18. Application Example – Photo OCR:

The Photo OCR Problem

Machine Learning Pipeline: A system with many stages/components, several of which may use machine learning.

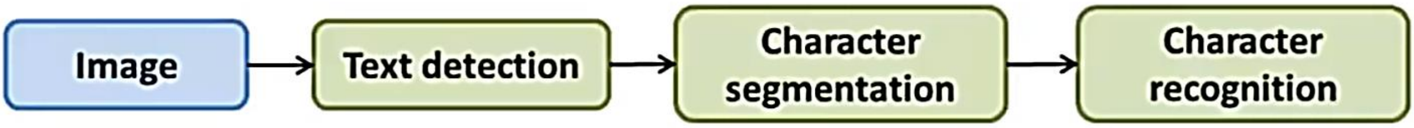
Photo OCR pipeline

1. Text detection

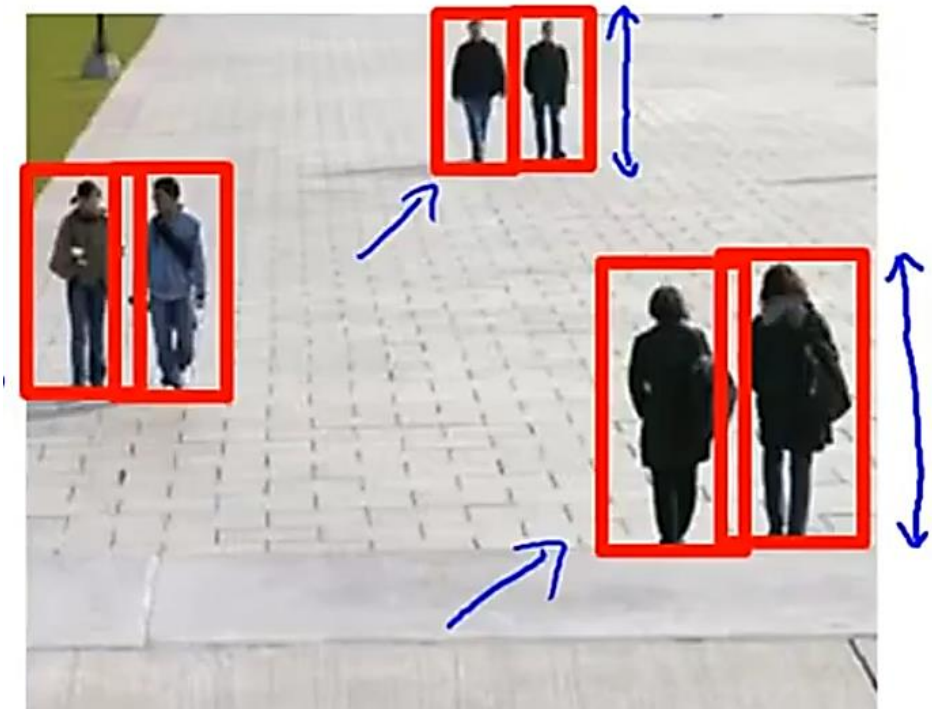
2. Character segmentation

3. Character classification

Photo OCR pipeline



Sliding Window: PEDESTRIAN DETECTION



Supervised learning for pedestrian detection

$x =$ pixels in 82x36 image patches

Positive examples $(y = 1)$

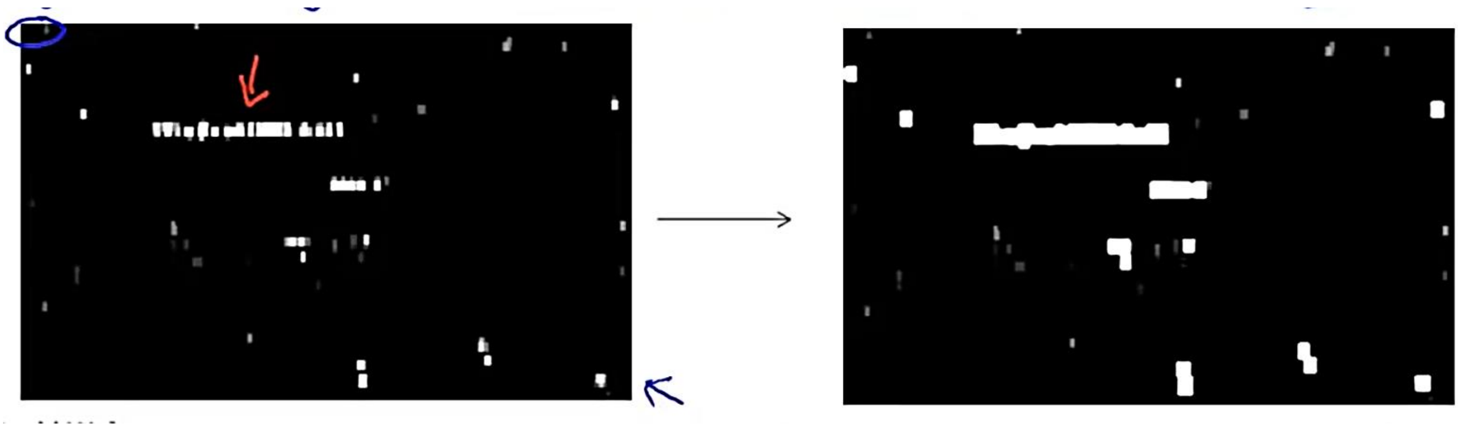
Negative examples $(y = 0)$

To detect the pedestrians, **small frames** are allowed to scan the whole image.

It is tried with many frames of different sizes and all the captured images are resized to a particular size and then that image is sent to Neural network to determine if there is a pedestrian or not

SLIDING WINDOW DETECTION

For text detection example:

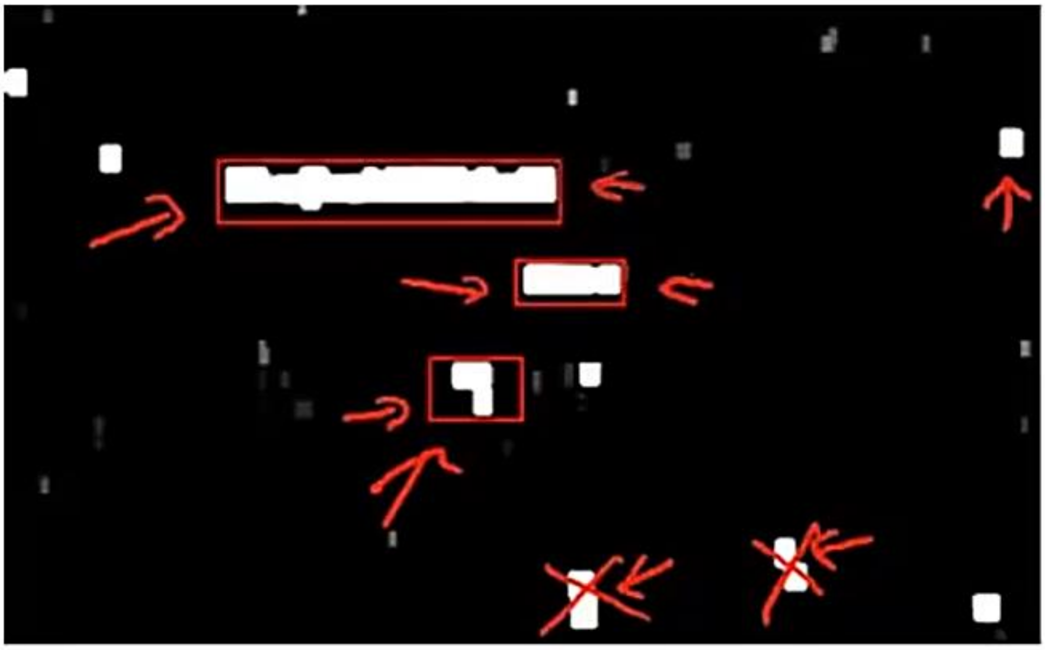


The white regions show where the text is detected

The grey regions show where there is a probability of text. The algo has lower confidence in those parts

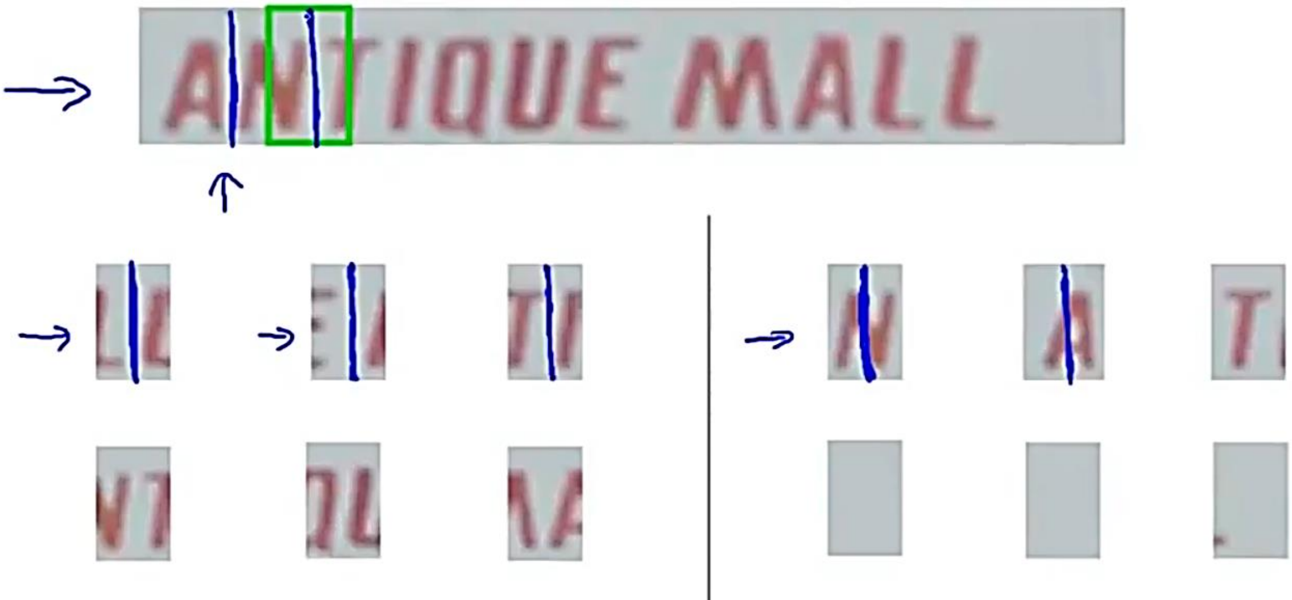
- In expansion (on the right), we ask if a pixel is in within 5 pixels of a white pixel, then that pixel is also made white pixel

Next we filter to only those white boxes, where the aspect ratio is likely to be suitable for text



We now cut out these regions from the image and use them in later stages of detection

1D Sliding window for character segmentation



Positive examples $(y = 1)$

Negative examples $(y = 0)$

Artificial Data Synthesis: to amplify the training set:

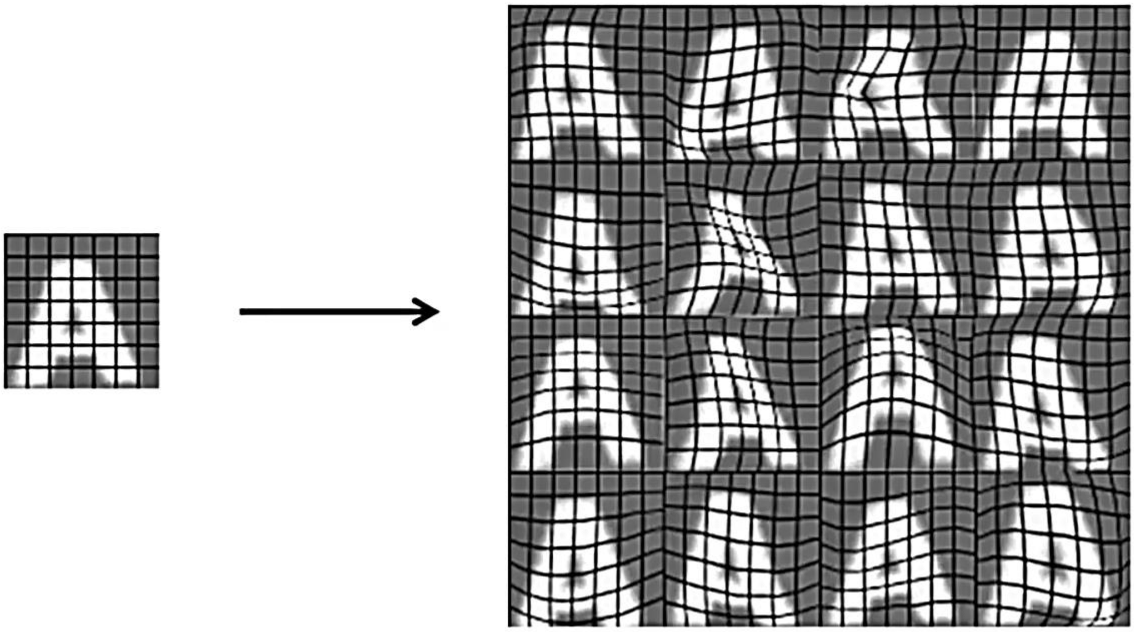
Artificial data synthesis for photo OCR

Real data

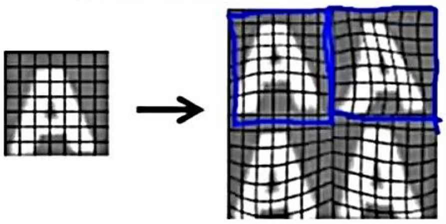
Synthetic data

Synthetic data is prepared by using different fonts and putting letters on different backgrounds

Synthesizing data by introducing distortions

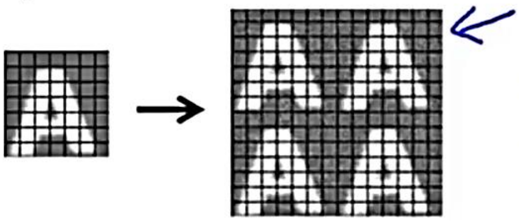


- Distortion introduced should be representation of the type of noise/distortions in the test set.



- Audio:
  Background noise,
  bad cellphone connection

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



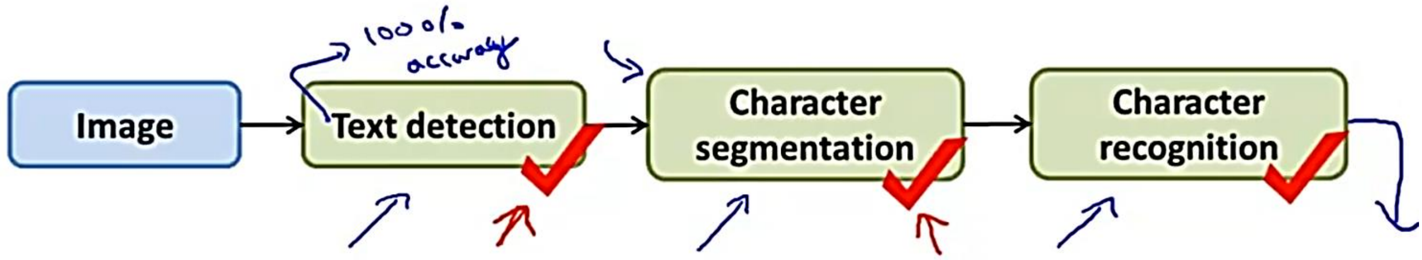
- $x_i =$ intensity (brightness) of pixel i
- $x_i \leftarrow x_i +$ random noise

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)

Ceiling Analysis: What part of the pipeline to work on next?

Estimating the errors due to each component (ceiling analysis)

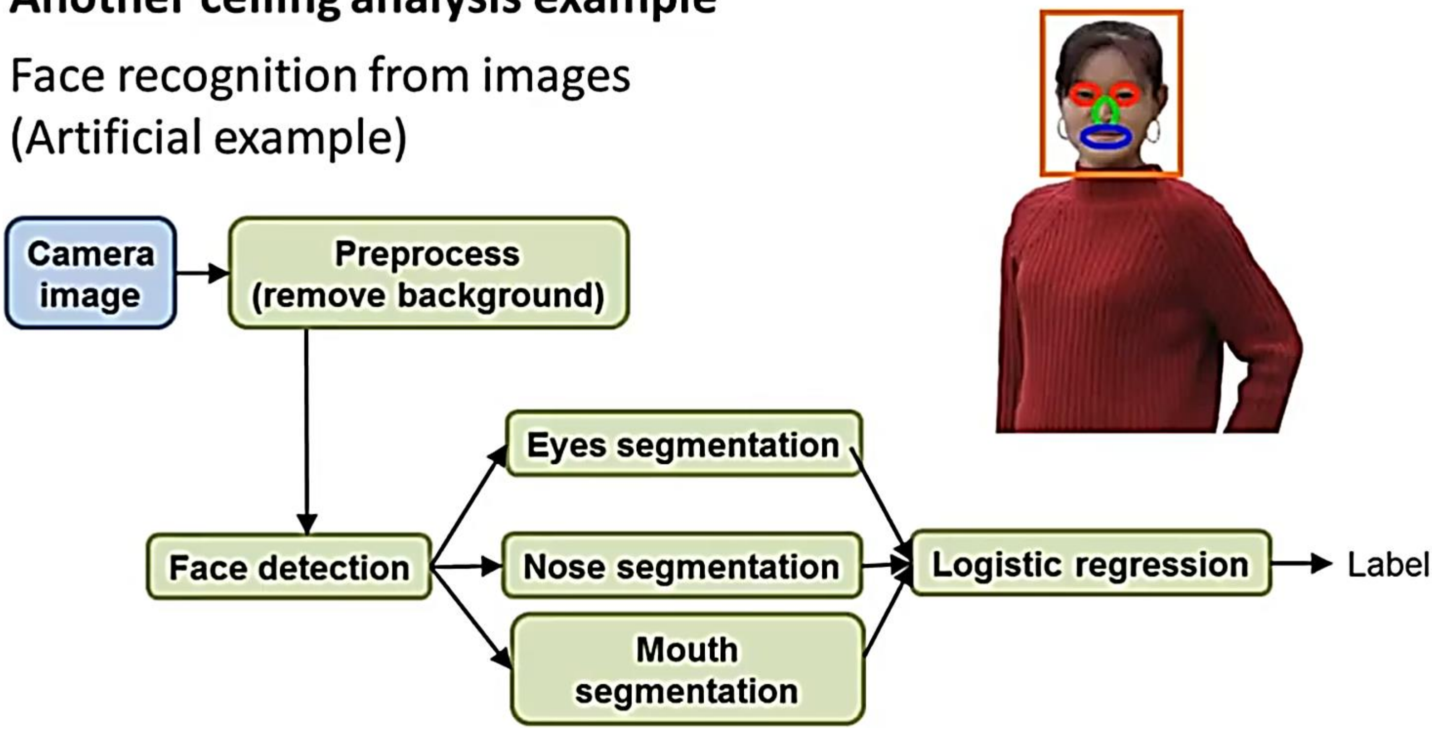


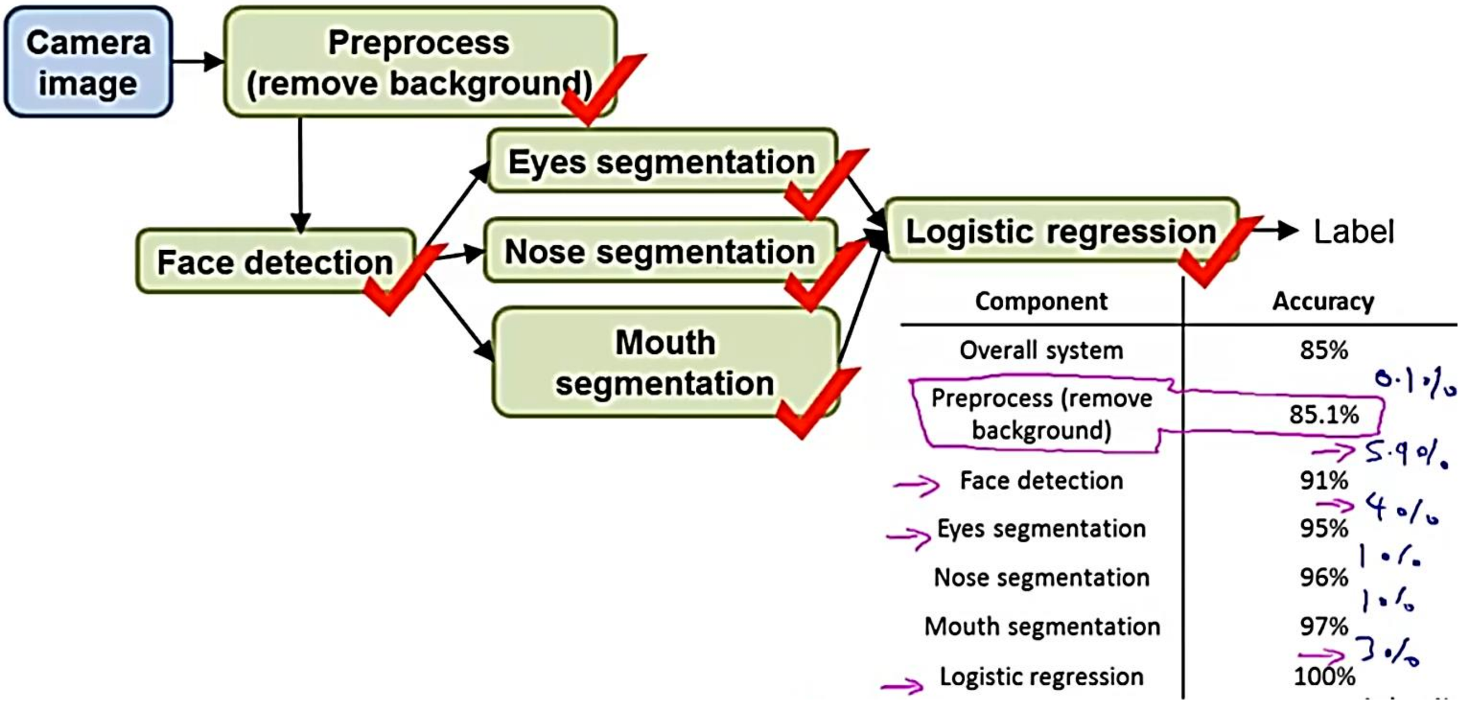
What part of the pipeline should you spend the most time trying to improve?

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

Another ceiling analysis example

Face recognition from images
(Artificial example)





3. What are the benefits of performing a ceiling analysis? Check all that apply.

- [x] It can help indicate that certain components of a system might not be worth a significant amount of work improving, because even if it had perfect performance its impact on the overall system may be small.
- [ ] If we have a low-performing component, the ceiling analysis can tell us if that component has a high bias problem or a high variance problem.
- [x] It helps us decide on allocation of resources in terms of which component in a machine learning pipeline to spend more effort on.
- [ ] It is a way of providing additional training data to the algorithm.

Summary: Main topics

$(x^{(i)}, y^{(i)})$

- Supervised Learning
  - Linear regression, logistic regression, neural networks, SVMs

- Unsupervised Learning $x^{(i)}$
  - K-means, PCA, Anomaly detection

- Special applications/special topics
  - Recommender systems, large scale machine learning.

- Advice on building a machine learning system
  - Bias/variance, regularization; deciding what to work on next: evaluation of learning algorithms, learning curves, error analysis, ceiling analysis.

Machine Learning by Stanford University on Coursera. Certificate earned at Friday, April 12, 2019 9:43 AM GMT

How should I list my course on my resume/C.V.?

We'd recommend "Machine Learning by Stanford University on Coursera. Certificate earned at Friday, April 12, 2019 9:43 AM GMT".

coursera.org/verify/4VW5AT4B38TZ