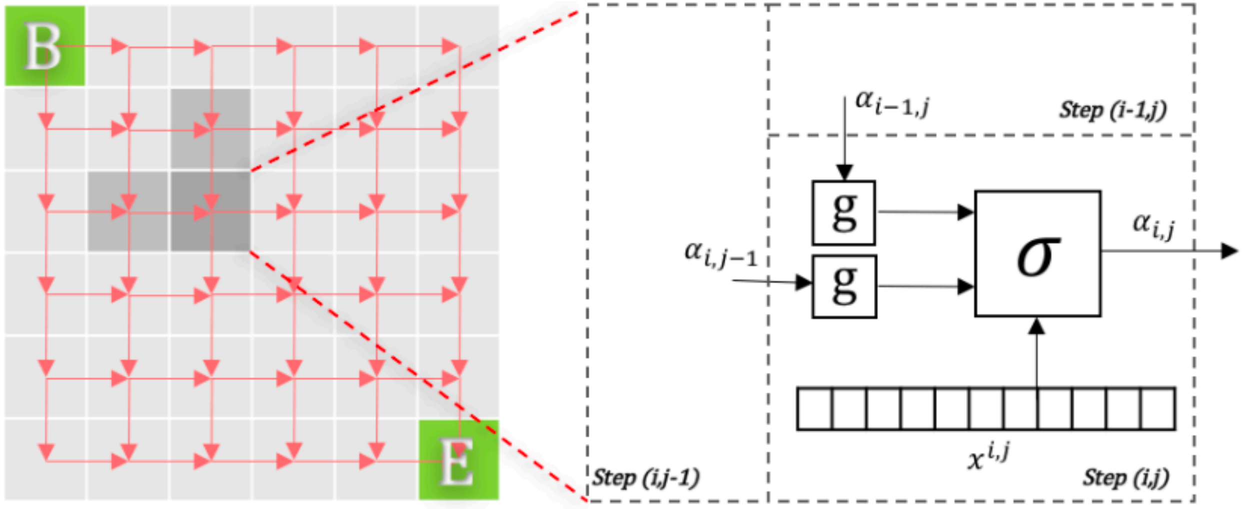
Problem	Method	Typical works
MSR	NEE	[5], [6], [7], [8]
MSR	QEE	[9], [10], [11], [?], [13]
SSR	PEE	[12], [14]
MSR	PEE	Ours

Method

Aims: transform a three-dimensional tensor $\boldsymbol{X}$ to a conditional probability distribution over multiple character sequences $P(\boldsymbol{Z}|\boldsymbol{X})$.

$$\mathbf{X} = \begin{pmatrix} x^{00} & x^{01} & \dots & x^{0W'} \\ x^{10} & x^{11} & \dots & x^{1W'} \\ \vdots & \vdots & \ddots & \vdots \\ x^{H'0} & x^{H'1} & \dots & x^{H'W'} \end{pmatrix}$$

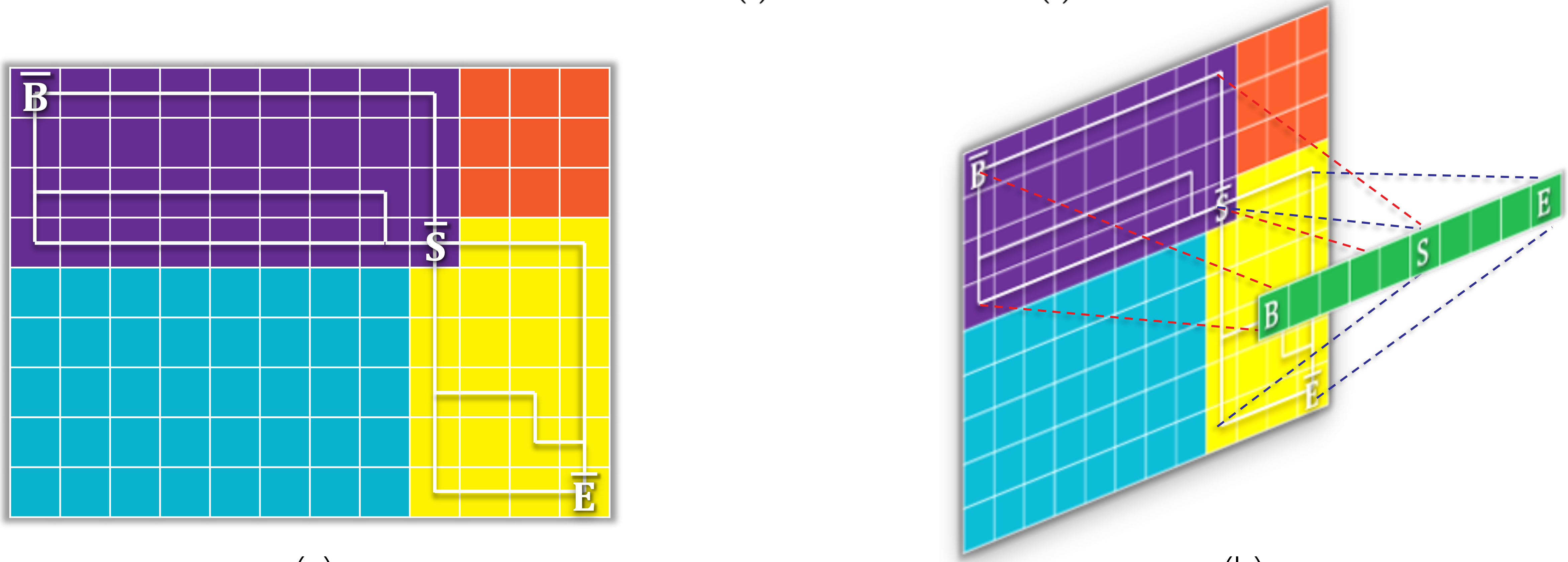
$$p(\mathbf{Z}|\mathbf{X}) \overset{def}{=} \frac{1}{N} \sum_{i=1}^{N} p(\mathbf{l}_i|\mathbf{X})$$



Path Generation

Method

Path search problem: $p(\mathbf{l}|\mathbf{X}) = \sum_{\bar{l}\in\mathcal{B}^{-1}(\mathbf{l})} p(\bar{l}|\mathbf{X}) = \sum_{\bar{l}\in\mathcal{B}^{-1}(\mathbf{l})} \prod_{t=0}^{|\bar{l}|-1} x_{\bar{l}_t}^{i_t,j_t}$



(a)

(b)

Forward and Backward Algorithms

Method-Forward

$$\alpha_{i,j}(s) \overset{def}{=} \sum_{\bar{l}\in\mathcal{B}^{-1}(\mathbf{l}'_{0:s})} \prod_{t=0}^{|\bar{l}|-1} x^{i_t,j_t}_{\bar{l}_t}$$

Define $\alpha_{i,j}(s)$ as the probability for $\bar{l}$ matching $l'_{0:s}$ at (i,j).

$$\begin{aligned}\alpha_{i,j}(s) &= \sigma(g(\alpha_{i,j-1},s), g(\alpha_{i-1,j},s)) \\ &= \lambda_1 g(\alpha_{i,j-1},s) + \lambda_2 g(\alpha_{i-1,j},s)\end{aligned}$$

λ_1, λ_2 are the hyper-parameters of linear function σ.

$$g(\alpha_{i,j},s) \overset{def}{=} (\alpha_{i,j}(s) + \alpha_{i,j}(s-1) + \eta\alpha_{i,j}(s-2))x^{i,j}_{l'_s}$$

$$\eta = \begin{cases} 0 & \text{if } \mathbf{l}'_s = \text{blank or } \mathbf{l}'_s = \mathbf{l}'_{s-2}, \\ 1 & \text{otherwise.} \end{cases}$$

The state transfer strategy:

- blank and any non-blank character
- any pair of distinct non-blank characters

$$p(\mathbf{l}|\mathbf{X}) = \alpha_{H',W'}(|\mathbf{l}'|-1) + \alpha_{H',W'}(|\mathbf{l}'|-2)$$

Answer Representation

Method-Backward

$$\beta_{i,j}(s) \overset{def}{=} \sum_{\bar{l}\in\mathcal{B}^{-1}(\mathbf{l}'_{s:|\mathbf{l}'|-1})} \prod_{t=1}^{|\bar{l}|-1} x^{i_t,j_t}_{\bar{l}_t}$$

Define $\beta_{i,j}(s)$ as the probability for $\bar{l}$ matching $l'_{s:|l'|-1}$ at (i,j) but not relying on $x^{i_0,j_0}_{\bar{l}_0}$

$$\beta_{i,j}(s) = \lambda_1 g'(\beta_{i,j+1}, s) + \lambda_2 g'(\beta_{i+1,j}, s)$$

$$g'(\beta_{i,j}, s) \overset{def}{=} \beta_{i,j}(s)x^{i,j}_{\mathbf{l}'_s} + \beta_{i,j}(s+1)x^{i,j}_{\mathbf{l}'_{s+1}} + \eta'\beta_{i,j}(s+2)x^{i,j}_{\mathbf{l}'_{s+2}}$$

$$\eta' = \begin{cases} 0 & \text{if } \mathbf{l}'_s = \text{blank or } \mathbf{l}'_s = \mathbf{l}'_{s+2}, \\ 1 & \text{otherwise.} \end{cases}$$

The state transfer strategy:

- blank and any non-blank character
- any pair of distinct non-blank characters

Method

Objective Function

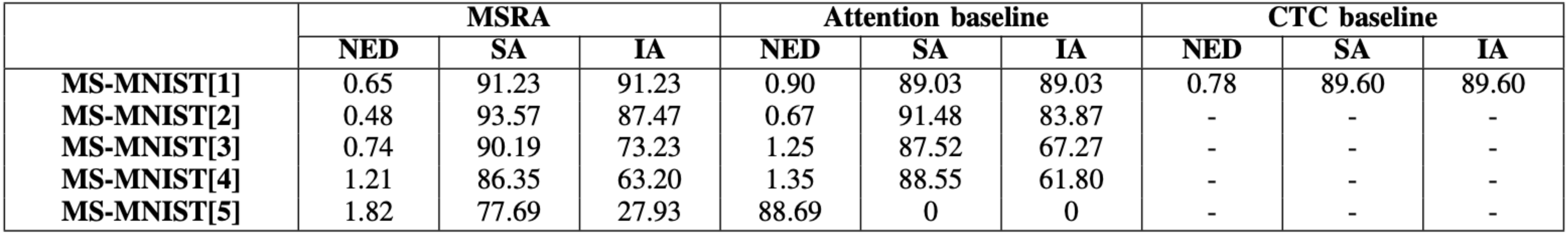
$$O = -\sum_{(\mathbf{X},\mathbf{Z})\in\mathcal{S}} \ln p(\mathbf{Z}|\mathbf{X})$$

$$O = -\sum_{(\mathbf{X},\mathbf{Z})\in\mathcal{S}} (\ln \sum_{i=1}^{N} p(\mathbf{l}_i|\mathbf{X}) - \ln N)$$

$$\frac{\partial p(\mathbf{l}|\mathbf{X})}{\partial x_k^{i,j}} = \frac{1}{x_k^{i,j}} \sum_{s\in lab(\mathbf{l},k)} \alpha_{i,j}(s)\beta_{i,j}(s)$$

$$\frac{\partial O}{\partial x_k^{i,j}} = -\frac{1}{x_k^{i,j}\sum_{t=1}^{n} p(\mathbf{l}_t|\mathbf{X})} \sum_{t=1}^{n} \sum_{s\in lab(\mathbf{l}_t,k)} \alpha_{i,j}(s)\beta_{i,j}(s)$$

Experiments

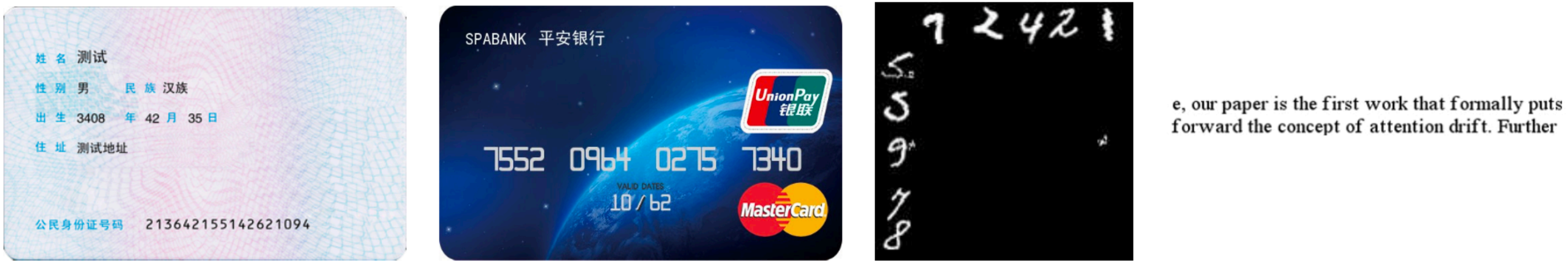


	MSRA			Attention baseline			CTC baseline		
	NED	SA	IA	NED	SA	IA	NED	SA	IA
MS-MNIST[1]	0.65	91.23	91.23	0.90	89.03	89.03	0.78	89.60	89.60
MS-MNIST[2]	0.48	93.57	87.47	0.67	91.48	83.87	-	-	-
MS-MNIST[3]	0.74	90.19	73.23	1.25	87.52	67.27	-	-	-
MS-MNIST[4]	1.21	86.35	63.20	1.35	88.55	61.80	-	-	-
MS-MNIST[5]	1.82	77.69	27.93	88.69	0	0	-	-	-

MS-MNIST

- NED(%): the normalized edit distance.
- SA(%): the sequence recognition accuracy.
- IA(%): the image recognition accuracy.

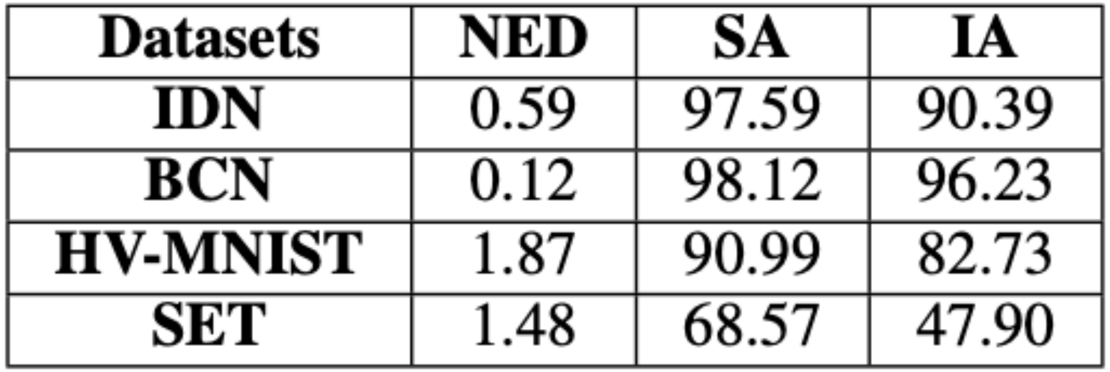
Experiments



e, our paper is the first work that formally puts forward the concept of attention drift. Further

(a) (b) (c) (d)

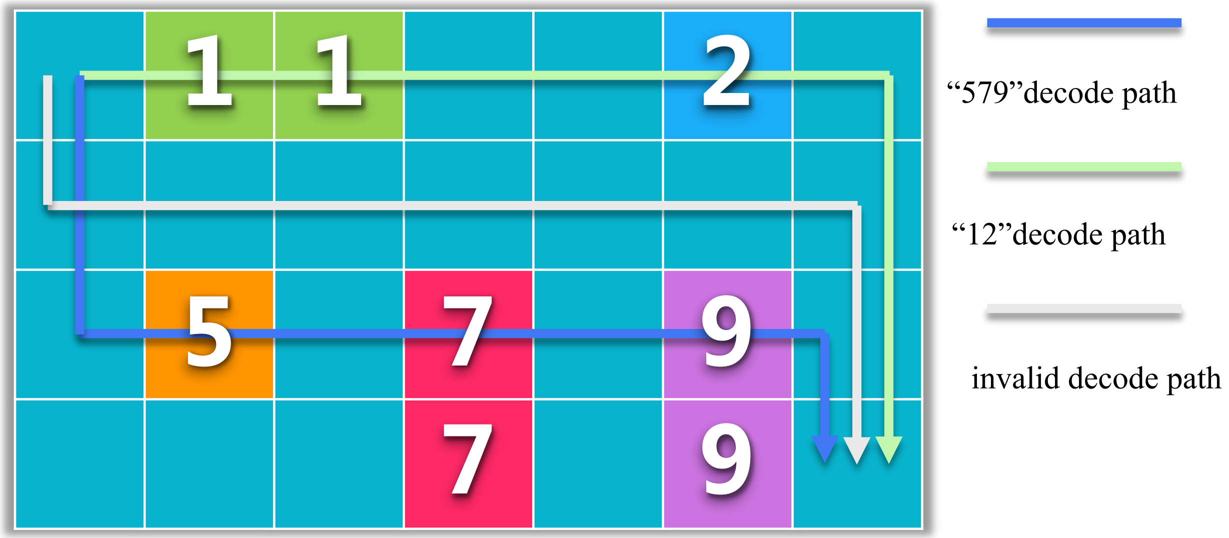
Real Application Scenarios



Datasets	NED	SA	IA
IDN	0.59	97.59	90.39
BCN	0.12	98.12	96.23
HV-MNIST	1.87	90.99	82.73
SET	1.48	68.57	47.90

Experiments

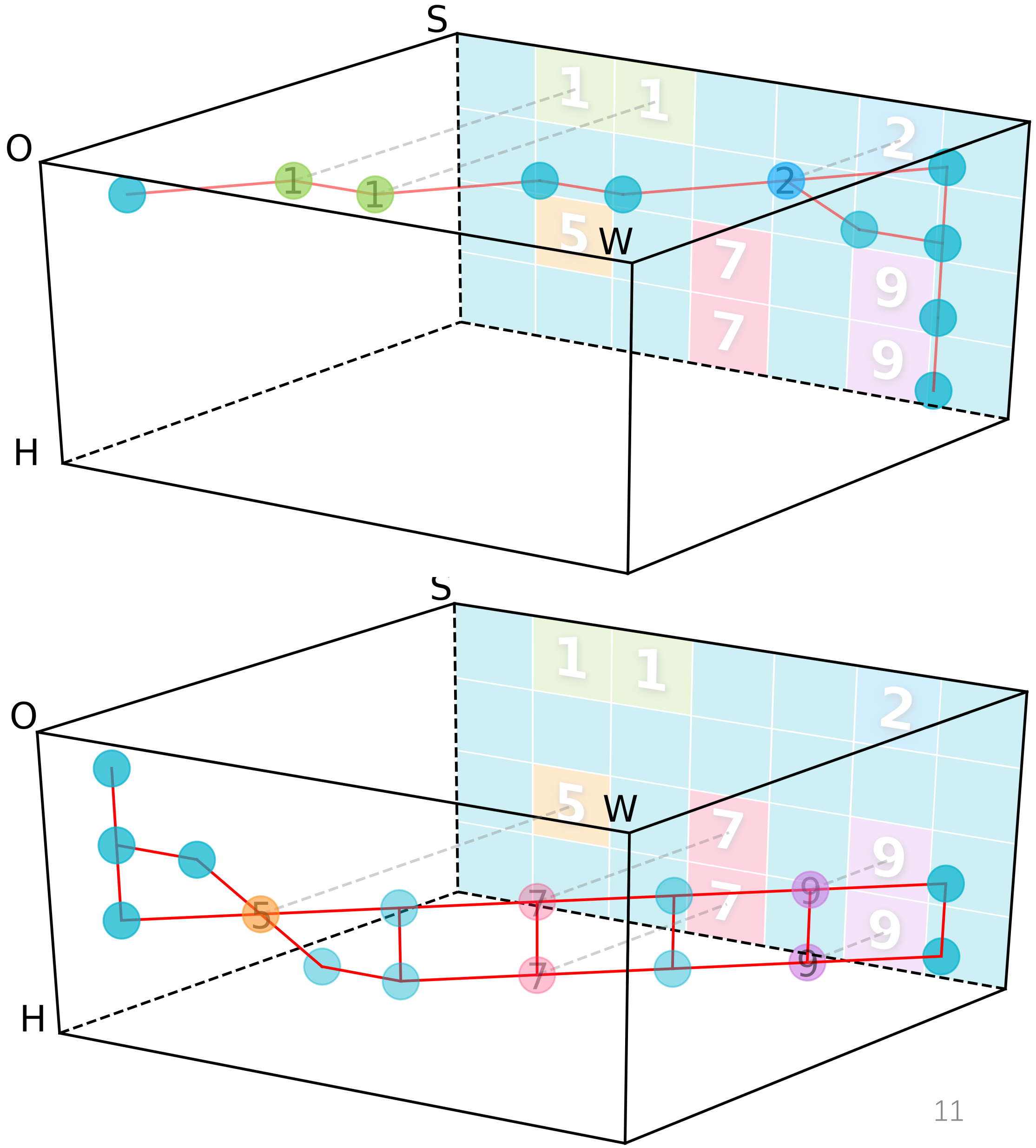
Decoding process demonstration



“579”decode path

“12”decode path

invalid decode path



Conclusion

- A new taxonomy of text recognition methods: NEE, QEE, PEE;
- A novel PEE method MSRA to solve MSR;
- Build up several datasets: MS-MNIST and real application scenarios
- Conduct extensive experiments on these datasets which show MSRA can effectively recognize multiple sequences from images and outper- forms two CTC/attention based baseline methods.