文本识别–MORAN

<Excerpt in index | 首页摘要>

MORAN: A Multi-Object Rectified Attention Network for Scene Text Recognition

KeyWords Plus: Scene text recognition optical character recognition

- **relevant blog**： MORAN不规则文本纠正：刷新多个OCR数据集最优算法
- **paper**： MORAN
- **coding**： Github

<The rest of contents | 余下全文>

Introduction

MORAN是一种文本识别算法，可以针对不规则文本进行处理

MORAN文本识别算法由矫正子网络**MORN**和识别子网络**ASRN**组成，在**MORN**中设计了一种新颖的**像素级弱监督学习机制用于不规则文本的形状纠正**，大大降低了不规则文本的识别难度。

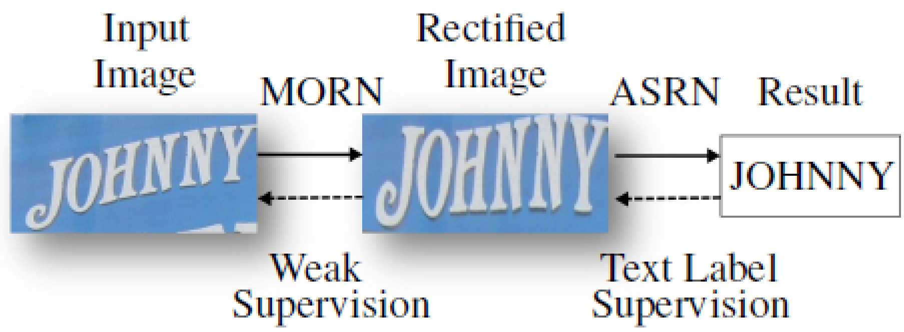


Figure 2. Overview of the MORAN. The MORAN contains a MORN and an ASRN. The image is rectified by the MORN and given to the ASRN. The dashed lines show the direction of gradient propagation, indicating that the two sub-networks are jointly trained.

The training of the **MORN** is guided by the **ASRN**, which requires only text labels. Without any `geometric-level or pixel-level supervision`, the MORN is trained in a `weak supervision way.`

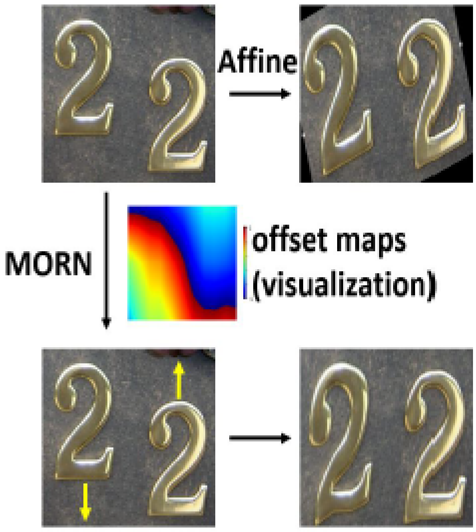
几个创新点和论文贡献：

1、propose the MORAN framework to recognize **irregular scene text.**

2、Trained in a **weak supervision** way, the subnetwork MORN is flexible. It is free of geometric constraints and can rectify images with complicated distortion.

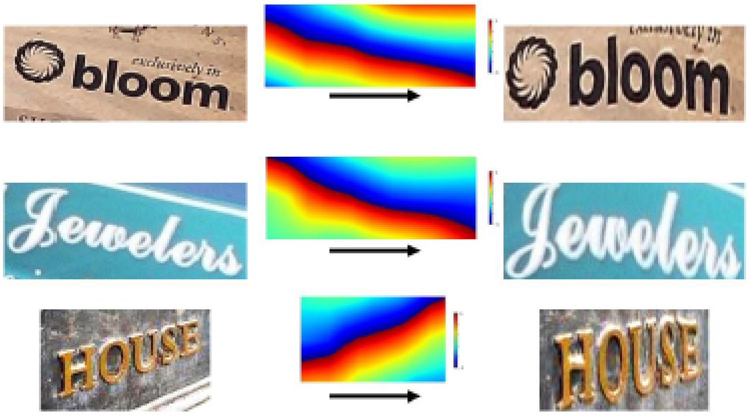
3、propose a **fractional pickup method** for the training of the attention-based decoder in the ASRN. To address noise perturbations, we expand the visual field of the MORAN, which further improves the sensitivity of the attentionbased decoder.

Multi-Object Rectification Network



Comparison of the **MORN and affine transformation.** The MORN is free of geometric constraints. The main direction of rectification predicted by the MORN for each character is indicated by a **yellow arrow.**

在**黄色和蓝色之间的像素补偿是0**，颜色的深浅程度代表着补偿的量级，矫正网络如下。



(a) Perspective texts

(b) curved texts

place a pooling layer before the convolutional layer to **avoid noise and reduce the amount of calculation.**

Table 1. Architecture of the MORN

Type	Configurations	Size
Input	-	1×32×100
MaxPooling	k2, s2	1×16×50
Convolution	maps:64, k3, s1, p1	64×16×50
MaxPooling	k2, s2	64×8×25
Convolution	maps:128, k3, s1, p1	128×8×25
MaxPooling	k2, s2	128×4×12
Convolution	maps:64, k3, s1, p1	64×4×12
Convolution	maps:16, k3, s1, p1	16×4×12
Convolution	maps:2, k3, s1, p1	2×4×12
MaxPooling	k2, s1	2×3×11
Tanh	-	2×3×11
Resize	-	2×32×100

Here k, s, p are kernel, stride and padding sizes, respectively. For example, $k3$ represents a 3×3 kernel size.

Similar to the offset maps, the grid contains two channels, which represent the xcoordinate and y-coordinate

如下图所示，通过补偿网络会产生一个**offset maps**，他有两个通道分别代表着x和y方向上的补偿信息，同时也会产生一个**basic grid**用于记录original positions of the pixels。最终的补偿网络计算如下：

$$offset'_{(c,i,j)} = offset_{(c,i,j)} + basic_{(c,i,j)}, c = 1, 2$$

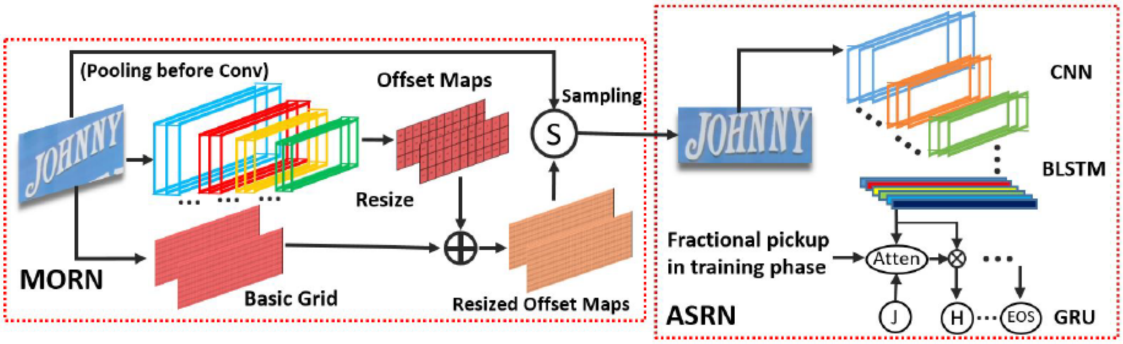


Figure 4. Overall structure of MORAN.

整体的MORAN如上图所示，左边为矫正网络，右边为识别网络

The **advantages of the MORN** are manifold：

1、**The rectified images are more readable** owing to the regular shape of the text and the reduced noise

2、The MORN is more flexible than the affine transformation. It is free of geometric constraints，which enables it **to rectify images using complicated transformations.**

3、The MORN is **more flexible than methods using a specific** number of regressing points

free of geometric constraints，which enables it **to rectify images using complicated transformations.**

4、The MORN does not **require extra labelling information** of character positions.

Attentionbased Sequence Recognition Network

ASRN网络框架如下图所示：

Table 2. Architecture of the ASRN

Type	Configurations	Size
Input	-	1×32×100
Convolution	maps:64, k3, s1, p1	64×32×100
MaxPooling	k2, s2	64×16×50
Convolution	maps:128, k3, s1, p1	128×16×50
MaxPooling	k2, s2	128×8×25
Convolution	maps:256, k3, s1, p1	256×8×25
Convolution	maps:256, k3, s1, p1	256×8×25
MaxPooling	k2, s2x1, p0x1	256×4×26
Convolution	maps:512, k3, s1, p1	512×4×26
Convolution	maps:512, k3, s1, p1	512×4×26
MaxPooling	k2, s2x1, p0x1	512×2×27
Convolution	maps:512, k2, s1	512×1×26
BLSTM	hidden unit:256	256×1×26
BLSTM	hidden unit:256	256×1×26
GRU	hidden unit:256	256×1×26

Here, k, s, p are kernel, stride and padding sizes, respectively. For example, $s2 \times 1$ represents a 2×1 stride size. "BLSTM" stands for bidirectional-LSTM. "GRU" is in attention-based decoder.

先经过pooling和卷积层之后再接blstm，Each convolutional layer is followed by a batch normalization layer and a ReLU layer.

The largest number of steps that the decoder **generates is T**. The decoder stops processing when it predicts an end-of-sequence token "EOS" [47]. At time step t, **output yt is:**

$$y_t = Softmax(W_{out}s_t + b_{out})$$

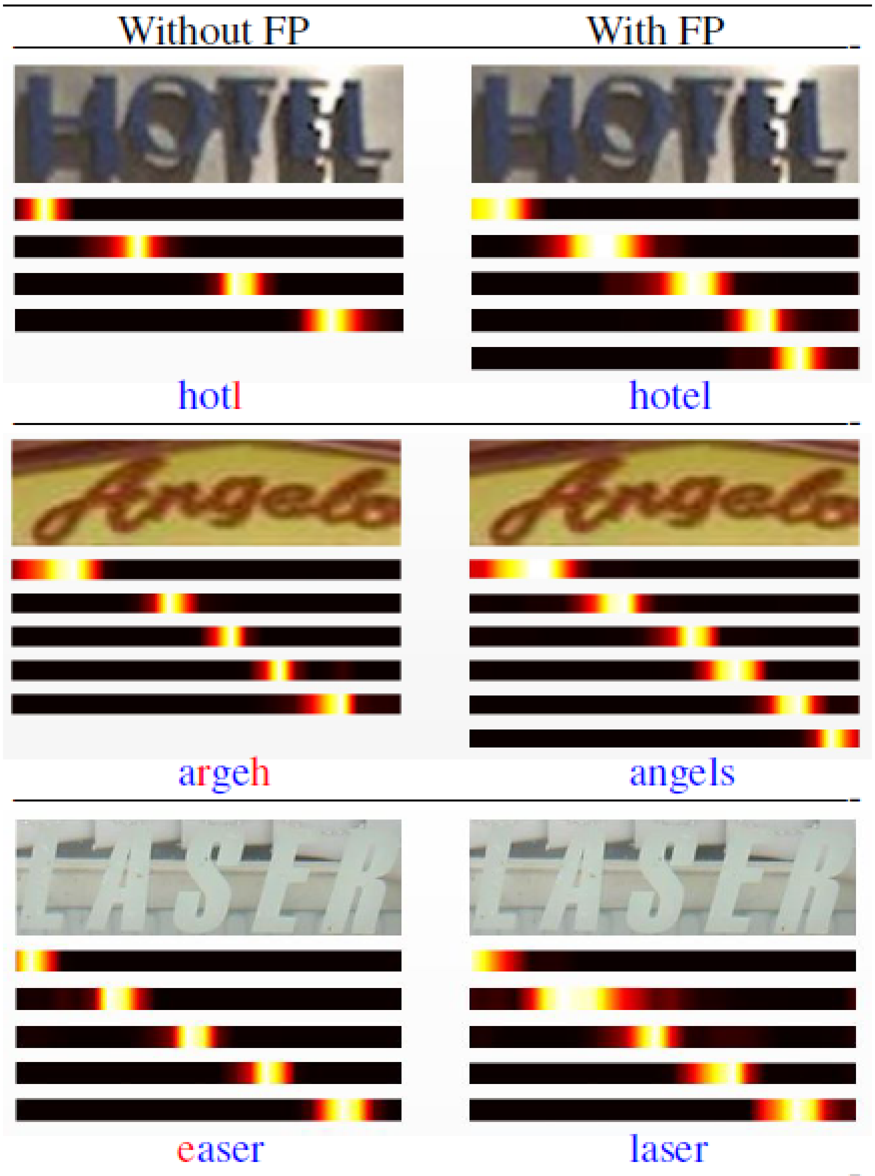
State s^t is computed as:

$$s_t = GRU(y_{prev}, g_t, s_{t-1})$$

Fractional Pickup

针对一些由于噪声干扰而产生的误测，如下图所示，该论文还提出了一种措施叫做**fractional pickup**

An attention-based decoder trained by fractional pickup method can **perceive adjacent characters**. The wider field of attention contributes to the **robustness of the MORAN.**



a pair of **attention weights** are selected and modified at every time step:

$$\begin{cases} \alpha'_{t,k} = \beta\alpha_{t,k} + (1-\beta)\alpha_{t,k+1} \\ \alpha'_{t,k+1} = (1-\beta)\alpha_{t,k} + \beta\alpha_{t,k+1} \end{cases}$$

主要有以下几个优点：

1、**Variation of Distribution**

因为参数是参考临近features，而且具有随机性，增强了参数$\alpha^{t,k}$,$\alpha^{t,k+1}$的鲁棒性，这就造成了即使对于同一张图片，每一个step产生的贡献可能不相同，所以容易**避免过拟合和增强编码的鲁棒性。**

2、**Shortcut of Forward Propagation**

for step k + 1 in the bidirectional-LSTM, a shortcut connecting to step k is created by fractional pickup. The shortcut retains some features of the previous step in the training phase, which is the interference to the forget gate in bidirectional-LSTM.

3、**Broader Visual Field**

Without fractional pickup, the error term of sequence feature vector h^k is

$$\delta_{h_k} = \delta_{g_t} \alpha_{t,k}$$

结果只和一个固定的参数相关，但是加入了fractional pickup以后，等式就变成了：

$$\delta_{h_k} = \delta_{g_t}(\beta\alpha_{t,k} + (1-\beta)\alpha_{t,k+1})$$

结果不仅与当前的feature相关，也与相邻的features相关，back-propagated gradients are able to **dynamically optimize** the decoder over a **broader range of neighbouring regions.**

Performance of the MORAN

Table 6. Comparison with STAR-Net.

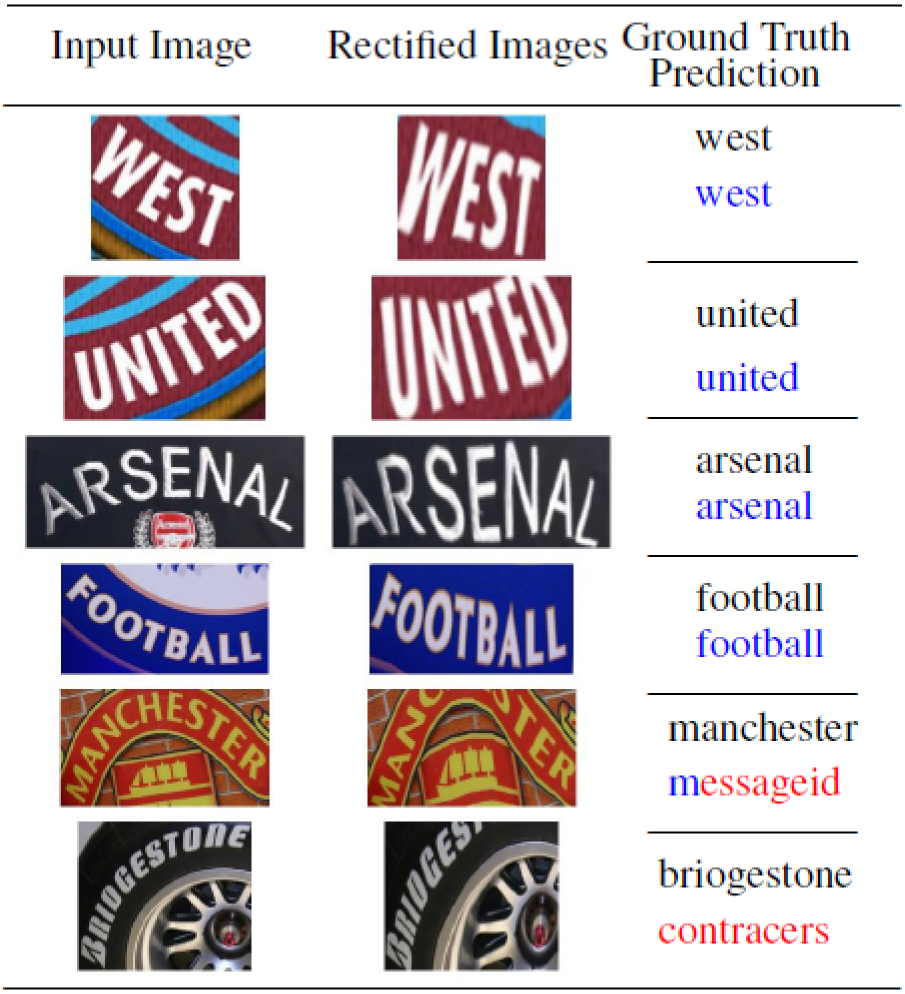
Method	IIIT5K	SVT	IC03	IC13	SVT-P
Liu et al. [28]	83.3	83.6	89.9	**89.1**	73.5
Ours	**87.5**	**83.9**	**92.5**	**89.1**	**74.6**

Table 8. Results on general benchmarks. "50" and "1k" are lexicon sizes. "Full" indicates the combined lexicon of all images in the benchmarks. "None" means lexicon-free.

Method	IIIT5K			SVT		IC03			IC13
	50	1k	None	50	None	50	Full	None	None
Almazán et al [1]	91.2	82.1	-	89.2	-	-	-	-	-
Yao et al. [52]	80.2	69.3	-	75.9	-	88.5	80.3	-	-
R.-Serrano et al. [38]	76.1	57.4	-	70.0	-	-	-	-	-
Jaderberg et al. [23]	-	-	-	86.1	-	96.2	91.5	-	-
Su and Lu [44]	-	-	-	83.0	-	92.0	82.0	-	-
Gordo [12]	93.3	86.6	-	91.8	-	-	-	-	-
Jaderberg et al. [21]	95.5	89.6	-	93.2	71.7	97.8	97.0	89.6	81.8
Jaderberg et al. [22]	97.1	92.7	-	95.4	80.7*	**98.7**	**98.6**	93.1*	90.8*
Shi, Bai, and Yao [41]	97.8	95.0	81.2	**97.5**	82.7	**98.7**	98.0	91.9	89.6
Shi et al. [42]	96.2	93.8	81.9	95.5	81.9	98.3	96.2	90.1	88.6
Lee and Osindero [27]	96.8	94.4	78.4	96.3	80.7	97.9	97.0	88.7	90.0
Liu et al. [28]	97.7	94.5	83.3	95.5	83.6	96.9	95.3	89.9	89.1
Yang et al. [51]	97.8	96.1	-	95.2	-	97.7	-	-	-
Yin et al. [54]	98.7	96.1	78.2	95.1	72.5	97.6	96.5	81.1	81.4
Cheng et al. [5]	98.9	96.8	83.7	95.7	82.2	98.5	96.7	91.5	89.4
Cheng et al. [6]	**99.6**	**98.1**	87.0	96.0	82.8	98.5	97.1	91.5	-
Ours	97.9	96.2	**91.2**	96.6	**88.3**	**98.7**	97.8	**95.0**	**92.4**

Limitation of the MORAN

because of complicated background, the MORAN will fail when the **curve angle is too large.**



Conclusion

The proposed framework involves two stages: **rectification and recognition.** First, a multiobject rectification network, which is free of geometric constraints and flexible enough to handle complicated deformations, was proposed to transform an image containing **irregular text into a more readable one.**The proposed MORAN is trained in **a weak-supervised way,** which requires **only images and the corresponding text labels.**

反馈与建议

- 微博：@柏林designer
- 邮箱：weijia_wu@yeah.net