Visual Object Recognition

Lecture 8: Performance Evaluation

Per-Erik Forssén, docent
Computer Vision Laboratory
Department of Electrical Engineering
Linköping University

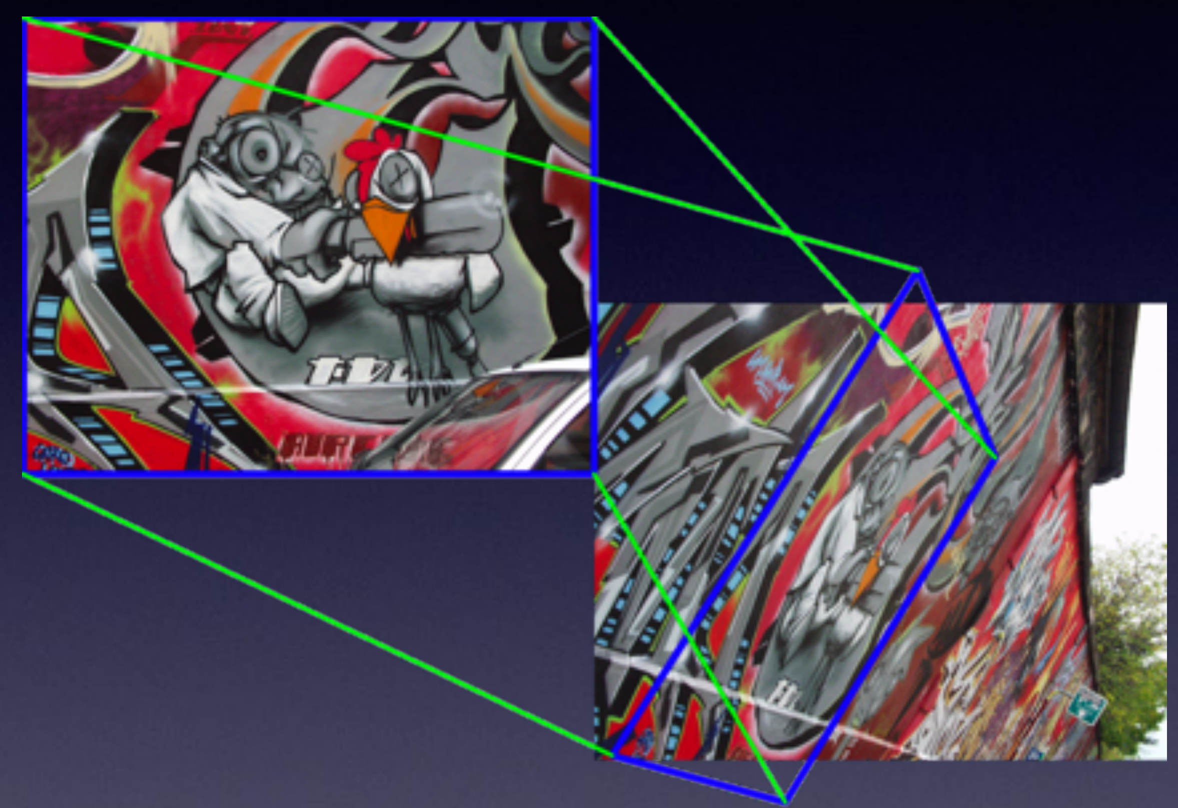
Lecture 8: Performance Evaluation

- Detector: **Repeatability Tests**
- Descriptor matching: **Inlier frequency curve**
- Classifier: **ROC and Precision-Recall curves**
- Discussion of exam and evaluation

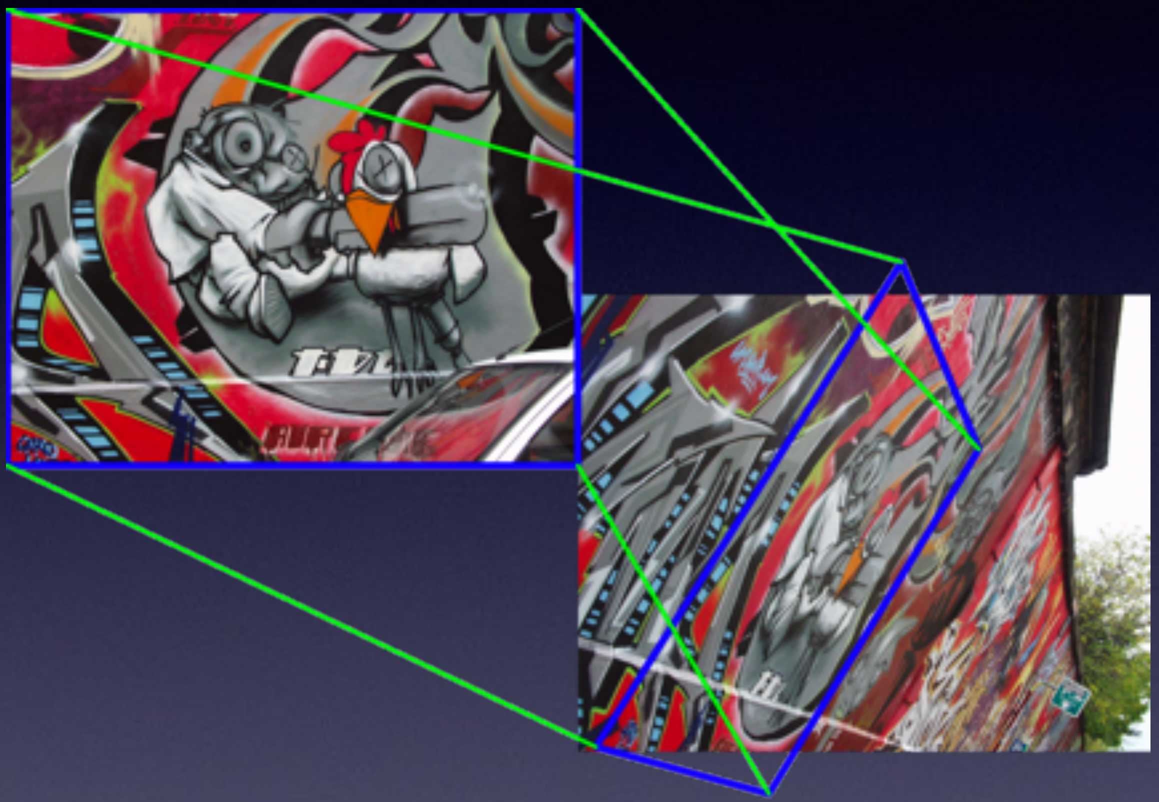
Repeatability Tests

- Used for evaluating feature detectors.
  E.g. Mikolajczyk et al. IJCV’06.

- Known geometric transformation between two views can be used to check if the same region is detected in two images.



Repeatability Tests



- Example: Homography

  A point **x** should be transformed to a point **x'** according to:

  $$\mathbf{x}' = \mathbf{H}\mathbf{x}$$

  In reality we detect regions: $\mathbf{x}^T\mathbf{C}\mathbf{x} \leq 0$

$$\mathbf{C} = \frac{1}{4}\begin{bmatrix} \mathbf{I}^{-1} & -\mathbf{I}^{-1}\mathbf{m} \\ -\mathbf{m}^T\mathbf{I}^{-1} & \mathbf{m}^T\mathbf{I}^{-1}\mathbf{m} - 4 \end{bmatrix}$$

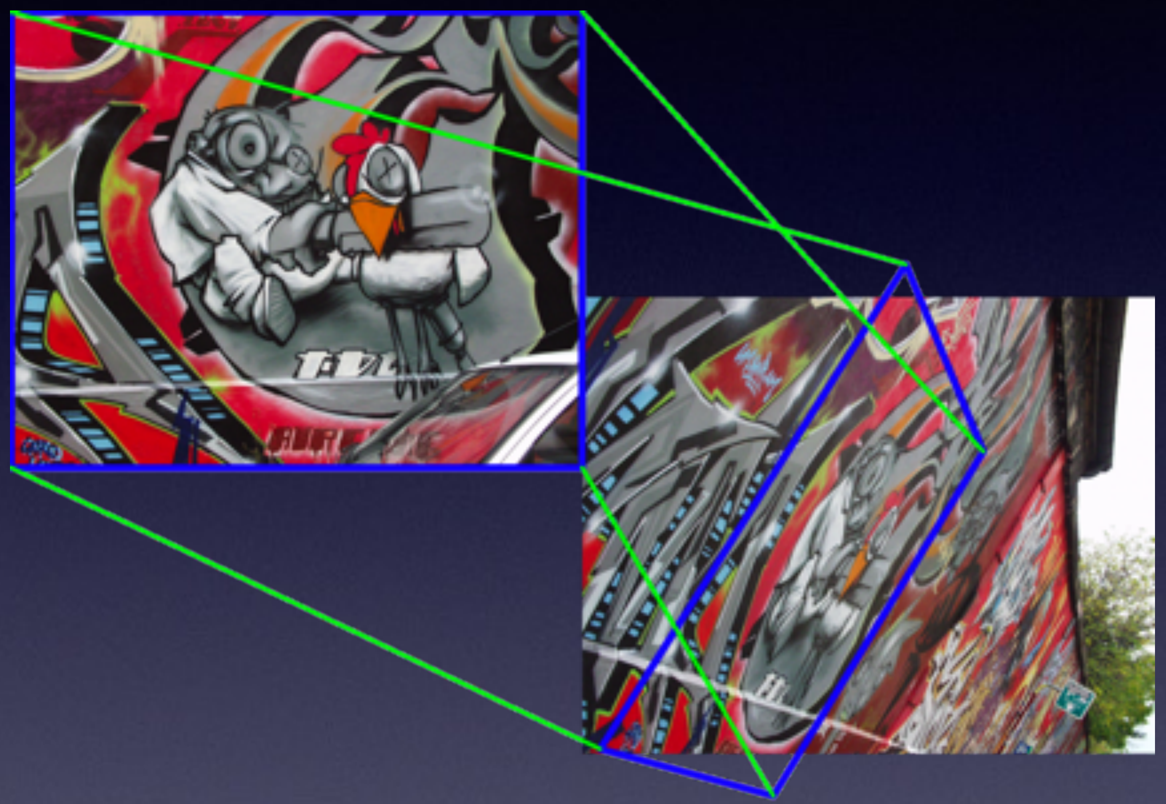
Repeatability Tests

- Example: Homography

  An elliptic region **C**(**m**,**l**) should be transformed to a region **C**'(**m**',**l**') according to:

  $$\mathbf{C}' = \mathbf{H}^{-T}\mathbf{C}\mathbf{H}^{-1}$$

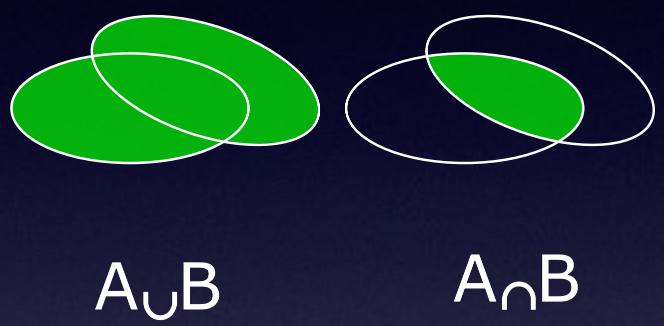
  Can be derived from perimeter equation: $\mathbf{x}^T\mathbf{C}\mathbf{x} = 0$ (transform **x** to **x'** and identify **C'**)



Repeatability Tests

1. Compute overlap error:

$$\epsilon = 1 - \frac{\text{area}(A \cap B)}{\text{area}(A \cup B)}$$

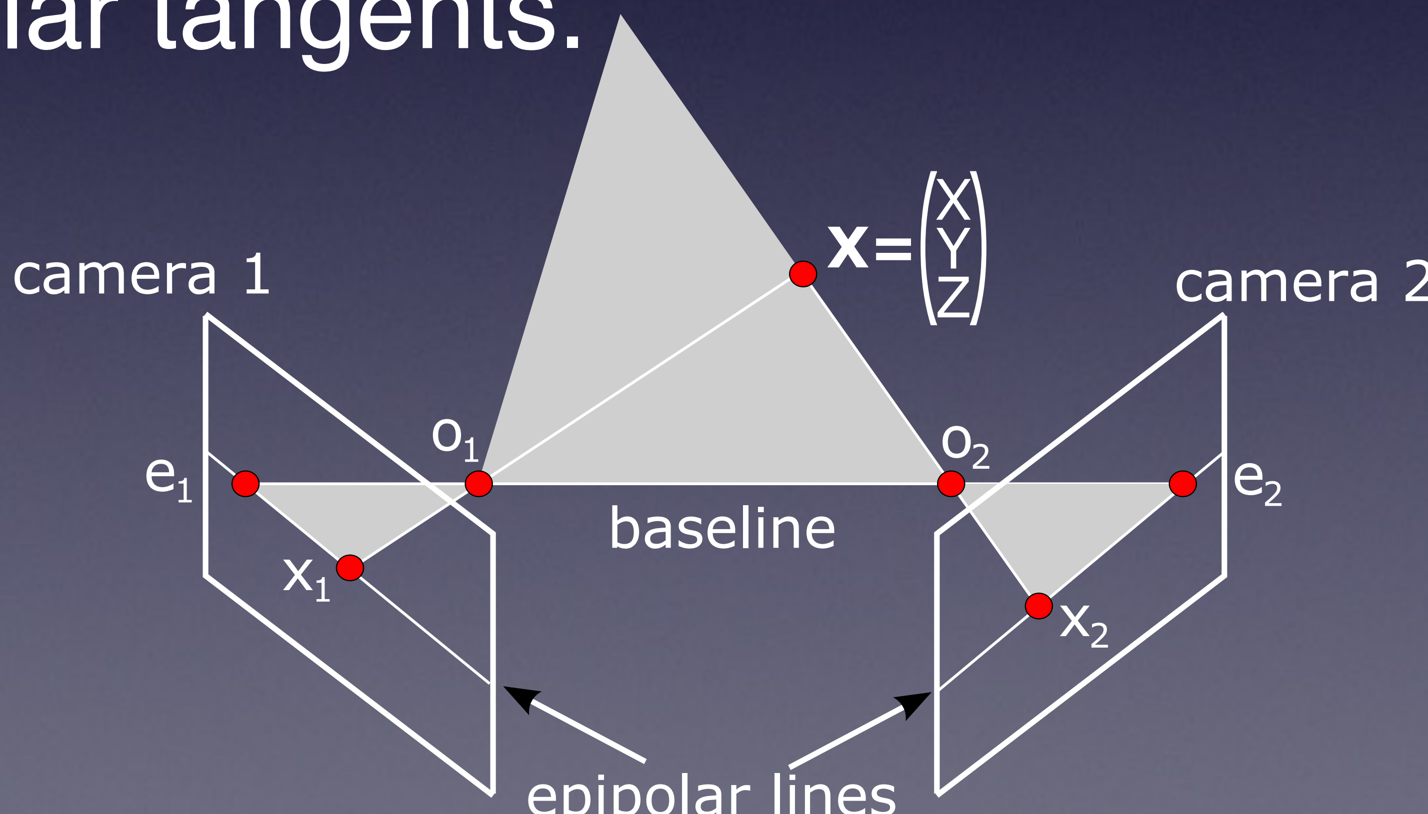


2. Assign 1-to-1 correspondences from image 1 to image 2. (Combinatorial problem if nested regions are detected)

3. repeatability = correspondences (with $\epsilon \leq \texttt{thr}$) divided by #features (in mutually visible region)

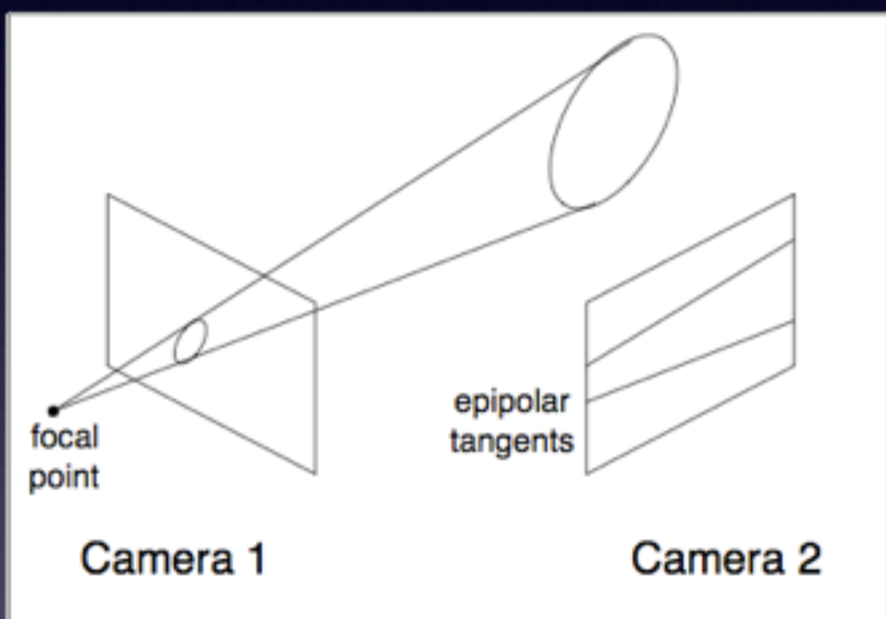
3D Repeatability Tests

- Using generalisation of overlap error to 3D correspondences (Forssén&Lowe ICCV'07)
- Using epipolar geometry, and specifically epipolar tangents.



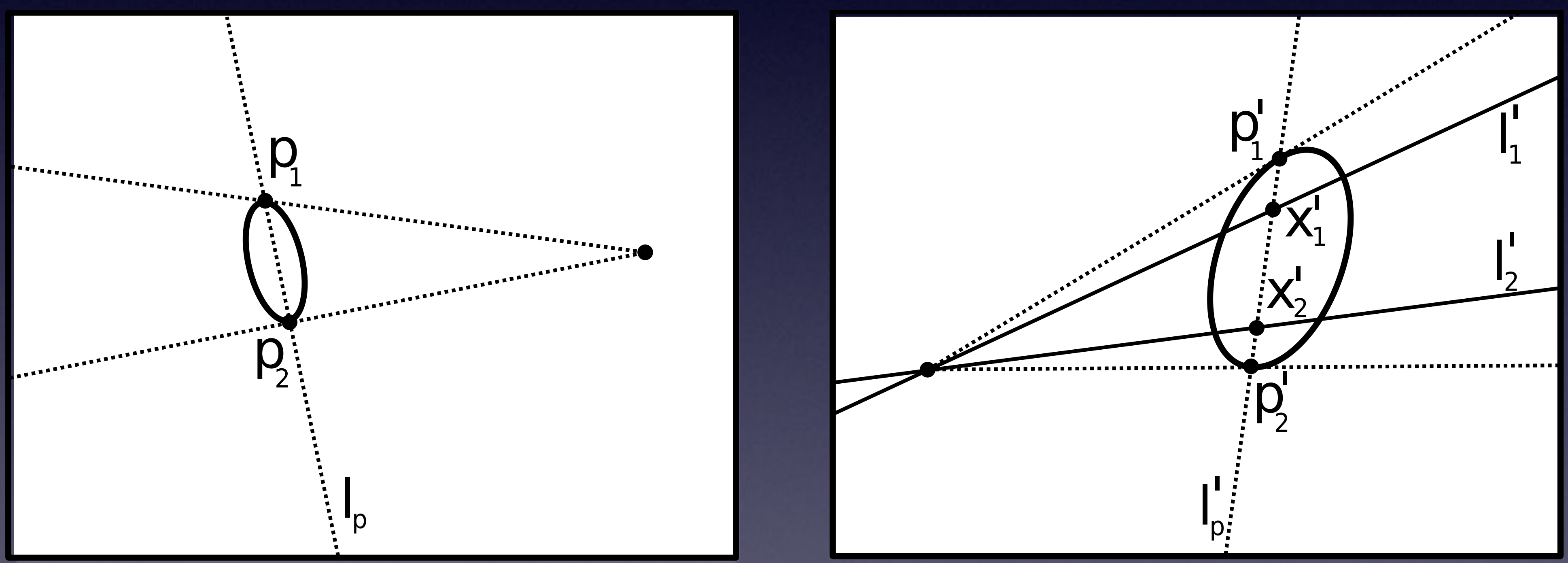
3D Repeatability Tests

- Epipolar tangents



3D Repeatability Tests

- Measure overlap of tangents and projected epipolar tangents.



$$\epsilon = 1 - \frac{\max(0, \min(x_h, p_h) - \max(x_l, p_l))}{\max(x_h, p_h) - \min(x_l, p_l)}$$

Repeatability Tests

- Repeatability measures probability that a feature will be detected again.

$$P(\text{detection}|\text{visibility})$$

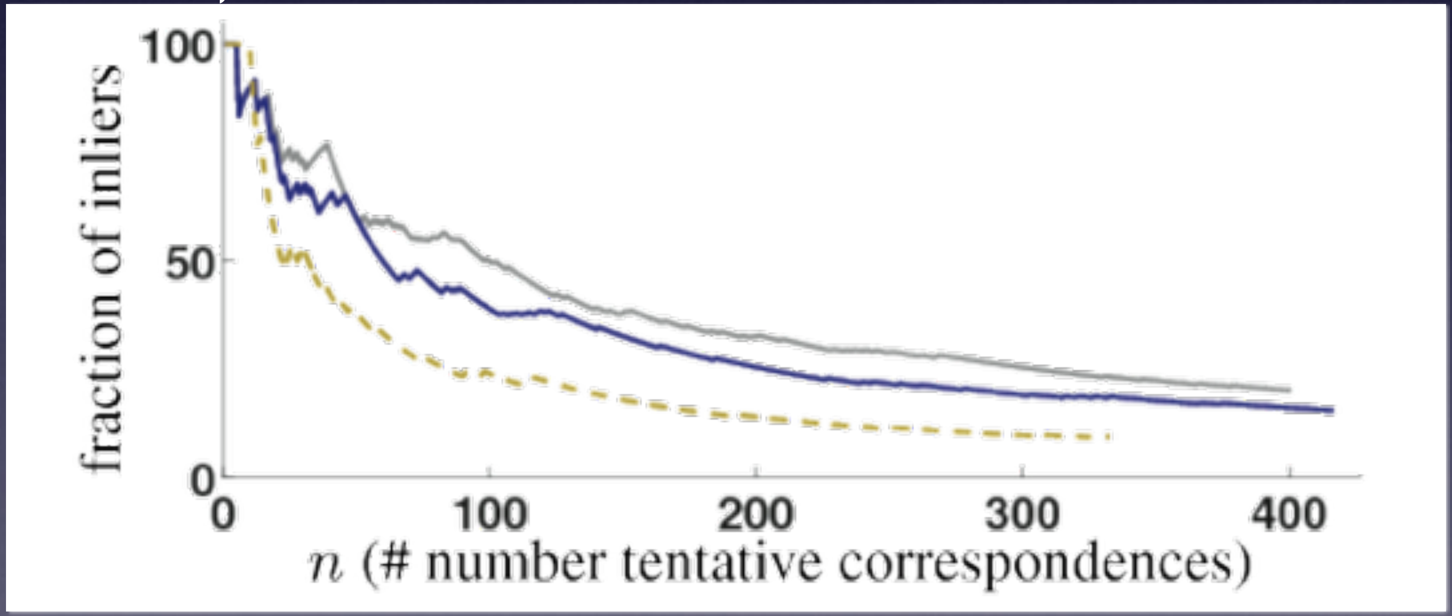
- Repeatability is not useful for non-rigid objects/categories. (As a geometric constraint is used.)

Correspondence Count

- A complementary statistic is to simply count the number of corresponding regions (skip division by number of detected features).
- Better for object recognition:
  If each feature match casts a vote, the probability of a cluster forming by chance is low, so outliers can be tolerated.
- Also: All hypothesis generation(HG)+verification schemes. HG costs only time.

Inlier Frequency Curve

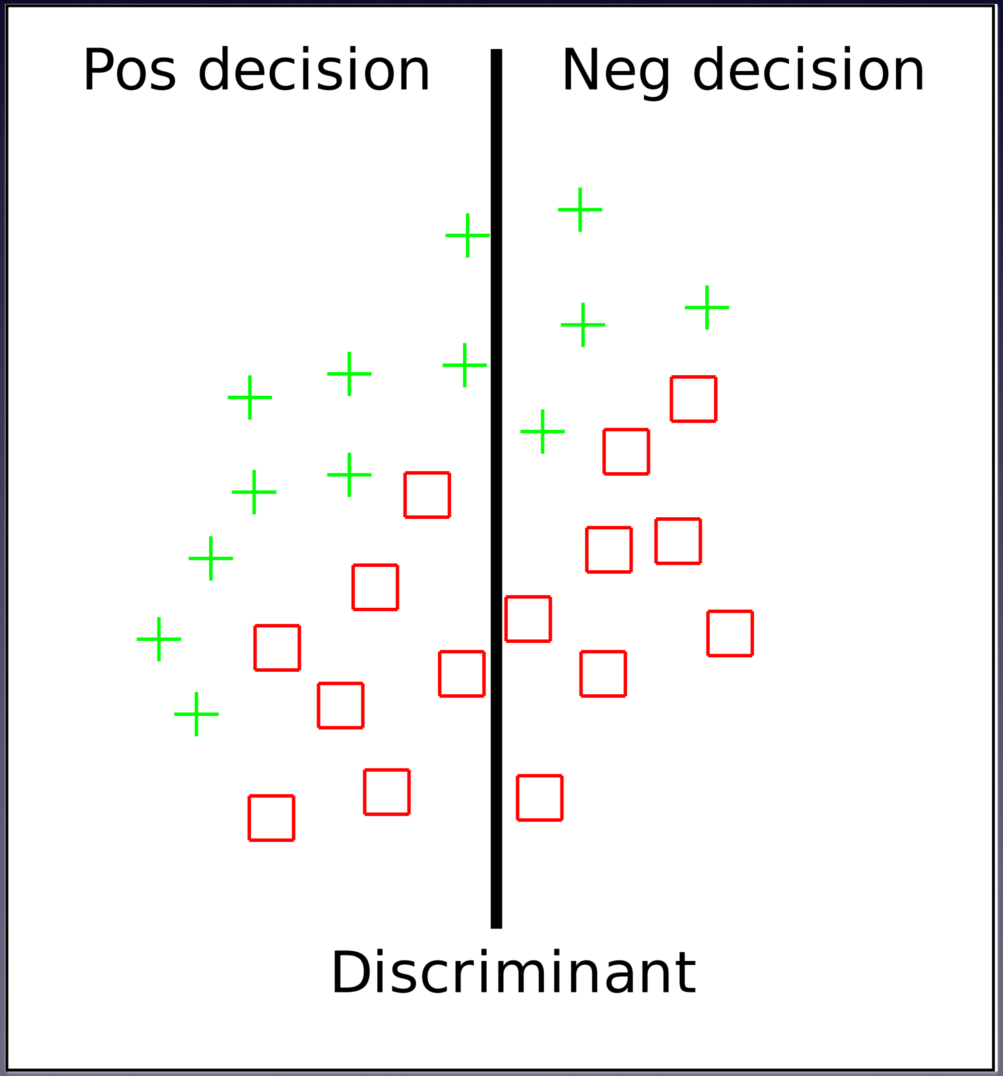
- Descriptor matching generates ordered *tentative correspondences*. When ground-truth is known, these can be evaluated with an *inlier frequency curve*, Chum&Matas, CVPR06.



- Good for RANSAC, and e.g. PROSAC (which uses the ranking).

ROC and PR curves

- Used for evaluating *binary classifiers* across a change of the discriminant.



ROC and PR curves

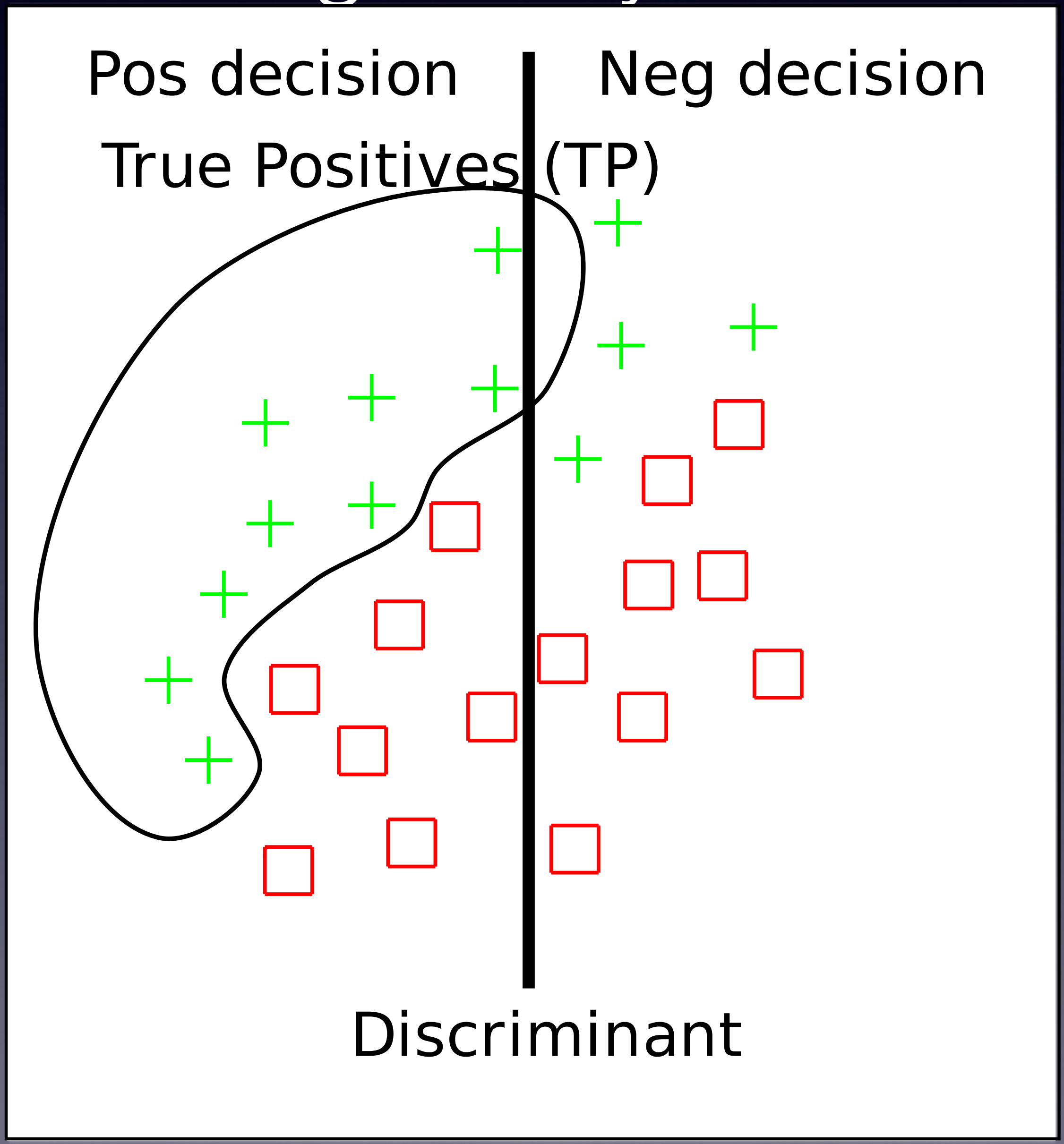
- Used for evaluating *binary classifiers* across a change of the discriminant.
- The optimal discriminant direction is often application independent, but the actual threshold is not.
- With ROC and PR curves, comparison can be done without committing to a specific discriminant.

ROC and PR curves

- Instead of a single performance measure we get a curve.
- Useful if criterion changes over time. E.g.
  1. Few false alarms might be most important.
  2. It might be very important not to miss a positive.
- To adapt, read curve in a different place.

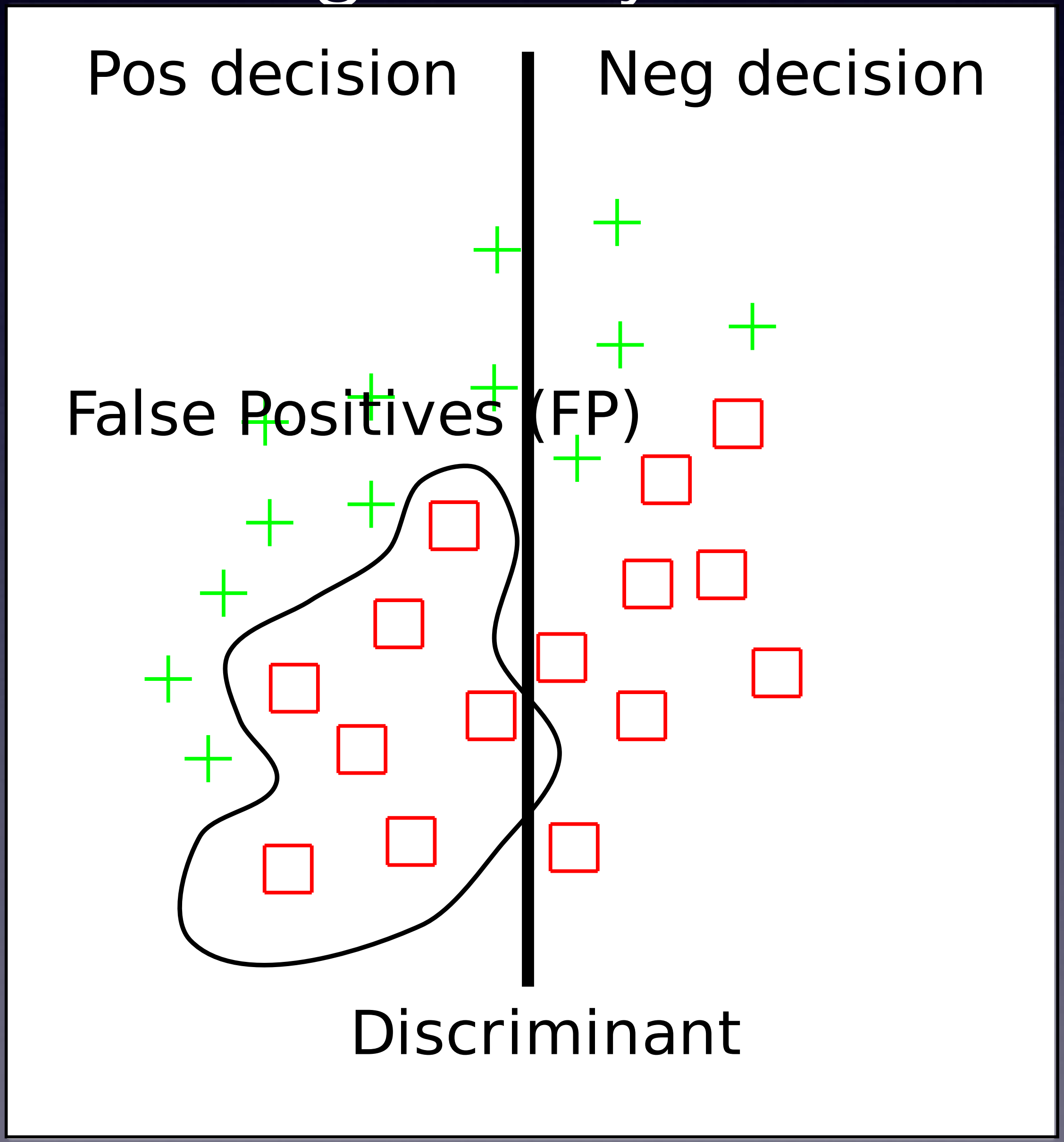
ROC and PR curves

- Used for evaluating binary classifiers.



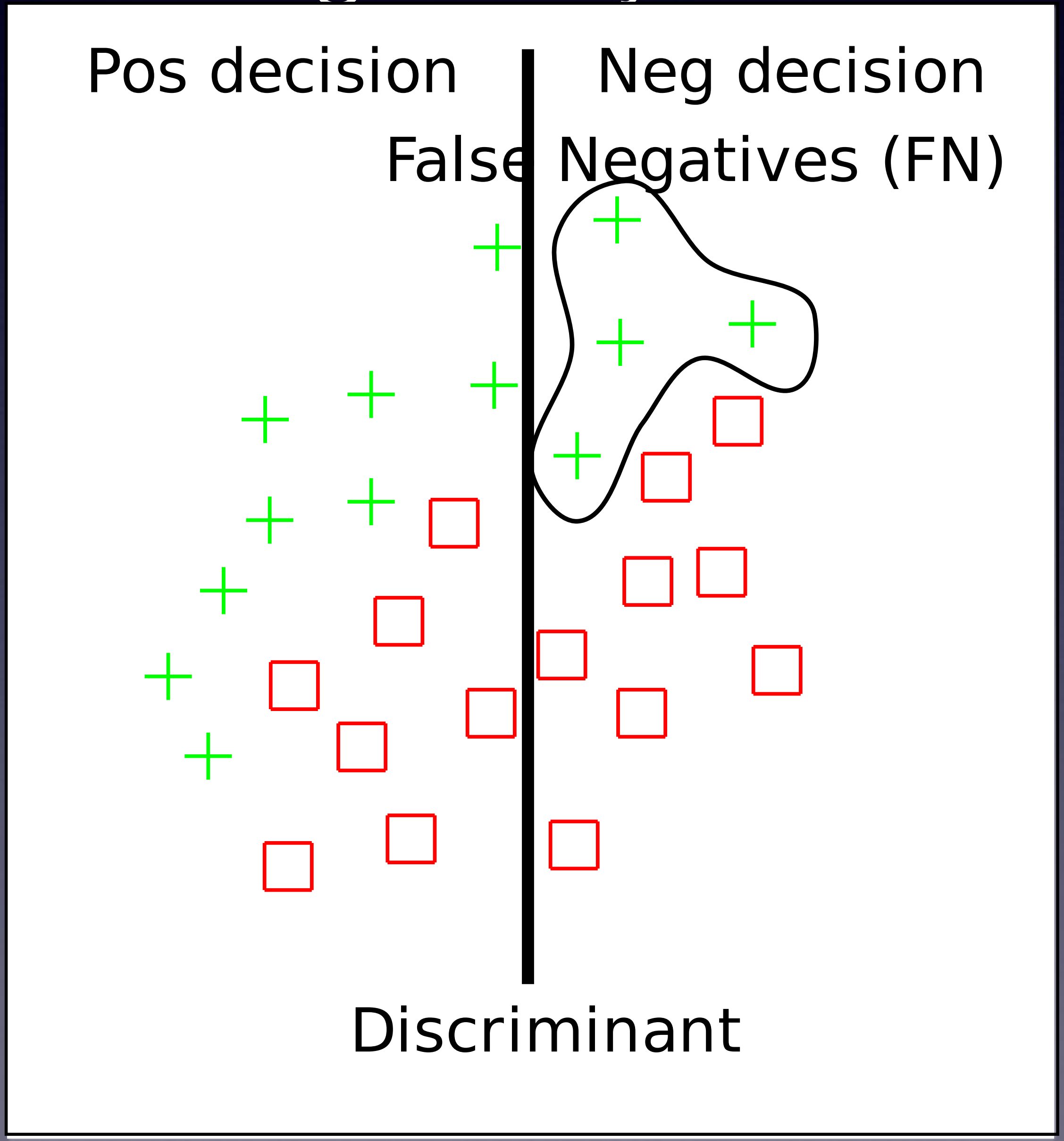
ROC and PR curves

- Used for evaluating binary classifiers.



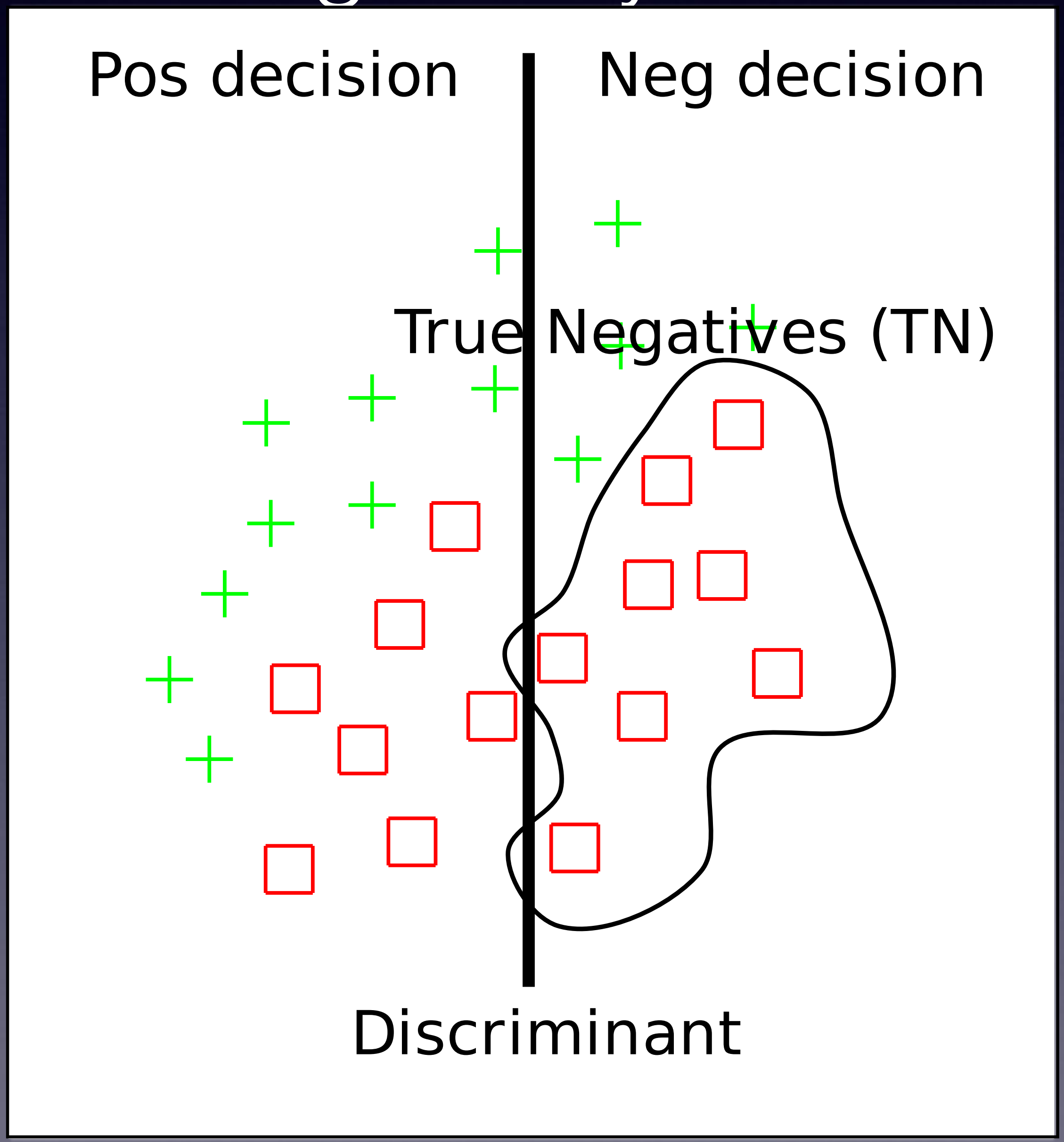
ROC and PR curves

- Used for evaluating binary classifiers.



ROC and PR curves

- Used for evaluating binary classifiers.



ROC curve

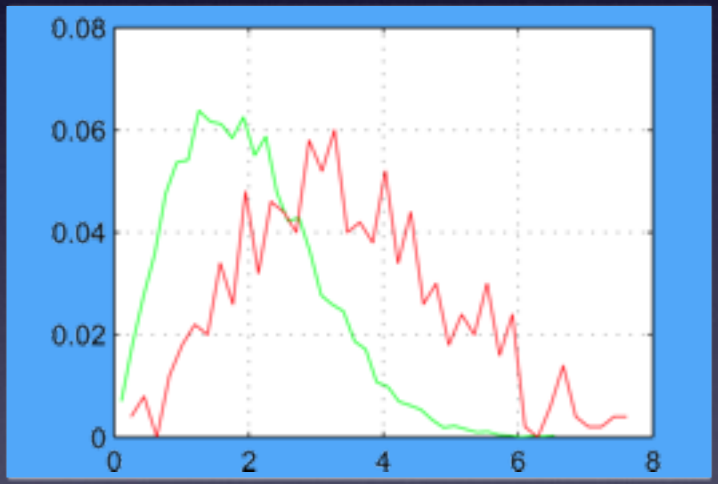
- Move discriminant, and plot True Positive Rate(TPR) against False Positive Rate(FPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

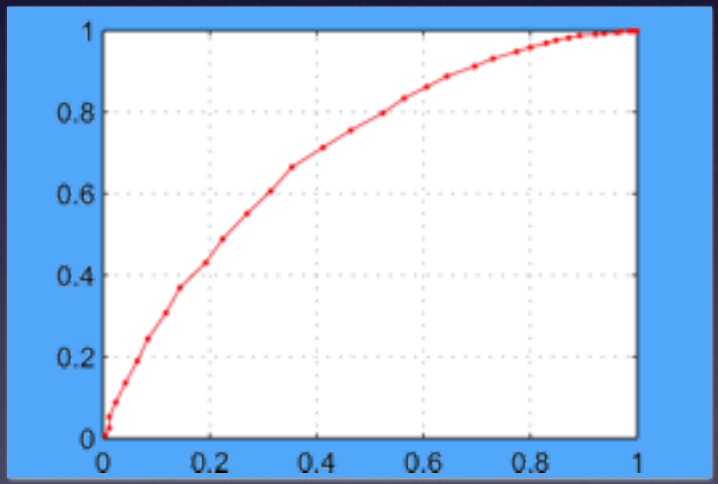
- Invariant to skewed datasets (bad).
  Since normalisation is done with actual number of positives and negatives.

ROC from histograms

- ROC curves can used for evaluating matching performance as well. By using error histograms for inlier&outlier sets.



$\Rightarrow$



- Discriminant moving from left to right.

$$\mathrm{TPR}(\epsilon) = \int_0^{\epsilon} p(\epsilon'|\texttt{inlier})d\epsilon' \quad \mathrm{FPR}(\epsilon) = \int_0^{\epsilon} p(\epsilon'|\texttt{outlier})d\epsilon'$$

Precision-Recall curve

- Move discriminant, and plot Precision against Recall

$$\mathtt{Precision} = \frac{\mathtt{TP}}{\mathtt{TP} + \mathtt{FP}} \qquad \mathtt{Recall} = \mathtt{TPR} = \frac{\mathtt{TP}}{\mathtt{TP} + \mathtt{FN}}$$

- Looks only at correctly reported positives. Better than ROC if positives are rare.

Precision-Recall curve

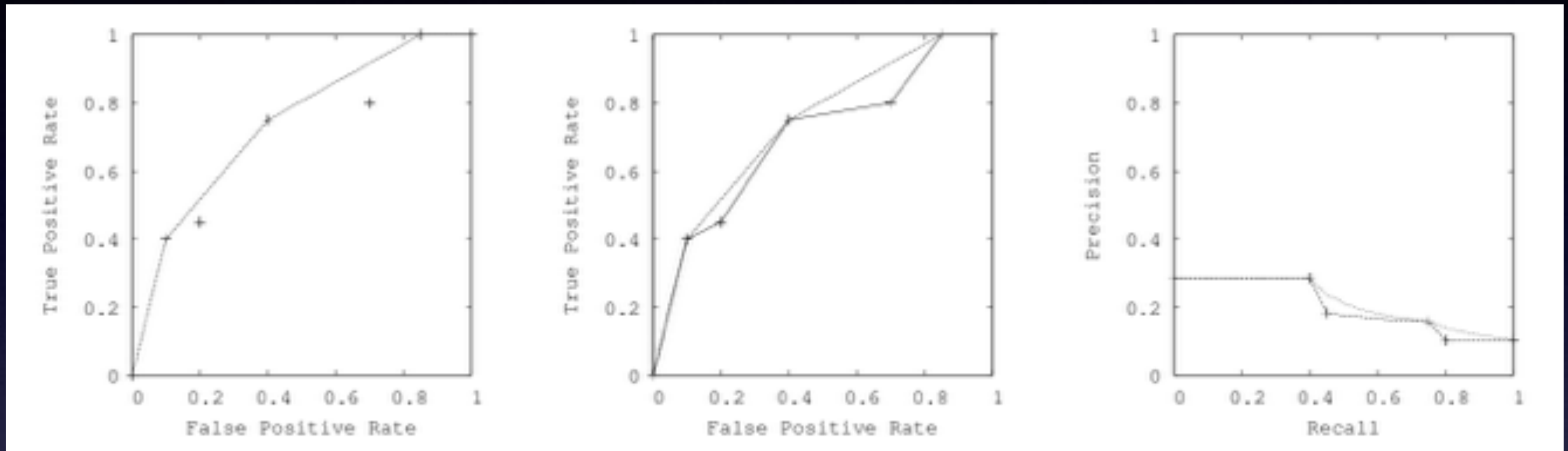
- TPR and FPR (used in ROC) are monotonic
  $\Rightarrow$ Linear interpolation between points on an ROC curve is reasonable.

- Conversion between ROC and PR is possible as |gt=0|=TP+FN and |gt=1|=FP+TN are constant and known.

	gt=0	gt=1
out=0	TP	FP
out=1	FN	TN

- Solve confusion table for each point, and compute the corresponding point on the other curve. Davis & Goadrich, ”The relationship between Precision-Recall and ROC curves”,ICML06

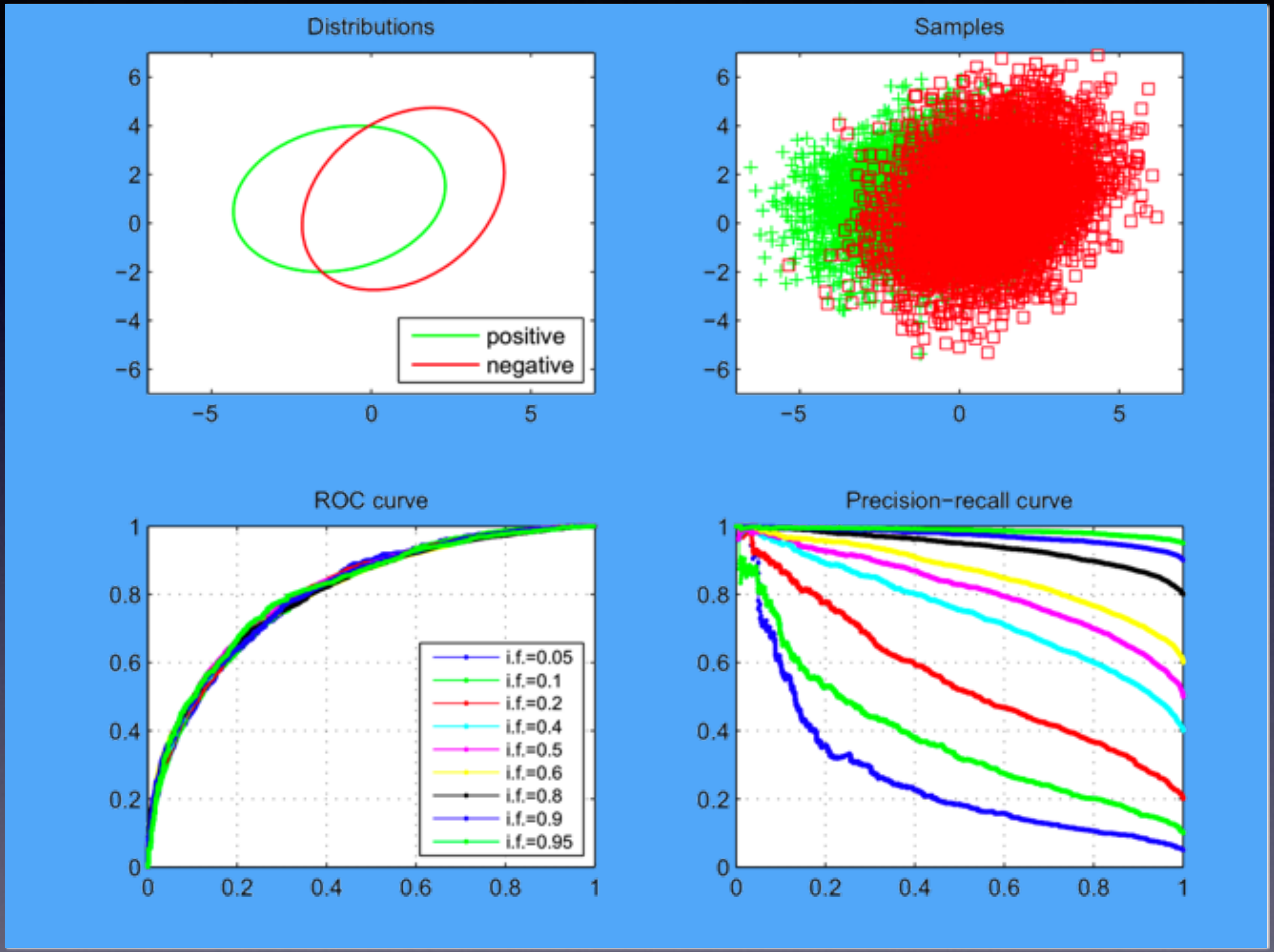
Precision-Recall curve



(a) Convex hull in ROC space (b) Curves in ROC space (c) Equivalent curves in PR space

- This suggests that the proper way to interpolate a PR curve is to linearly interpolate in ROC space, and then transfer the result.
- A linear interpolation of PR may be impossible to attain.

ROC vs PR curves



- Curves show different positive/negative sample ratios.

For classifier output

- Recognition algorithms often output class probability estimates. E.g. for four classes:

- For each class we can compute a PR curve by assigning to class k if $v_k > t$, and letting t go from 0 to 1.

F-scores

- Precision and recall combined into a single measure (using the harmonic mean)

$$F_1 = \frac{2}{1/P + 1/R} = \frac{2PR}{R+P}$$

F-scores

- Precision and recall combined into a single measure (using the harmonic mean)

$$F_1 = \frac{2}{1/P + 1/R} = \frac{2PR}{R + P}$$

- Weighted F-scores

$$F_k = (1 + k^2)\frac{PR}{R + k^2 P}$$

- All computed at a specific detection threshold

Summarization

- If a quality measure is to be used in optimization, a single measure is better than a curve.
- A common way to summarize ROC (and PR) is to look at *area under the curve* (AUC). Also called *average precision* for a PR curve.
- Another option for ROC is the point of *equal error rate* (EER), i.e. where FPR=1-TPR
- AUC is in general better than EER as AUC considers the whole curve.

Summary

- For detection and matching, both inlier frequency and total number matters.
- Use ROC and PR curves in classification to avoid committing to a threshold.
- PR curves are better for skewed datasets
- For optimisation, area under a curve is a useful summary

Exam

- Written exam format:
  - **In total 16 questions**
  - **Example:**
    Explain when a Precision-Recall should be used instead of a ROC curve, and why.
  - **Questions will be based on:**
    1. The slides from all eight lectures
    2. The seven articles

Exam

- Times for the written exam:

April 16, 9-11 >2people

April 29, 13-15 >6people

Discussion

- Questions/comments on today’s paper:

  Russakovsky and Deng et al., ”ImageNet Large Scale Visual Recognition Challenge”, **ArXiV15**