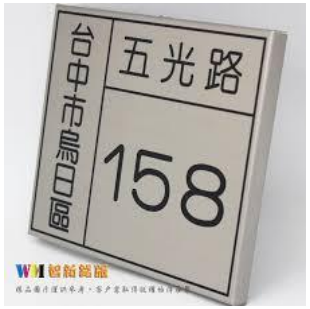
HUAWEI

# Day 8 文字识别（OCR）关键技术

# 华为OCR使用场景介绍

## 通用类





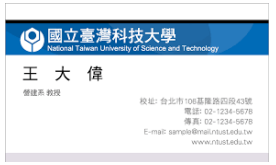



## 单据类





## 证件类





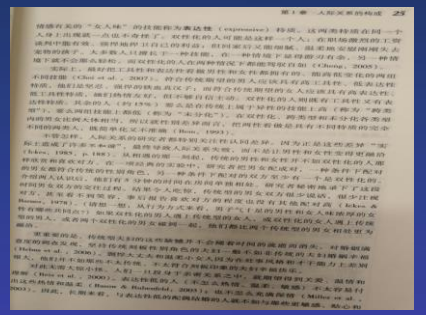

電子護照

# 文字定位与识别难点

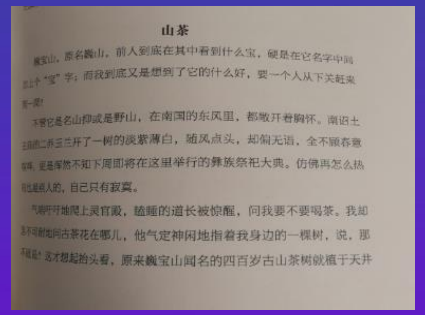
小尺度文字


多尺度文字

与背景融为一体、极大宽高比



密集文字




扭曲


倾斜

## 难点与挑战

### 小尺度文字

- 在通用场景中，文字存在占比极小情况，文字宽高与图像宽高不成比例，对文字定位检测造成困扰

### 多尺度文字

- 在同一种场景中，文字大小、尺寸不集中，相差很大，给文字定位检测带来了很大的挑战

### 复杂背景

- 与背景融为一体，容易当做背景漏检造成识别失败

### 密集文字

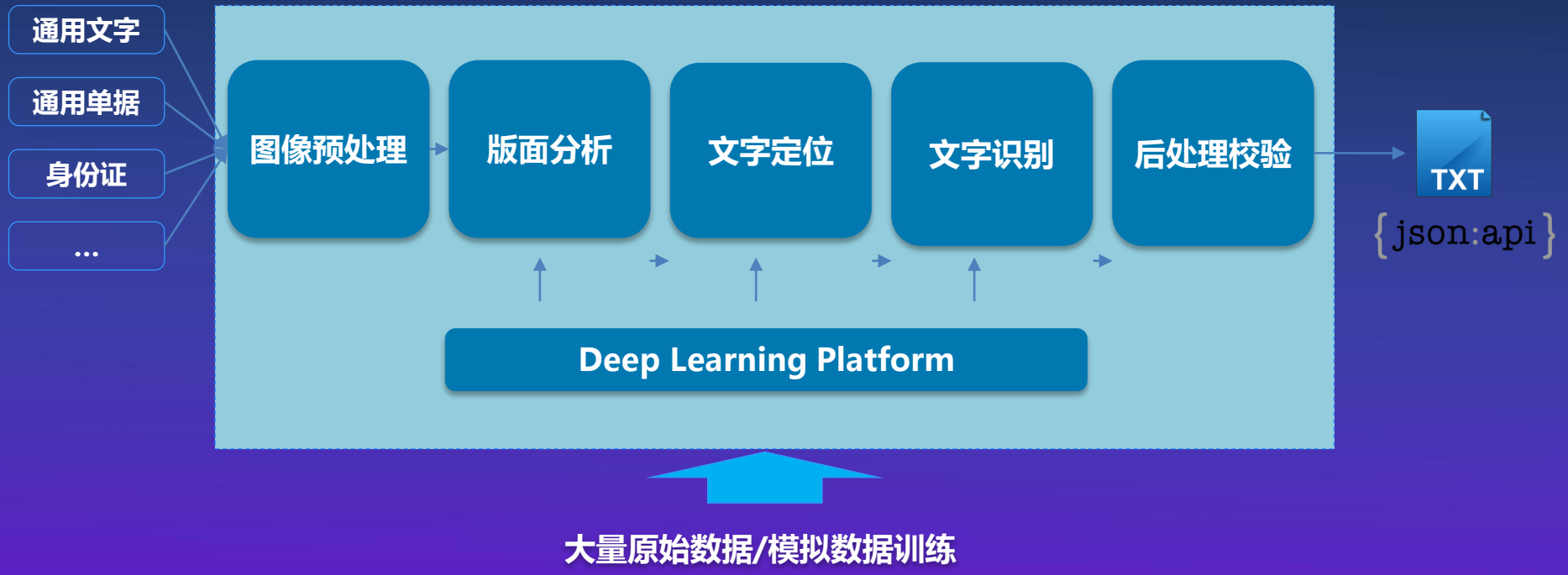
- 在单据场景中，文字通常铺满整个版面，且伴随倾斜角度以及可能的错行、重叠等复杂情况
- 目前文字检测识别算法大都只能做到按行检测识别，很多算法并不然很好处理解决此场景

### 扭曲、倾斜

- 扭曲、倾斜文字在画框截取图像块识别时容易与其他文字块重叠，从而导致识别出错

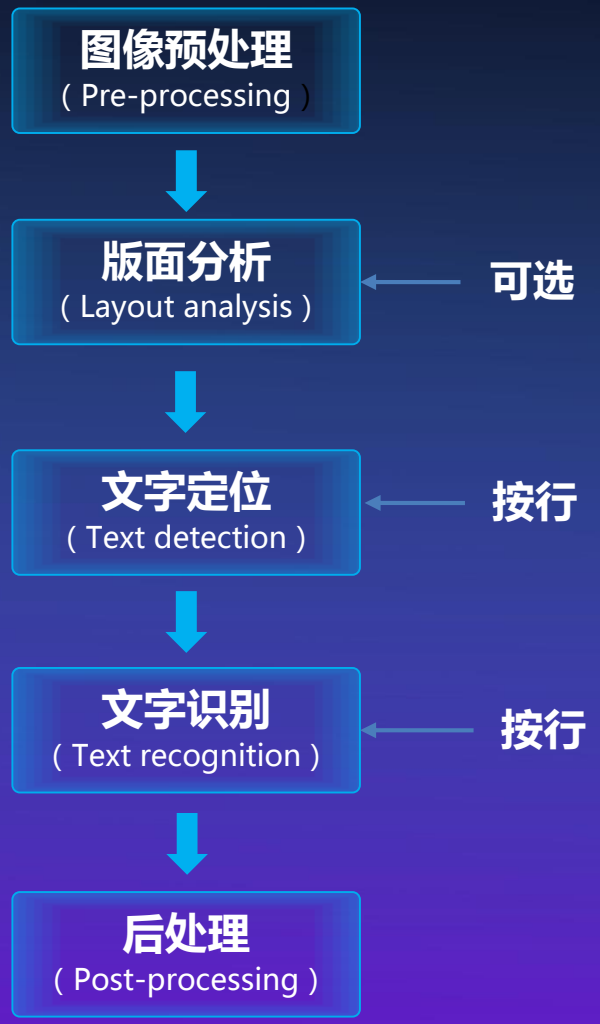
# 华为云OCR关键技术

融合多种先进深度学习技术，提供通用的单据、证件、通用文字识别等服务，**高精度、鲁棒、自适应。**
基于华为众多合作伙伴的长期实践，成功应用于各类场景，经受住各种复杂场景考验。



# 华为OCR识别关键技术

OCR单据识别关键技术




## 适用场景与效果

多种单据类型

- 智能版面分析，适应不同表单结构，有/无表格线

不同图像质量

- 自动矫正图像，如模糊、倾斜、扭曲等

盖章

- 智能去除各种形状的盖章干扰，提取覆盖文字

错行

- 适用文字与表格线倾斜交叉的场景，消除线条干扰

文字重叠

- 适用相邻文字重叠、文字行交错的场景，例如机打发票等

中英文混合

- 自动识别中文、英文、符号等，针对专业领域字符优化

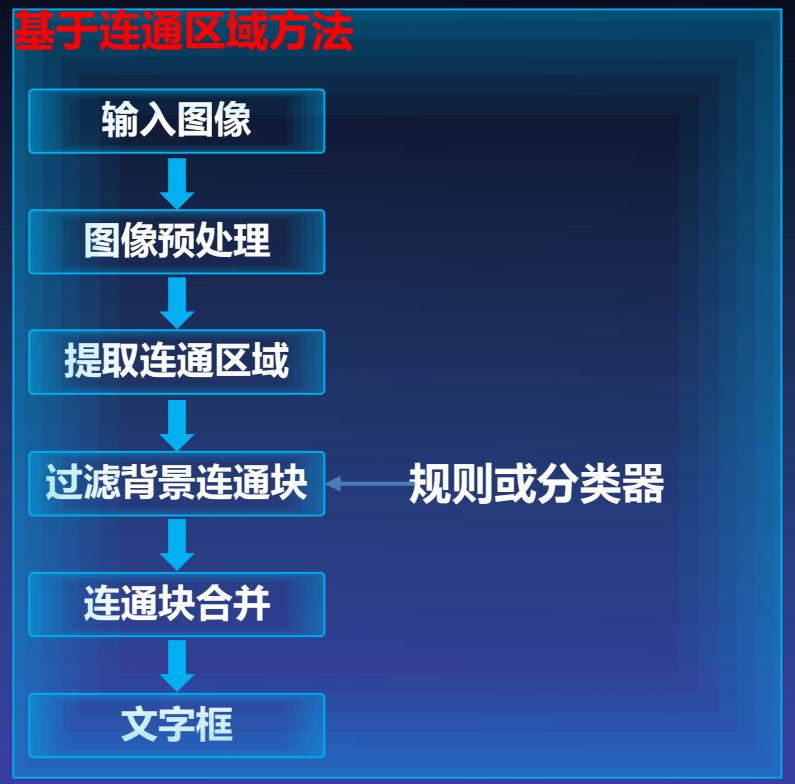
# 文字定位


✓ 文字定位是文字识别的前提条件，要解决的问题是如何在杂乱无序、千奇百怪的不同场景中准确地定位出文字的位置。由于不同场景背景的复杂性、光照的多变性以及字体的不可预测性等原因，文字定位面临着极大的挑战。文字定位的发展历程与绝大多数的计算机视觉任务相似，传统的检测方法包括基于连通域的方法，以及基于滑动窗口的方法，自2014年起，基于深度学习的方法逐渐成为主流方法。

# 文字定位

- **基于传统方法(~2014年)**
  - 基于连通区域(MSER, SWT...)
  - 基于滑动窗
- **基于深度学习（2014年~）**
  - 基于侯选框
    - TextBoxes: A fast text detector with a single deep neural network
    - Detecting Oriented Text in Natural Images by Linking Segments
  - 基于分割
    - Multi-oriented text detection with fully convolutional networks
  - 基于融合
    - Deep Direct Regression for Multi-Oriented Scene Text Detection

# 文字定位：传统定位方法





●**基于连通区域方法**：该方法认为图像中的文字一般都是作为连通域出现，这类方法一般可以分为连通域提取和文字连通域判别两个环节。其中，连通域提取环节会将图像中的所有连通域提取出来作为备选，文字连通域判别环节则会根据分类器或者启发式规则判断提取的连通域到底是文字连通域还是背景连通域，常见方法有如基于MSER（Maximally Stable Extrernal Regions）最大稳定极值区域检测算法。

●**基于滑动窗方法**：该方法将目标文字作为一类特殊的目标，使用类似于目标检测的方法对图像中的文字进行定位检测。首先从每一个滑动窗口中提取若干特征，这里特征提取常使用如(Histogram of Oriented Gradients)等特征，提取完特征后会将提取的特征送入事先训练好的分类器判断当前滑动窗口中是否包含文字，最后需要通过二值化来精确定位文本的位置。

# 传统文字定位：基于MSER稳定极值区域

●MSER（Maximally Stable Extrernal Regions）最大稳定极值区域检测算法： 从场景图像中检测到文字块，并分割出来进行OCR识别
●对光照、模糊、低对比度等有一定的鲁棒性
●不足：只能针对背景单一场景，对稍复杂场景效果不佳

| | | |
|---|---|---|
| 3 | 2 | 2 |
| 2 | 3 | 1 |
| 1 | 2 | 3 |

图像

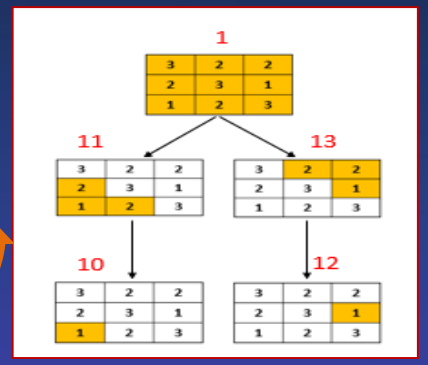
| | | |
|---|---|---|
| 3 | 2 | 2 |
| 2 | 3 | 1 |
| 1 | 2 | 3 |

ER栈

| ID | 点 | Gray level | Grow history |
|---|---|---|---|
| 0 | | 256 | |
| 1 | (1,2)(0,1)(0,2)(2,0)(2,1)(1,0)(0,0)(2,2)(1,1) | 3 | 13 |

Grow History

| ID | 点 | Gray Level | Parent | Child |
|---|---|---|---|---|
| 10 | (2,0) | 1 | | 11 |
| 11 | (2,0)(2,1)(1,0) | 2 | 10 | |
| 12 | (1,2) | 1 | | 13 |
| 13 | (1,2)(0,1)(0,2) | 2 | 12 | |




✓MSER的基本原理是对一幅灰度图像（灰度值为0～255）取阈值进行二值化处理，阈值从0到255依次递增。阈值的递增类似于分水岭算法中的水面的上升，随着水面的上升，有一些较低的丘陵会被淹没，如果从天空往下看，则大地分为陆地和水域两个部分，这类似于二值图像。在得到的所有二值图像中，图像中的某些连通区域变化很小，甚至没有变化，则该区域就被称为最大稳定极值区域。

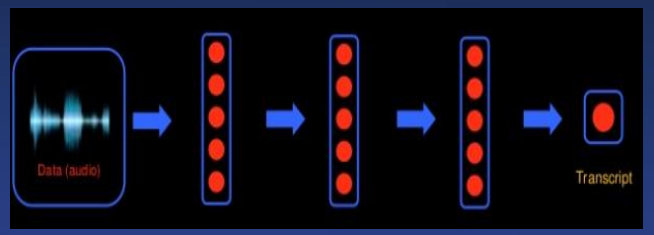


```
识别结果为：HANDICAPPED
PARKING
*SPECIAL PLATE
REQUIRED
UNAUTHORIZED
VEHICLES
MAY BE TOWED
AT OWNERS
EXPENSE
```

# 深度学习在很多领域取得了历史性的突破，证明了实效性

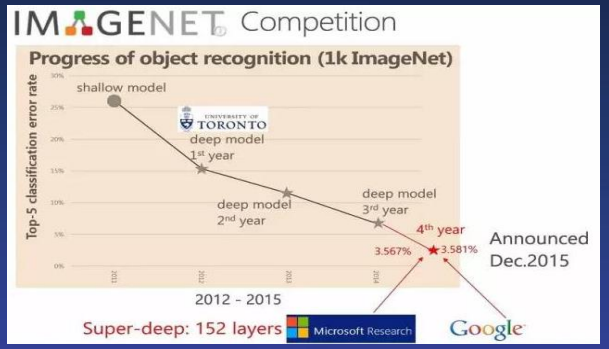
## 被认为已经解决或接近解决的问题

### 语音识别



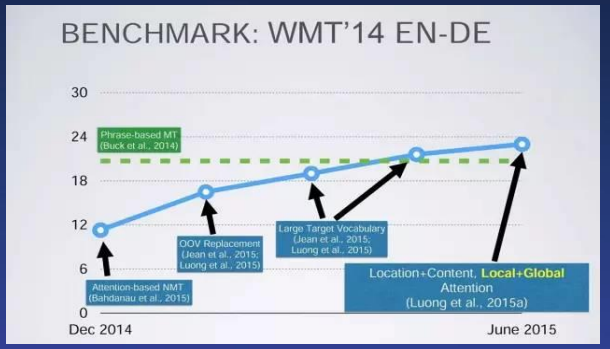

业界最好水平：
CER 3-5%, WER 9-11%，均已达到甚至超过人的平均能力

### 静态图像识别




在一些场景限定领域已经超越人类

### 翻译




准确度和自然度上还没有超越人类

## 正在解决的问题（非完全列表）

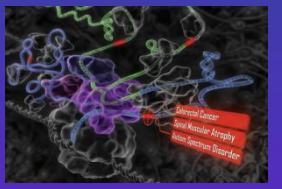
自然语言对话

用户画像

行为建模控制

搜索和推荐

图像语义



个人健康

# 深度学习在OCR识别领域的提升

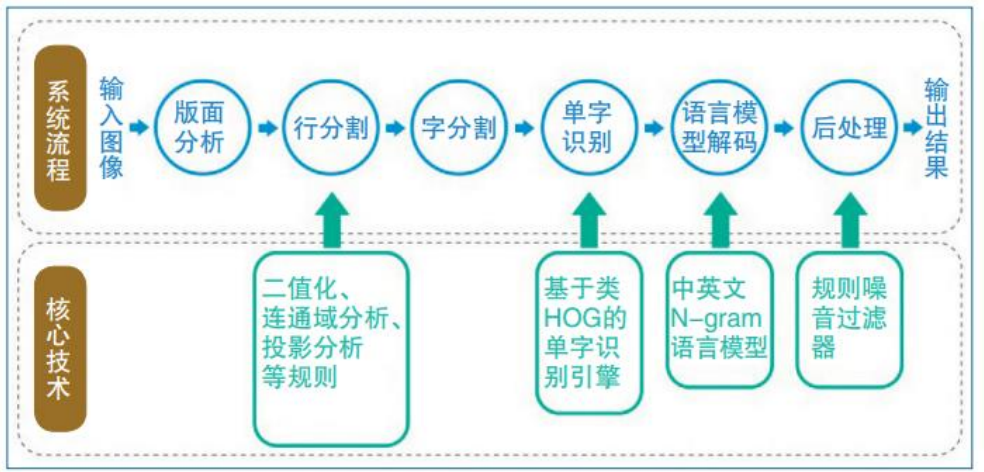



图3　经典的光学字符识别系统流程和技术框架

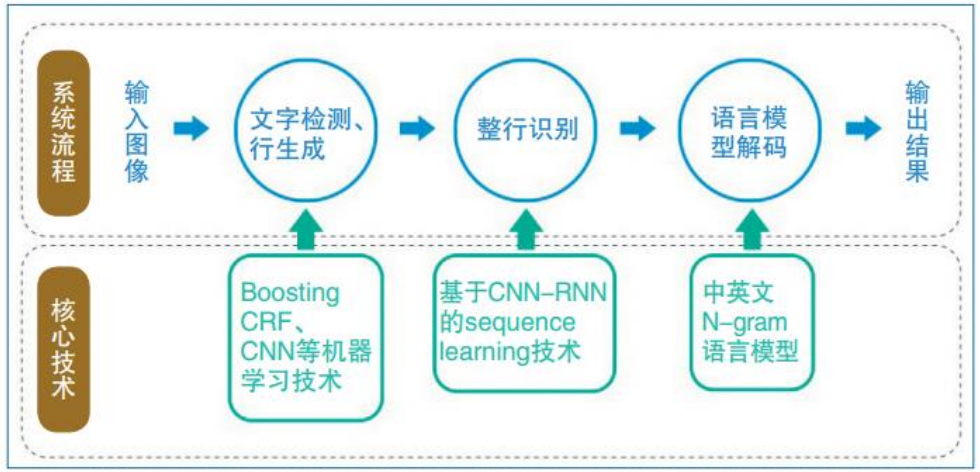
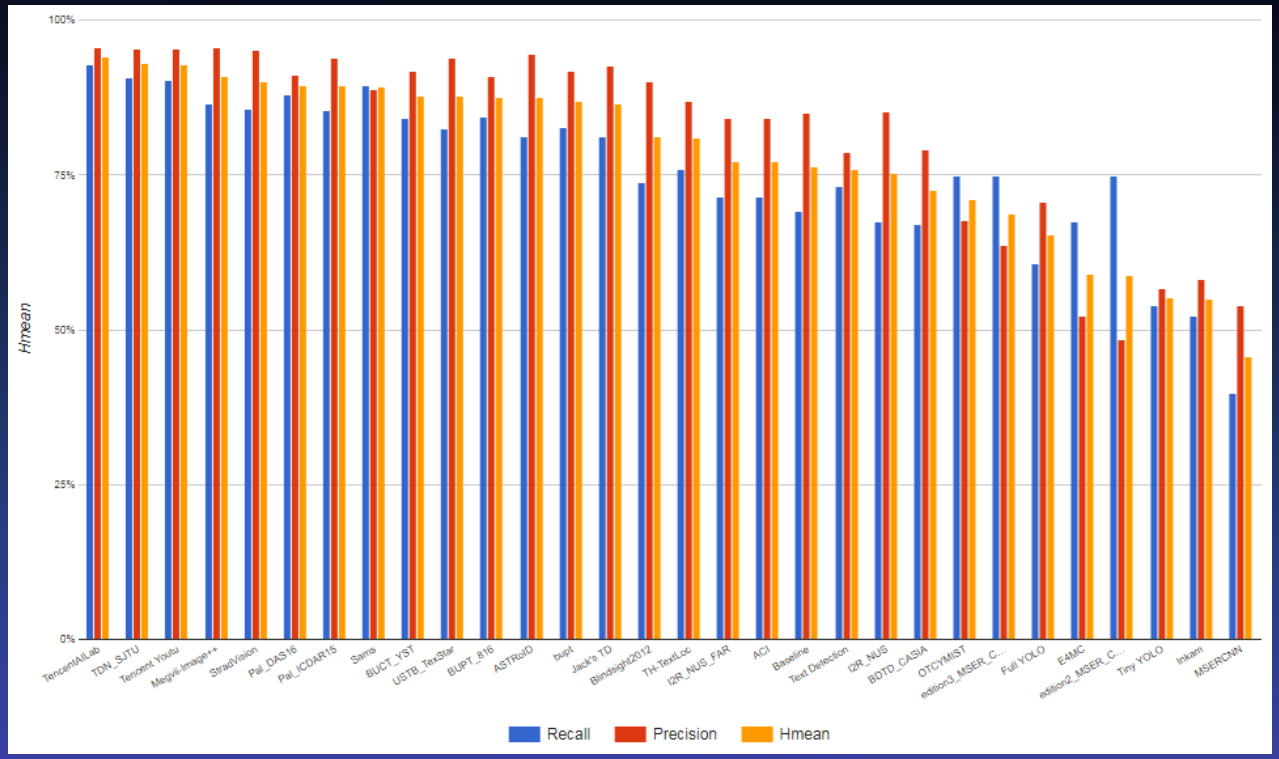


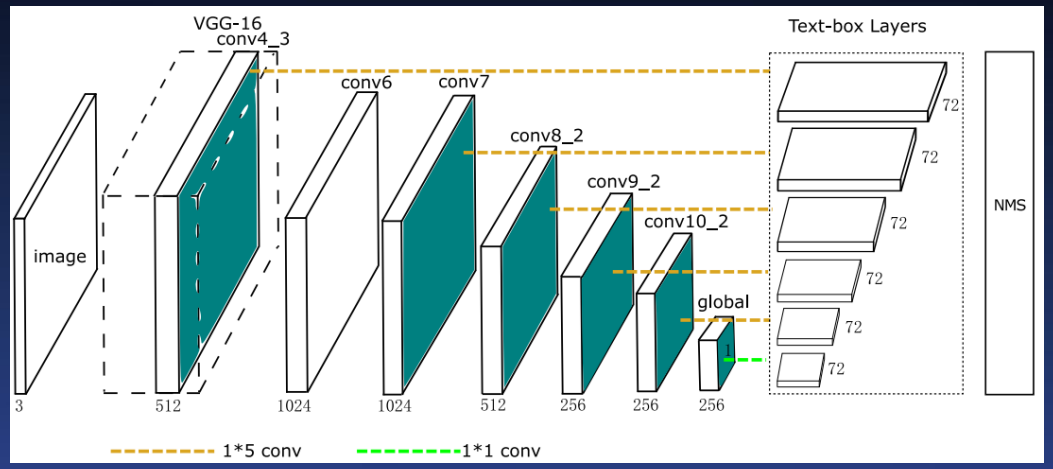

图4　基于CNN-RNN的序列光学字符识别流程




ICDAR2013数据集竞赛结果

目前，在OCR文字定位与识别领域，主流方法都是基于深度学习方法。如在ICDAR2013自然场景上已经能达到94.08%的f-measure准确率，前20名已经见不到传统方法的影子了。
http://rrc.cvc.uab.es/?ch=1&com=evaluation&task=1&f=1&e=1

# 基于候选框深度学习文字定位方法


TextBoxes网络

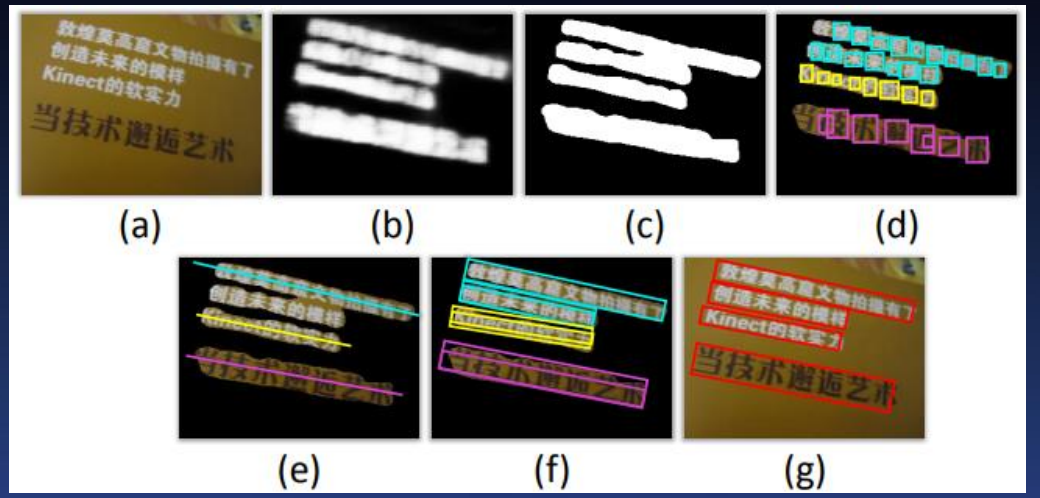
在检测领域，基于候选框（region proposal）的深度卷积神经网络算法是主流方法之一，如基于Region Proposal两阶段的Faster RCNN。在基于深度卷积网络的物体检测框架中，每一层卷积相当于一层滤波器，多层滤波器的串联可以使最有效的特征被传递到最后，再根据文字目标与物体目标的差异，针对性地改进检测任务的实现机制，能够获得良好的文本检测效果。目前基于候选框深度学习网络有如TextBoxes， SegLink。

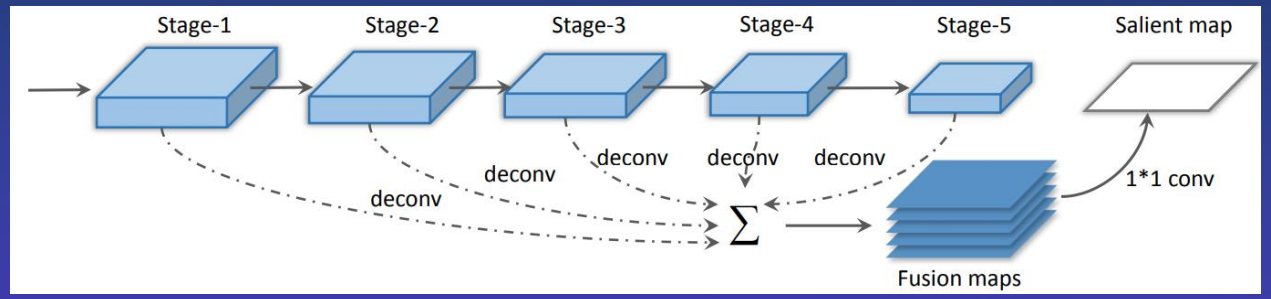
Liao, M.; Shi, B.; Bai, X.; Wang, X.; and Liu, W. 2016a. Textboxes: A fast text detector with a single deep neural network.

B. Shi et al. Detecting Oriented Text in Natural Images by Linking Segments. IEEE CVPR, 2017

# 基于语义分割的FCN文字定位方法

- 将文字定位问题结转化成语义分割问题；
- 采用FCN网络构，优势：
  - ✓ 同时包含local和global context information；
  - ✓ 端到端的训练方式；
  - ✓ 去掉了全连接层，高效pixel labeling




Text-Block FCN

前5个conv stages来自于VGG-16；
每一个conv stages后接一个deconv层；
获取Fusion maps

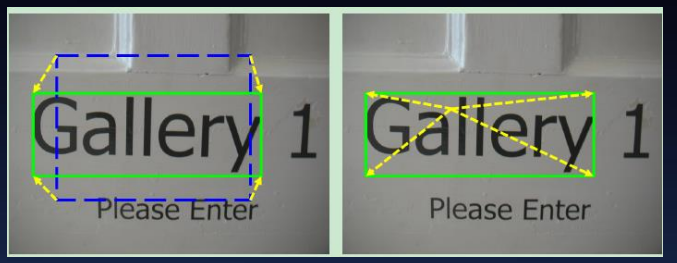
算法流程：
（a）Input
（b）Text-block FCN预测到的文字区域的salient map
（c）Text block
（d）字符元素提取：MSER
（e）方向估计
（f）文字行提取
（g）Output

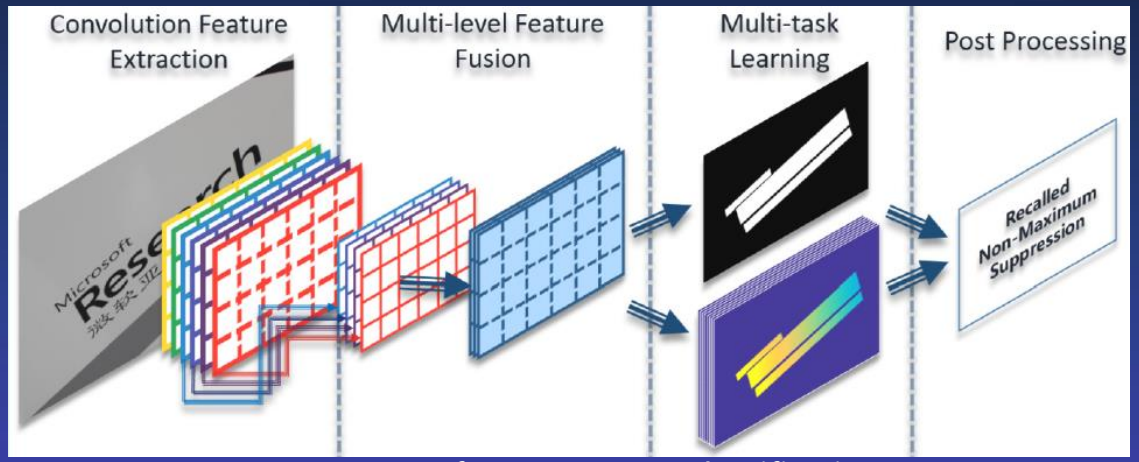
Zhang, Zheng, et al. "Multi-oriented text detection with fully convolutional networks." CVPR. 2016.

# 基于融合的深度网络定位方法

直接 vs 间接回归：

- 间接回归：基于生成的候选框进行回归, e.g., Faster-RCNN, SSD, etc.
  但不支持多方向文字检测。
- 直接回归：从目标点直接回归，更适合用来生成不规则四边形。




感受野大于图片大小，确保能定位到长文字块

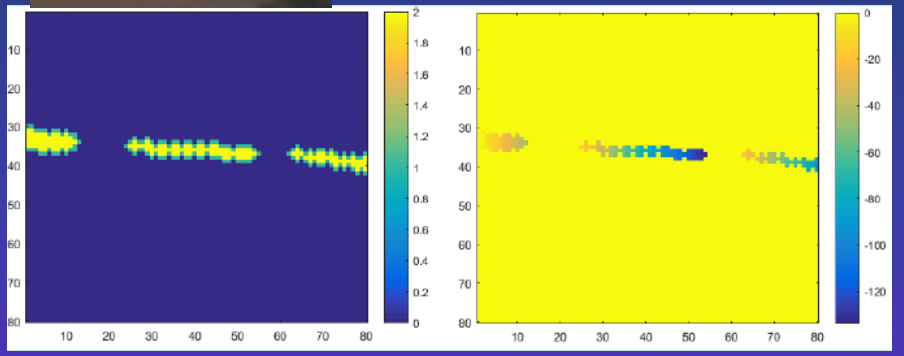
refer to FCN

classification & regression

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{loc} \cdot \mathcal{L}_{loc}$$

Classification task: hing loss
Regression task: smooth L1 loss，8 outputs

文字区域的中心线部分作为正样本，这样可以更好区分文字和非文字背景



Ground truth for classification(left) and regression(right)

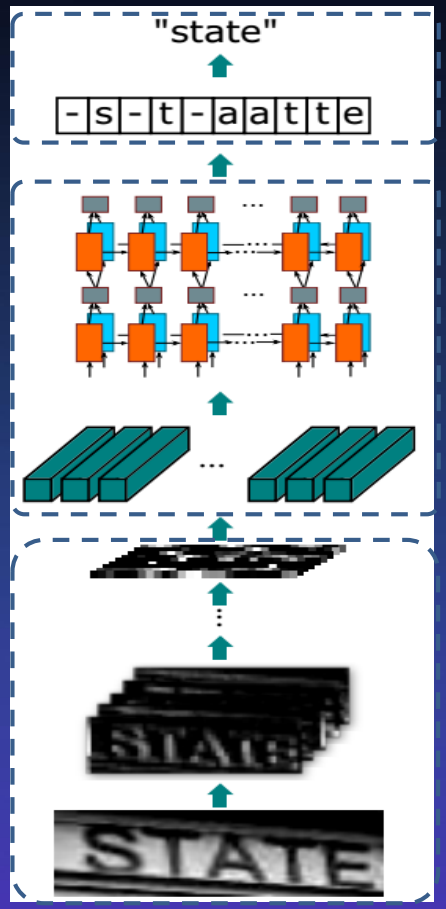
W. He, etc., Deep Direct Regression for Multi-Oriented Scene Text Detection. ICCV, 2017

# 文字识别

CTC

RNN

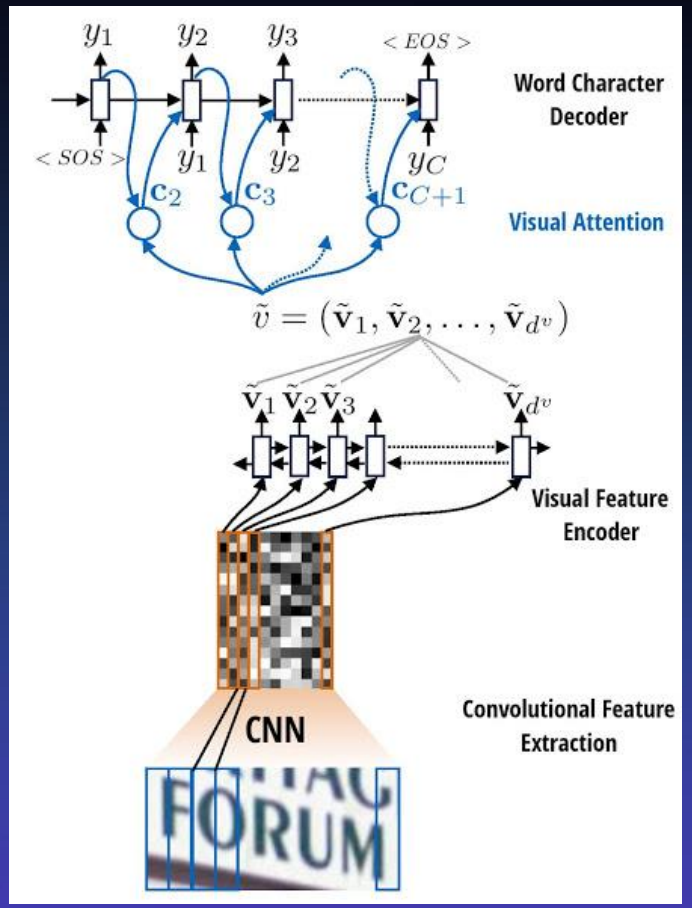
CNN




Attention

RNN

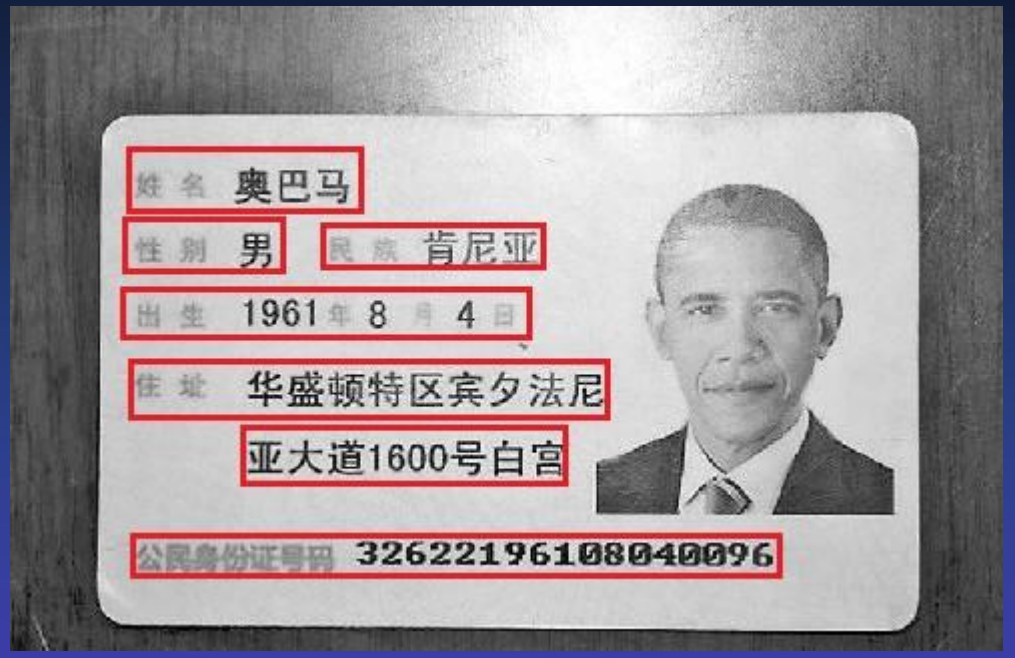
CNN



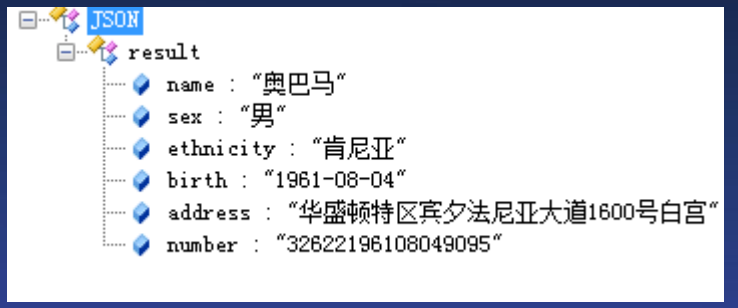

- 在定位出文字行后，对文字行进行识别目前主流方法都是基于CNN+RNN的深度学习网络，不同的是在RNN层后的解码器，主流方法有两种，一种是基于CTC，一种基于Attention注意力机制。

# 后处理：结构化数据提取

身份证




结构化json数据输出




- 后处理结构化数据提取主要是从实际业务出发，从图像识别出来的所有文字行信息中提取所需要的关键键值对信息，如上图身份证为提取对应的姓名、性别等关键键值对信息。
- 华为云OCR基于自身业务出发，探索融合了多种先进深度学习技术，基于华为众多合作伙伴的长期实践，成功应用于各类场景，从预处理到结构化数据提取，以精度高、稳定、安全等特性经受住各种复杂场景考验。

# THANK YOU