TRECVID-2005 Low-level (camera motion) feature task

Wessel Kraaij

TNO

&

Tzveta Ianeva

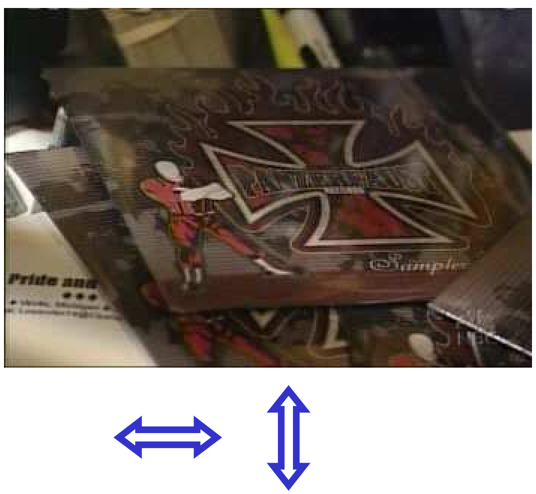
NIST

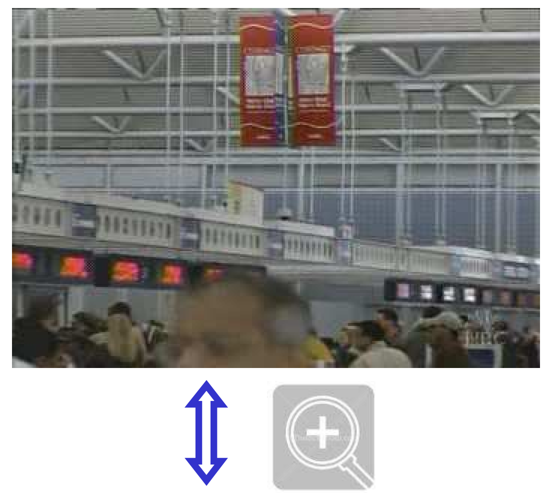
Task definition

- ØTRECVID 2005 pilot task
- ØAbility to detect camera movement features:
  - q Pan (left or right) or track
  - q Tilt (up or down) or boom
  - q Zoom (in or out) or dolly

Task definition ...

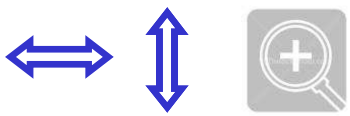
- Camera movement features are usually combined
  - Pan & Tilt
  - Pan & Zoom
  - Tilt & Zoom





Task definition ...

- Pan & Tilt & Zoom

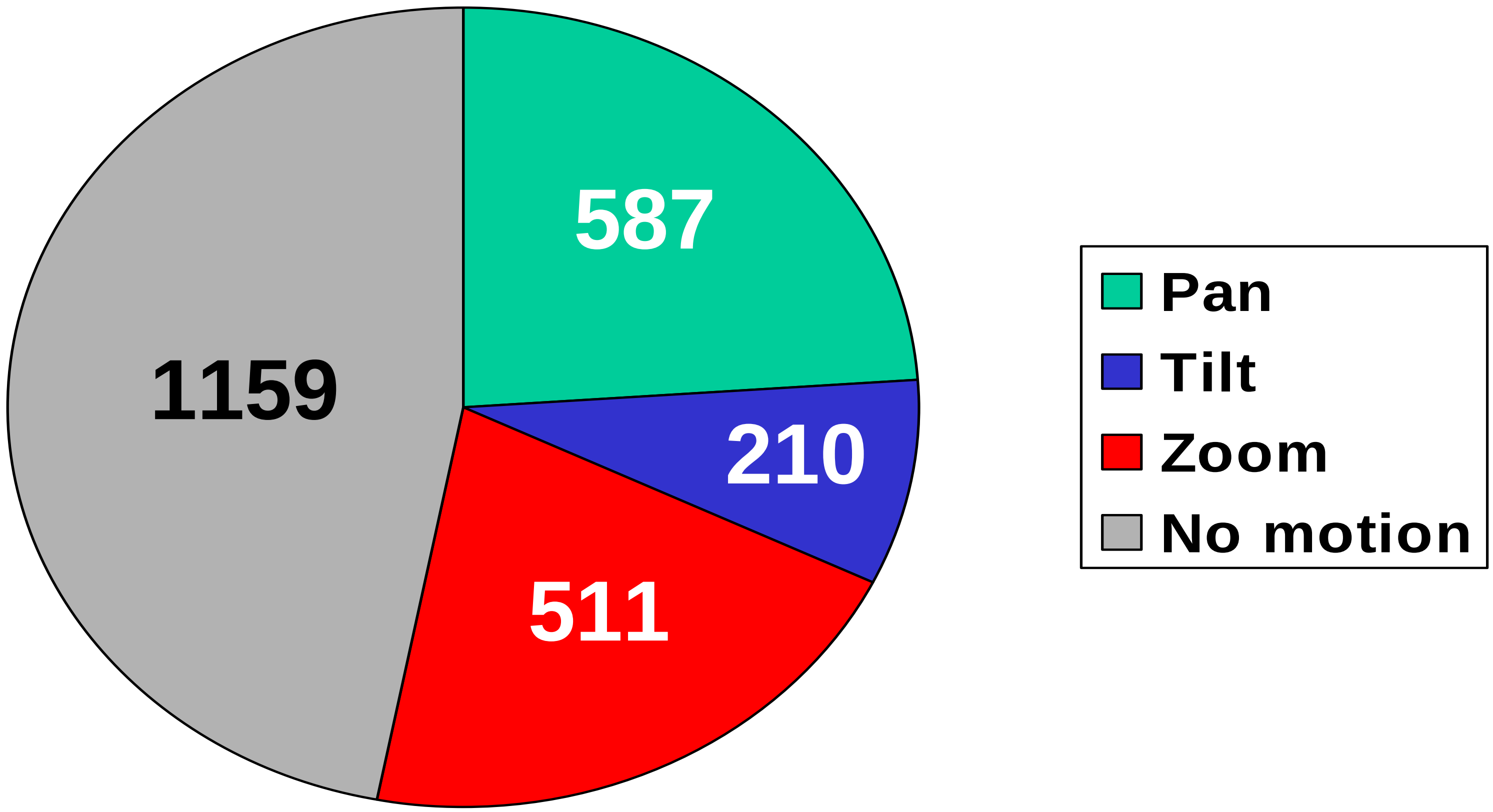


- Submissions provide complete judgments for test set by specifying all shots identified as positive by the system
- No Training data provided by NIST
- Tool to create development data developed by Werner Bailer at Joanneum Researh

Ground truth creation at NIST

- Watch randomly chosen subset of test data (~5000 shots)
- Keep only shots with "clear" examples of (no) motion (~2226)
- No-motion shots seem to more clearly exhibit no motion than shots with motion features exhibit motionŁ *#FP will tend to be small, #FN will tend to be high*
- Define test subset for each feature by combining
  - shots exhibiting the feature
  - shots exhibiting no motion (same for all features)
- No adjustments to subset sizes or true:false ratios
  - Pan 587:1159
  - Tilt 210:1159
  - Zoom 511:1159

Truth data distribution (number of shots)



Truth and evaluation issues

- **Why feature groups?**
  - Perceptual limits in truth creation
- Cost of creating truth data
  - Many shots with lots of small camera movement – not what's wanted when user asks for a "pan", etc.
- **Implications of test set construction on measures**
  - Lack of randomness makes generalization hard
  - Varying true:false ratios make precision harder for tilt than pan and zoom
  - Greater clarity of no-motion shots would make false positive less likely then false negatives and higher precision easier to achieve than higher recall

No motion shots

Truth data costly to create – lot's of shaky shots

Hard to judge

Not what a user wants

12 Participating Groups

Carnegie Mellon University (CMU) - USA

City University of Hong Kong (CUHK) - China

Fudan University (FUDAN) - China

Institute for Infocomm Research (IIR) - Singapore

JOANNEUM RESEARCH (Joanneum) - Austria

KDDI & R&D Laboratories, Inc. (KDDI) - Japan

LaBRI (LaBRI) - France

Tsinghua University (Tsinghua) - China

University of Central Florida / University of Modena (UCF) – USA/Italy

University of Iowa (Uiowa) - USA

University of Marburg (MARBURG) - Germany

Univ. of Amsterdam & TNO (UvA) - Netherlands

NIST baseline runs

- Ø All features true for all shots (TrueForAllShots)
- Ø Random run with true distribution of Pan, Tilt, Zoom as in truth data (TruthDataDistrib)
- Ø Features randomly true/false for each shot (Random)

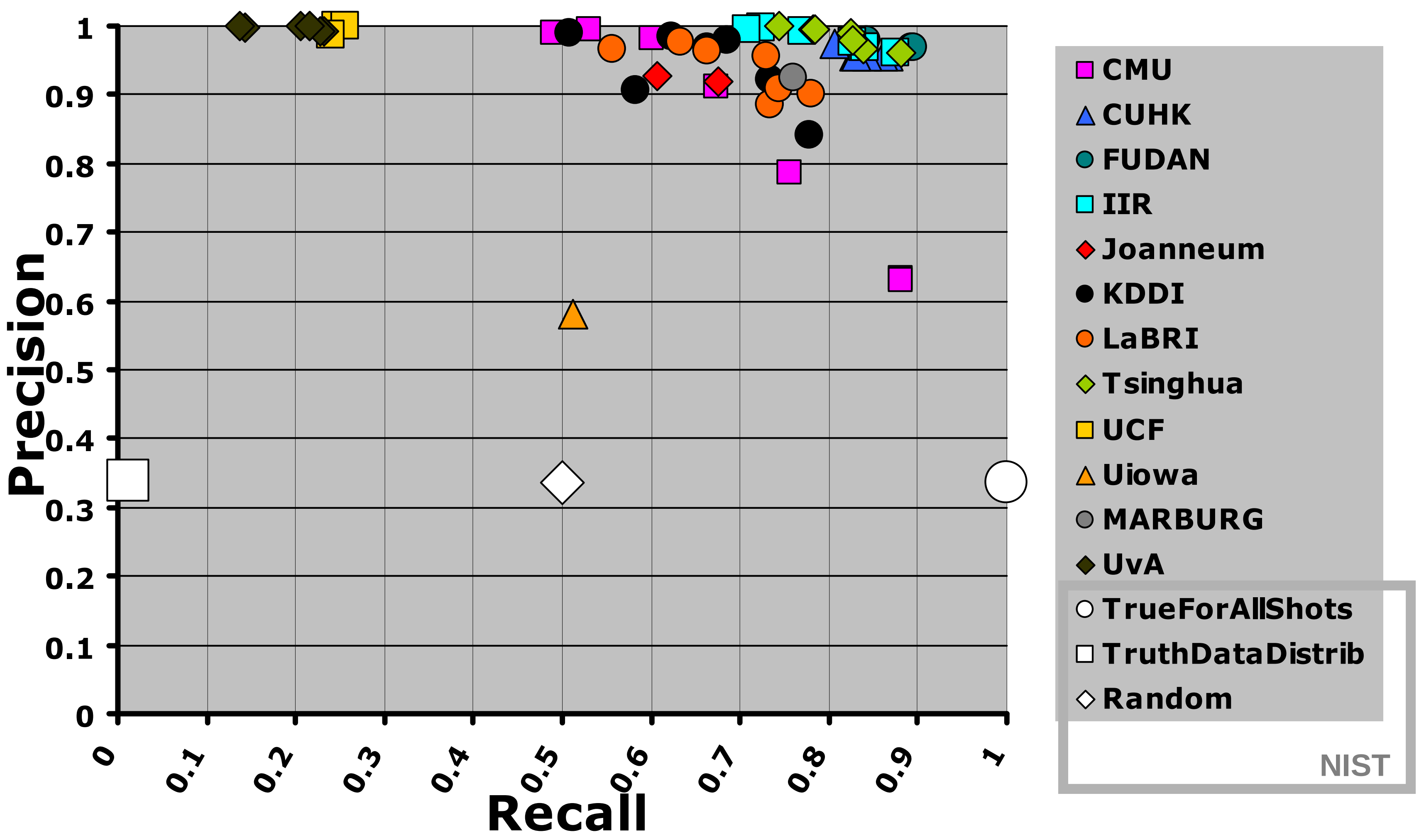
Evaluation Measures

$$\text{Precision} = \frac{\#\ \text{True positives}}{\#\ \text{True positives} + \#\ \text{False positives}}$$

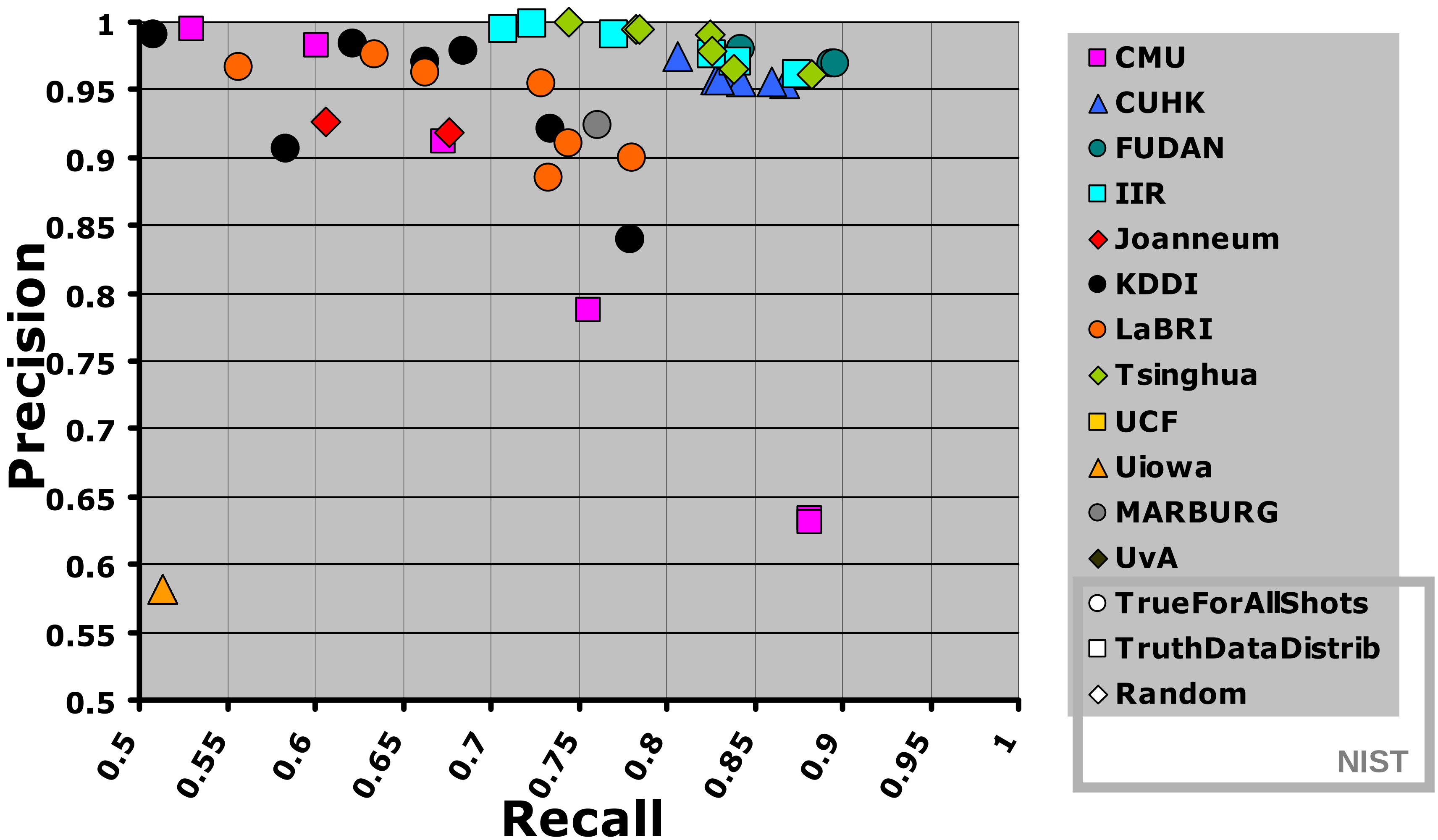
$$\text{Recall} = \frac{\#\ \text{True positives}}{\#\ \text{True positives} + \#\ \text{False negatives}}$$

Given the imbalance in class properties, it's easier to achieve a high precision than a high recall. The use of $F_{\beta=1}$ seems not appropriate

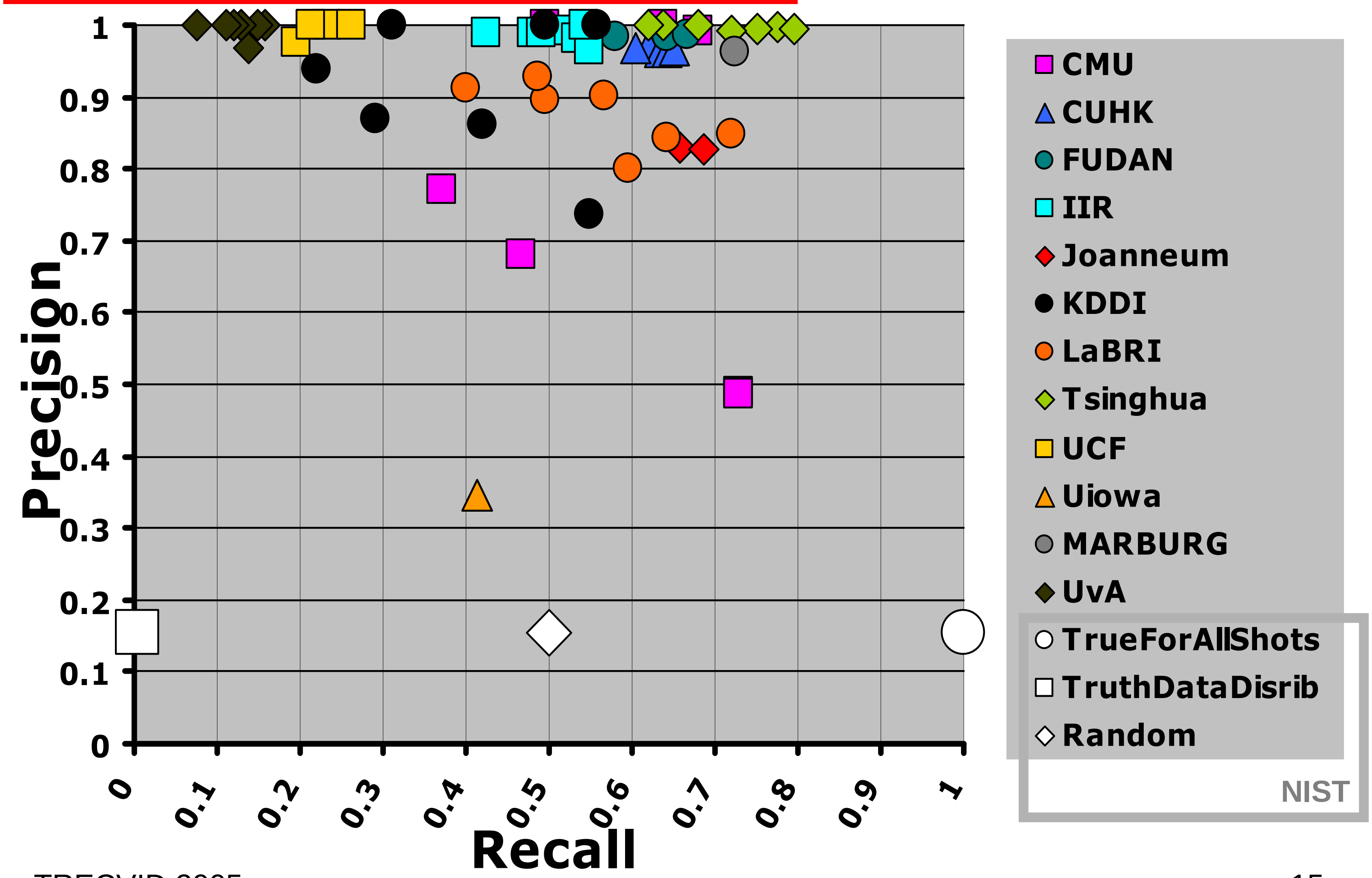
Pan: recall and precision by system



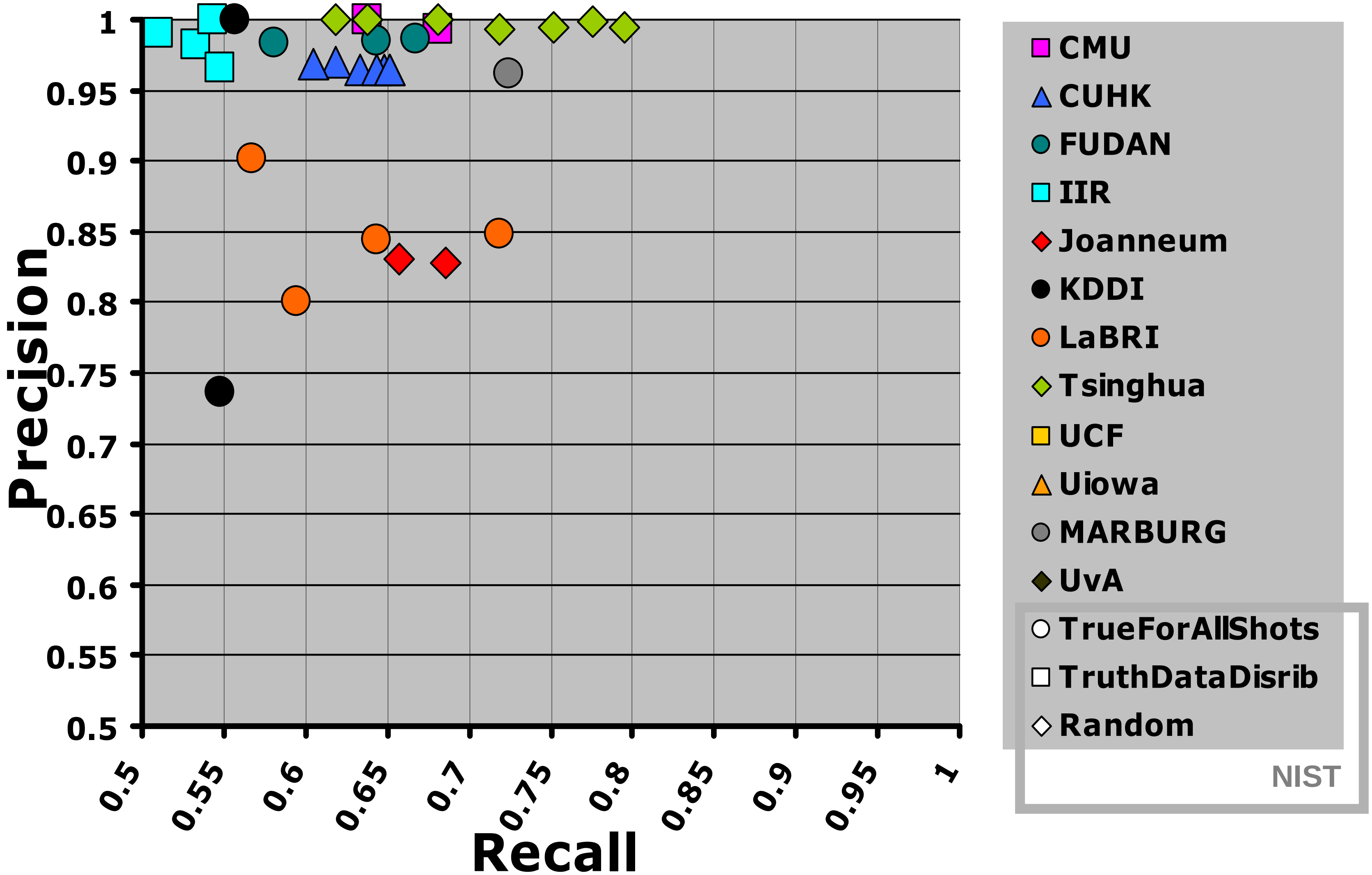
Pan: recall and precision by system (zoomed)



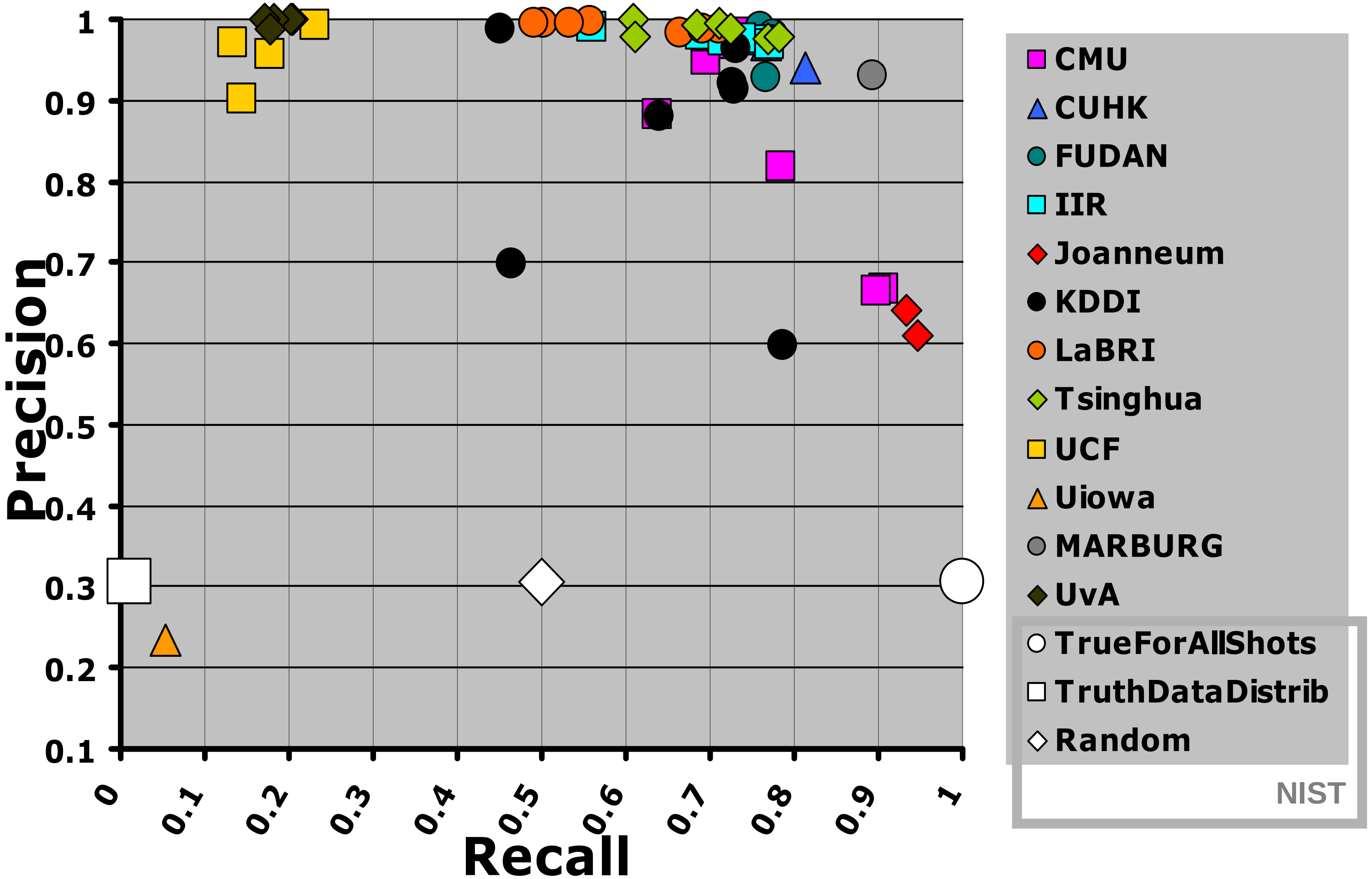
Tilt: recall and precision by system



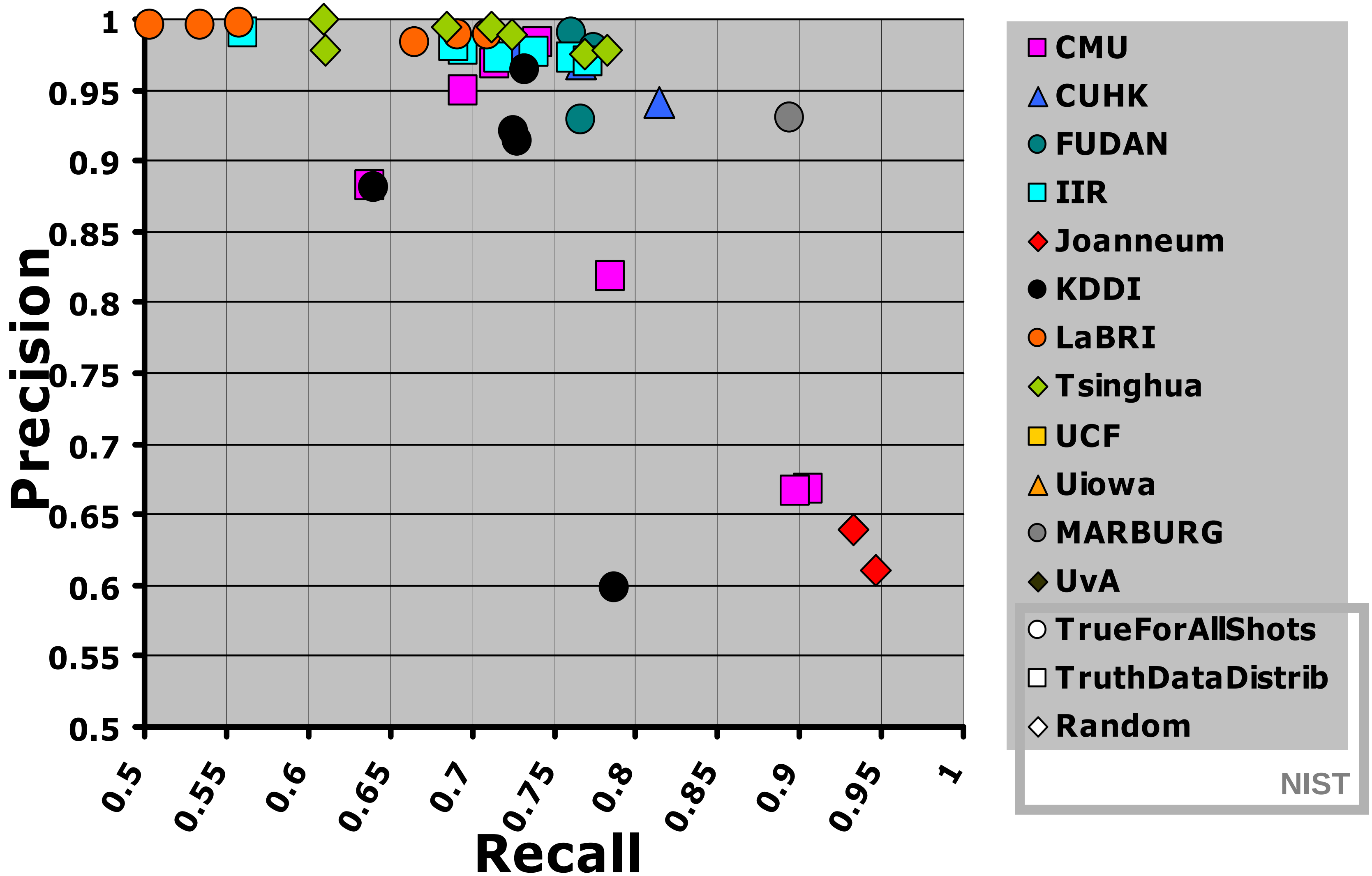
Tilt: recall and precision by system (zoomed)



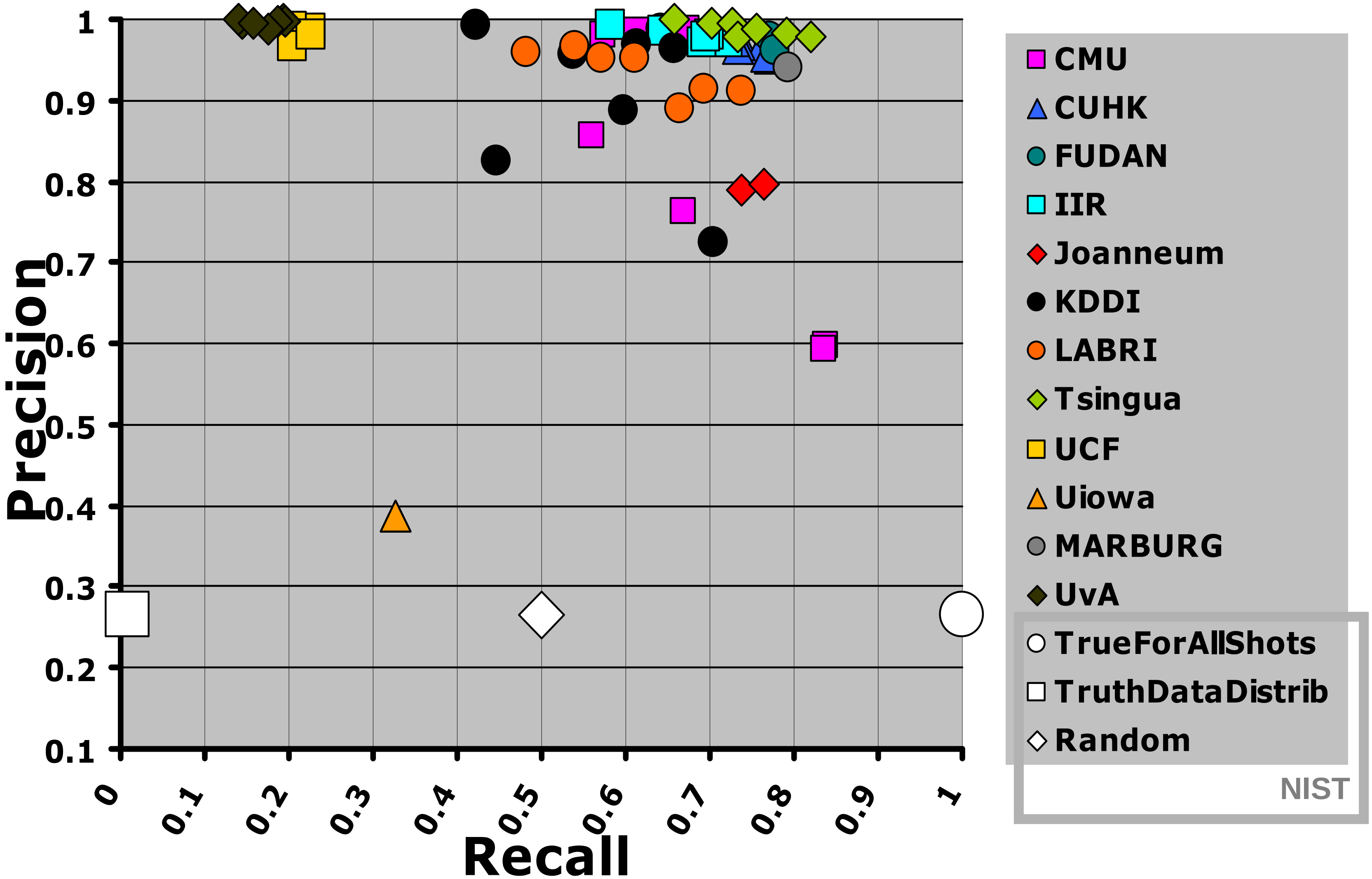
Zoom: recall and precision by system



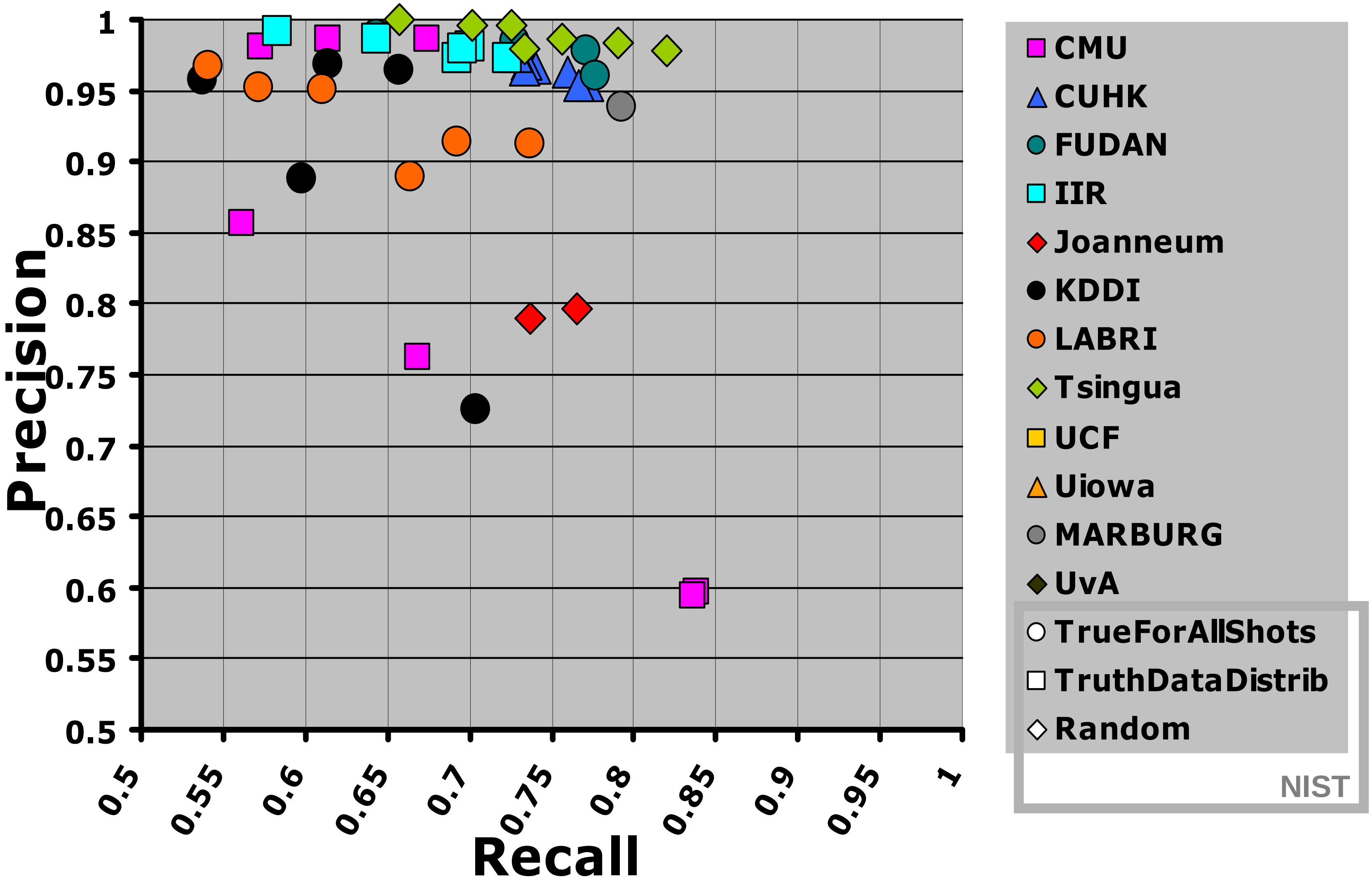
Zoom: recall and precision by system (zoomed)



Mean recall and precision over all 3 features by system



Mean recall and precision over all 3 features by system (zoomed)

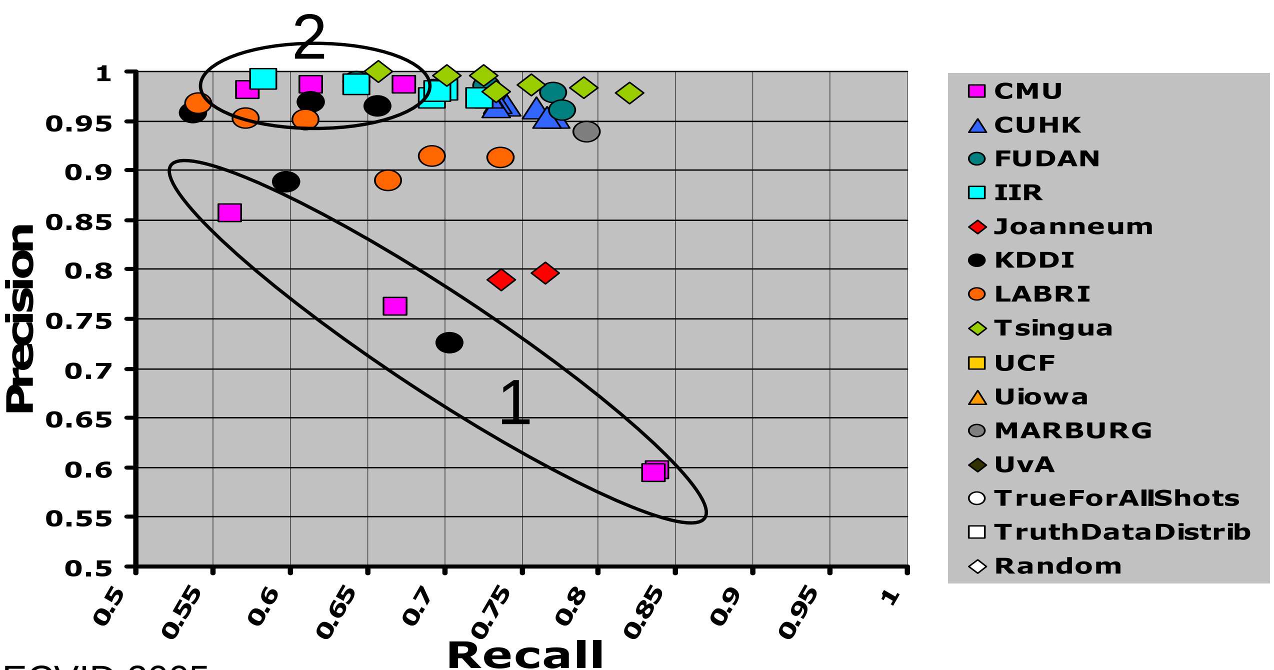


General points

- NIST did not provide training data: some training data was available from other sources and some training data was produced by participants
- Input:
  - MPEG motion vectors: optimal for compression, not optimal for modeling real motion
  - Frame to frame motion analysis
- Distinguish “jitter” from intended motion

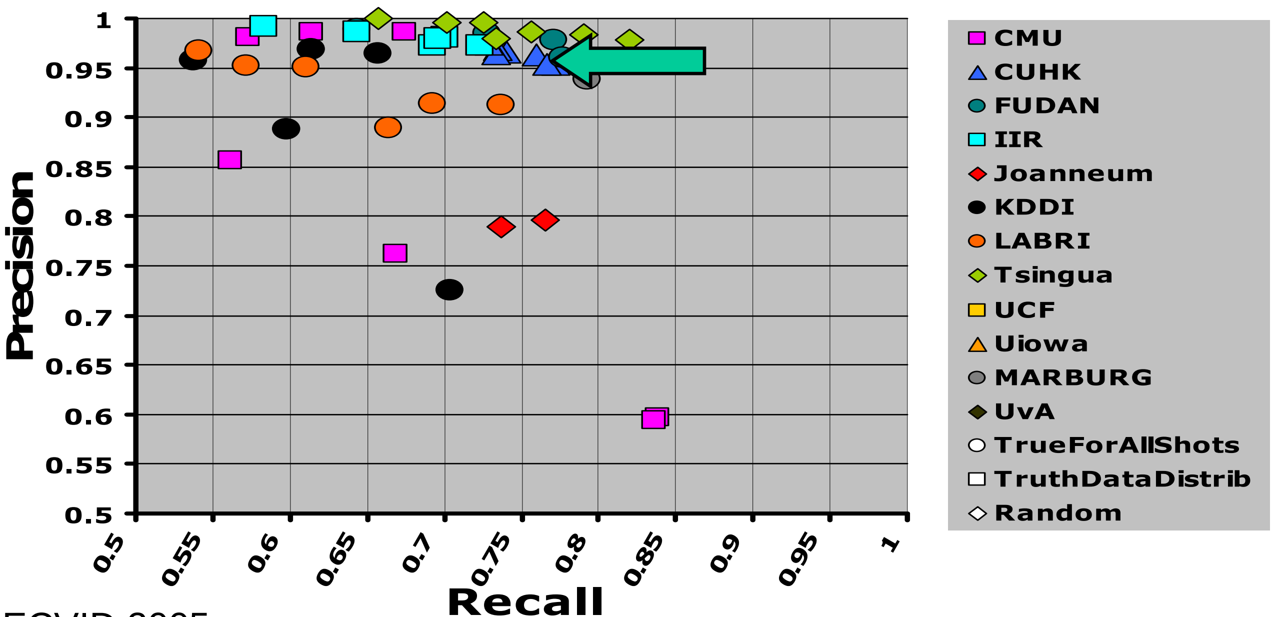
CMU

- Approach
  1. Probabilistic model (fitted using EM) based on MPEG motion vectors
  2. Optical Flow model: extract the most consistent motion from the optical flows (frame to frame)



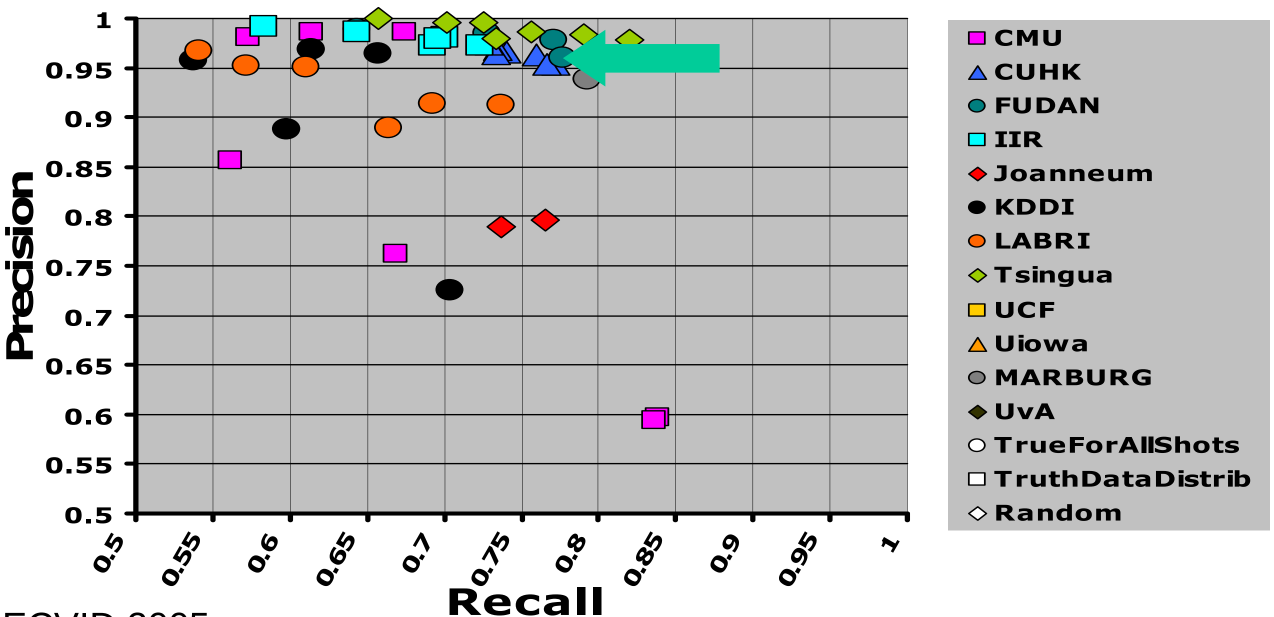
CUHK

- Approach
  - Motion features extracted from tracking image features in consecutive frames
  - Estimation of 6 parameter affine model, transformation in p,t,z vector for each set of adjacent frames
  - Rule based motion classification using empirical thresholds
  - Interesting failure analysis



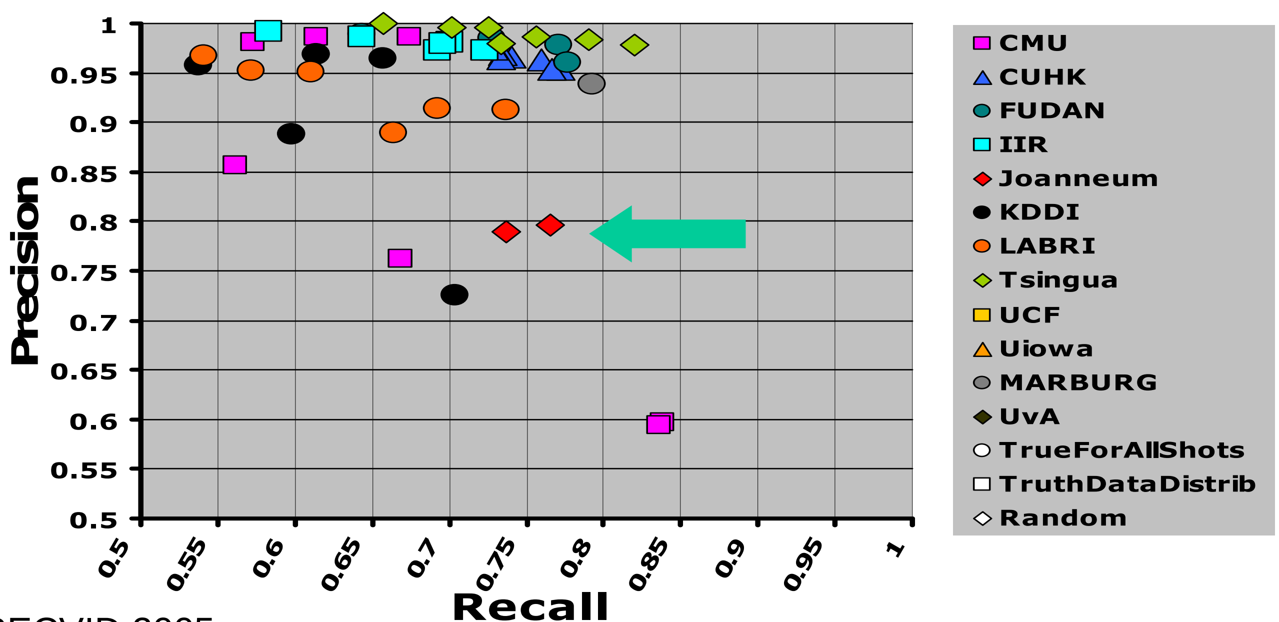
Fudan

- Approach
  - Motion vectors from MPEG,SVM, motion accumulation method to filter out imperceptable movements
  - Filter method seems to decrease precision though…



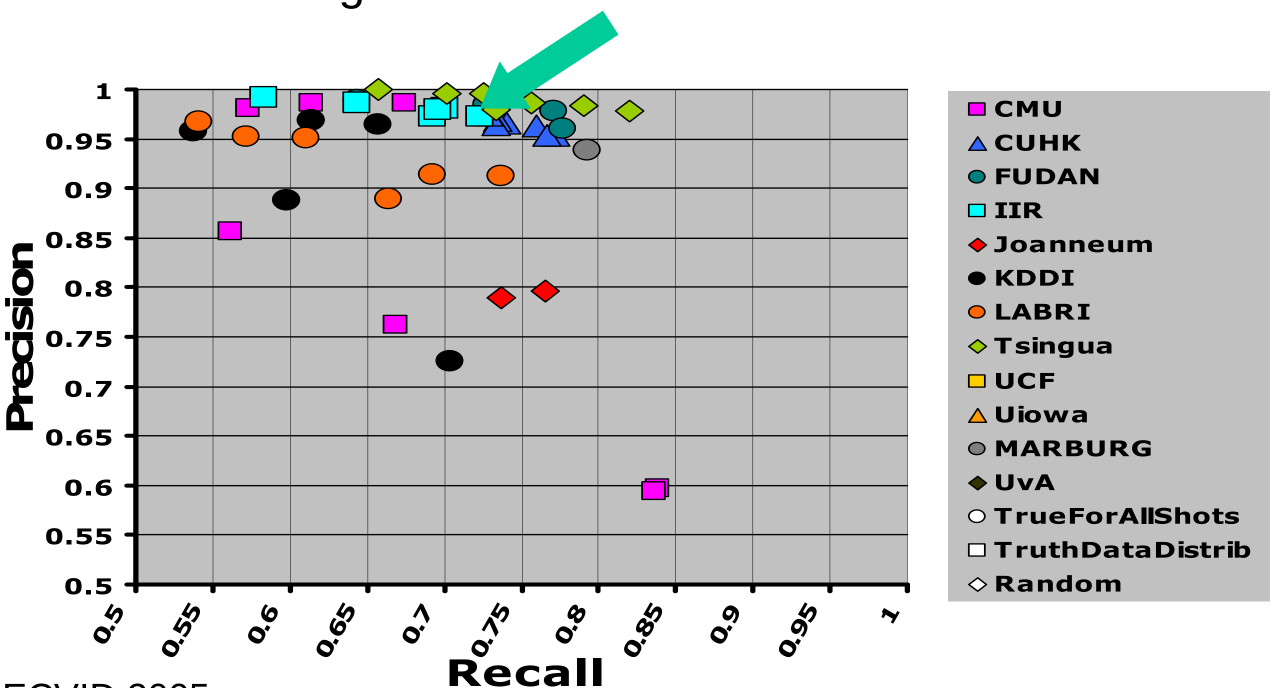
Joanneum
- presentation follows -

- Approach
  - Developed a training set , problems with annotation..
  - Feature tracking, clustering trajectories, dominant cluster selection, camera motion detection, thresholding



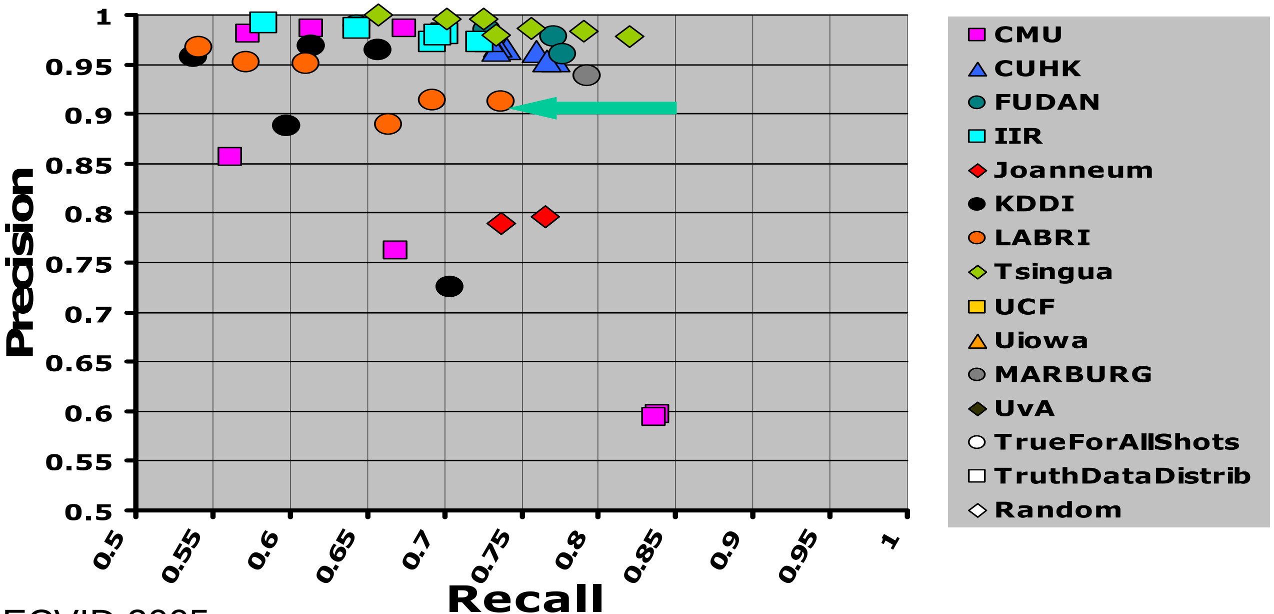
IIR

- Approach
  - Annotated 24 video files
  - Estimated affine camera model based on MPEG motion vectors
  - Transformation of model parametersŁ   series of p,t,z values for each shot
  - Rule based classification of series using accumulation and thresholding



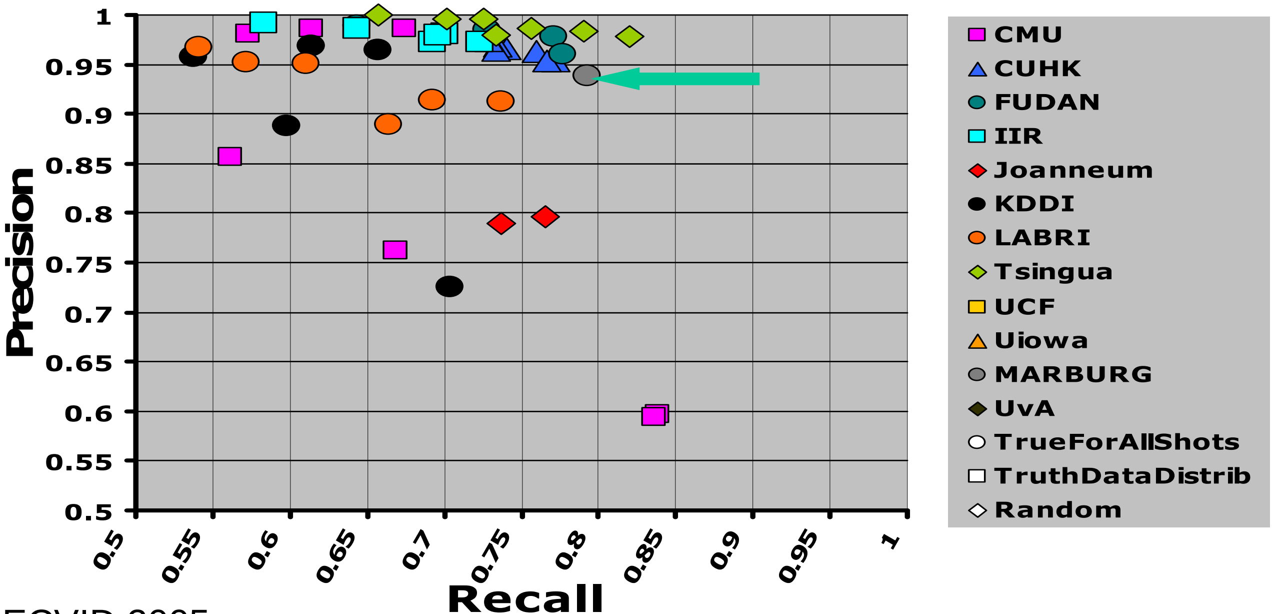
LaBRI
- presentation follows -

- Approach
  - Mpeg motion vector inputŁ 6 parameter affine model
  - Jitter suppression (statistical significance test)
  - Subshot segmentation (homogeneous motion)
  - Motion classification (using “a few annotated videos”)



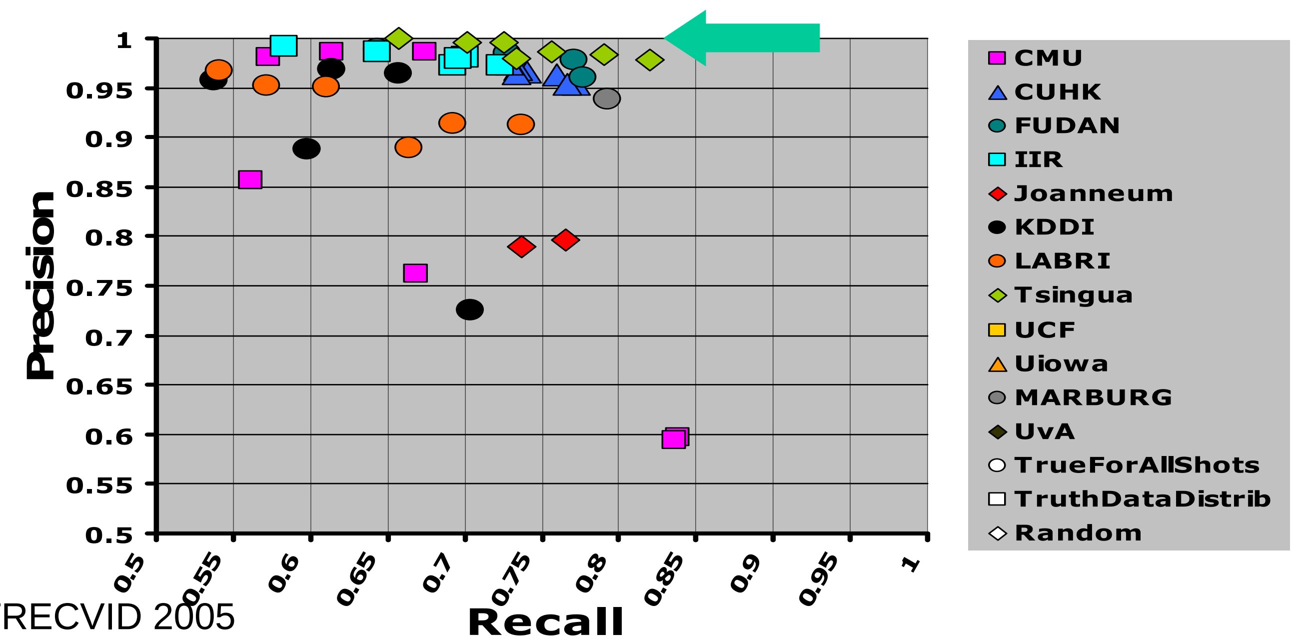
Marburg

- Approach
  - 3D camera model estimated from MPEG motion vectors
  - Cleaning necessary, + exlusion of center, frame border
  - Optimal thresholds estimated on tv2005 training set



Tsinghua

- Approach
  - Motion vector selection based spatial features, separating camera motion from object motion and accidental motion
  - 4 parameter camera model (Iterative Least Squares) parameter estimation
  - Rule based classification (FSA), using a range of thresholds for: 1.Continuous (speed) and noticable, 2,Minumum duration 3.Uninterrupted 4.Noticable in case in combination with other camera movement



Observations

- This is clearly an easier task than the HLF task, though a high recall is hard to achieve.
- Truth data costly to create – lot's of shaky shots
  - Many hard to judge
  - Many not really what a user wants when s/he asks for a "pan" etc.
- Hard to generalize from small, constructed test subset to larger, more realistic test set
- Given the definition of our task and test set characteristics, F measure not appropriate
- Concentrate on within-feature system comparisons