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] School of Computer Science, Fudan University

[2] Hikvision Research Institute, China

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

Motivation

- We address a challenging problem: recognizing multiple text sequences from an image by pure end-to-end learning.
  - Multiple text sequences recognition (MSR). Each image may contain multiple text sequences of different content, location and orientation.
  - Pure end-to-end (PEE) learning. Each training image is annotated with only text transcripts.
- Most existing works cannot handle this problem. Some of them use both text transcripts and text locations in a non-end-to-end (NEE) or quasi-end-to-end (QEE) way. Some of them are PEE method but for single text sequence recognition problem.
- We develop a novel PEE method MSRA to solve the MSR problem, in which the model is trained with only sequence-level text transcripts.

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

Fig 1. Examples of the MSR problem. (a)-(d) are 4 types of multi-sequence scenarios. Each sequence is bounded by a green box with the arrow indicating text orientation.

Multiple Sequence Recognition Approach (MSRA)

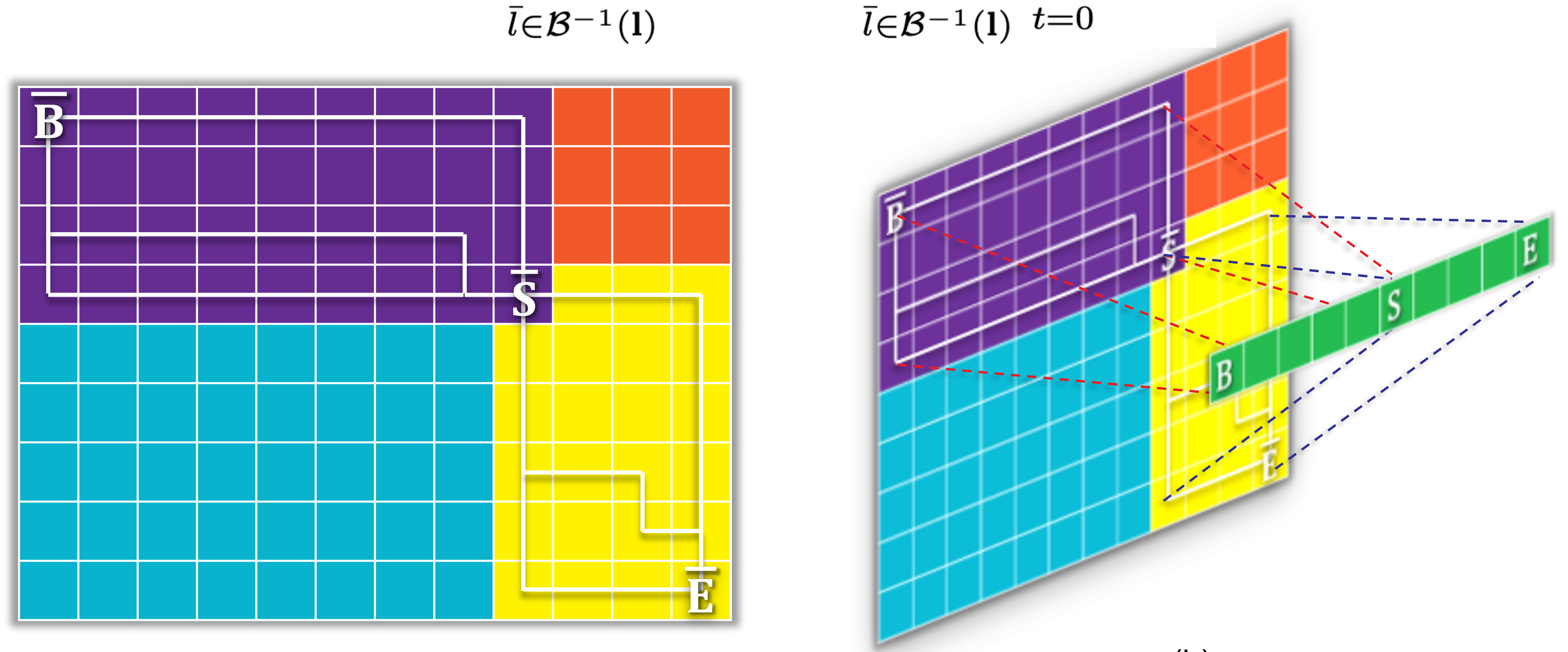
- MSRA aims to transform a three-dimensional tensor $\boldsymbol{X}$ to a conditional probability distribution over multiple character sequences $P(\boldsymbol{Z}|\boldsymbol{X})$.

$$\mathbf{X} = \begin{pmatrix} x^{00} & x^{01} & \dots & x^{0W'} \\ x^{10} & x^{11} & \dots & x^{1W'} \\ \vdots & \vdots & \ddots & \vdots \\ x^{H'0} & x^{H'1} & \dots & x^{H'W'} \end{pmatrix} \qquad p(\mathbf{Z}|\mathbf{X}) \overset{def}{=} \frac{1}{N}\sum_{i=1}^{N} p(\mathbf{l}_i|\mathbf{X})$$

$\boldsymbol{Z}$ is denoted as a set of text sequences l_i which is obtained by using the many-to-one $\mathcal{B}$-mapping strategy for path $\bar{l}$ on the two-dimensional probability distribution $\boldsymbol{X}$.

- The evaluation of $\mathrm{P}(l|\boldsymbol{X})$ turns to solve the two-dimensional probability path $\bar{l}$ search problem over $\boldsymbol{X}$.

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



(a) (b)

Fig 2. The illustration of the forward and backward algorithms matching the s position of l' at $\bar{S}(i,j)$. The dark purple area represents the path search area of the forward algorithm, where the white paths $\bar{l}$ from $\bar{B}$ to $\bar{S}$ are all solutions satisfying $\mathcal{B}(\bar{l}) = l'_{0:s}$. The yellow area represents the path search area of the backward algorithm, where the paths from $\bar{S}$ to $\bar{E}$ satisfying $\mathcal{B}(\bar{l}) = l'_{s:|l'|-1}$.

- Prefix sub-path search problem can be iteratively calculated with a dynamic programming algorithm.

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

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

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

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

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

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

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

For representing the non-text areas, adding blanks to the beginning and the end and inserting blanks between each pair of neighboring characters of l to get l'.

- Objective Function

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

Similar to $\alpha_{i,j}(s)$, $\beta_{i,j}(s)$ is defined as the probability for $\bar{l}$ matching $l'_{s:|l'|-1}$ at (i,j) but not relying on $x_{\bar{l}_0}^{i_0,j_0}$ and calculated by the backward algorithm. The gradient of the objective function can be obtained based on them where $lab(l,k) = \{s : l'_s = k\}$.

Experiments

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

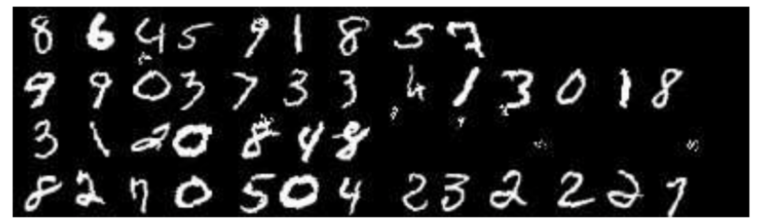


Fig 3. Sample of MS-MNIST[4].

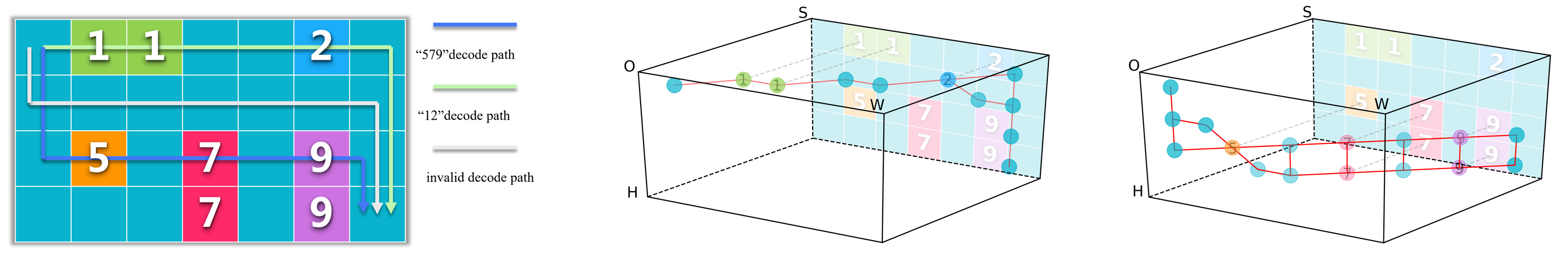
- Recognition results on MS-MNIST datasets

	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	-	-	-

- Recognition results on real application scenarios datasets

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

Fig 4. Samples of four more challenging datasets: (a) IDN, (b) BCN, (c) HV-MNIST, and (d) SET.



(a) (b) (c)

Fig 5. Decoding process demonstration on the the learnt maximum probability matrix of $\boldsymbol{X}$ and the matching paths for decoding text sequences in α space.

Conclusion

Our contribution can be summarized as below:

- A new taxonomy of text recognition methods: NEE, QEE, PEE;
- A novel PEE method MSRA to solve MSR;
- Build up several datasets and conduct extensive experiments on them;