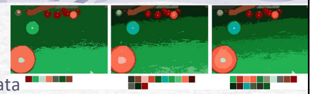
Learning & Evaluation

- Data / Experience
  - Unsupervised
  - Supervised
- Performance Evaluation
  - Ground Truth
  - Metrics
  - Training vs. Validation

Machine learning

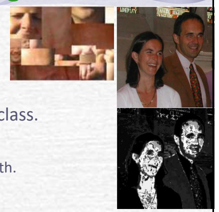
- An algorithm that can learn from data.
  - Data / Experience
  - Some task
    - Often classification
  - Performance improved by data
    - Measure accuracy or error rate
- Data / Experience
  - Unsupervised / Unlabelled.
  - Supervised / Labelled.

Unsupervised machine learning



- k-means.
  - Data – Unlabelled image data
  - Task – Grouping pixels to reduce the number of colours
  - Performance – Very application/image dependent
    - Measure accuracy or error rate
- Gaussian Mixture Model
  - Data – Unlabelled video data.
  - Task – Identification of moving objects.
  - Performance
    - Accuracy based on hand labelled group truth.

Supervised machine learning



- Back-projection.
  - Data – Labelled sample pixel values
  - Task – Classifying pixels as a particular class.
  - Performance
    - Accuracy based on hand labelled group truth.
- Statistical Pattern Recognition / Support Vector Machines
  - Data – Labelled samples (feature vectors).
  - Task – Recognition of objects/shapes.
  - Performance
    - Accuracy based on hand labelled group truth.

Evaluating Computer Vision Performance

- Does it work in real time?
  - What is real time?
  - Does it matter?
- What is the correct answer?
  - Ground truth.
  - How do we generate ground truth?
- How do we assess how well we have done?
  - Success/failure metrics?
  - Overlap?
  - What about time?

Performance – Ground Truth

- What should any computer vision system achieve?
  - The correct answer is easy for some tasks. e.g.
    - Is this picture me?
    - Determine the state of the traffic lights.
  - For some tasks it is harder.
    - Classify the object in this image.
    - Count the people.
    - Drive from A to B.

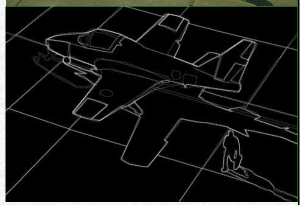



Performance – Ground Truth

- What should any computer vision technique achieve?
  - The correct answer is easy for some techniques. e.g.
    - Find skin pixels
  - For some techniques it is harder.
    - Find the edges.
  - The usefulness of a technique
    - depends on the application…


Ground Truth – Example

- How should an image be segmented?
  - Significant edges/regions?
- Impossible to get agreement?
- Berkeley segmentation dataset
  - 1,000 images
  - 12,000 segmentations (colour & grayscale)
  - Hand-labelled (30 subjects)
- Is this a good way to evaluate a segmentation technique?

From the public "Berkeley Segmentation dataset"

Ground Truth – Bounding boxes

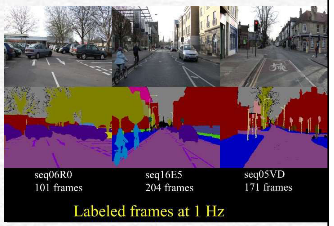
- Boxes incorporate parts of other objects/background.
- Labour intensive for video.
- CAVIAR:
  - Shopping centre (Lisbon)
  - Bounding boxes for all people.
  - Some have head, hands, and feet annotated.
  - Manually annotated.

From the public dataset CAVIAR

9

Ground Truth – Pixel based

- Better validity.
- Very expensive to create.
- CAMVID:
  - 32 classes
  - 10 minutes @30/15Hz
  - 700 frames
  - Manually annotated.
  - 2nd person inspection.



From the public dataset CAMVID

10

Performance – Simple Metrics

- Compute
  - TP
  - TN
  - FP
  - FN
- and then compute…

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{Total\ Samples}$$

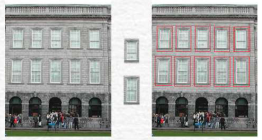
$$Specifcity = \frac{TN}{FP + TN}$$

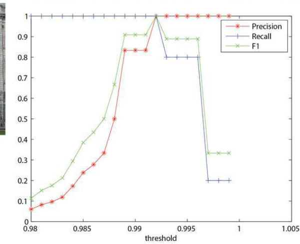
$$F_\beta = (1 + \beta^2).\frac{Precision.Recall}{(\beta^2.Precision) + Recall}$$

11

Performance – Tuning

- We can tune performance by altering parameters
  - For example altering the threshold used for accepting a match in a template matching example

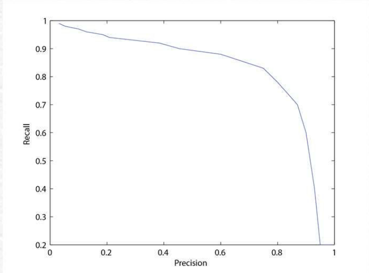




12

Performance – Precision-Recall Curves

- Alternative (common) visualisation
- Select the parameters which result in the point on the PR curve which is closest to P=1.0, R=1.0



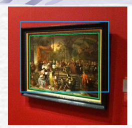
13

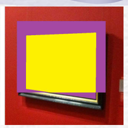
Performance – Overlap & mAP

- Intersection over Union (IoU)

$$IoU(Detected, GroundTruth) = \frac{Detected \cap GroundTruth}{Detected \cup GroundTruth}$$

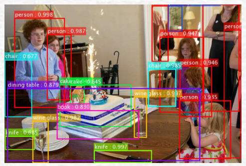
  - Values of IoU > 0.5 are often considered good matches.





- When considering multiple classes…
  - Mean Average Precision (mAP)

$$mAP = \frac{1}{|classes|} \sum_{c \in classes} \frac{\#TP(c)}{\#TP(c) + \#FP(c)}$$



ResNet result on COCO dataset (2015)

14

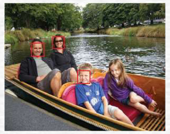
Performance – Confusion Matrix

- It is helpful to know how things are misclassified
  - Actual class
  - Determined class

Result \ Ground Truth	Walking	Running	Standing	Dancing	Hopping
Walking	0.9	0.0	0.0	0.0	0.0
Running	0.0	0.7	0.0	0.1	0.1
Standing	0.1	0.0	1.0	0.0	0.0
Dancing	0.0	0.2	0.0	0.4	0.0
Hopping	0.0	0.1	0.0	0.5	0.9

- If considering a binary classification
  - Confusion matrix shows TP, FP, FN, TN

	A	Not A
A	TP	FP
Not A	FN	TN



	Face	Not Face
Face	3	0
Not Face	1	65532

15

Training & Validation

- Typically dataset are divided into two parts:
  - Training set.
  - Validation/Testing set.
- This can be
  - Static
  - Random
  - Leave p-out cross validation
    - Leave p samples out – for testing. Train on the rest.
    - Do multiple rounds of training and testing using different partitions.
    - Average results.
    - Intended to determine how the results will generalise.

16

Outcome

- So what does testing tell me?
  - How well the system/technique works on the test data (images/videos)…
  - Actually, how close the system/technique gets to the manually created ground truth…
- How representative is the test set?
  - What conditions matter?
  - Can I ever be sure that a system will work in general?

17

Summary

- Learning (Data – Task – Performance)
  - Unsupervised
  - Supervised
- Ground Truth
  - Expensive to obtain.
  - Often unclear what is the right answer.
- Metrics
  - Precision, Recall, Accuracy
  - Tuning
  - IoU, mAP
- Training vs. Validation
  - Leave p-out cross validation

18