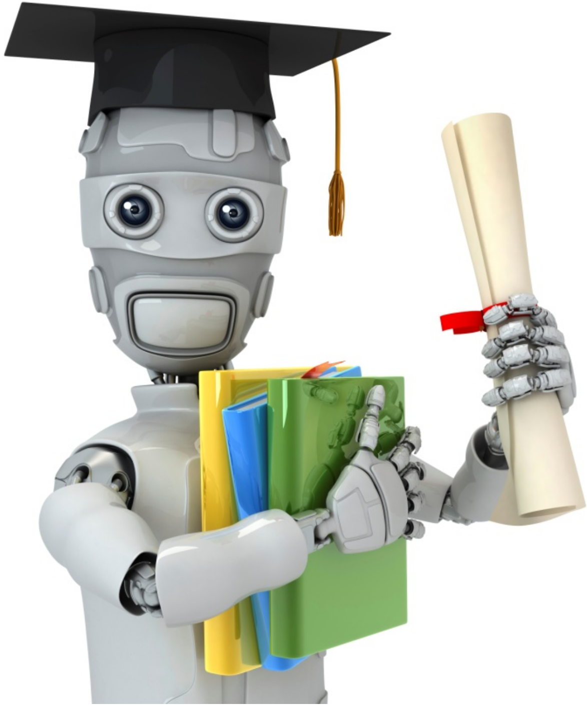


Machine Learning

Application example: Photo OCR

Problem description and pipeline

photo optical character recognition

这种照片OCR技术 主要解决的问题是让计算机 读出照片中拍到的文字信息

The Photo OCR problem

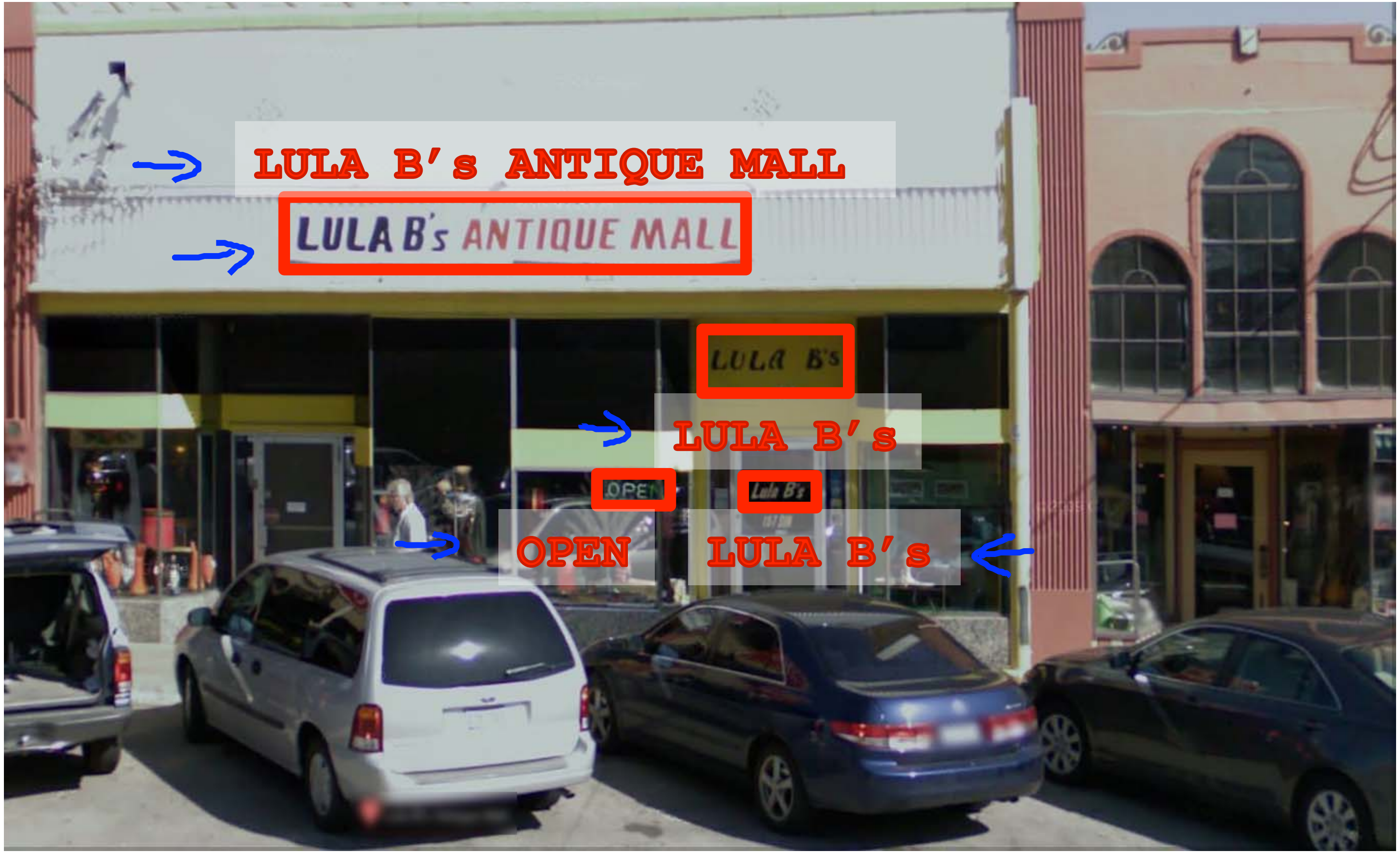
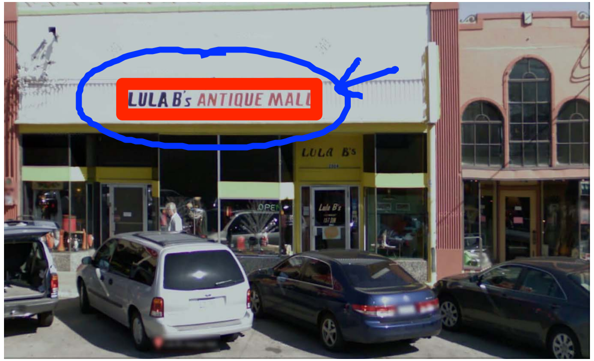
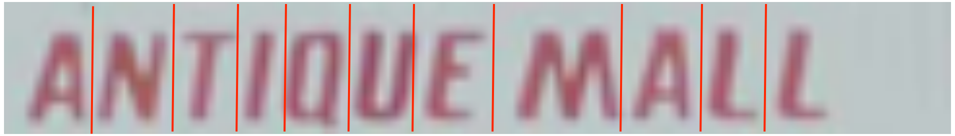


Photo OCR pipeline

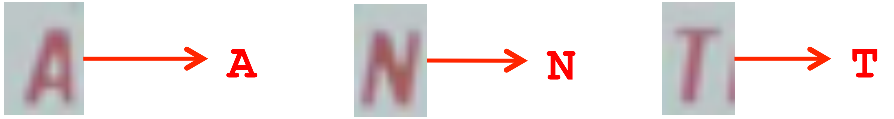
→ 1. Text detection



→ 2. Character segmentation

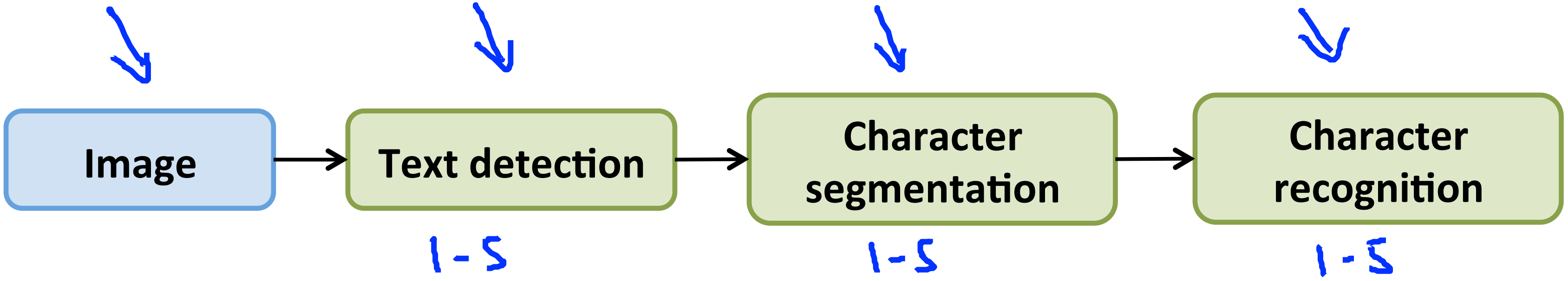


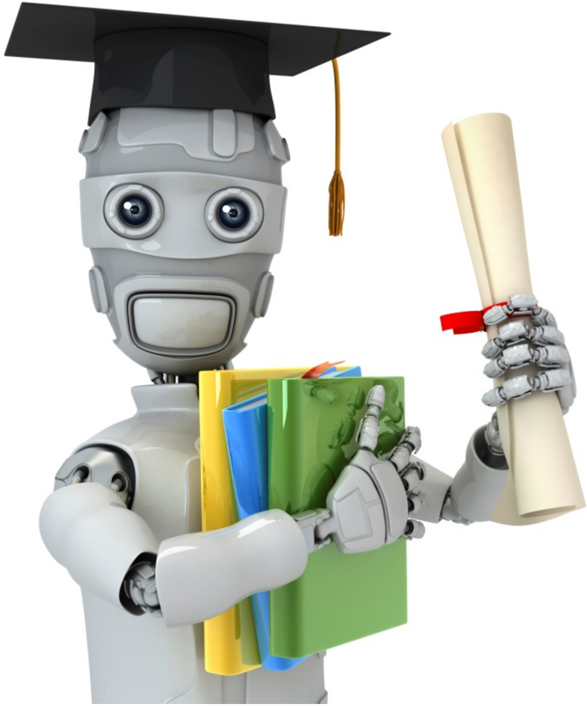
→ 3. Character classification



C1eaning → Cleaning

Photo OCR pipeline



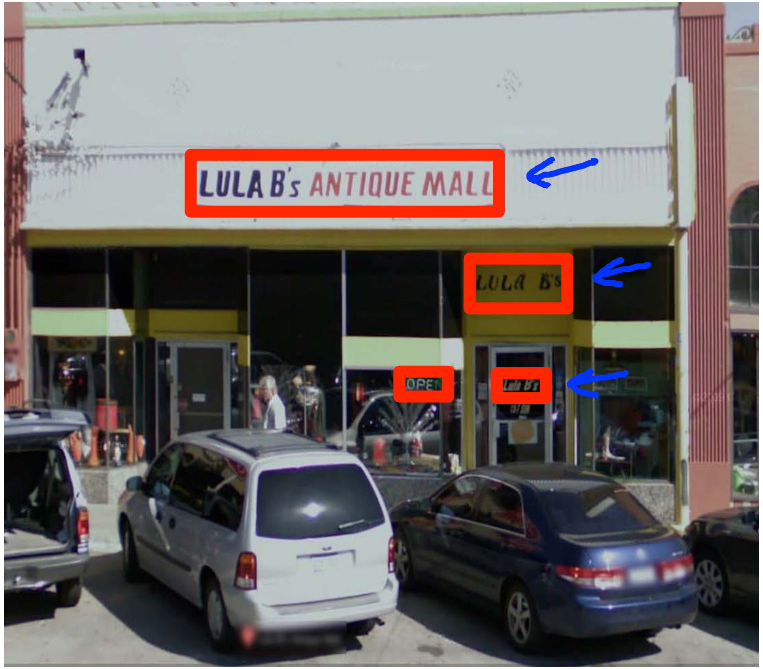


Machine Learning

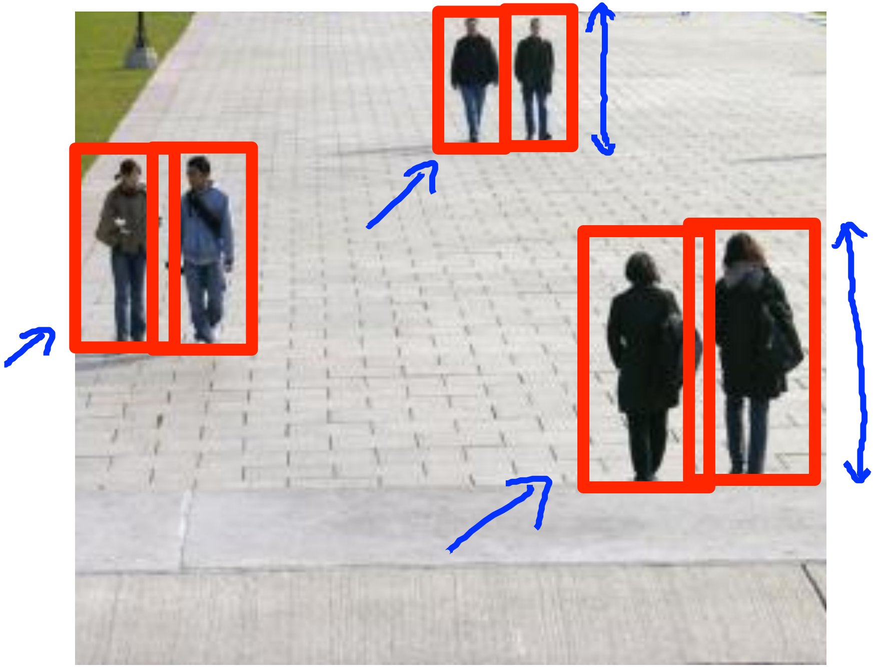
Application example: Photo OCR

Sliding windows

Text detection



Pedestrian detection



Andrew Ng

我们把 宽高比标准化到 82比36这样一个比例

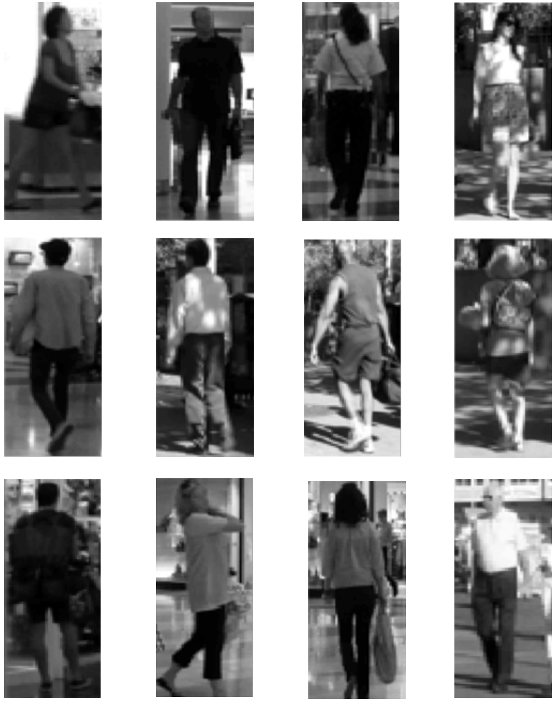
接下来我们要做的 就是到街上去收集一大堆正负训练样本 这些是82×36大小的
有行人的 图像样本 而这些样本里没有行人

Supervised learning for pedestrian detection

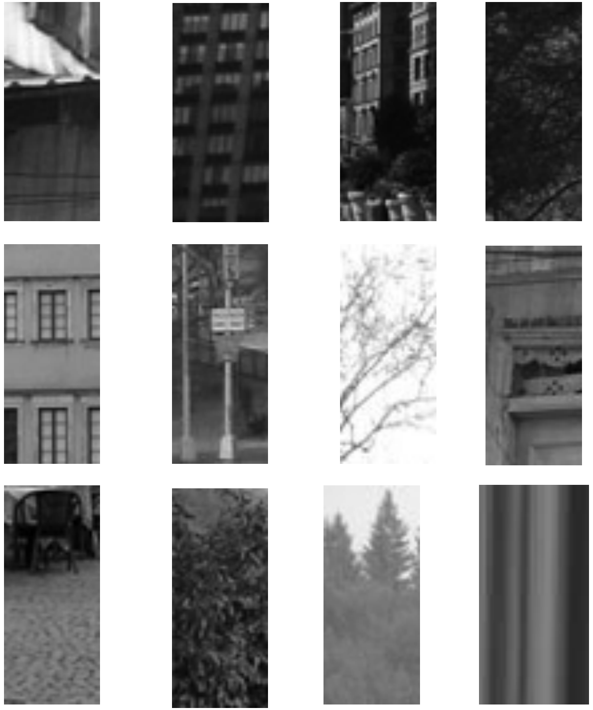
1,000

10,000

$x =$ pixels in 82x36 image patches



Positive examples $(y = 1)$



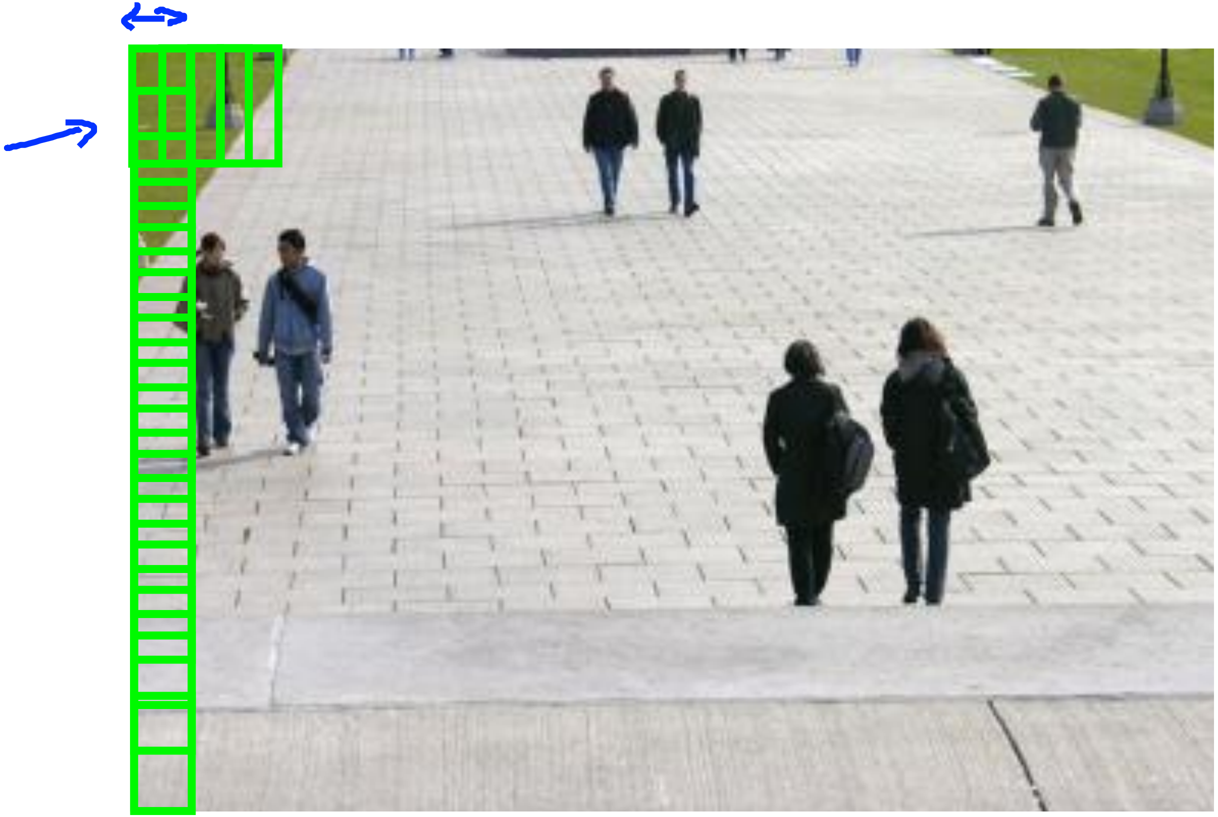
Negative examples $(y = 0)$

Andrew Ng

你每次滑动窗口的 大小是一个参数 通常被称为 步长(step size) 有时也称为 步幅参数(stride parameter)

Sliding window detection

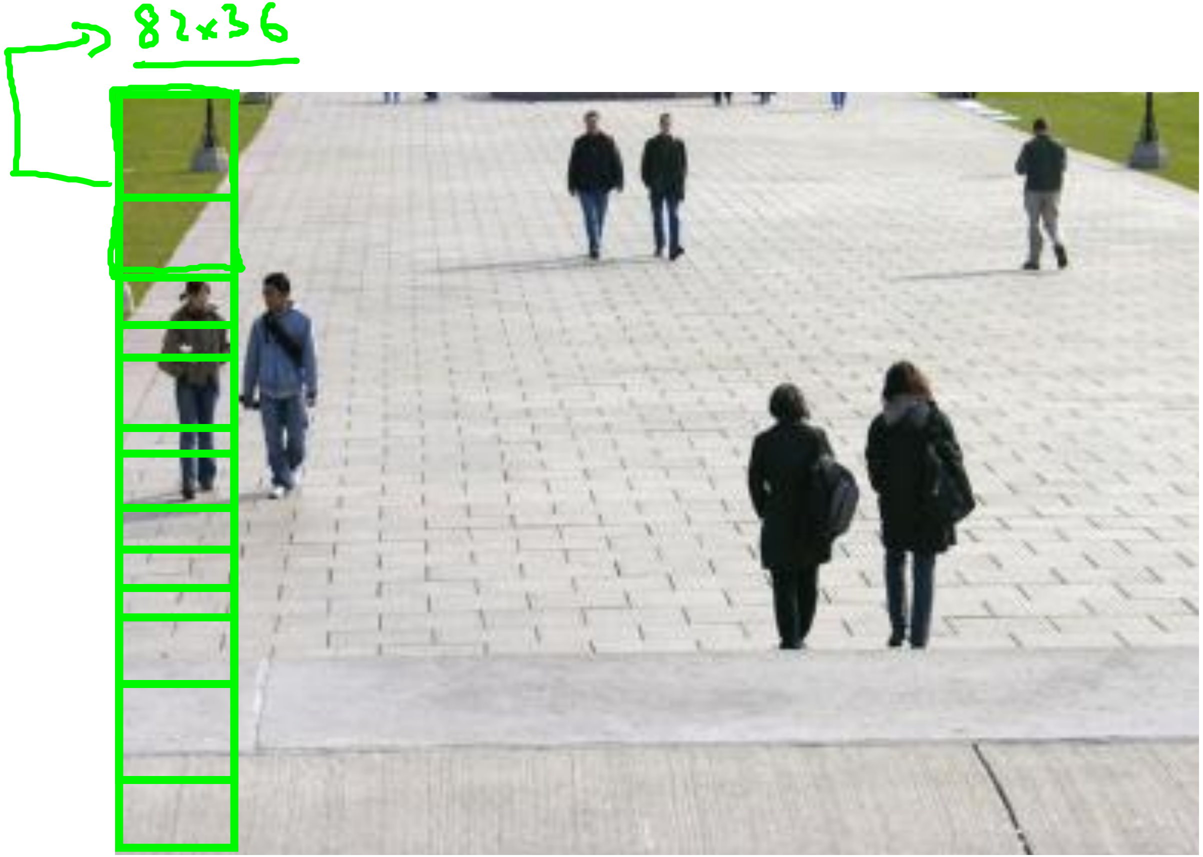
step-size /stride



Andrew Ng

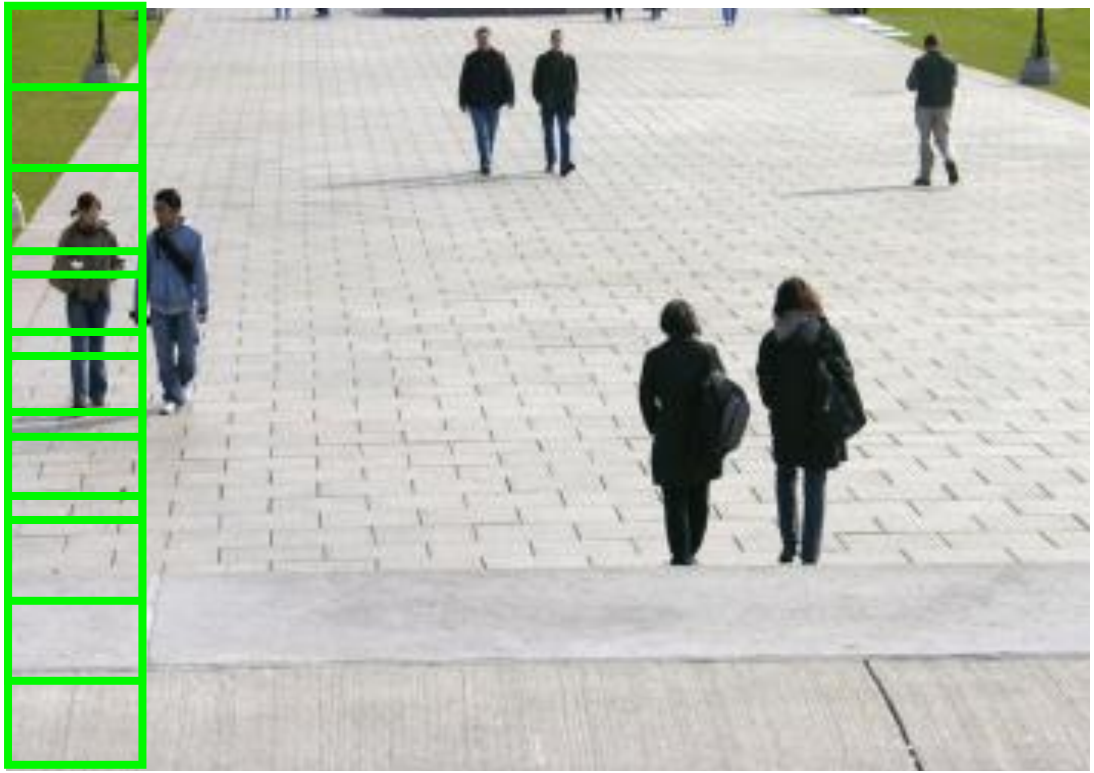
但这个矩形 是非常小的 只能探测到某种尺寸的行人 接下来我们要做的 是看看更大的图像块

Sliding window detection

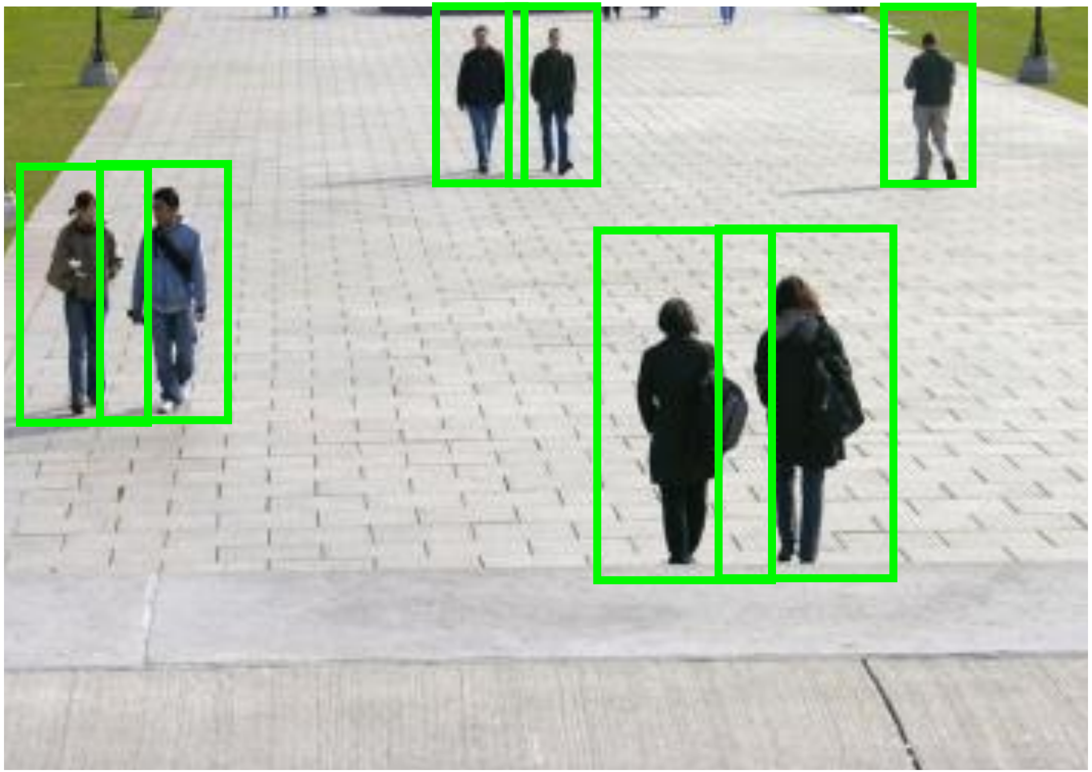


Andrew Ng

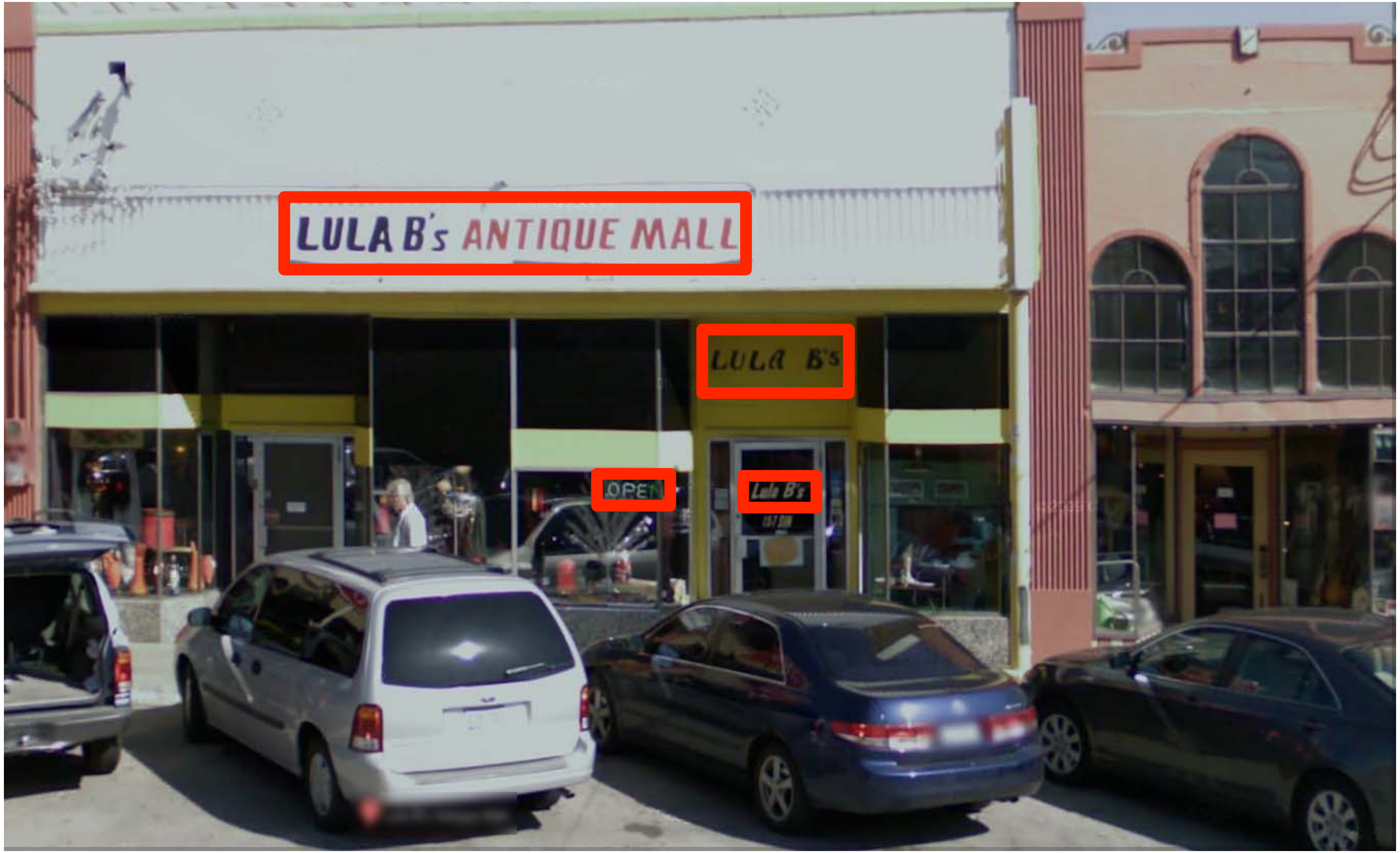
Sliding window detection



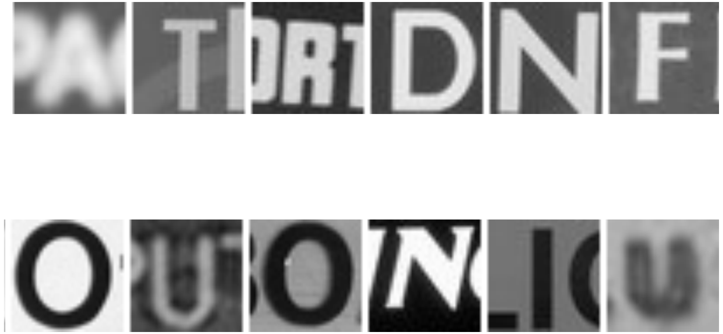
Sliding window detection



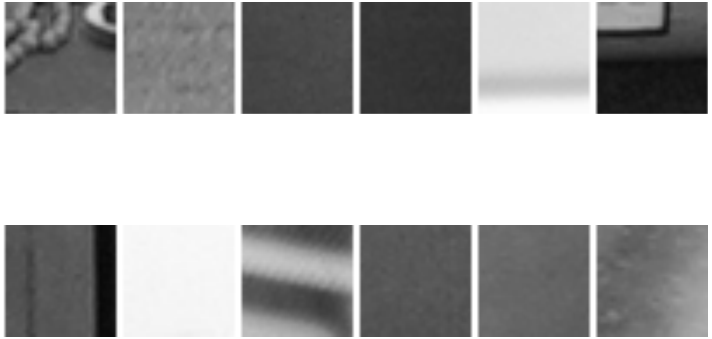
Text detection



Text detection

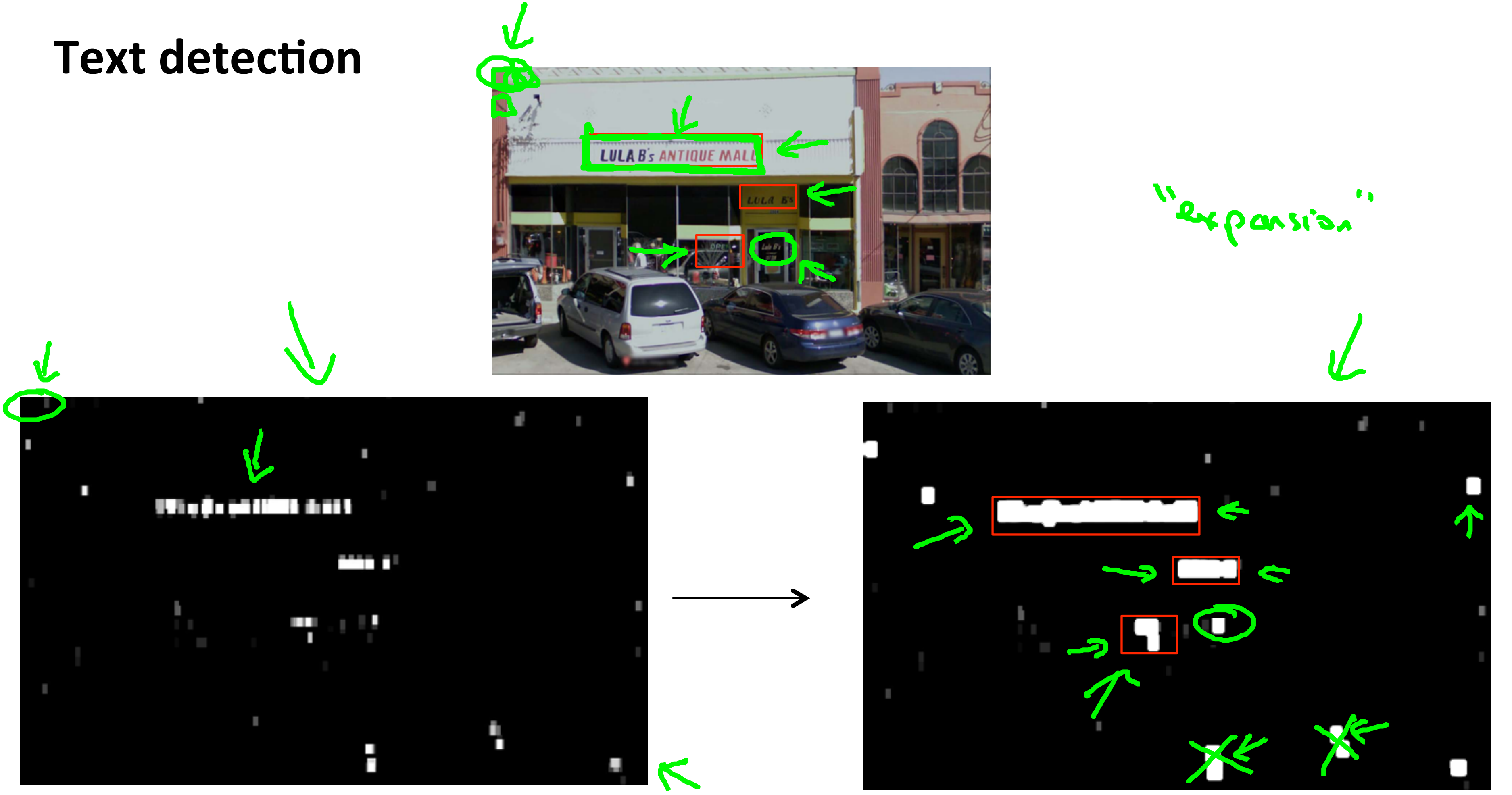


Positive examples $(\underline{y = 1})$



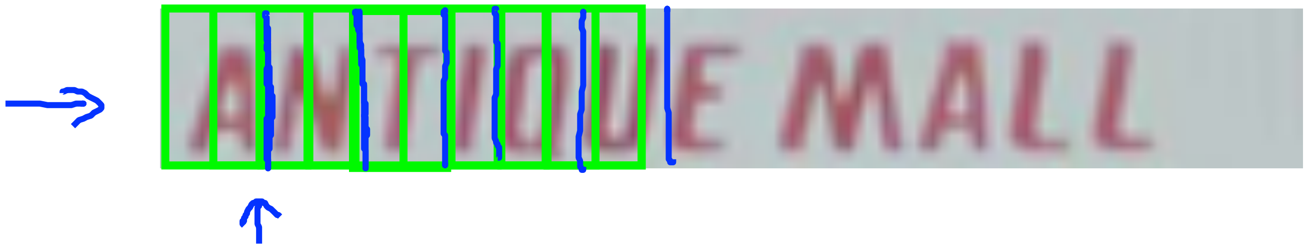
Negative examples $(\underline{y = 0})$

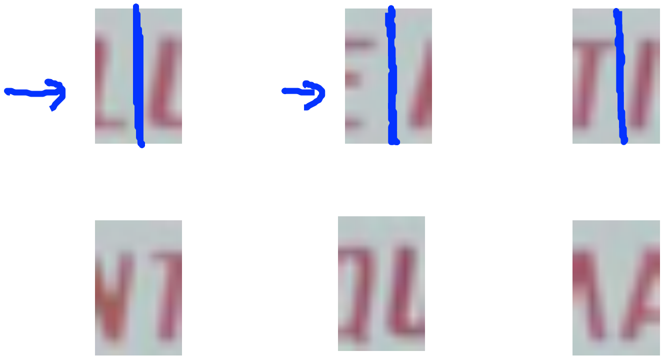
Text detection



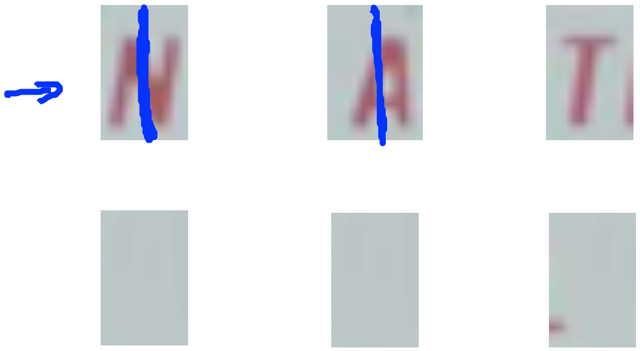
[David Wu]

1D Sliding window for character segmentation





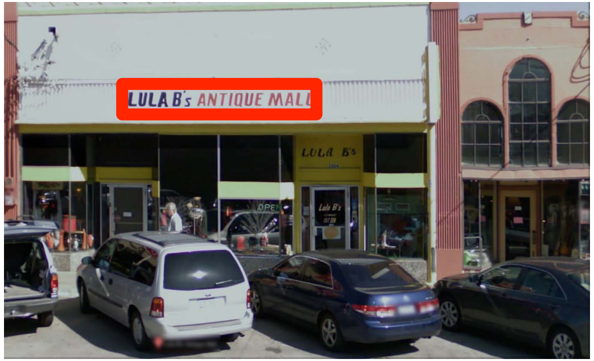
Positive examples $(y = 1)$



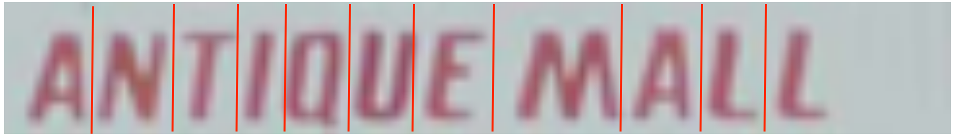
Negative examples $(y = 0)$

Photo OCR pipeline

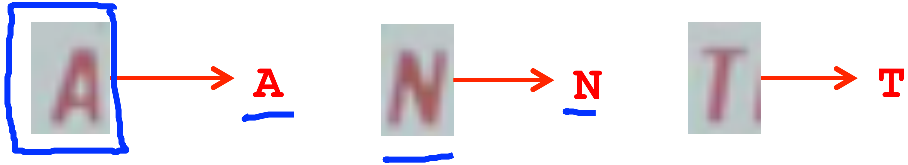
1. Text detection

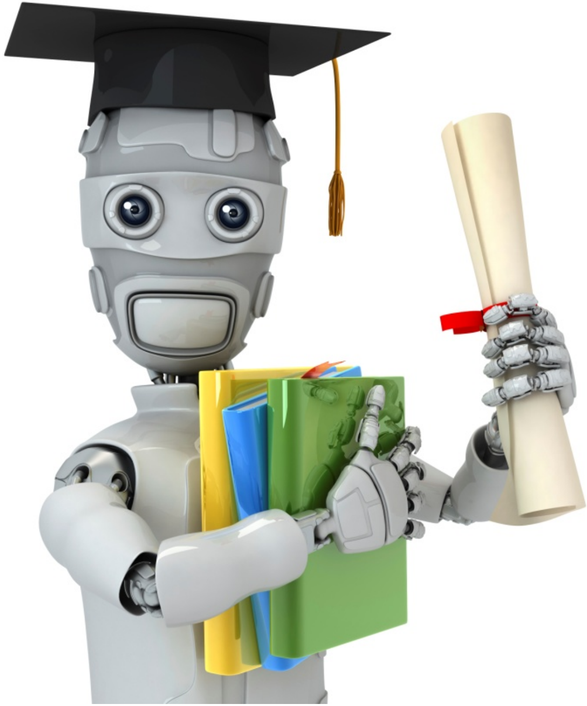


2. Character segmentation



3. Character classification





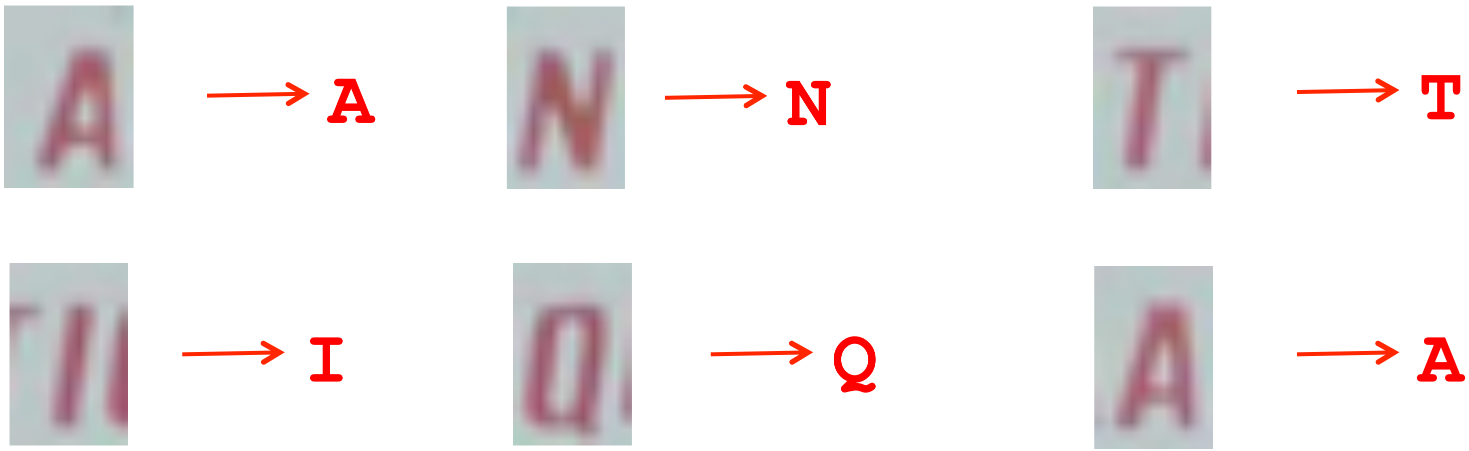
Machine Learning

Application example: Photo OCR

Getting lots of data: Artificial data synthesis

I've seen over and over that one of the most reliable ways to get a high performance machine learning system is to take a low bias learning algorithm and to train it on a massive training set.

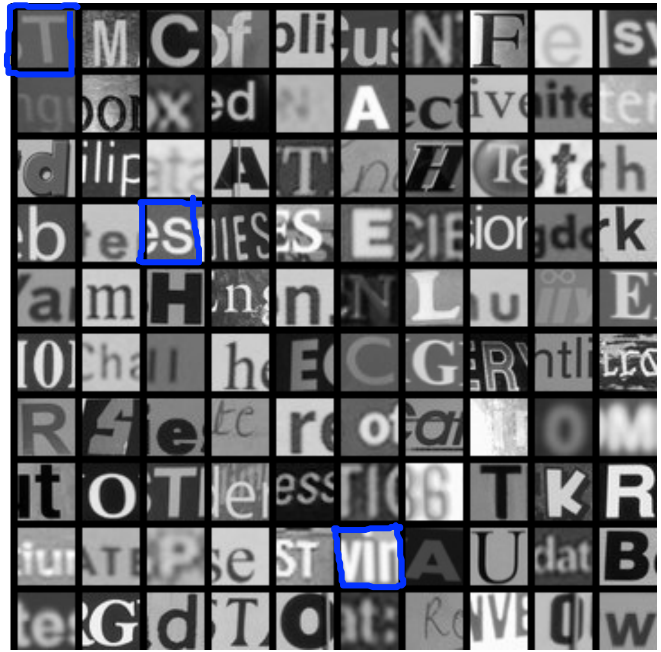
Character recognition



(下頁)
我们怎样获得一个更大的训练集呢
现代计算机通常都有一个 很大的字体库
如果你使用一个文字处理软件
你就会有所有这些字体
或者别的很多很多字体

(下頁)如果你想要获得 更多的训练样本
其中一种方法是你可以
采集同一个字符的不同种字体
然后将这些字符 加上不同的随机背景
如果你这样做的话 你现在就有了一个
关于字母C的训练样本 做完这些工作以后
你就得到了一个合成后的训练集

Artificial data synthesis for photo OCR



Real data

Abcdefg

Abcdefg

Abcdefg

Abcdefg

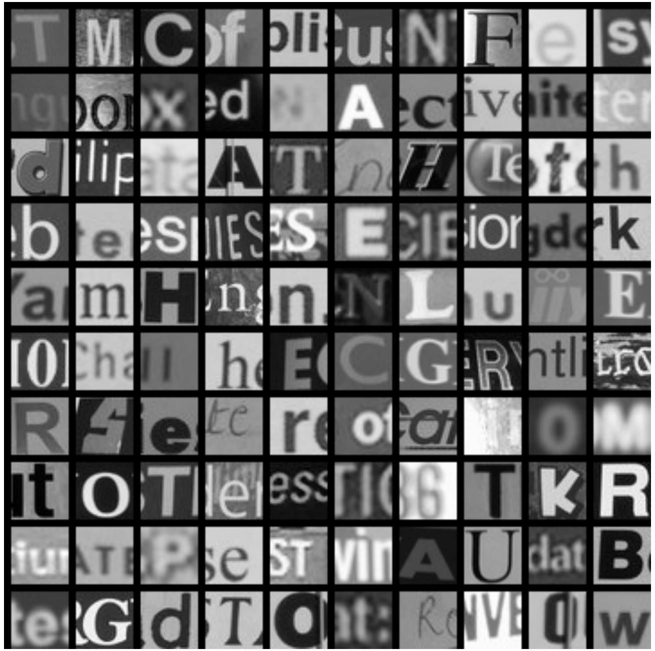
Abcdefg

[Adam Coates and Tao Wang]

右图中每一个图像实际上都是一个合成后的图像(即字體加背景)
然后应用某个模糊操作 模糊的意思是 让图像变形

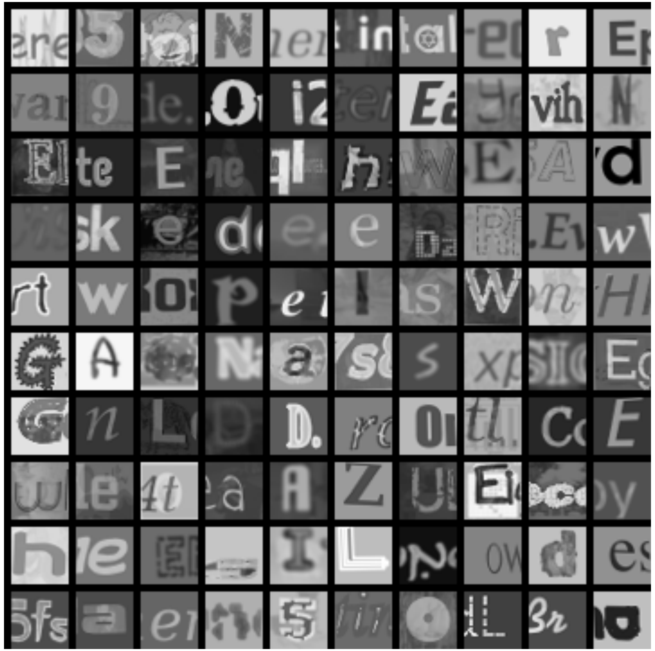
因此通过使用合成的数据 你实际上已经获得了
无限的训练样本 这就是人工数据合成

Artificial data synthesis for photo OCR



Real data

这便是一种人工数据合成的实例
你基本上是在 完全创造新的数据
完全创造新的图像数据



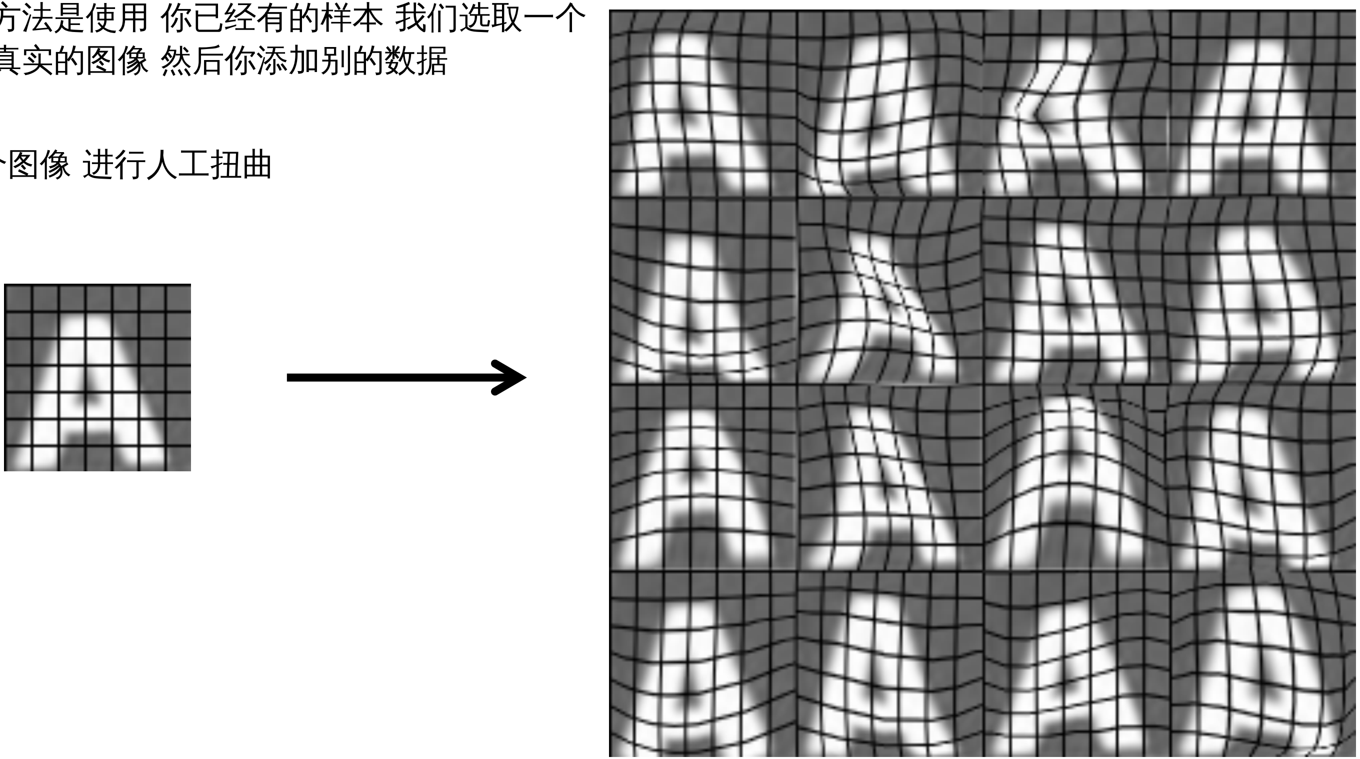
Synthetic data

[Adam Coates and Tao Wang]

Synthesizing data by introducing distortions

人工数据合成的 第二种方法是使用 你已经有的样本 我们选取一个
真实的样本 也许是一个真实的图像 然后你添加别的数据
来扩大你的训练集

取出这个字母A 取出这个图像 进行人工扭曲



[Adam Coates and Tao Wang]

Synthesizing data by introducing distortions: Speech recognition

Original audio: ←

Audio on bad cellphone connection

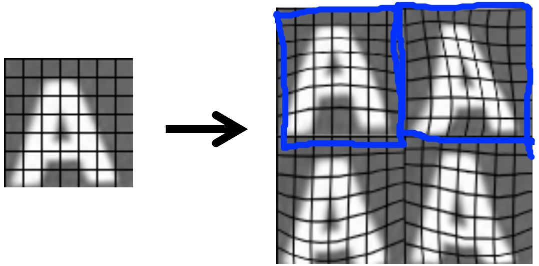
Noisy background: Crowd

Noisy background: Machinery

[www.pdsounds.org]

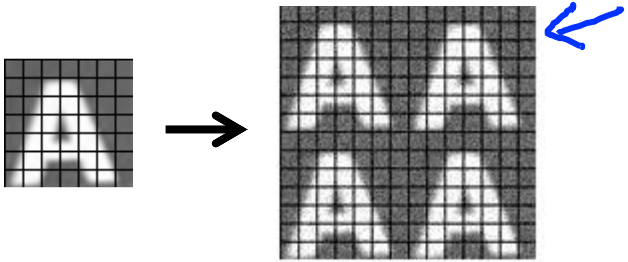
Synthesizing data by introducing distortions

- Distortion introduced should be representation of the type of noise/distortions in the test set.



Audio:
Background noise,
bad cellphone connection

- Usually does not help to add purely random/meaningless noise to your data.



- $x_i =$ intensity (brightness) of pixel i
- $x_i \leftarrow x_i +$ random noise

[Adam Coates and Tao Wang]

Discussion on getting more data

1. Make sure you have a low bias classifier before expending the effort. (Plot learning curves). E.g. keep increasing the number of features/number of hidden units in neural network until you have a low bias classifier.
2. "How much work would it be to get <u>10x</u> as much data as we currently have?"
   - Artificial data synthesis
   - Collect/label it yourself → #hours?
   - "Crowd source" (E.g. Amazon Mechanical Turk)

m = 1,000

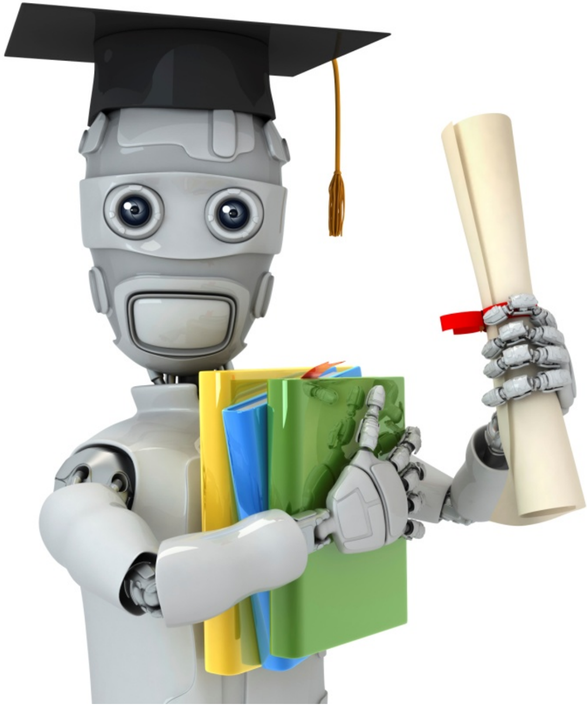
→ 10 secs / example ←

m = 10,000

现在已经有一些网站 或者一些服务机构 能让你通过网络 雇一些人替你完成 标记大量训练数据的工作 通常都很廉价

Discussion on getting more data

1. Make sure you have a low bias classifier before expending the effort. (Plot learning curves). E.g. keep increasing the number of features/number of hidden units in neural network until you have a low bias classifier.
2. "How much work would it be to get 10x as much data as we currently have?"
   - Artificial data synthesis
   - Collect/label it yourself
   - "Crowd source" (E.g. Amazon Mechanical Turk)

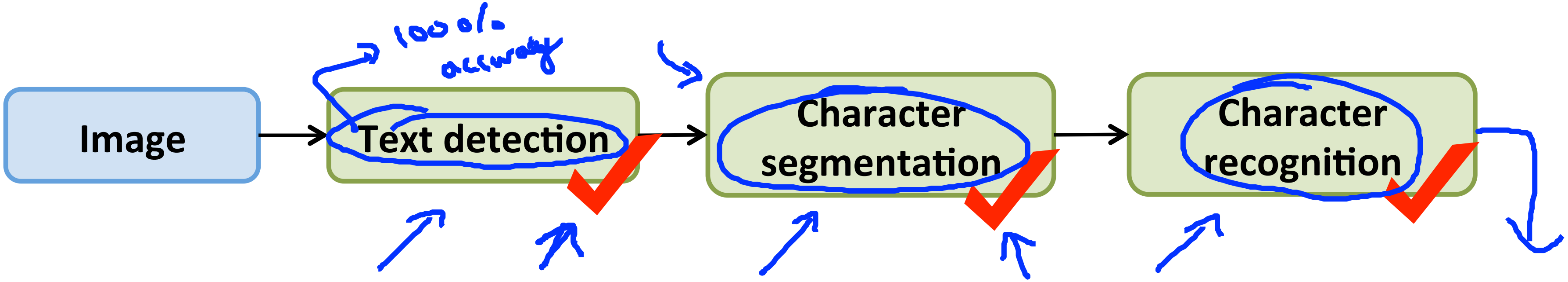


Machine Learning

Application example: Photo OCR

Ceiling analysis: What part of the pipeline to work on next

Estimating the errors due to each component (ceiling analysis)



What part of the pipeline should you spend the most time trying to improve?

Component	Accuracy
Overall system	72%
Text detection	89%
Character segmentation	90%
Character recognition	100%

↓ 17%

↓ 1%

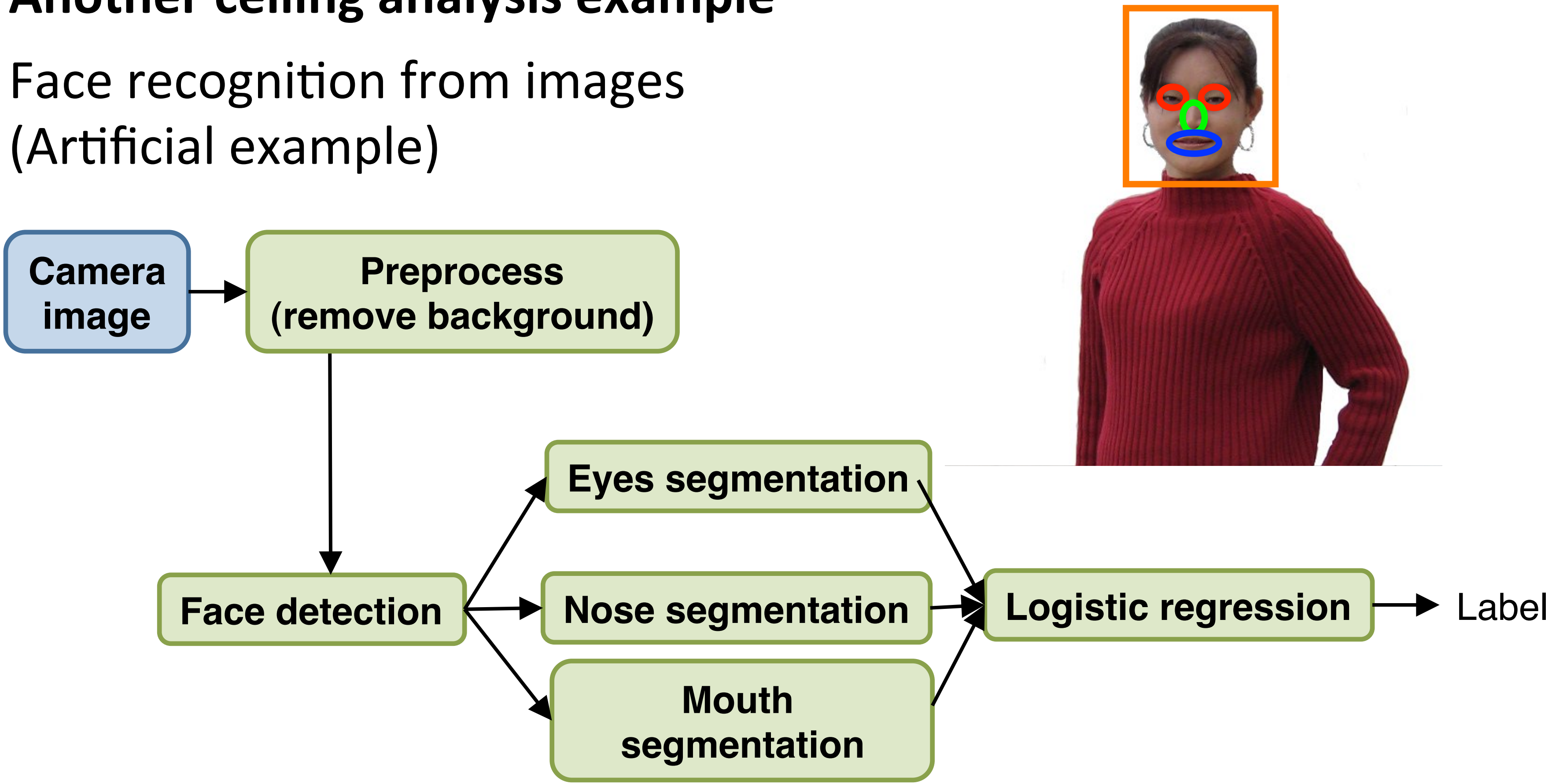
↓ 10%

进行上限分析的 一个好处是 我们现在就知道了 如果对每一个模块进行改善 它们各自的上升空间是多大

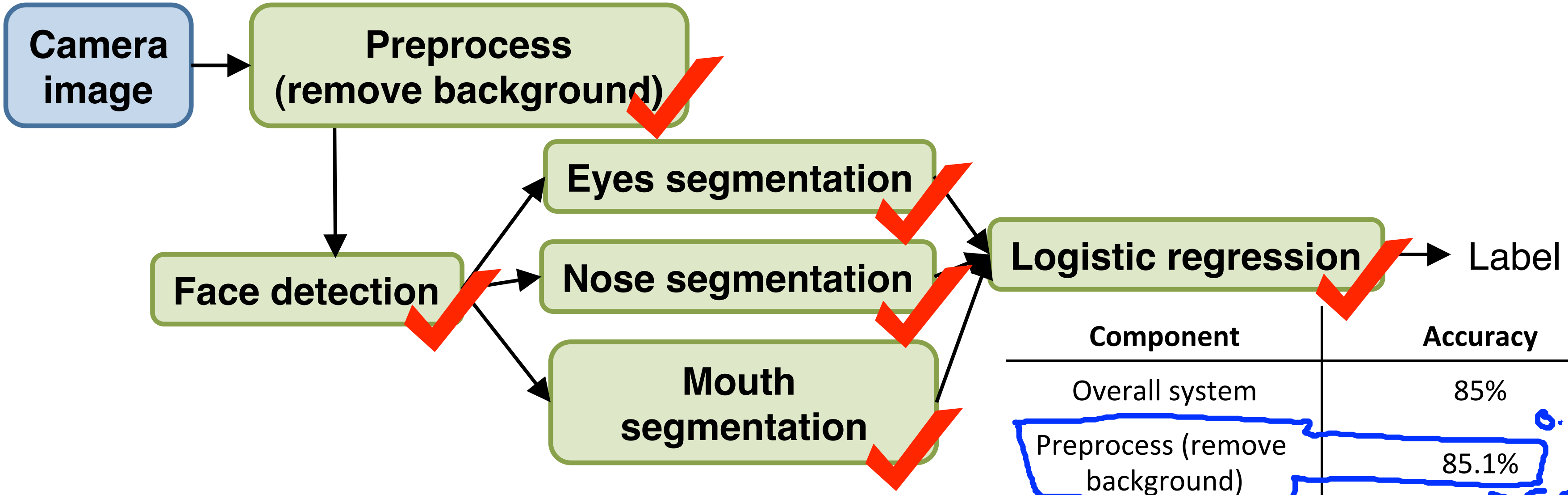
来为你决定 该把劲儿往哪儿使 该提高哪个模块的效果

Another ceiling analysis example

Face recognition from images
(Artificial example)



Another ceiling analysis example



Component	Accuracy	
Overall system	85%	
Preprocess (remove background)	85.1%	0.1%
→ Face detection	91%	→ 5.9%
→ Eyes segmentation	95%	→ 4%
Nose segmentation	96%	1%
Mouth segmentation	97%	1%
→ Logistic regression	100%	→ 3%

Andrew Ng