TREC 2003 Video Retrieval Evaluation
Overview

Coordinators: Alan Smeaton

Centre for Digital Video Processing

Dublin City University

Wessel Kraaij

Department of Data Interpretation

Information Systems Division

TNO TPD

NIST: Paul Over

Retrieval Group

Information Access Division

Information Technology Laboratory

National Institute of Standards and Technology

Origins

- Problem:
  - Rapidly growing quantities of digital video
  - Increasing research in content-based retrieval from digital video
  - But no common basis for evaluation/comparison of approaches
- Approach:
  - Find as much video data as possible and make it available to the community of researchers
  - Use the data to build an open, metrics-based evaluation in the Cranfield/TREC tradition
  - Invite participation and see what happens…

Goals

- Promote progress in content-based retrieval from large amounts of digital video
- Answer some questions:
  - How can systems achieve such retrieval (in collaboration with a human)?
  - How can one reliably benchmark such systems?

Evolution… 2001

- TREC 2001 Video retrieval track
- Data: 11 hrs (OpenVideo, NIST)
- 2 Tasks:
  - Shot boundary determination
  - Search
    - Fully automatic
    - Interactive
- Participating groups: 12

Evolution… 2002

- TREC 2002 Video retrieval track
- Data: 73 hrs (Prelinger Archive)
- 3 Tasks:
  - Shot boundary determination
  - High-level feature extraction (10)
  - Search (manual and interactive)
- Participating groups: 17
- New:
  - Common shot reference defines unit of retrieval
  - Common key frames
  - Shared features, ASR output provided by LIMSI

Evolution… 2003

- TRECVID Workshop
- Data: 133 hrs (1998 ABC/CNN news + C-SPAN)
- 4 Tasks:
  - Shot boundary determination
  - High-level feature extraction (17)
  - Story segmentation and classification
  - Search (manual and interactive)
- Participating groups: 24
- New:
  - Common annotation effort
  - Advisory committee

Advisory committee

- John Eakins (University of Northumbria at Newcastle)
- Peter Enser (University of Brighton)
- Alex Hauptmann (CMU)
- Annemieke de Jong (Netherlands Institute for Sound & Vision)
- Michael Lew (Leiden Insitute of Advanced Computer Science)
- Georges Quenot (CLIPS-IMAG Laboratory)
- John Smith (IBM)
- Richard Wright (BBC)

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	
Carnegie Mellon Univ. (US)			X	X
CLIPS-IMAG (FR)	X		X	
CWI Amsterdam / Univ. of Twente (NL)			X	X
Dublin City University (Irl)		X		X
Fudan Univ. (China)	X	X	X	X
FX-Pal (US)	X			
IBM Research (US)	X	X	X	X
Imperial College London (UK)	X		X	X
Indiana University (US)				X
Institut Eurecom (FR)			X	
KDDI (JP)	X	X		
KU Leuven (BE)	X			
Mediamill/U Amsterdam (NL)				X
National Univ. Singapore (Sing.)		X		X
Ramon Llull Univ. (ES)	X			
RMIT University (Aus)	X			
StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

Shot Boundary Detection task

- SBD is an enabling function for almost all content-based operations on digital video, so its important;
- (Still) not a new problem, but a challenge because of gradual transitions and false positives caused by photo flashes, rapid camera movement, object movement, etc.;
- Task is to identify transitions and determine whether each is “cut”, “dissolve”, “fadeout/in” or “other”;
- TRECVID2003 dataset is slightly (10%) larger than 2002 but has many more (78%) shot transitions;

Shot Boundary Detection task

- Manually created ground truth of 3,734 transitions (thanks again to Jonathan Lasko) with 70.7% hard cuts, 20.2% dissolves, 3.1% fades and 5.9% other … very similar ratios to 2002;
- Up to 10 submissions per group, measured using precision and recall, with a bit of flexibility for matching gradual transitions;
- Most participating groups use their 10 submissions to "tweak" some parameter;

14 Groups in Shot Boundary Detection

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	
Carnegie Mellon Univ. (US)			X	X
CLIPS-IMAG (FR)	X		X	
CWI Amsterdam / Univ. of Twente (NL)			X	X
Dublin City University (Irl)		X		X
Fudan Univ. (China)	X	X	X	X
FX-Pal (US)	X			
IBM Research (US)	X	X	X	X
Imperial College London (UK)	X		X	X
Indiana University (US)				X
Institut Eurecom (FR)			X	
KDDI (JP)	X	X		
KU Leuven (BE)	X			
Mediamill/U Amsterdam (NL)				X
National Univ. Singapore (Sing.)		X		X
Ramon Llull Univ. (ES)	X			
RMIT University (Aus)	X			
StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

What do the results look like ?

Evaluation Measures

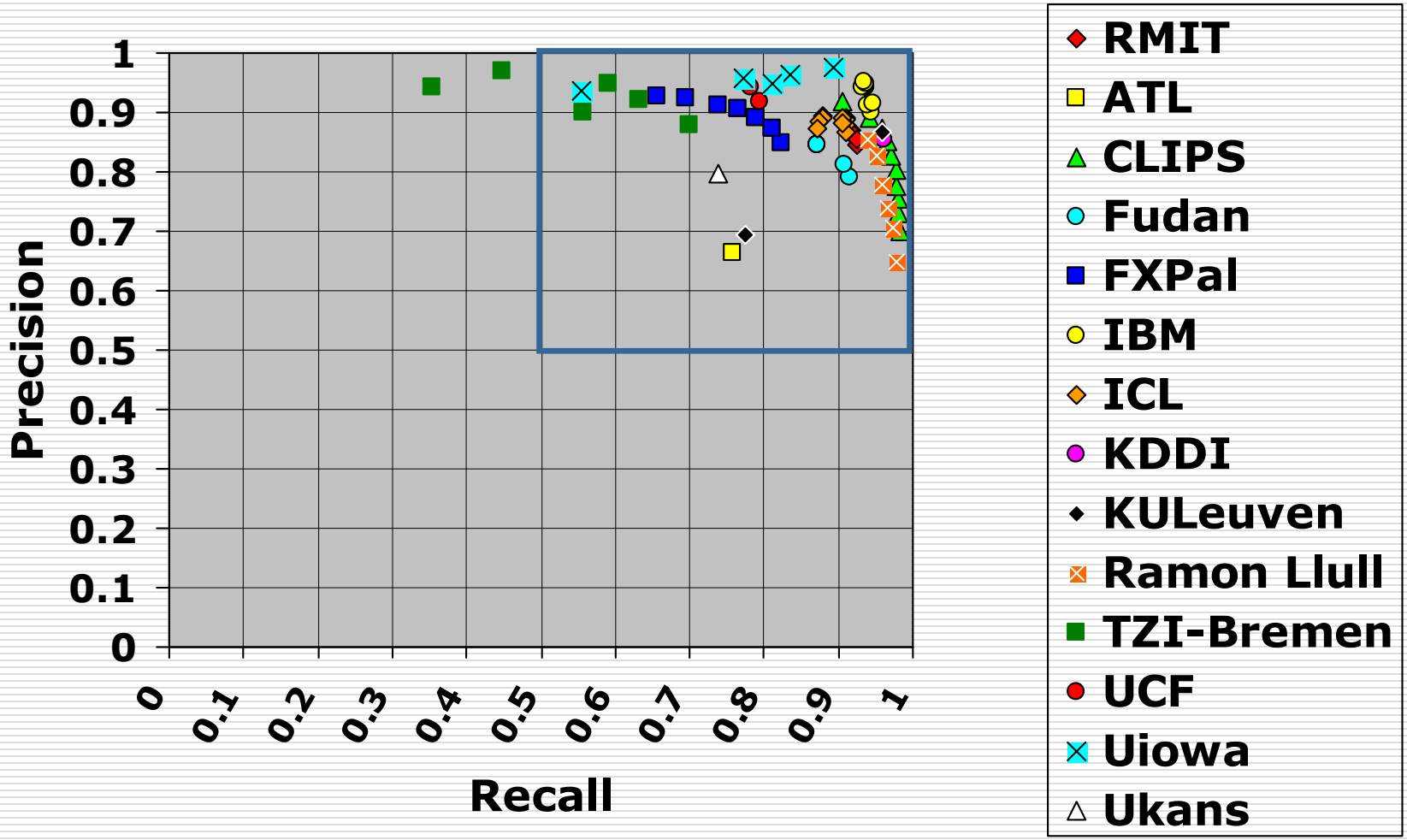
$$\text{Precision} = \frac{\#\ \text{Transitions Correctly Reported}}{\#\ \text{Transitions Reported}}$$

$$\text{Recall} = \frac{\#\ \text{Transitions Correctly Reported}}{\#\ \text{Transitions in Reference}}$$

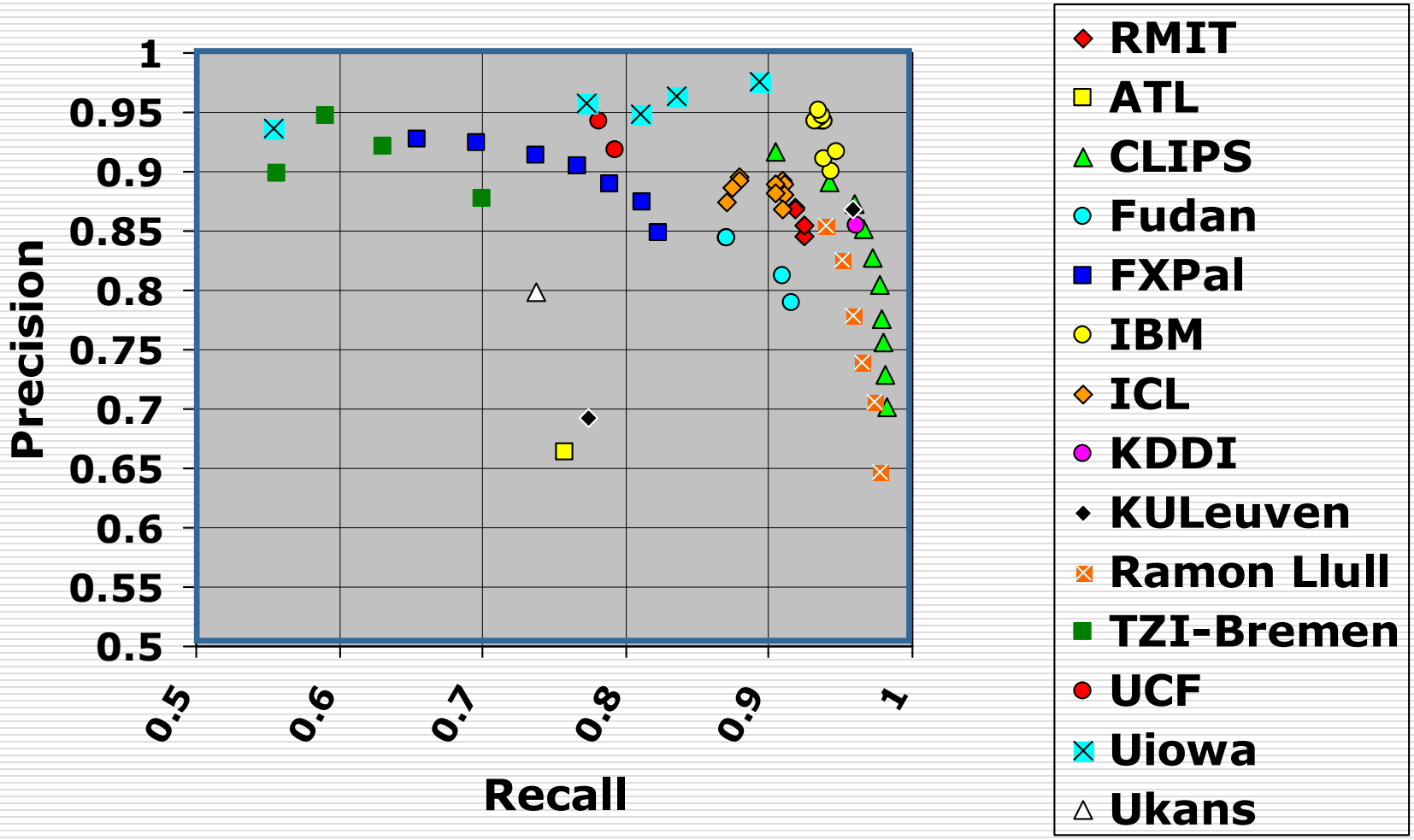
$$\text{Frame Precision} = \frac{\#\ \text{Frames Correctly Reported in Detected Transitions}}{\#\ \text{Frames reported in Detected Transitions}}$$

$$\text{Frame Recall} = \frac{\#\ \text{Frames Correctly Reported in Detected Transitions}}{\#\ \text{Frames in Reference Data for Detected Transitions}}$$

Recall and precision for cuts

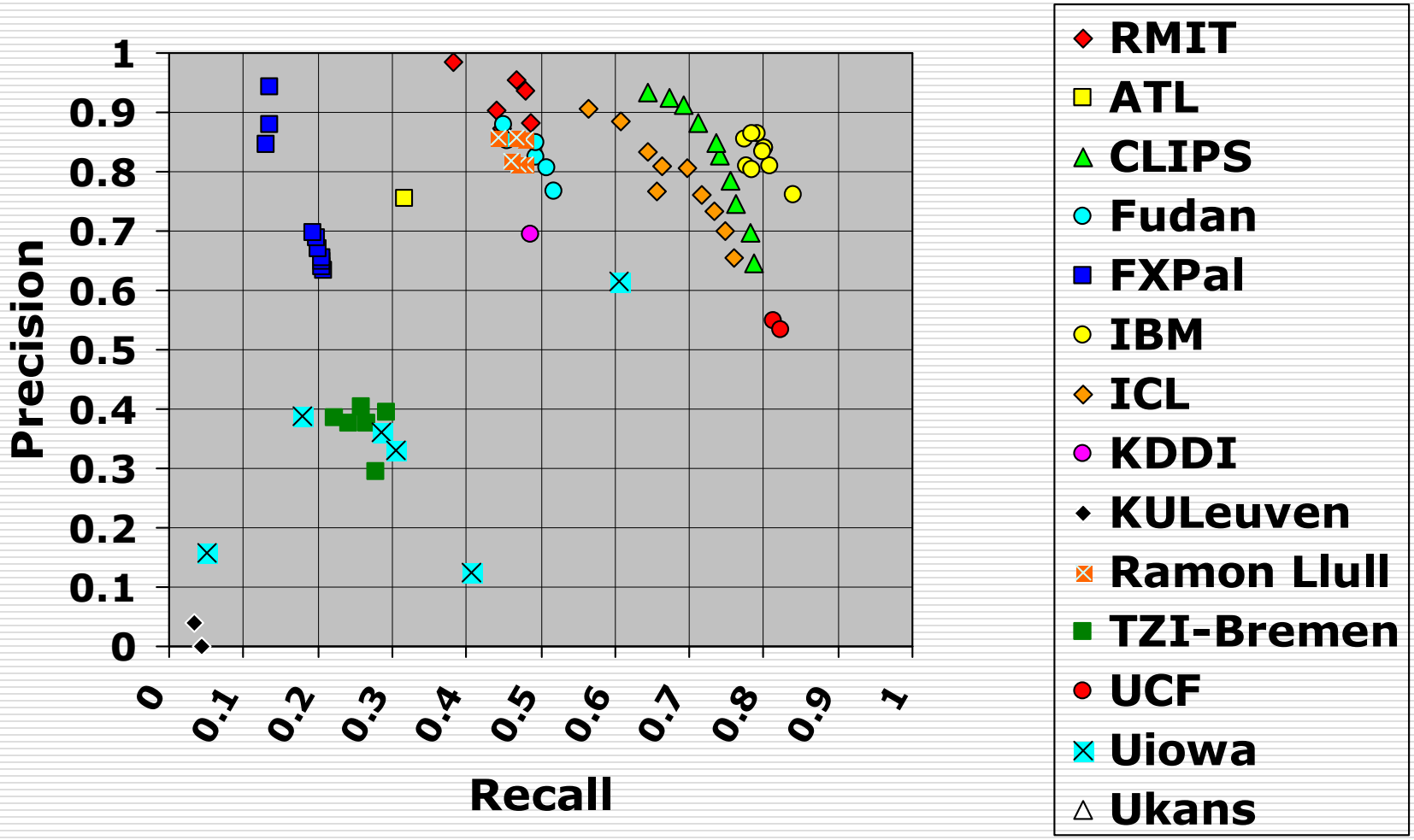


Recall and precision for cuts (zoomed)

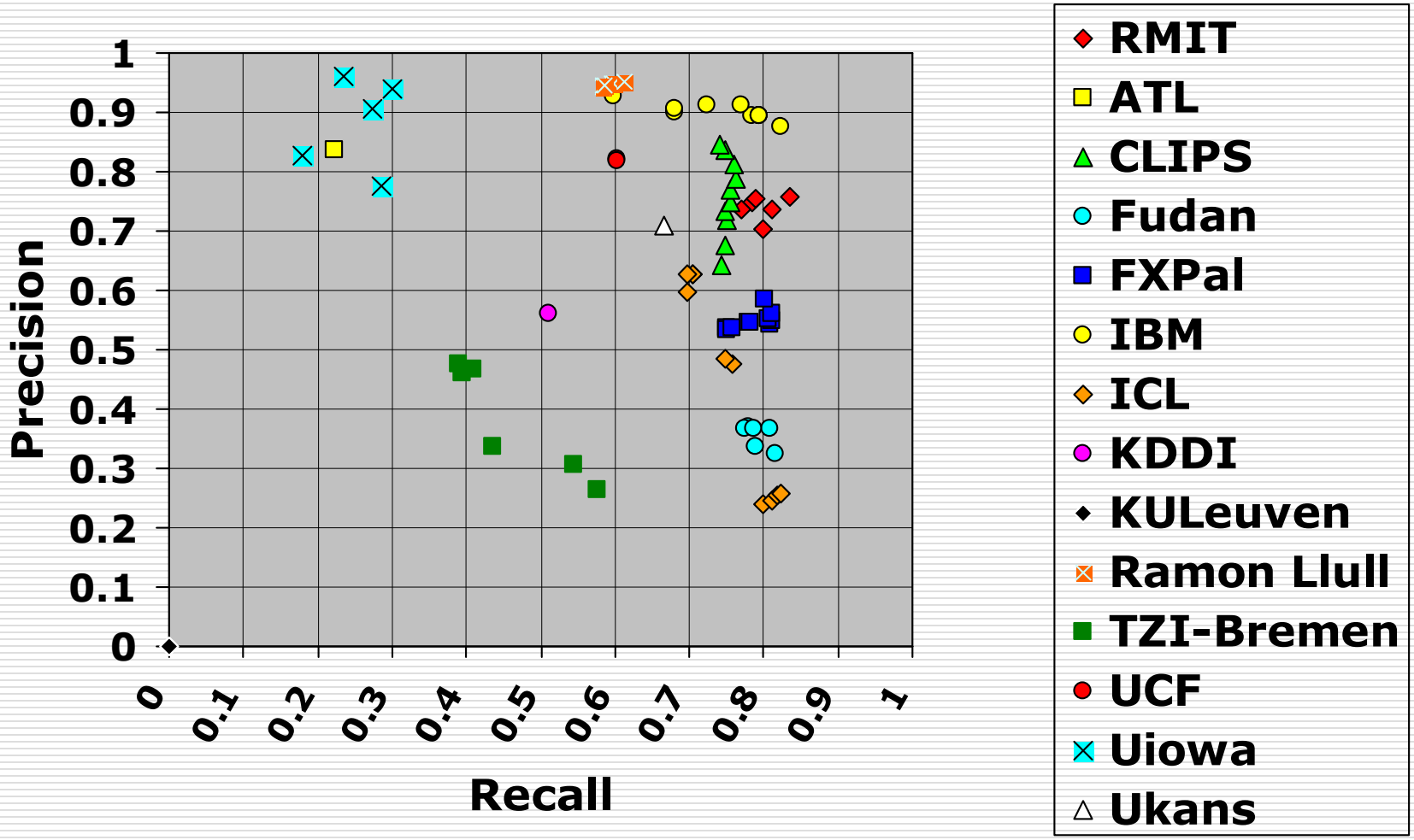


… and for Gradual Transitions …

Recall and precision for gradual transitions



Frame-recall & -precision for GTs



So, who did what ?

The approaches….

24 Participating Groups

Accenture Technology Laboratories:

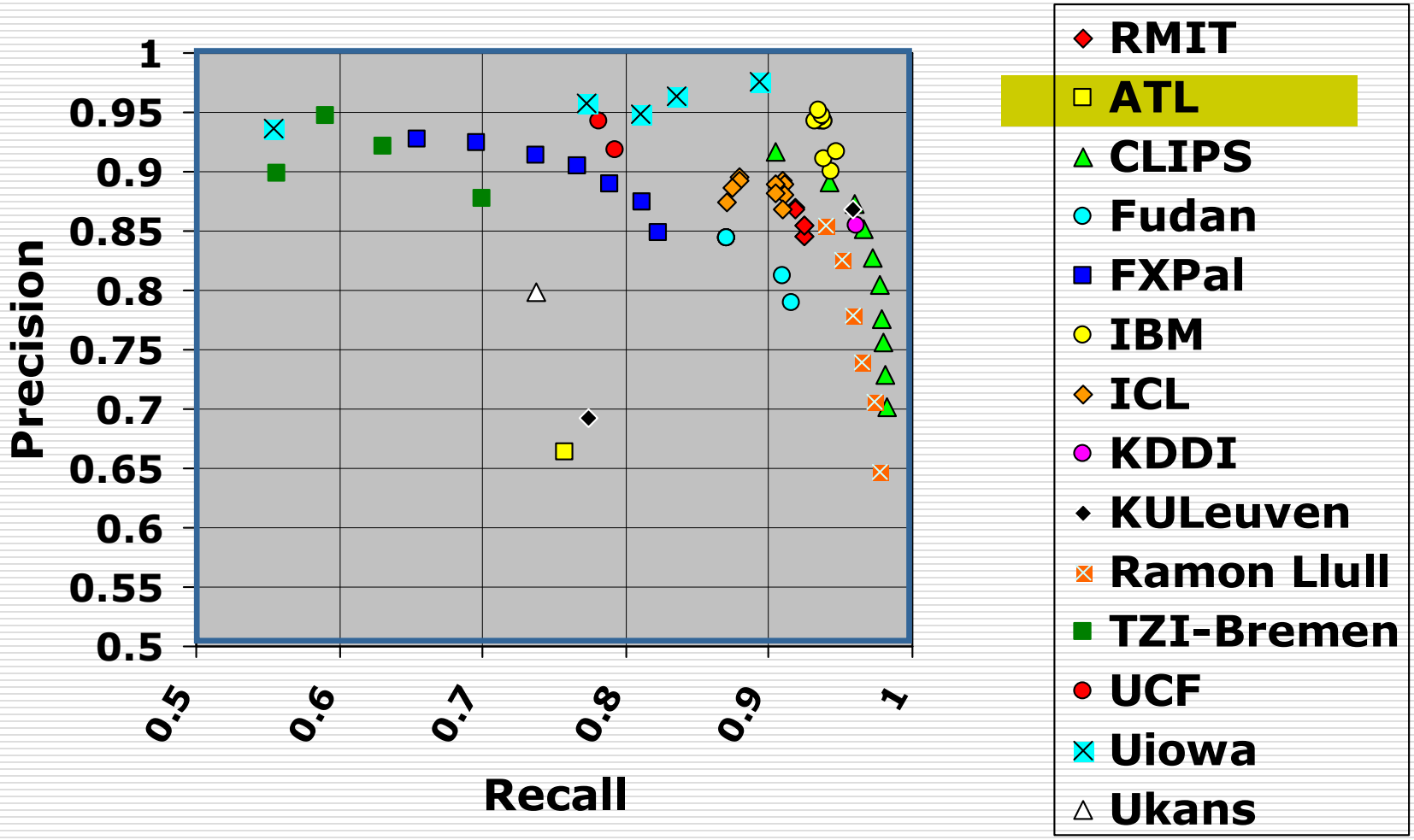
Extract I-frames from encoded stream;

Compute 3 Chi-square values across 3 separate histograms … global intensity, row intensity and column intensity and apply threshold, then combine;

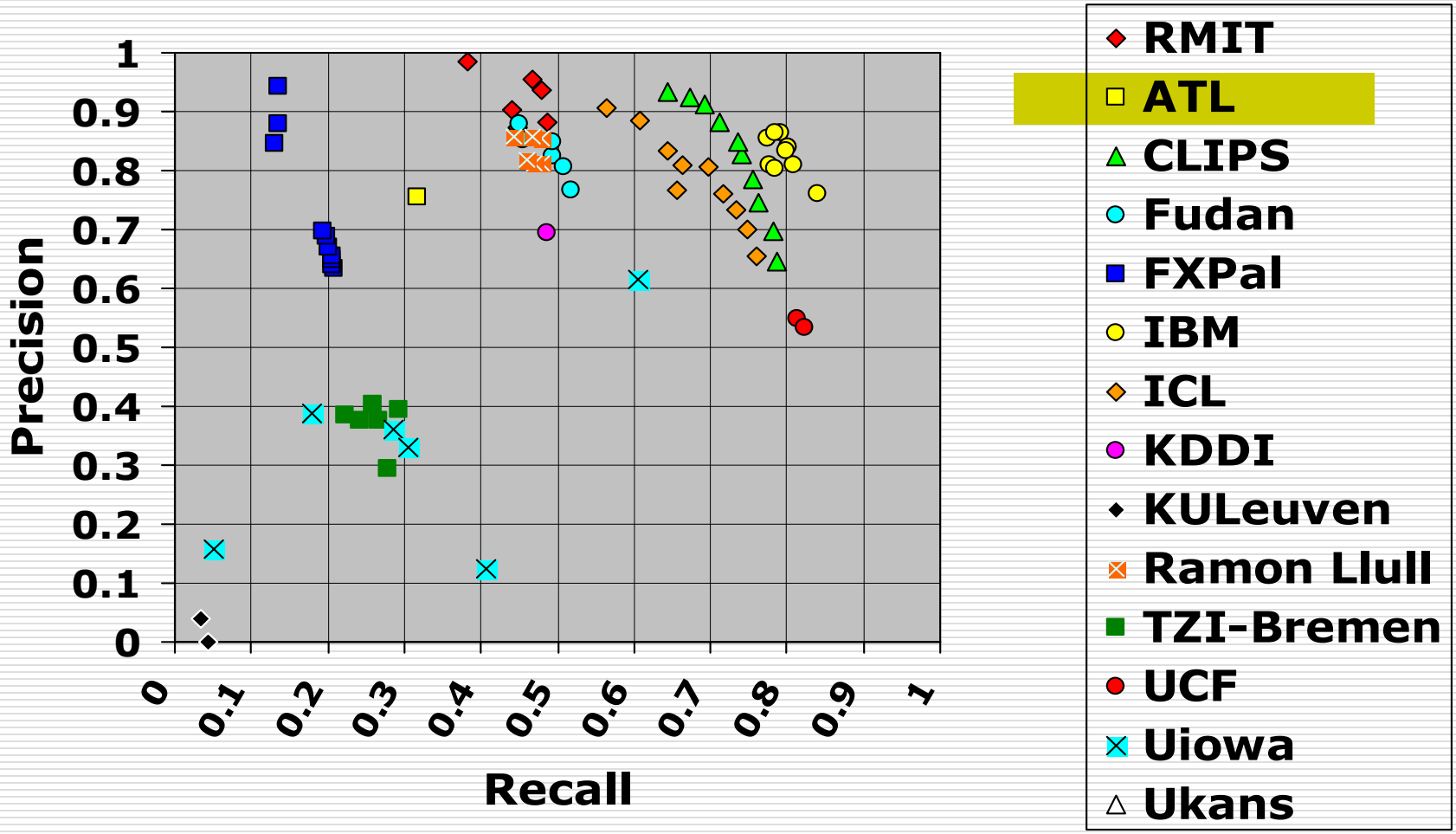
This gives indicator location and is followed by frame decoding and fine-grained examination;

Univ. of Iowa (US)	X	X			
Univ. of Kansas (US)	X				
Univ. of North Carolina (US)				X	
Univ. Oulu/VTT (FI)			X	X	

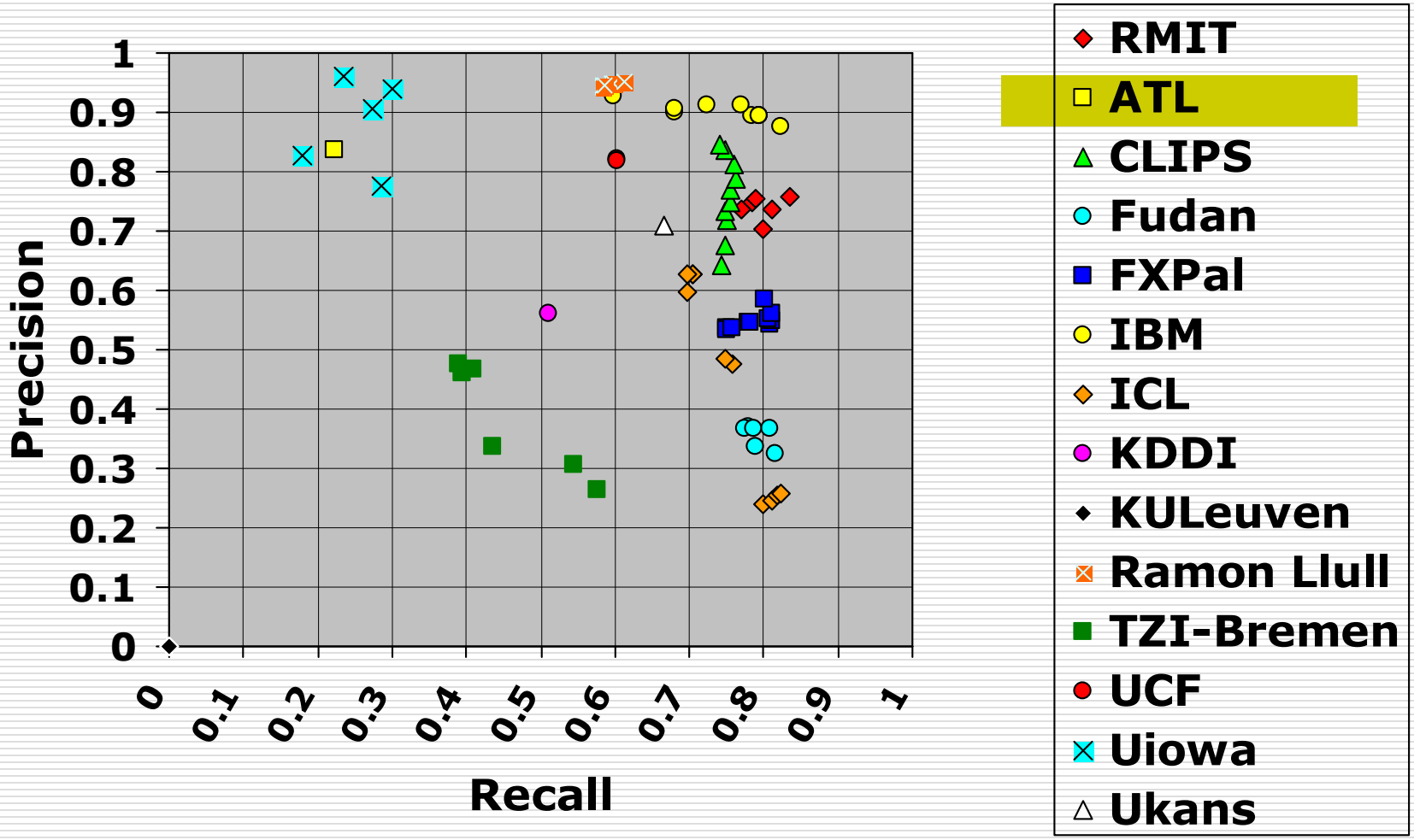
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

CLIPS-IMAG:

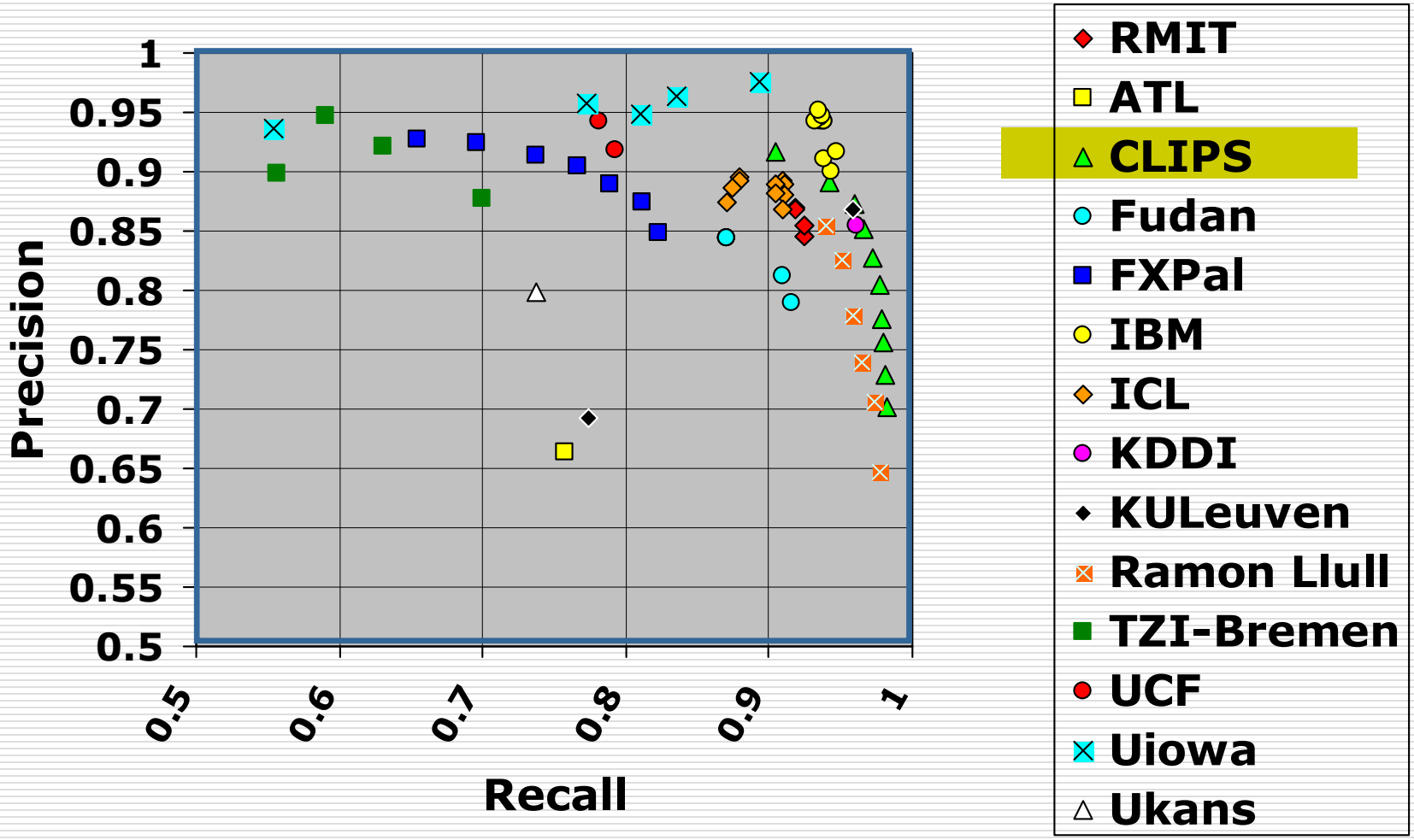
Based on image differences with motion compensation which uses optical flow as a pre-process and direct detection of dissolves;

Same as used in TV2001 and TV2002 with little modification;

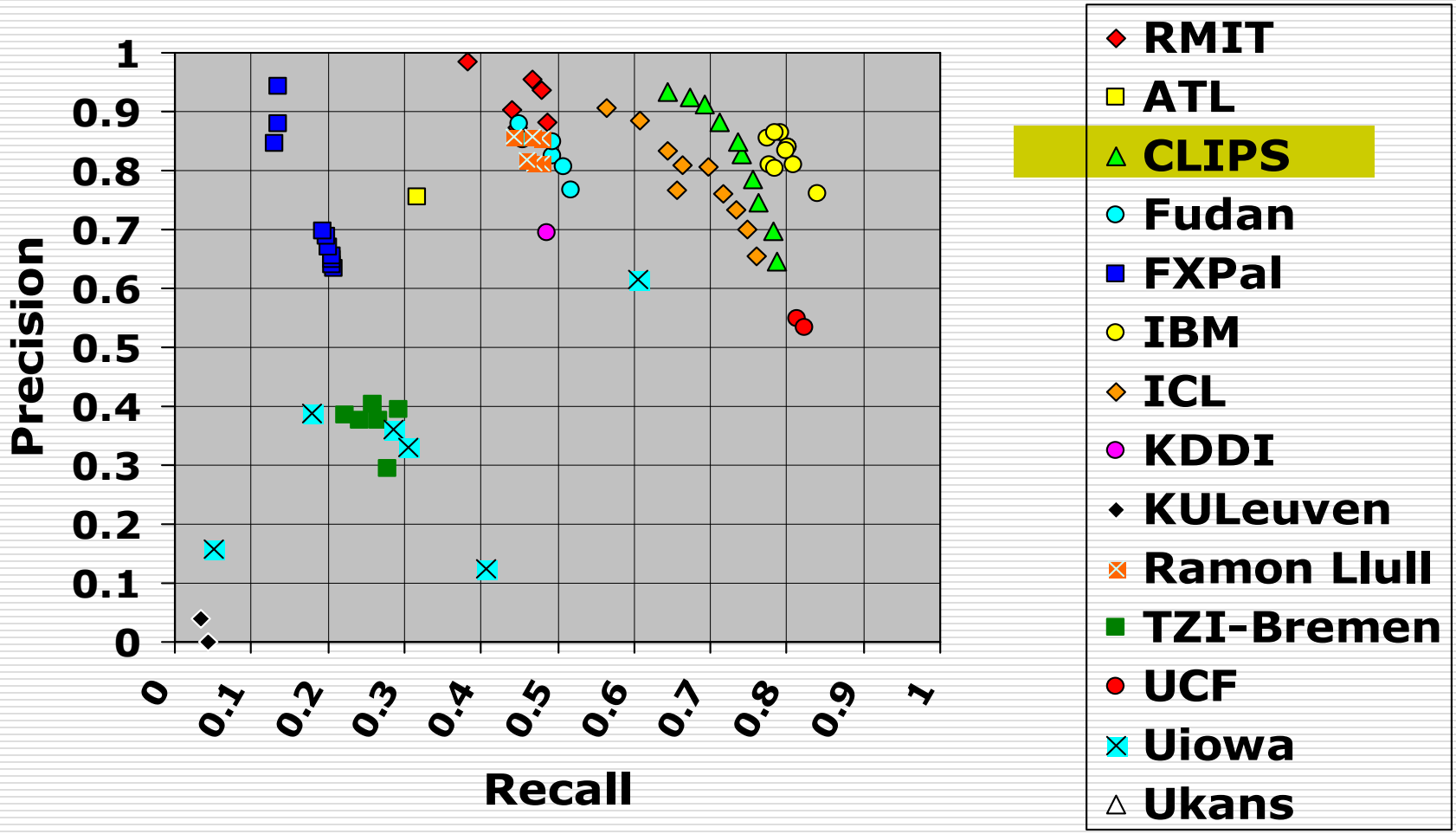
Also includes direct detection of camera flashes;

StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

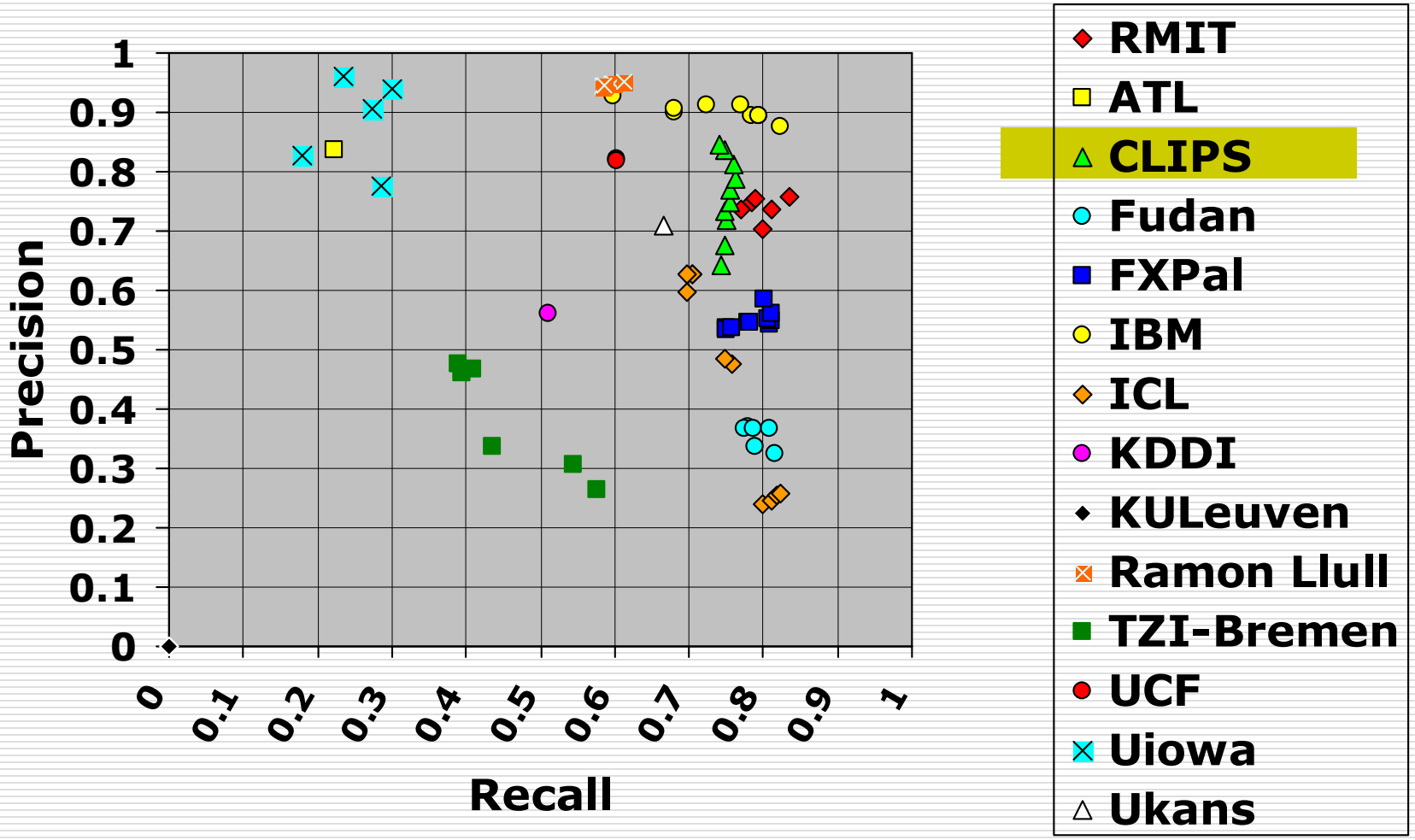
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

Fudan University:

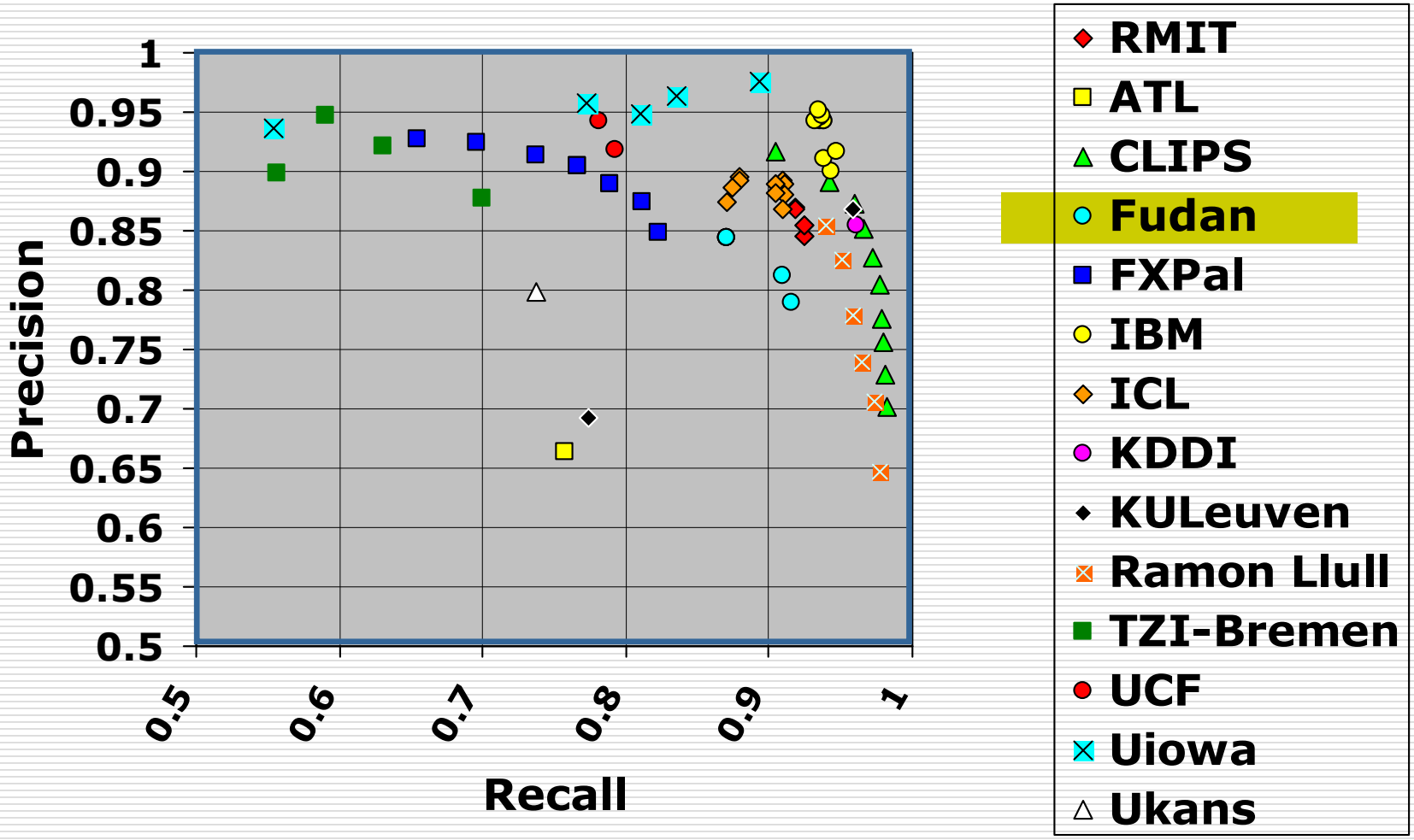
Reused TV2002 SBD approach based on frame-frame comparison using luminance difference and colour histogram similarity;

Adaptive thresholding

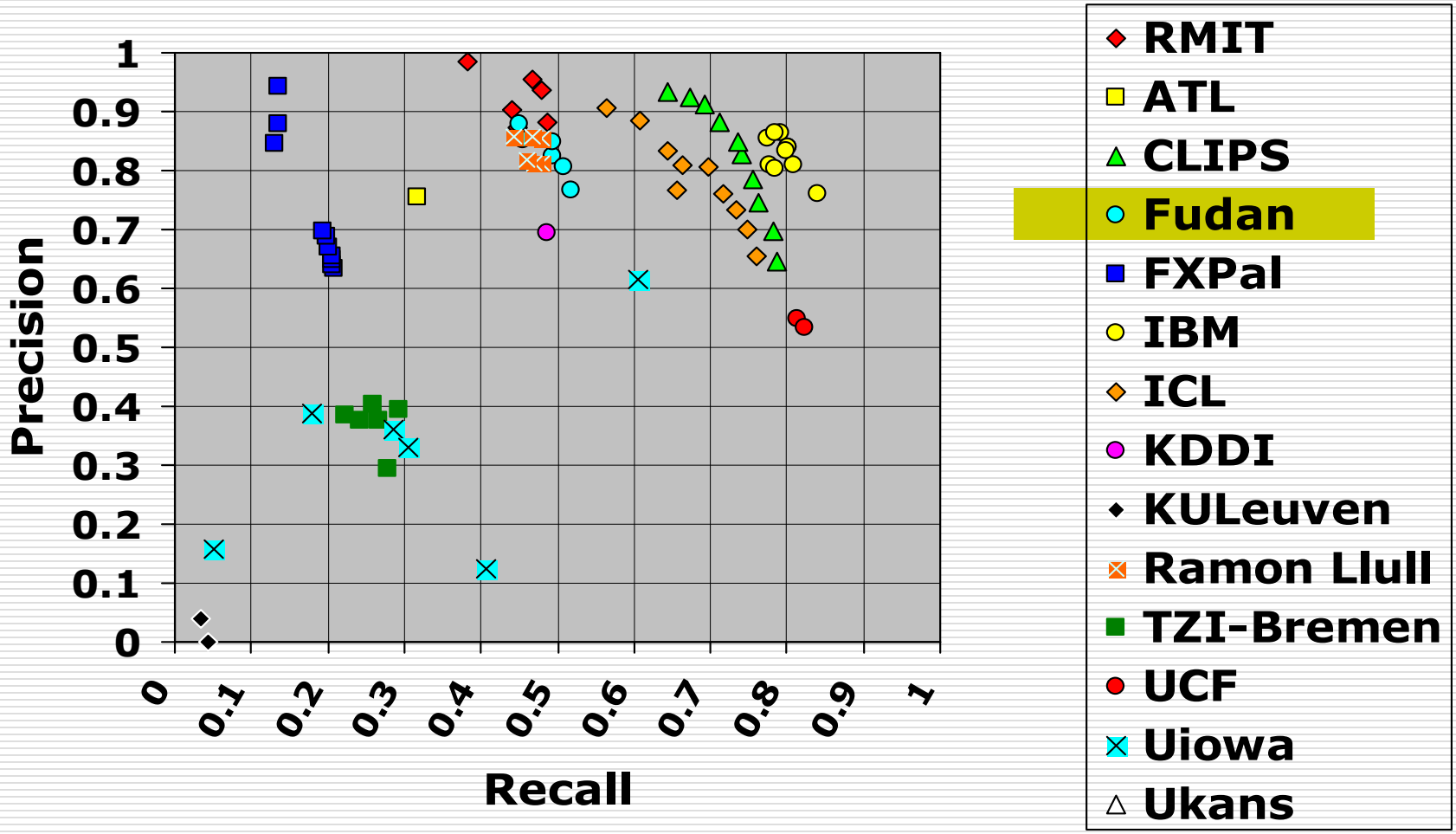
Detection of camera flashes;

GTs are searched seeking a black frame to determine whether they are fades, else dissolves;

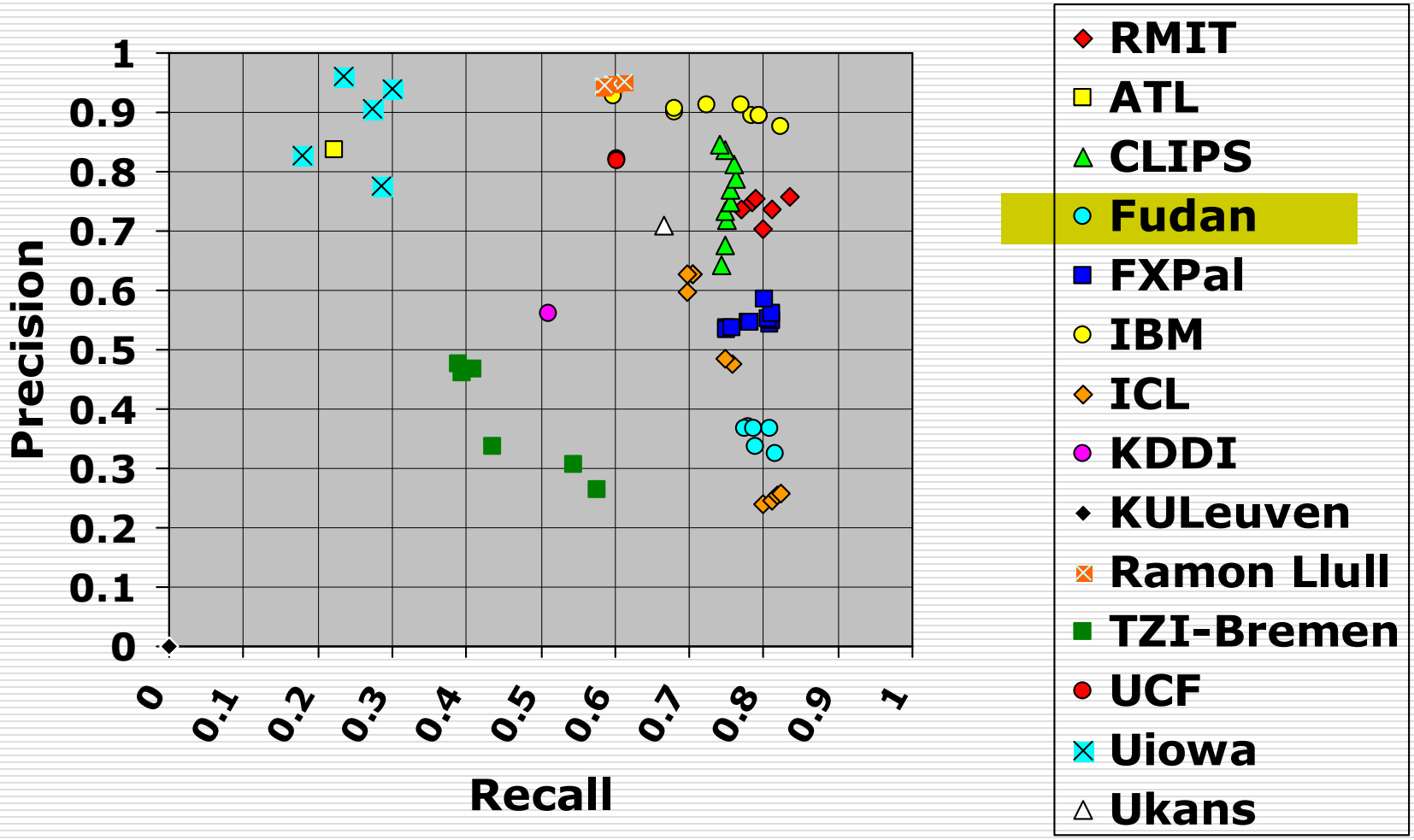
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

FX-PAL:

For each frame compute self-similarity against all in a window of past and future frames, as well as cross-similarity between past & future frames;

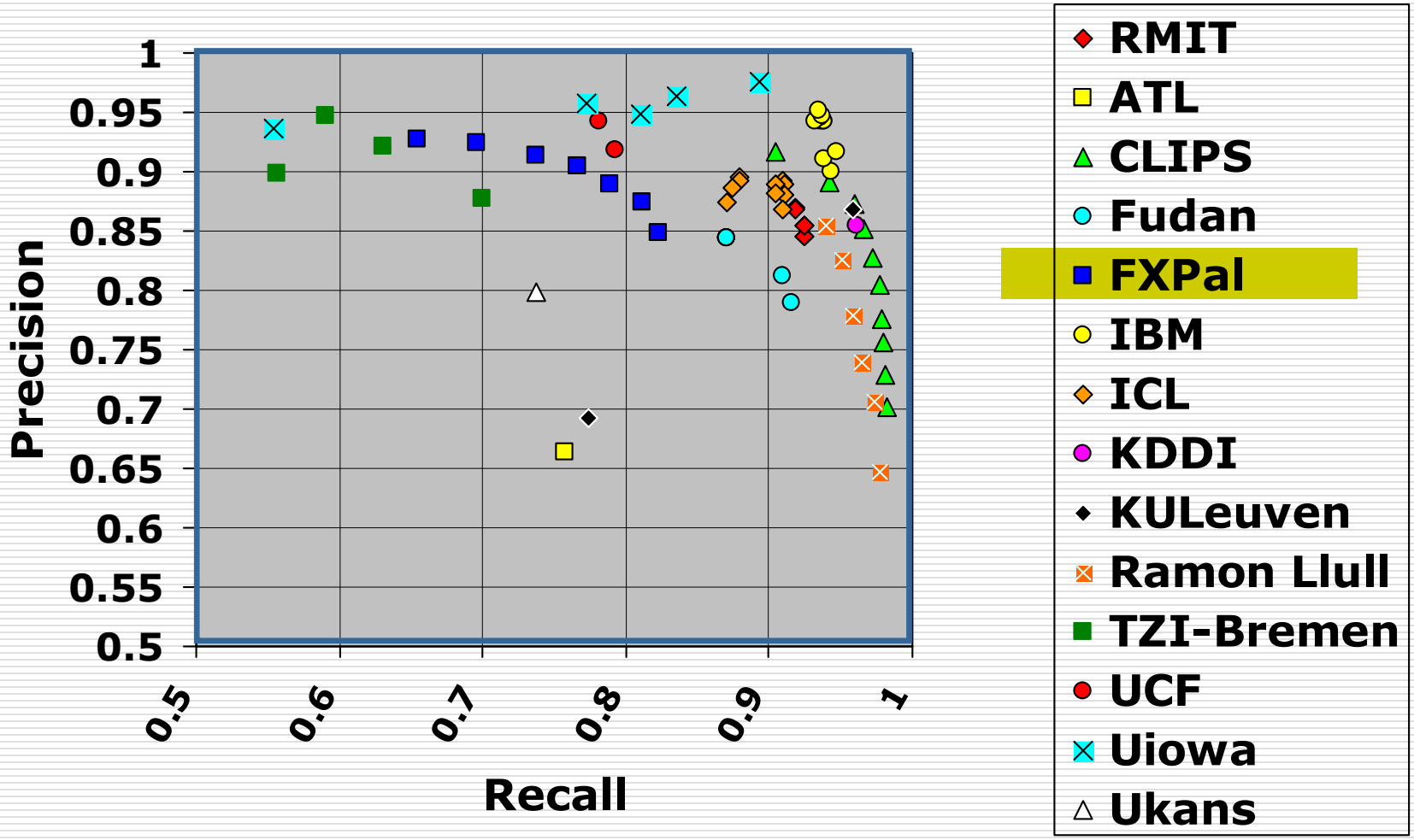
Generates a similarity matrix and examine characteristics of this matrix to indicate cuts and GTs;

Includes a clever way to reduce computation costs;

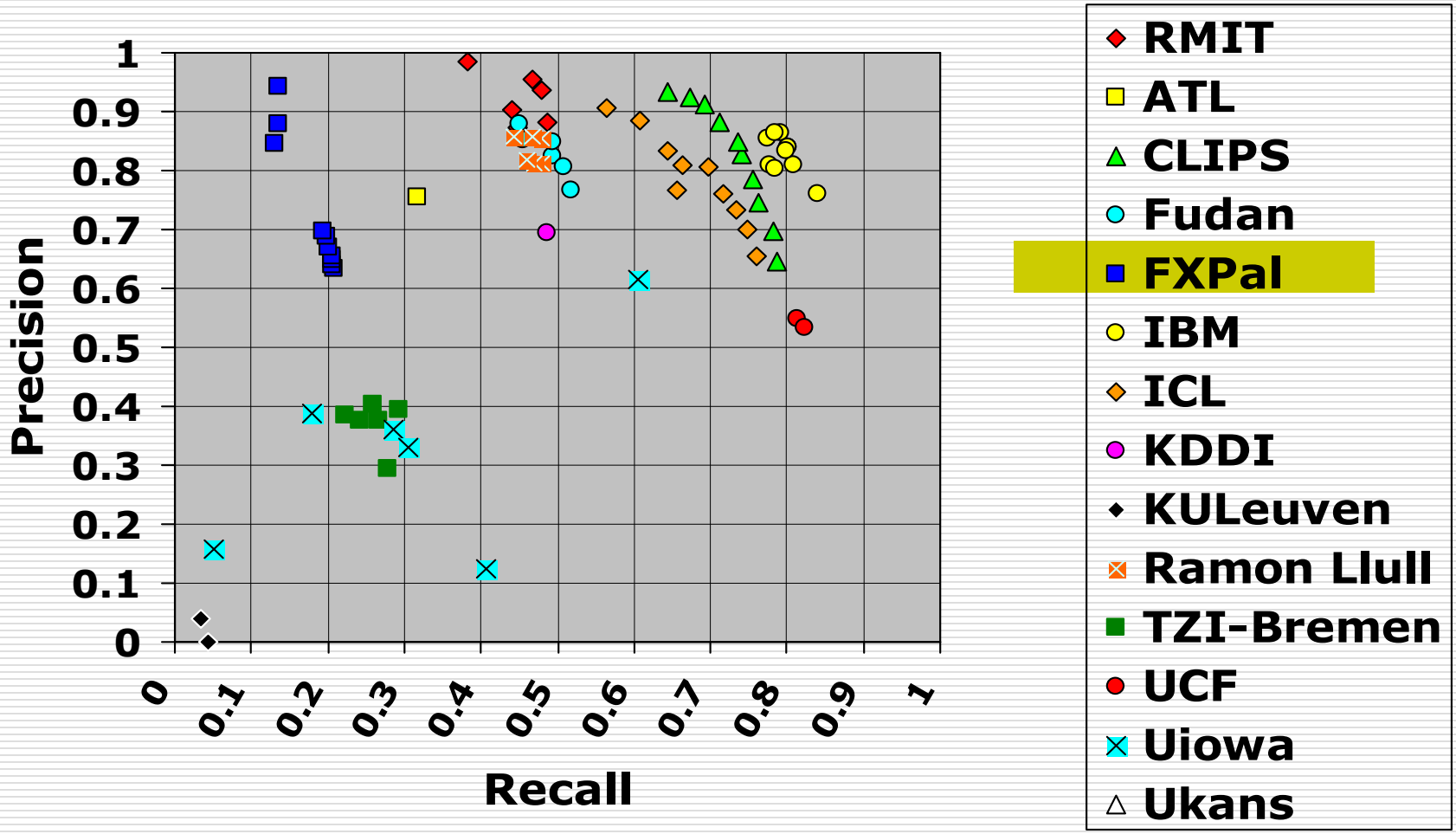
Presentation to follow;

Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

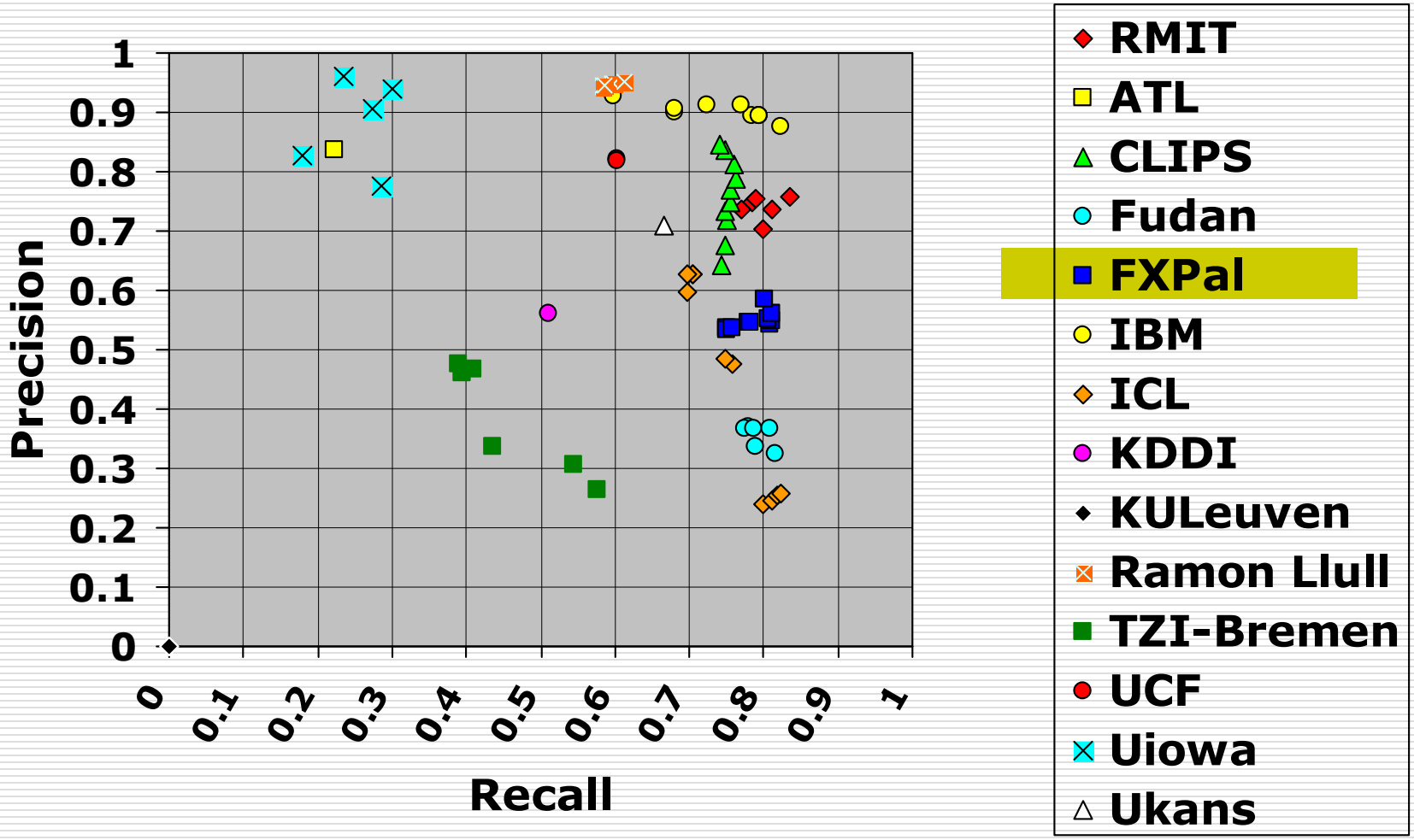
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

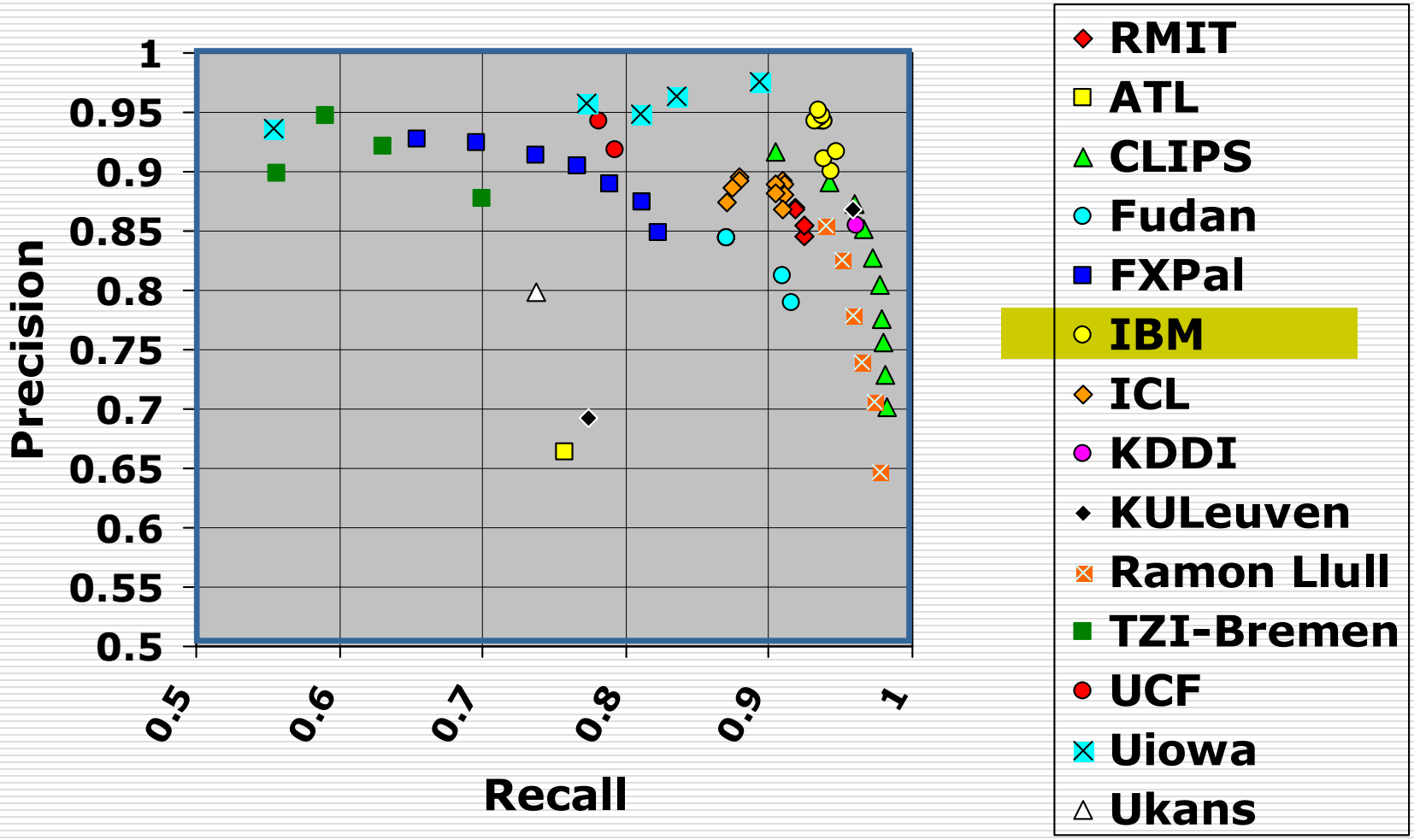
IBM Research:

Used SBD from CueVideo system

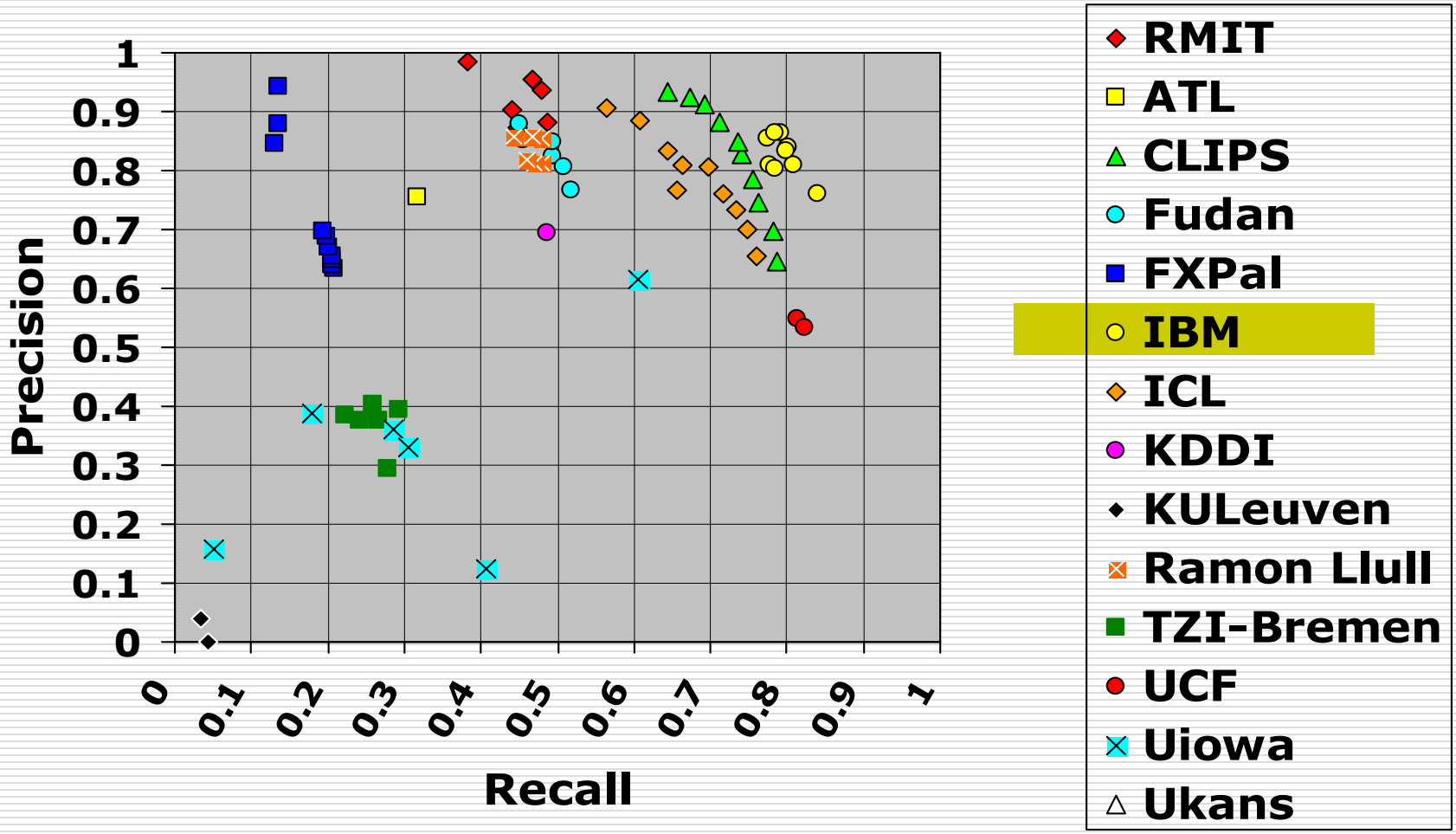
Presentation to follow

KDDI (JP)	X	X		
KU Leuven (BE)	X			
Mediamill/U Amsterdam (NL)				X
National Univ. Singapore (Sing.)		X		X
Ramon Llull Univ. (ES)	X			
RMIT University (Aus)	X			
StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

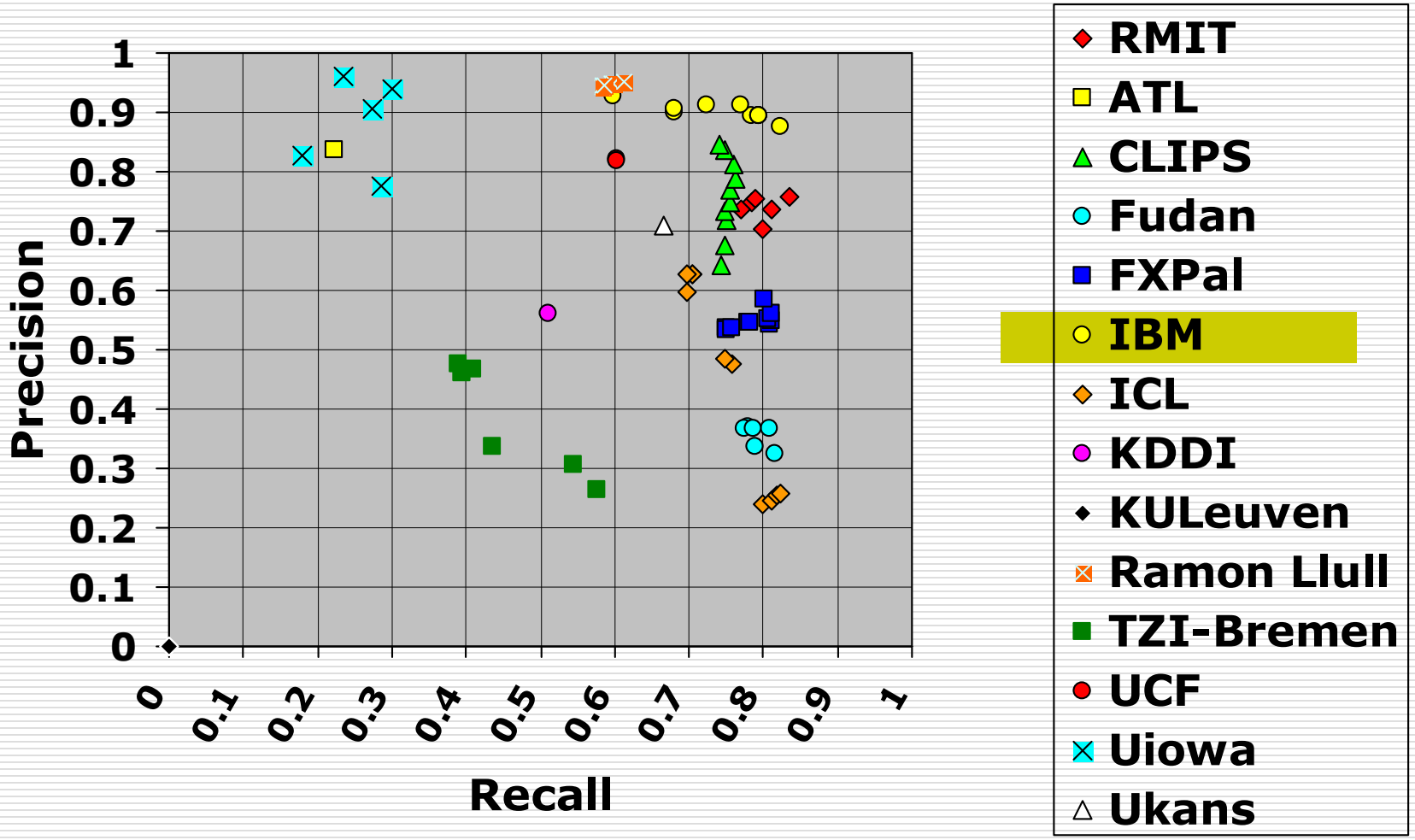
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

Imperial College London:

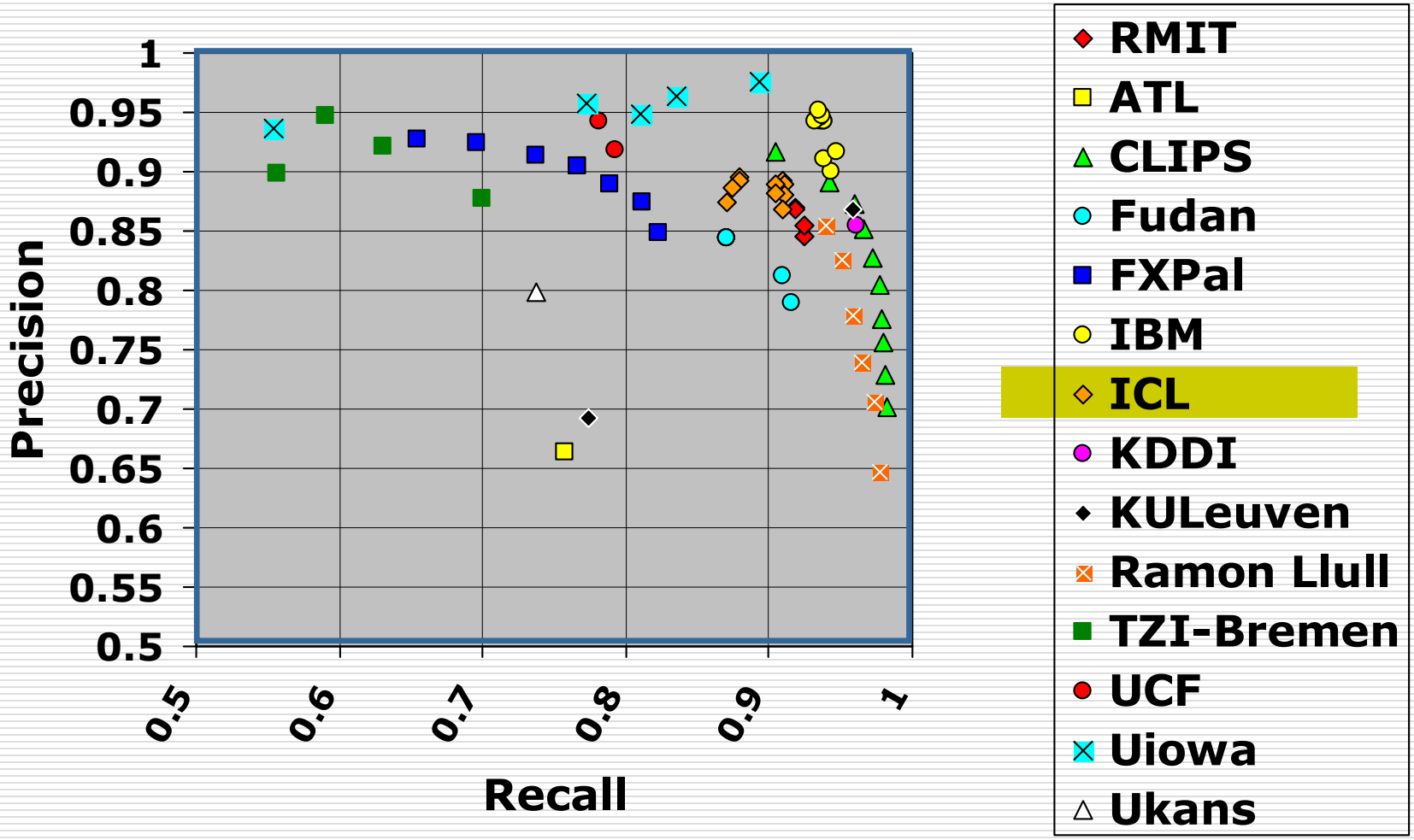
Colour histogram similarity of adjacent frames with a constant similarity threshold;

Same as TV2002 and showing tradeoff of P vs. R as threshold varies;

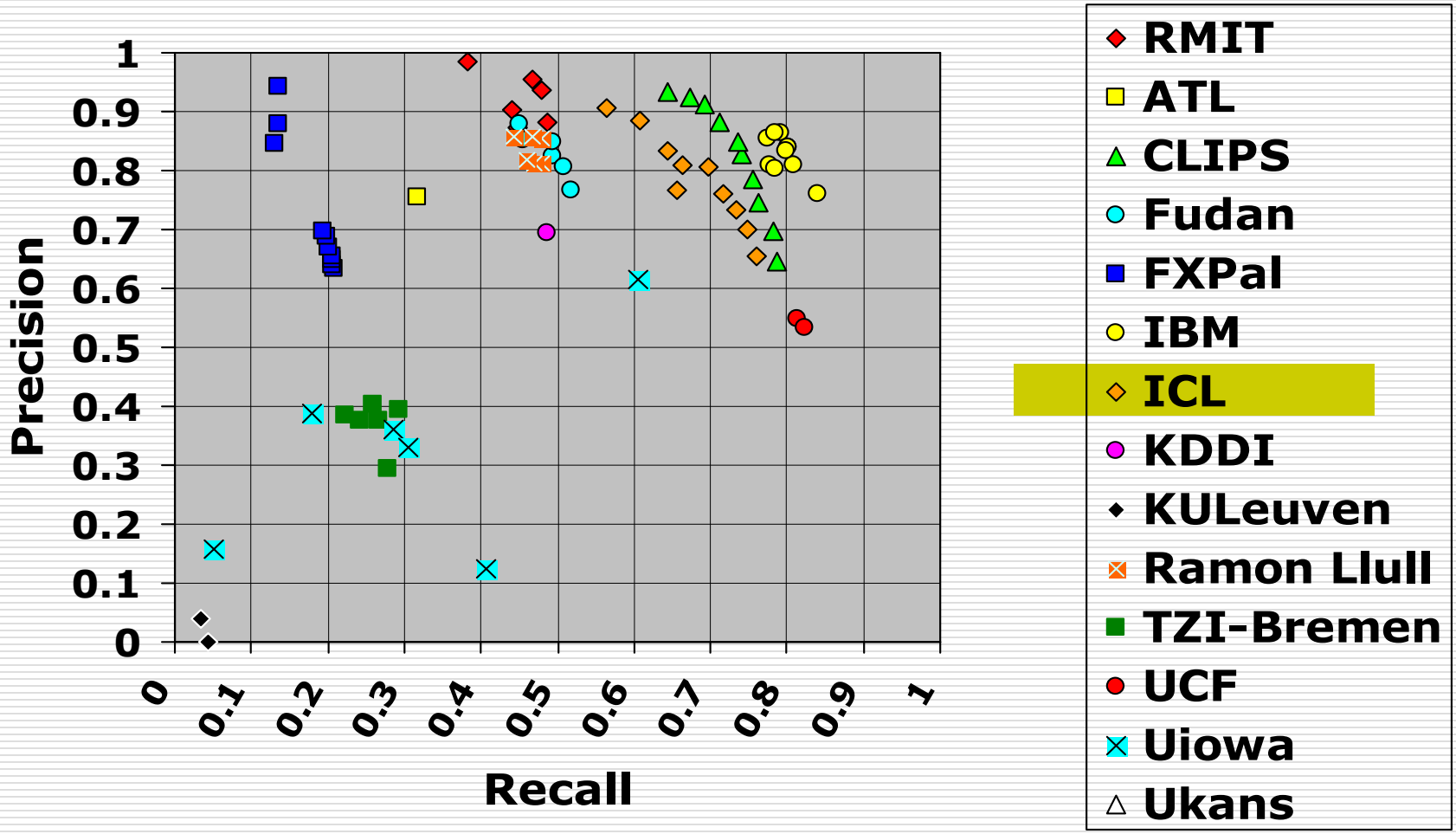
Good performance for simple approach;

StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

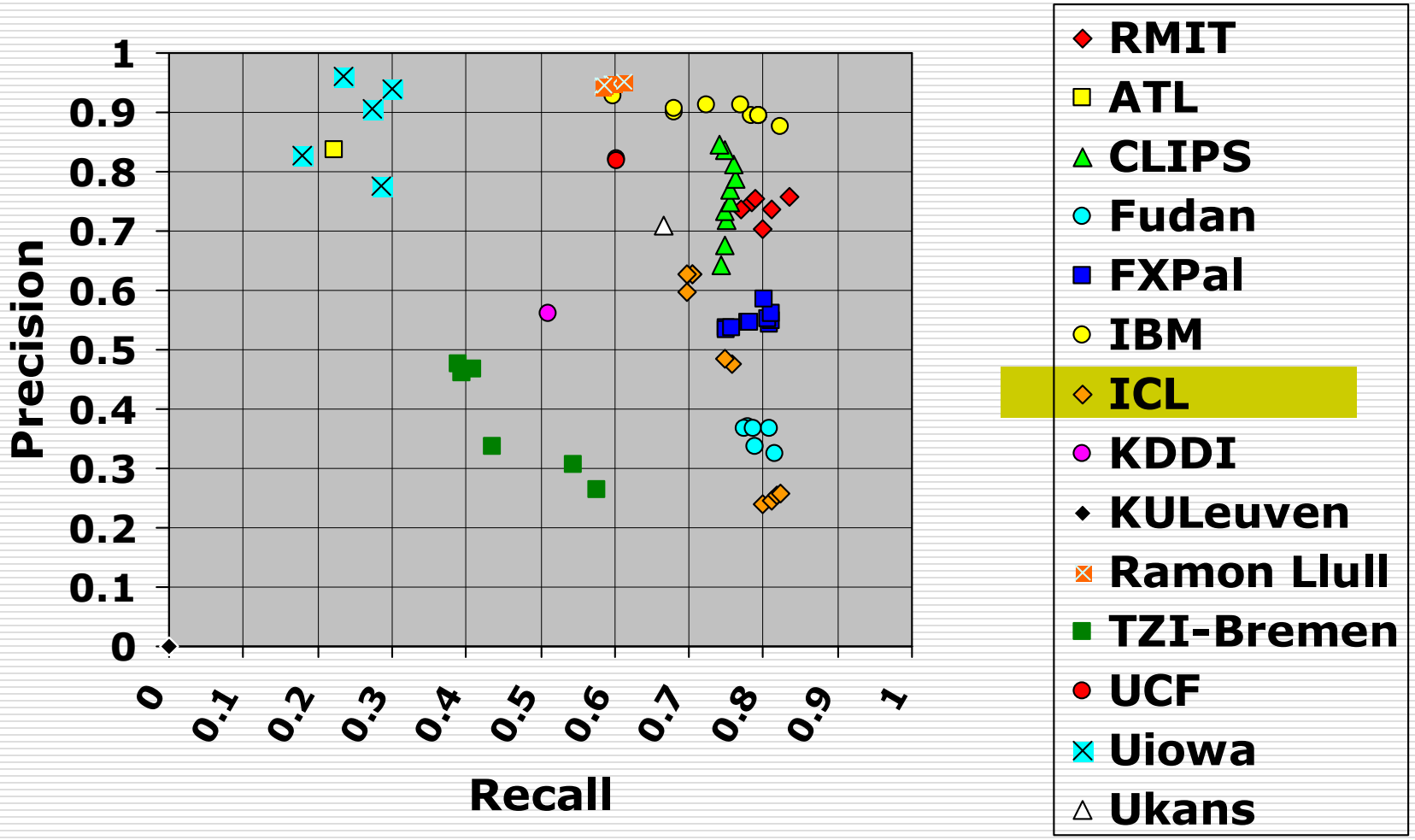
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

KDDI:

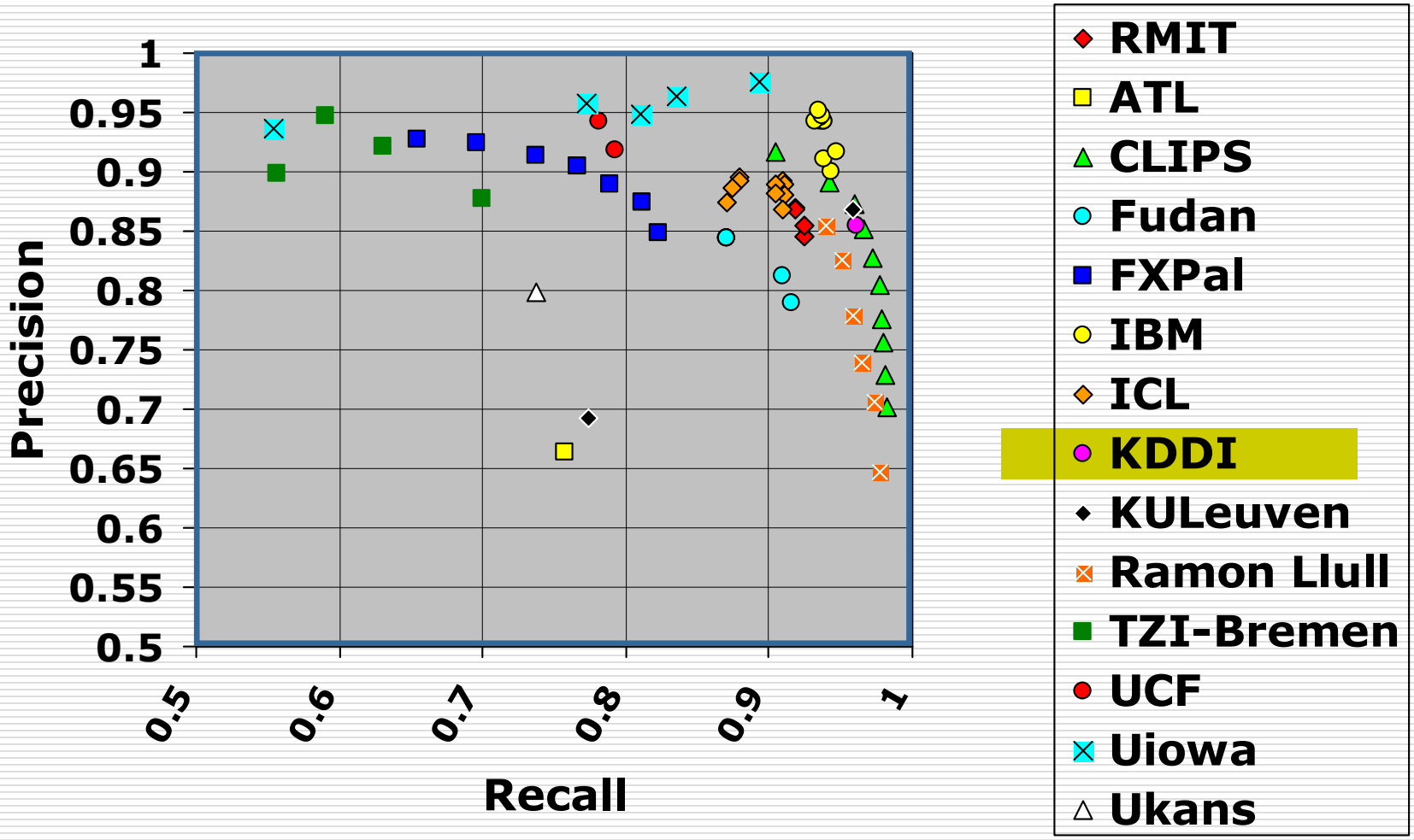
For cuts, preprocess the encoded MPEG-1 stream to locate high inter-frame differences using motion vectors then decode likely frames and test for luminance and chrominance differences;

For dissolves, detect gradual changing over time using DCT activity data;

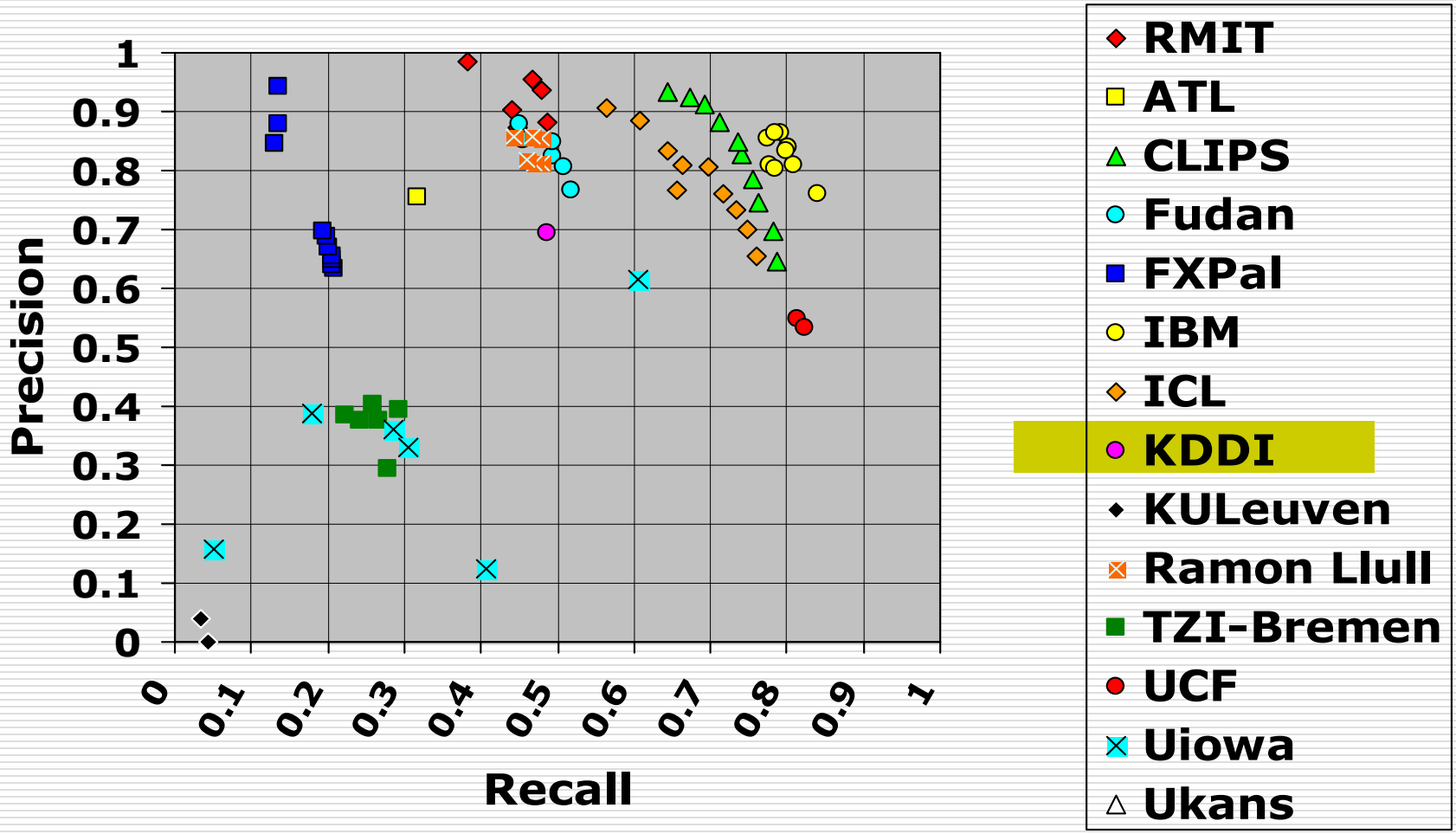
Specific detection looking for wipes, and for camera flashes;

Because it processes encoded stream, 24x real time on PC;

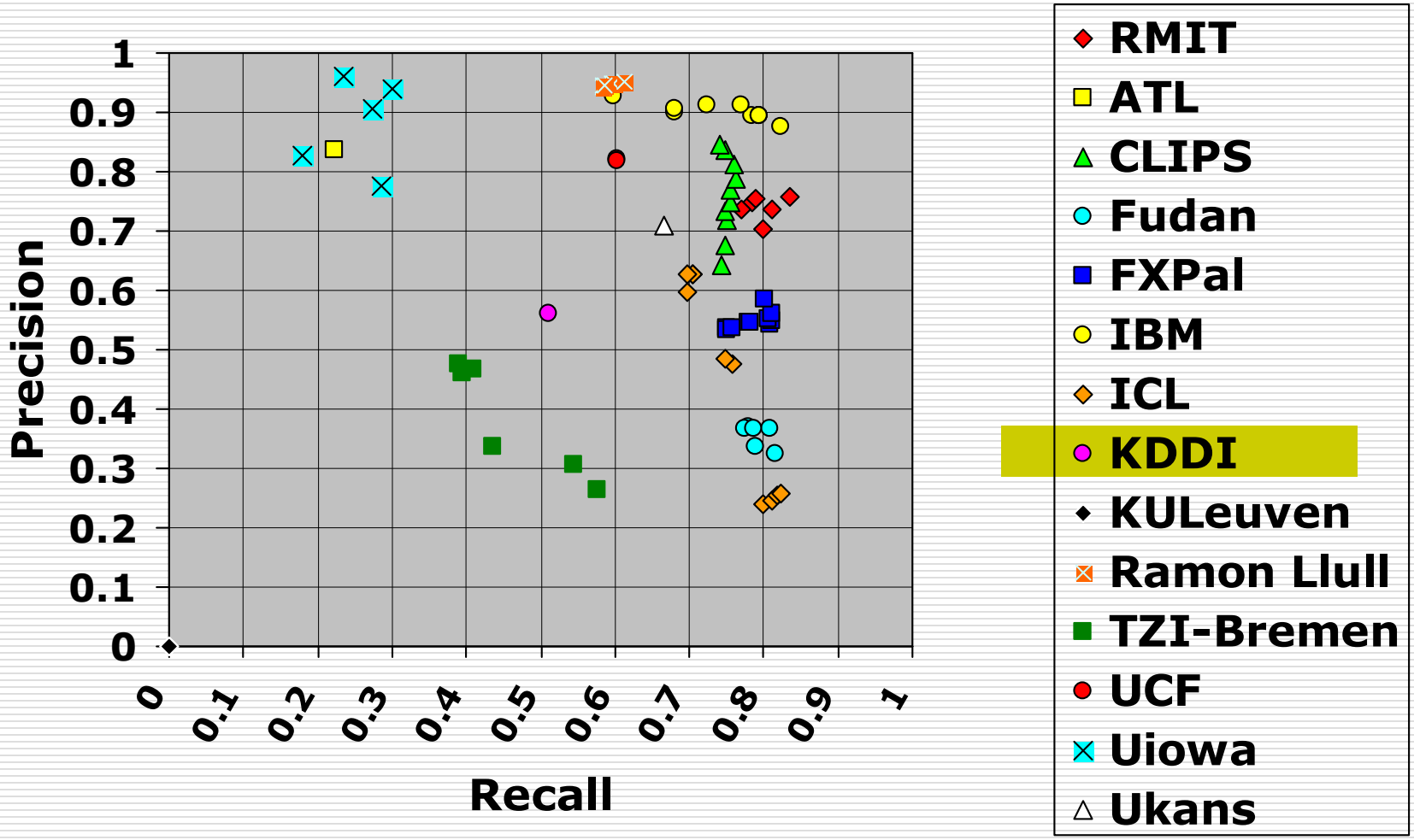
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

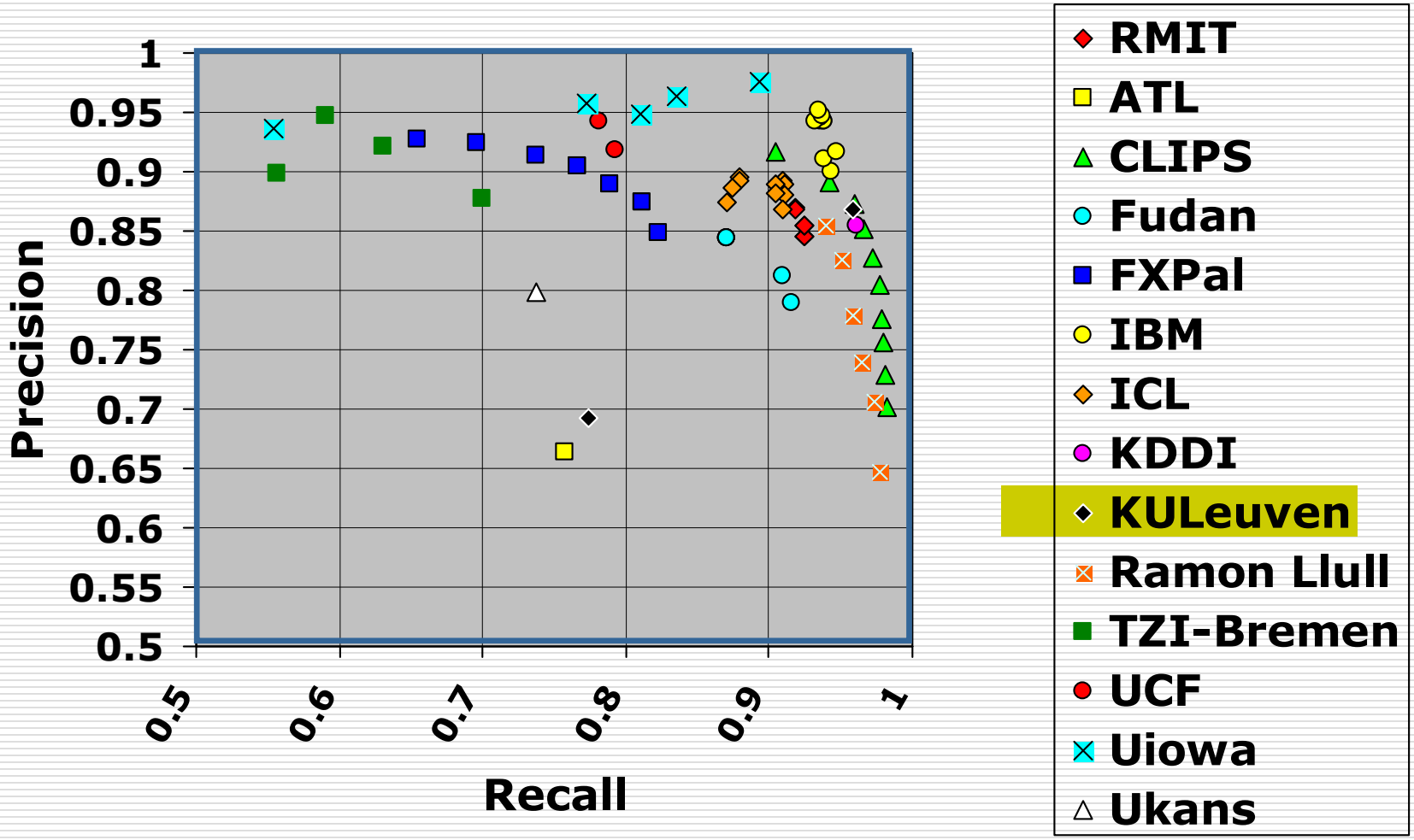
KU Leuven:

Adaptive thresholding on the average intensity differences between adjacent frames;

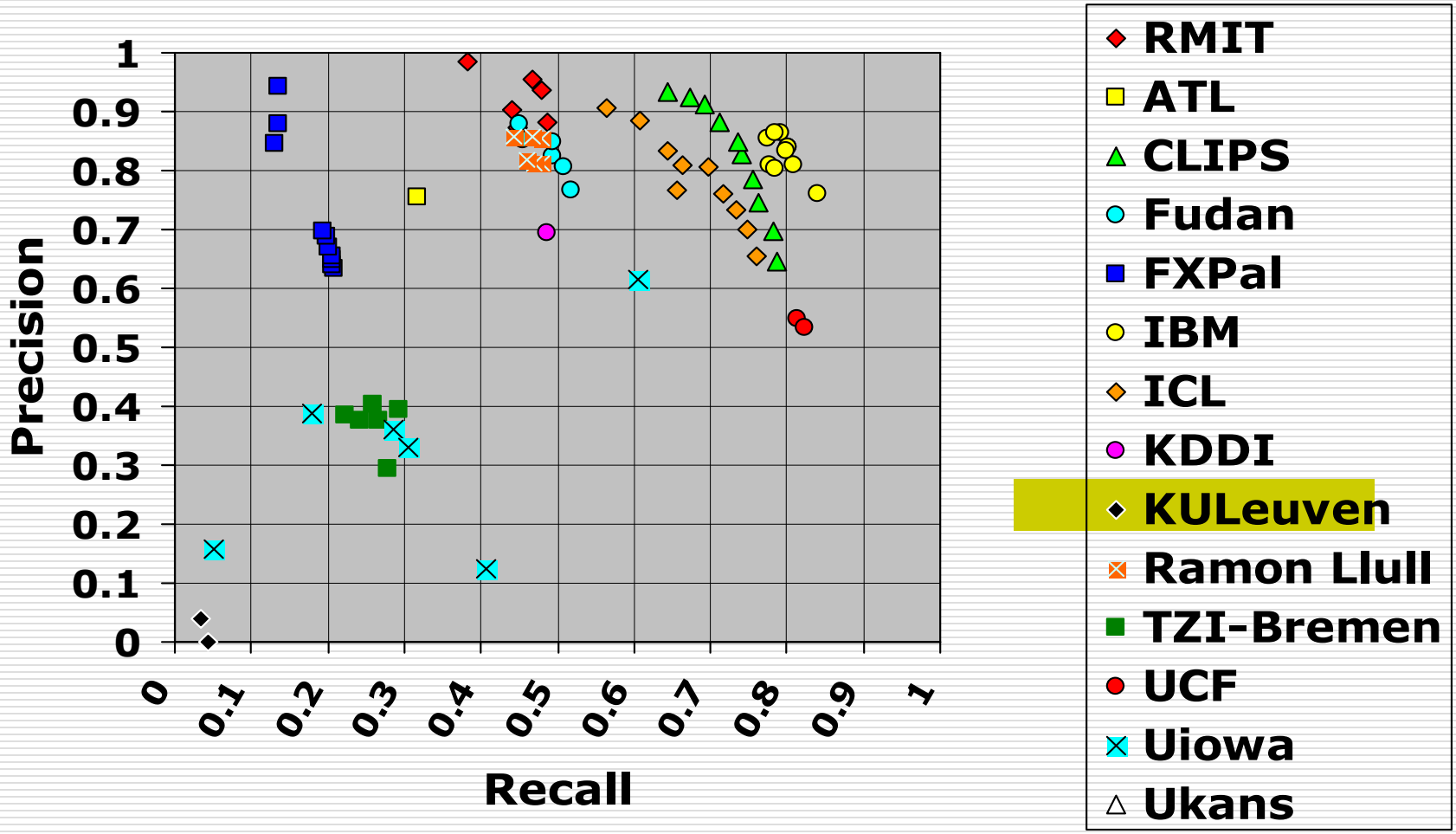
Includes motion compensation which computes an affine transformation between consecutive frames;

RMIT University (Aus)	X			
StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

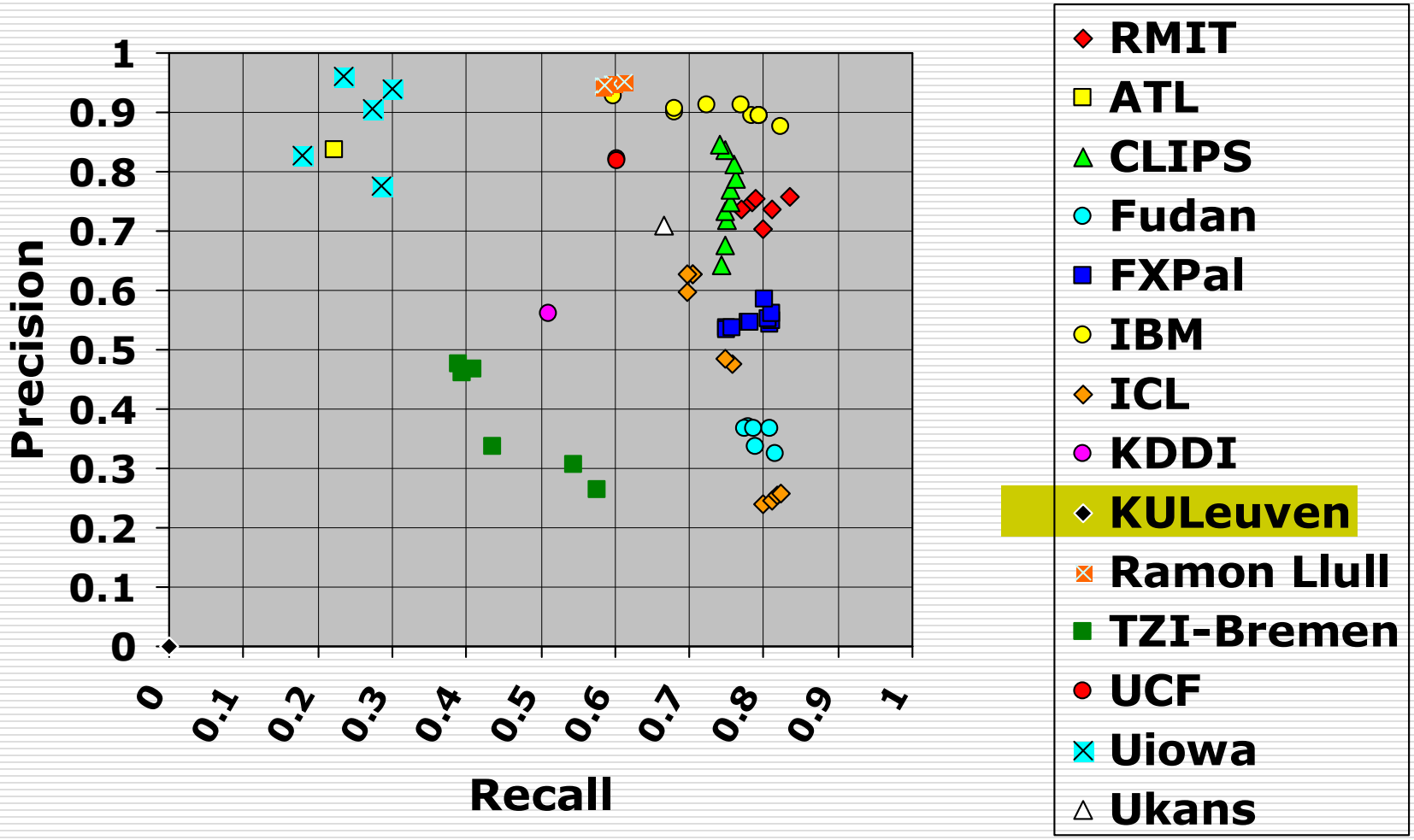
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

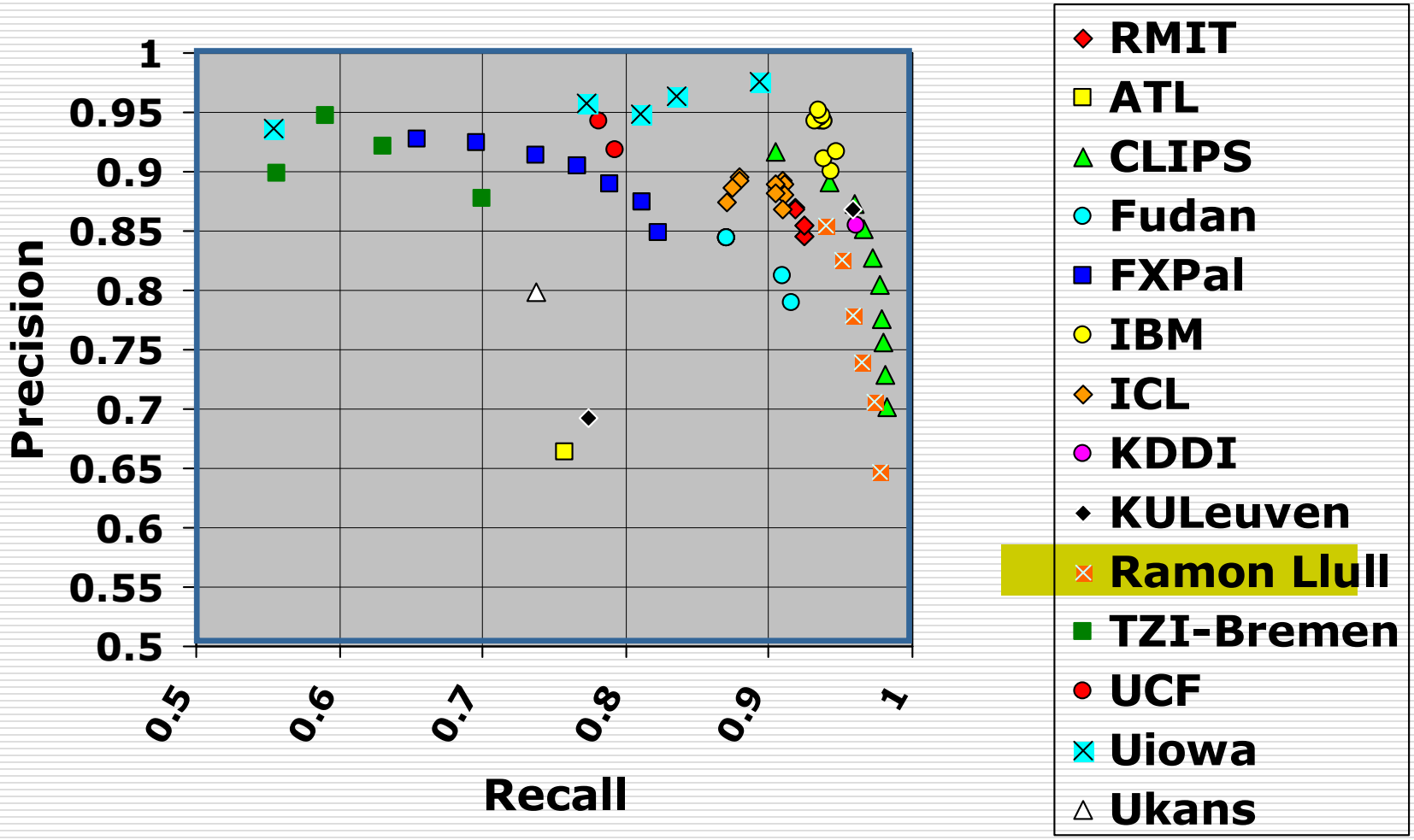
Ramon Llull University:

Global colour histogram differences as a measure of discontinuity is used to detect cuts;

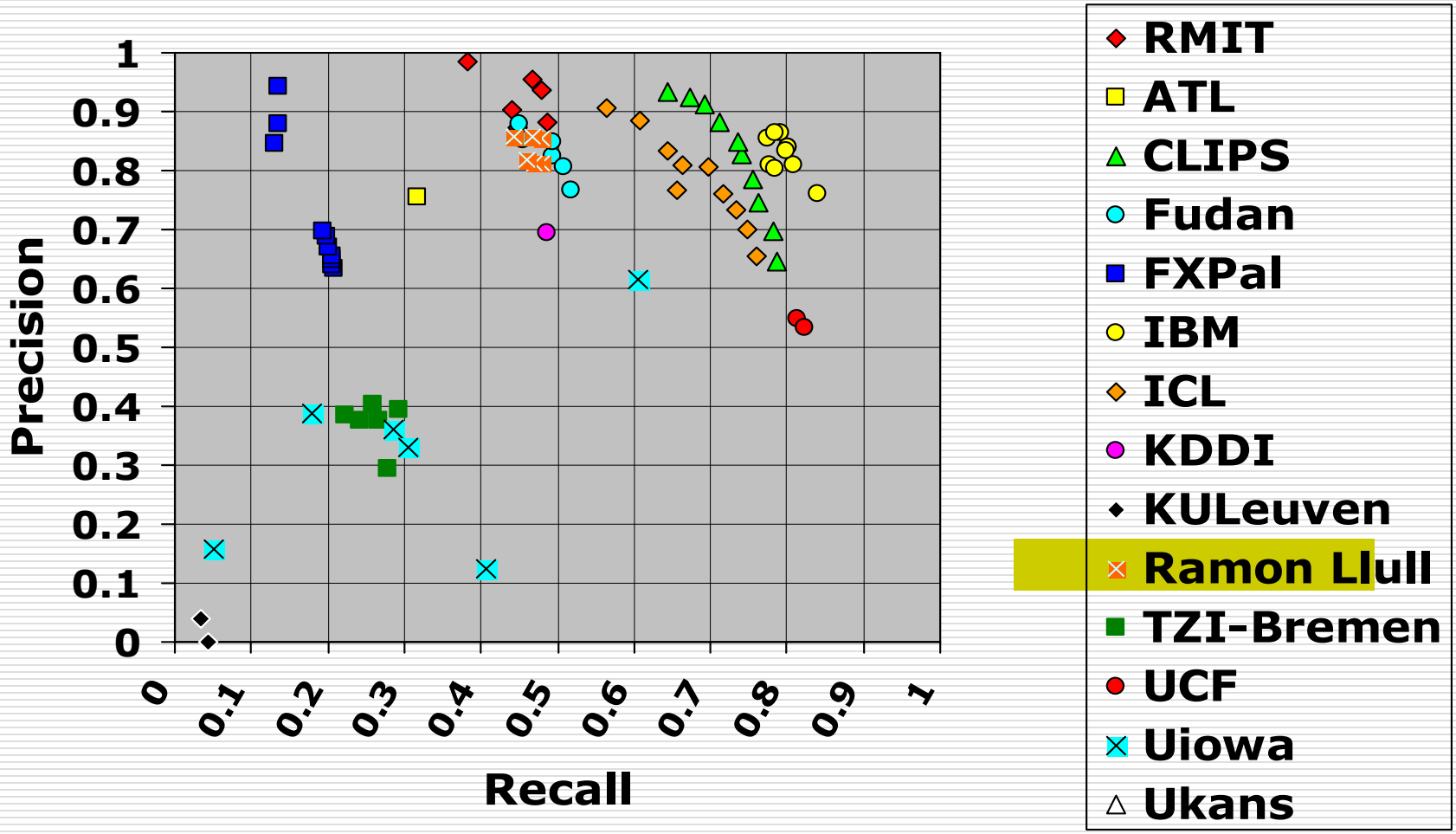
For GTs, a method to account for linear colour variation of images across the duration of the GT, with specific treatment of moving objects during the GT which can distort this;

Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

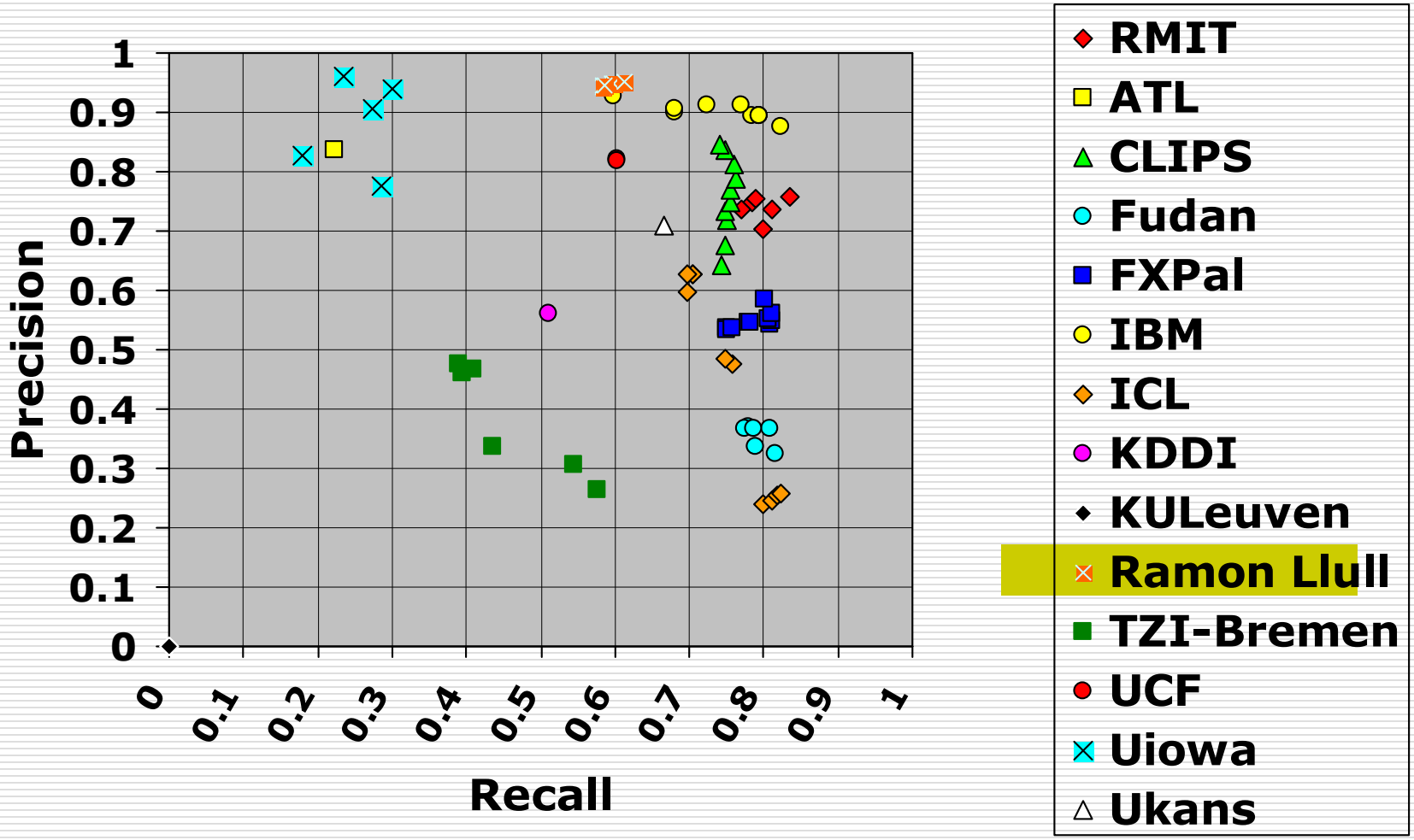
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

RMIT University:

Target GTs;

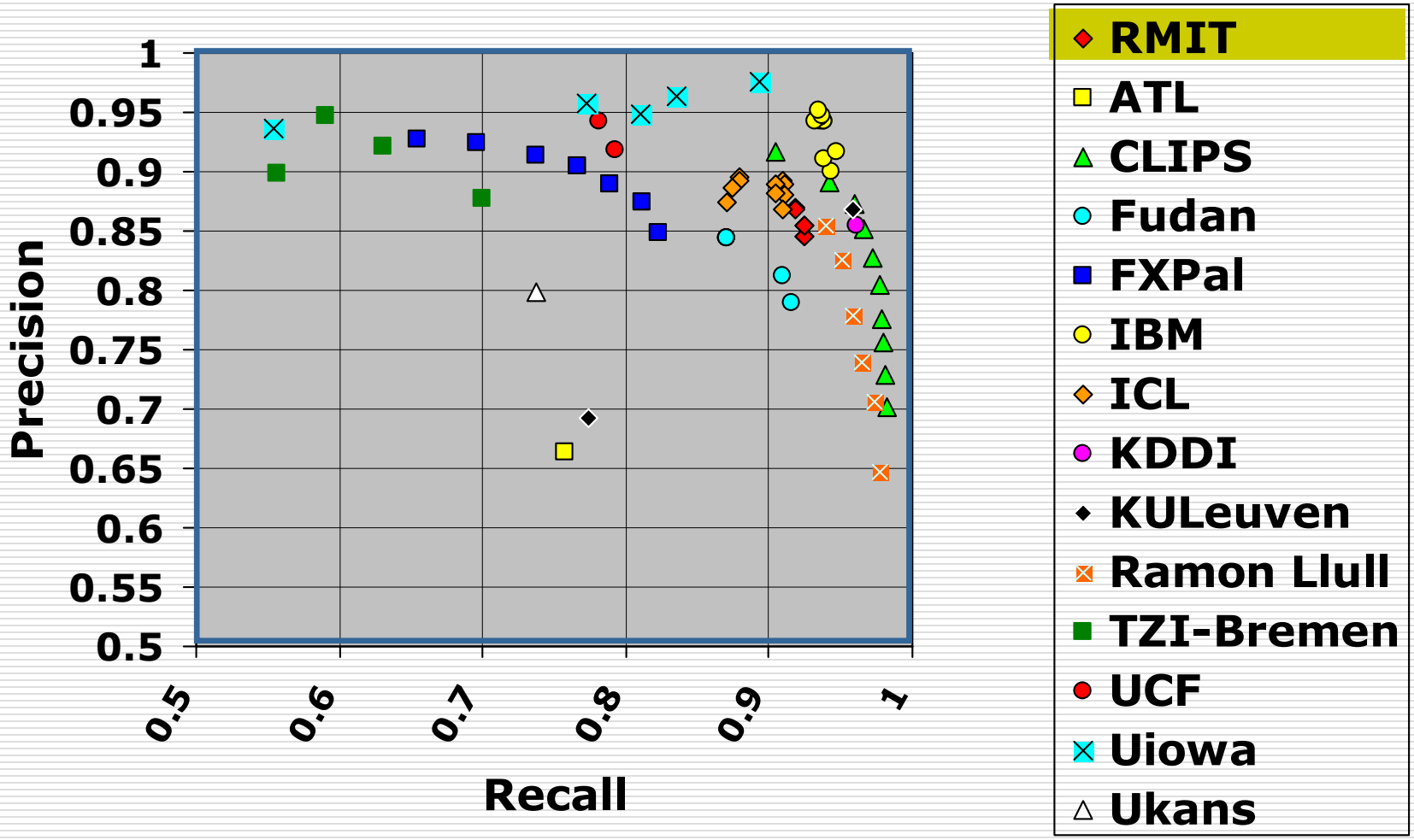
Using a moving window of (200) frames, use current frame as a QBE against all in the window with a 6-frame DMZ around current frame;

Based on frame-frame similarity and adaptive thresholding;

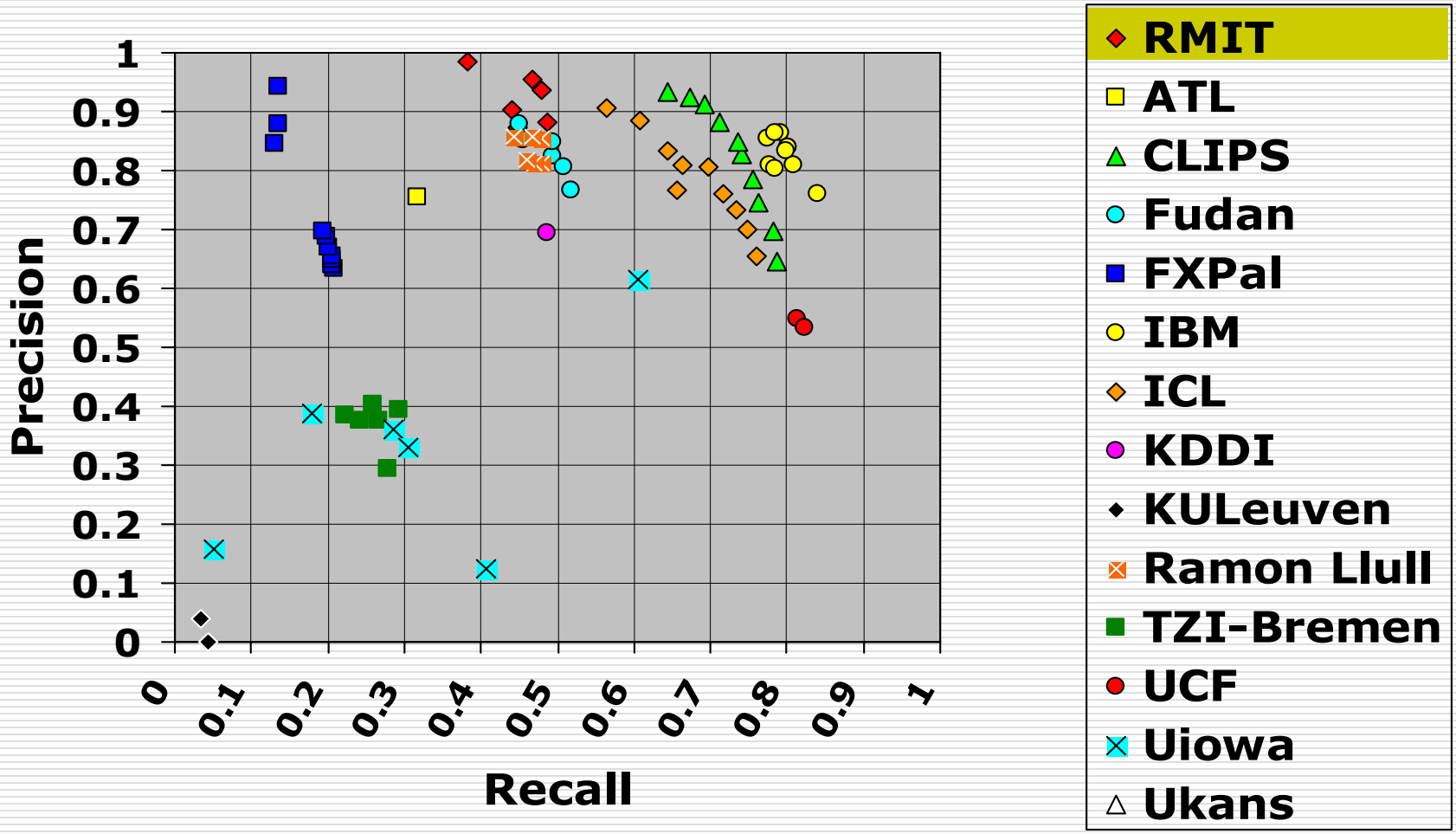
A refinement on TV2002;

Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

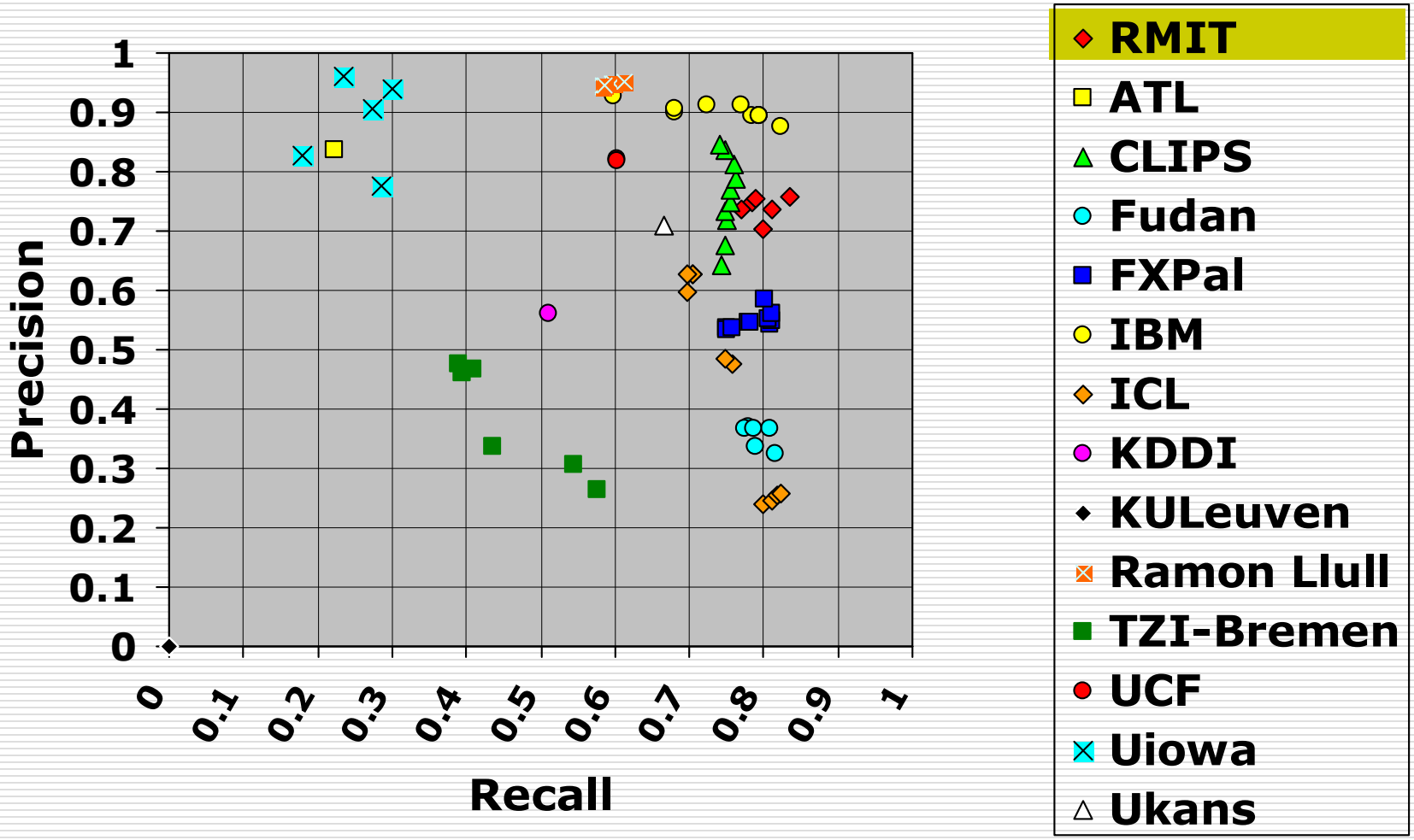
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

University of Bremen:

Combination of 3 approaches:

-changes in image luminance;

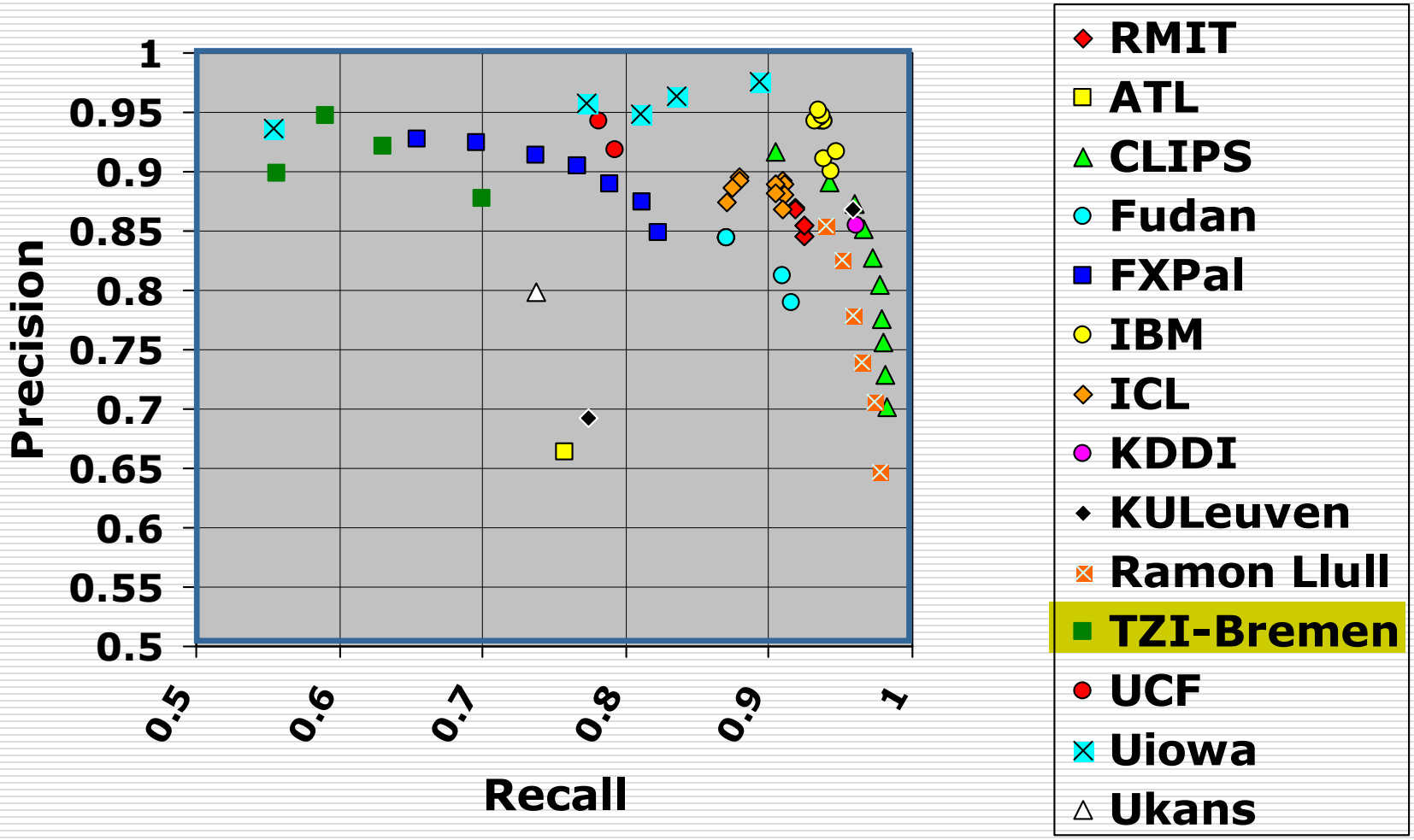
-gray level histogram differences;

-FFT feature extraction;

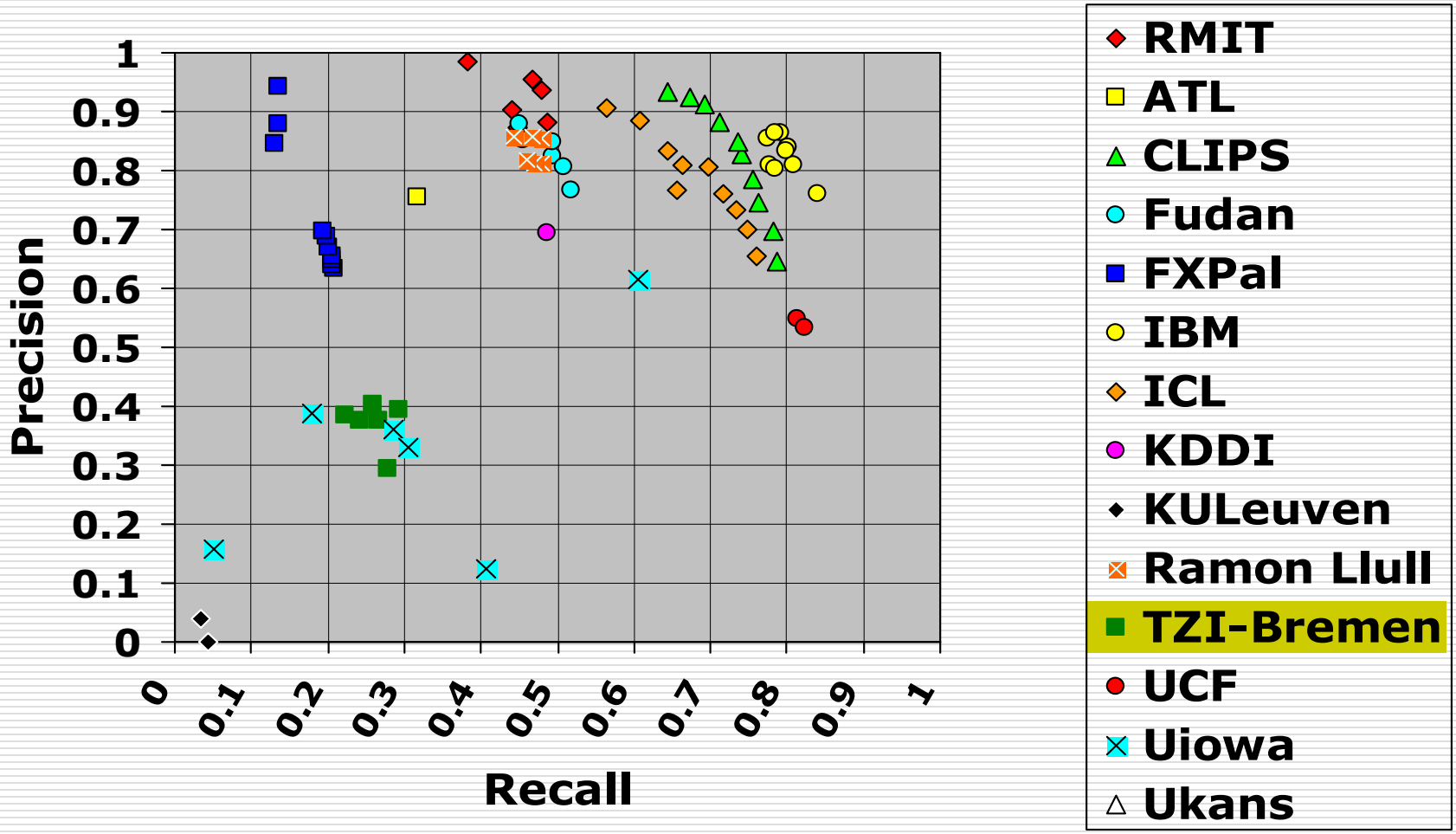
Combined, with adaptive thresholding;

StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

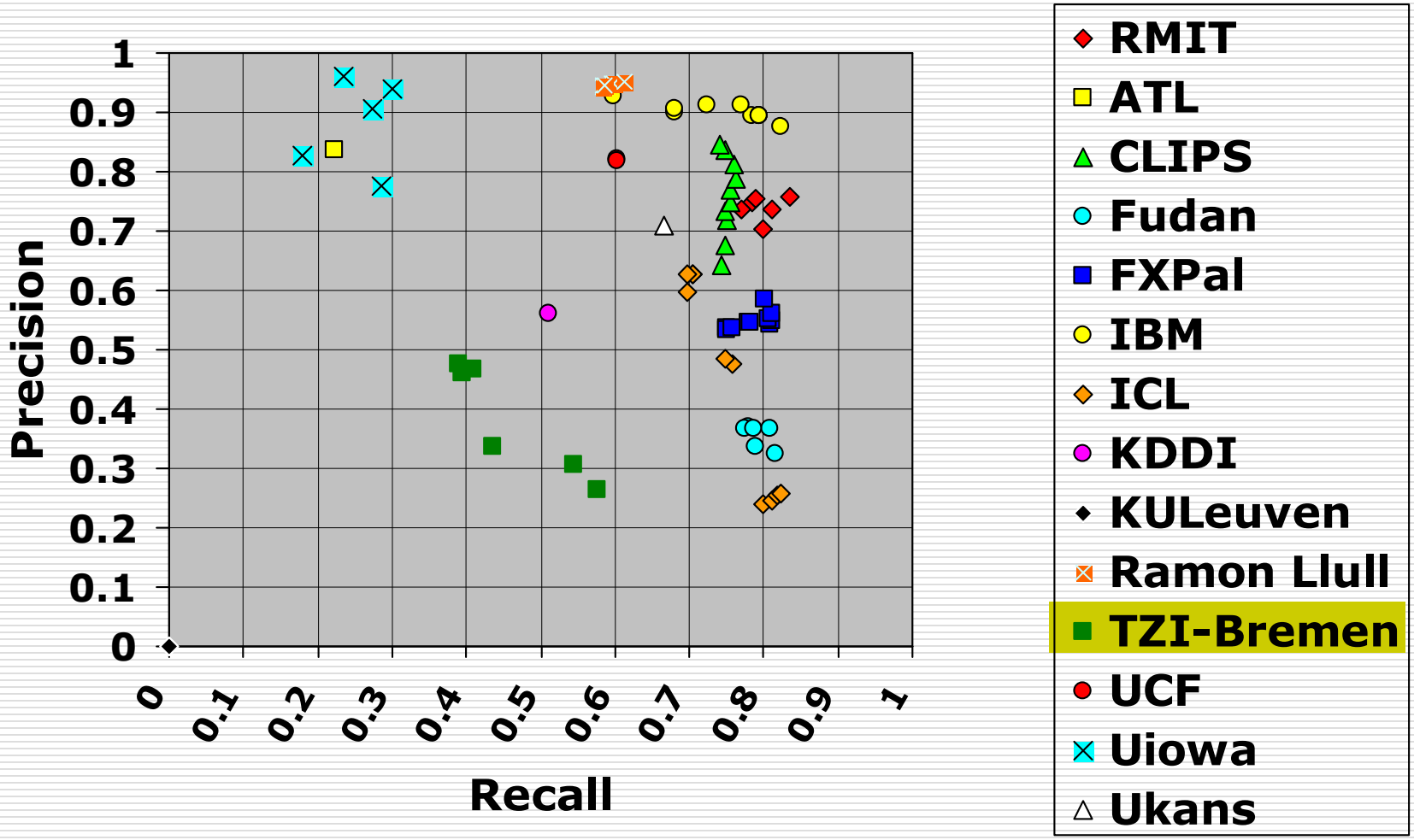
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

University of Central Florida:

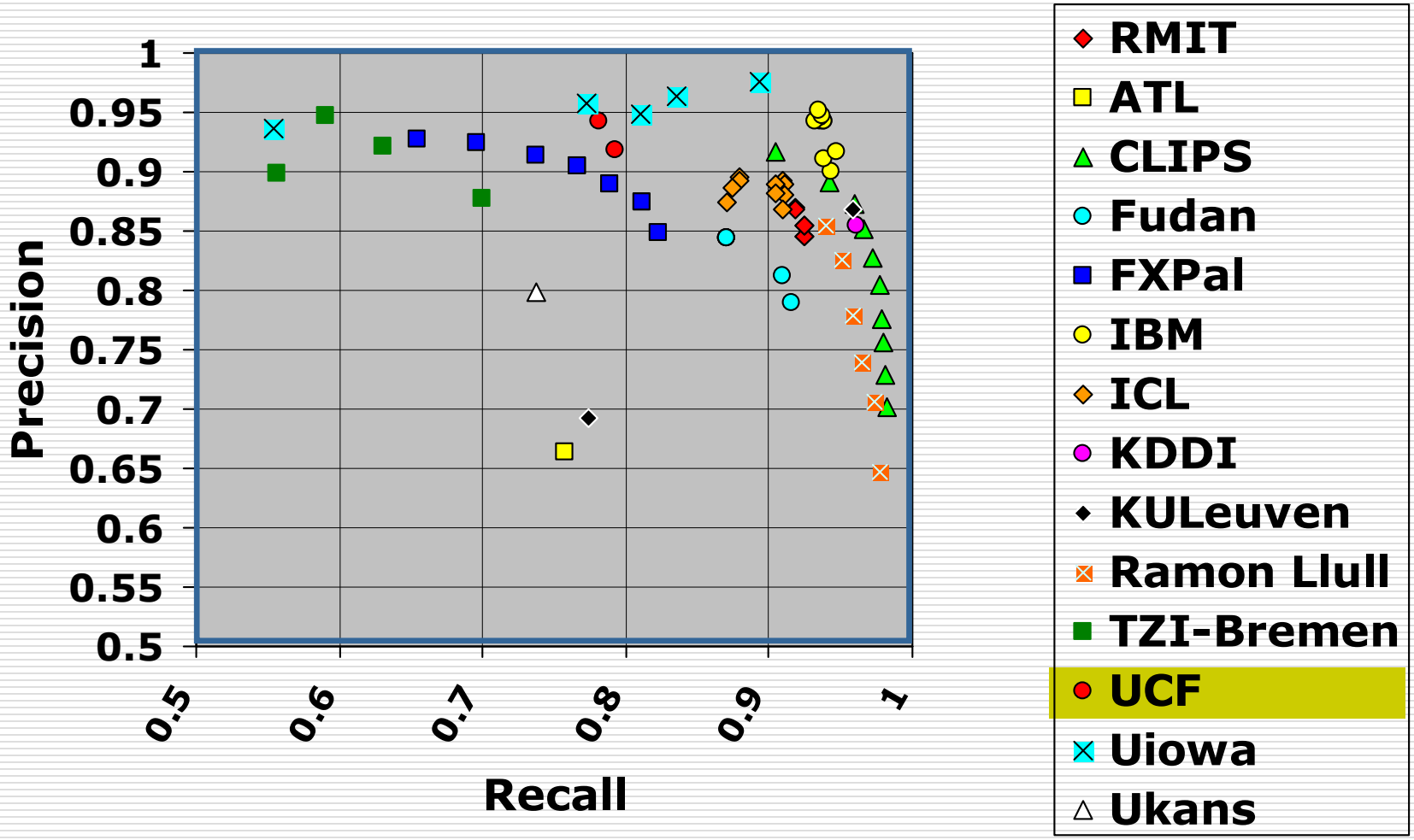
Colour histogram intersection of frames with sub-sampling of video at 5fps;

This gives approximate location of shot bounds, followed by fine-grained frame-frame comparison using 24-bin colour histogram;

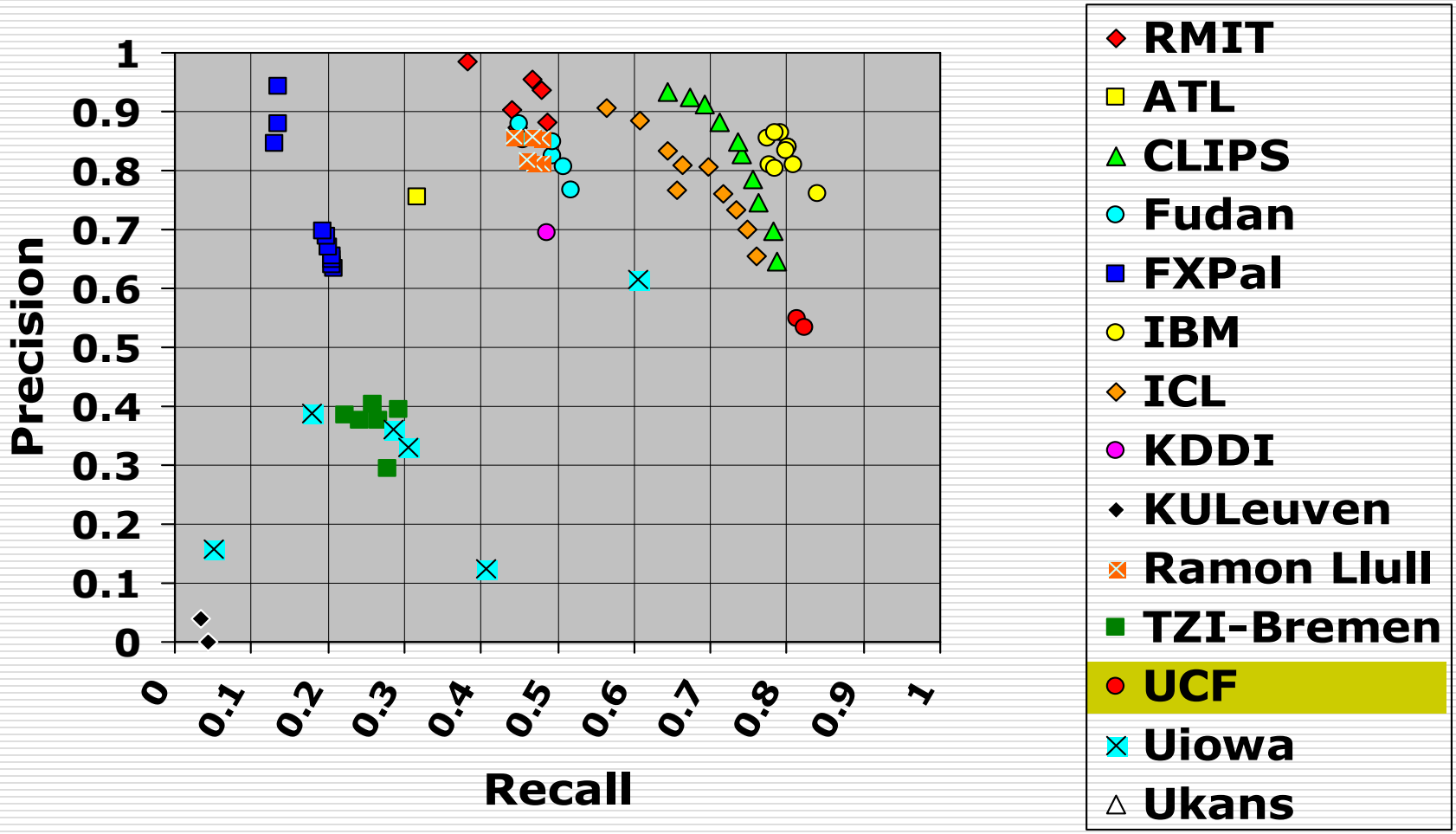
Post-processing to detect abrupt changes in illumination (camera flashes);

Also determined transition types;

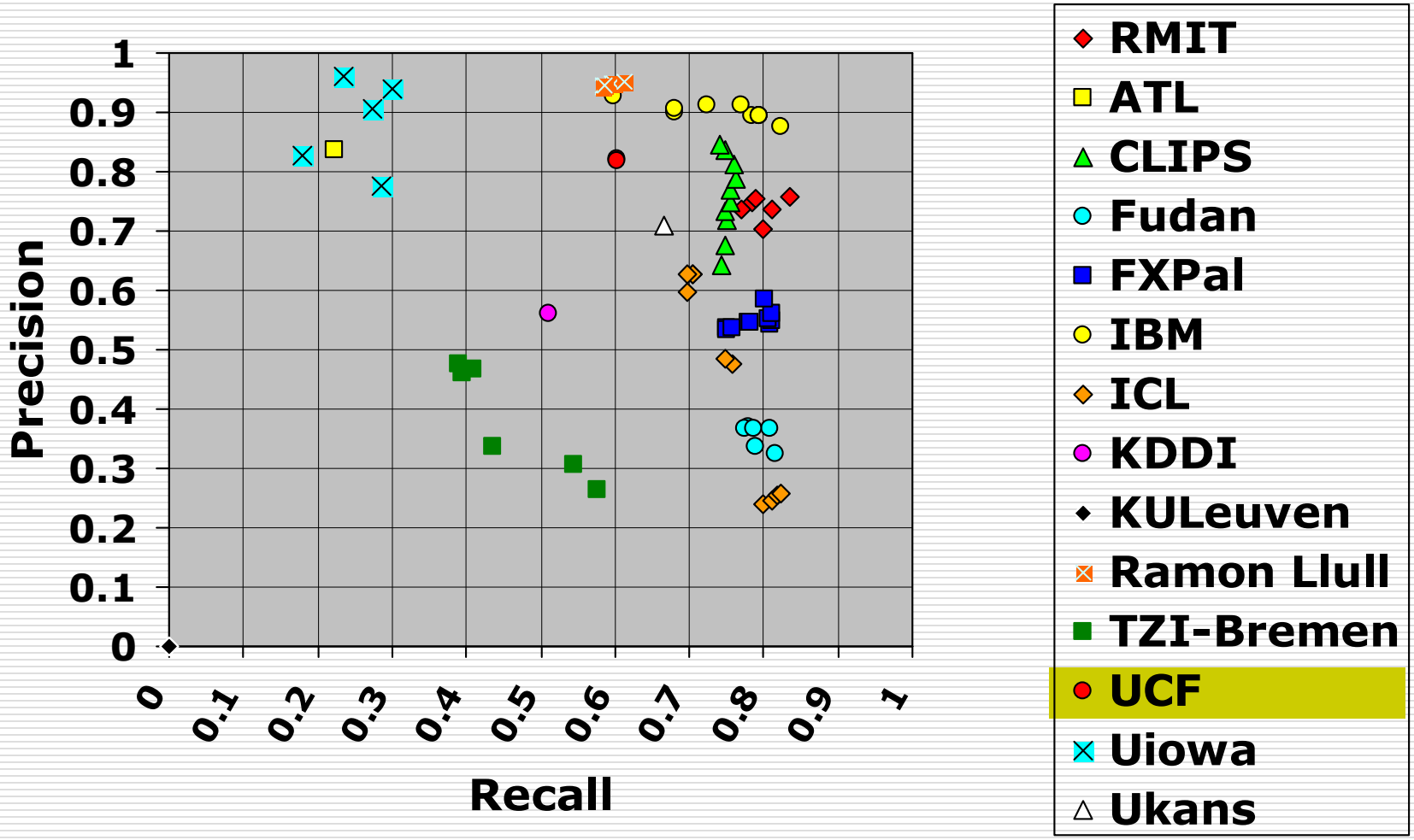
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



24 Participating Groups

University of Iowa:

Comparison of adjacent frames based on

-512-bin global colour histogram

-60x60 pixel thumbnail vs. thumbnail based on pixel/pixel

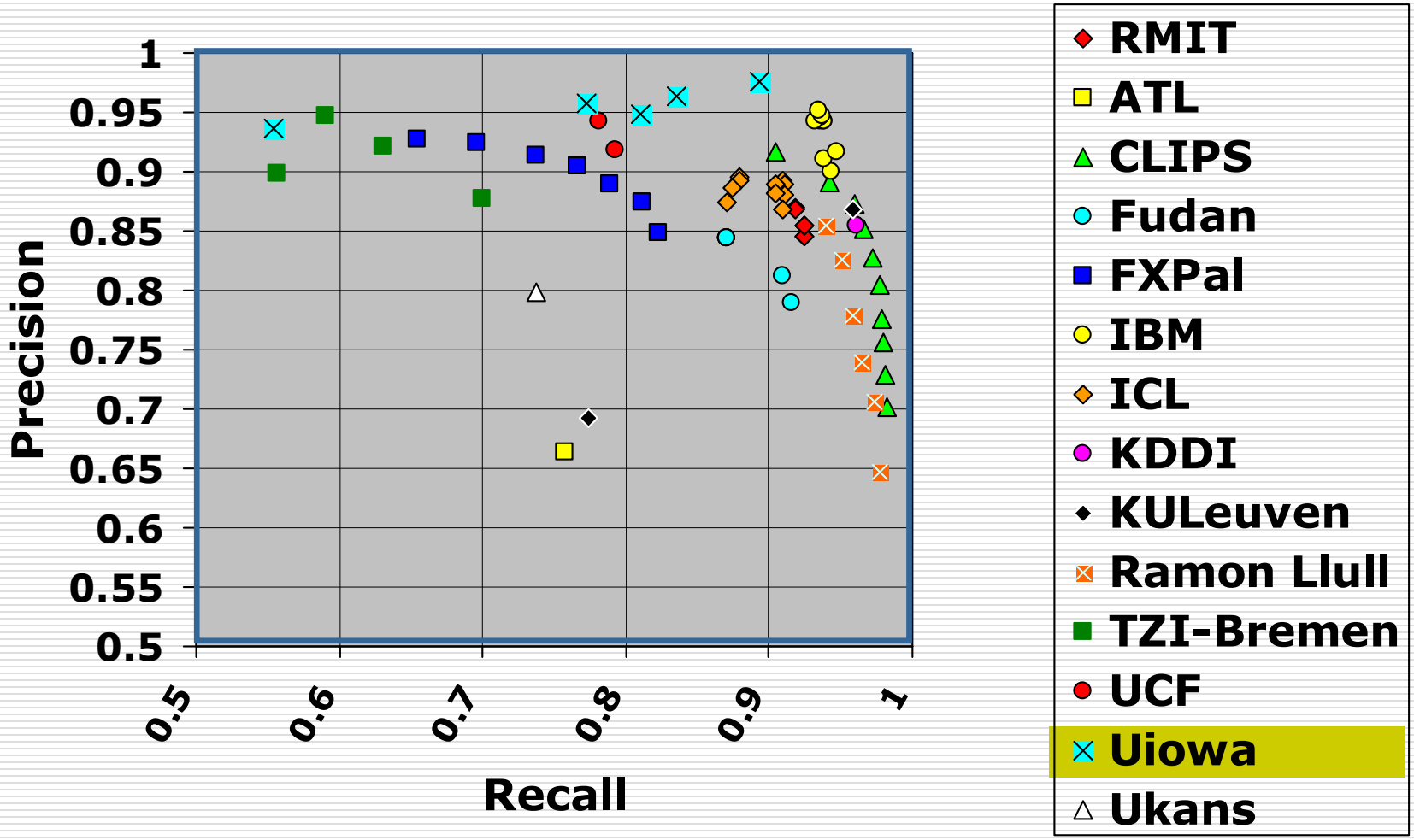
-Sobel filtering and detected edge differences

and then Boolean and arithmetic product combinations of these;

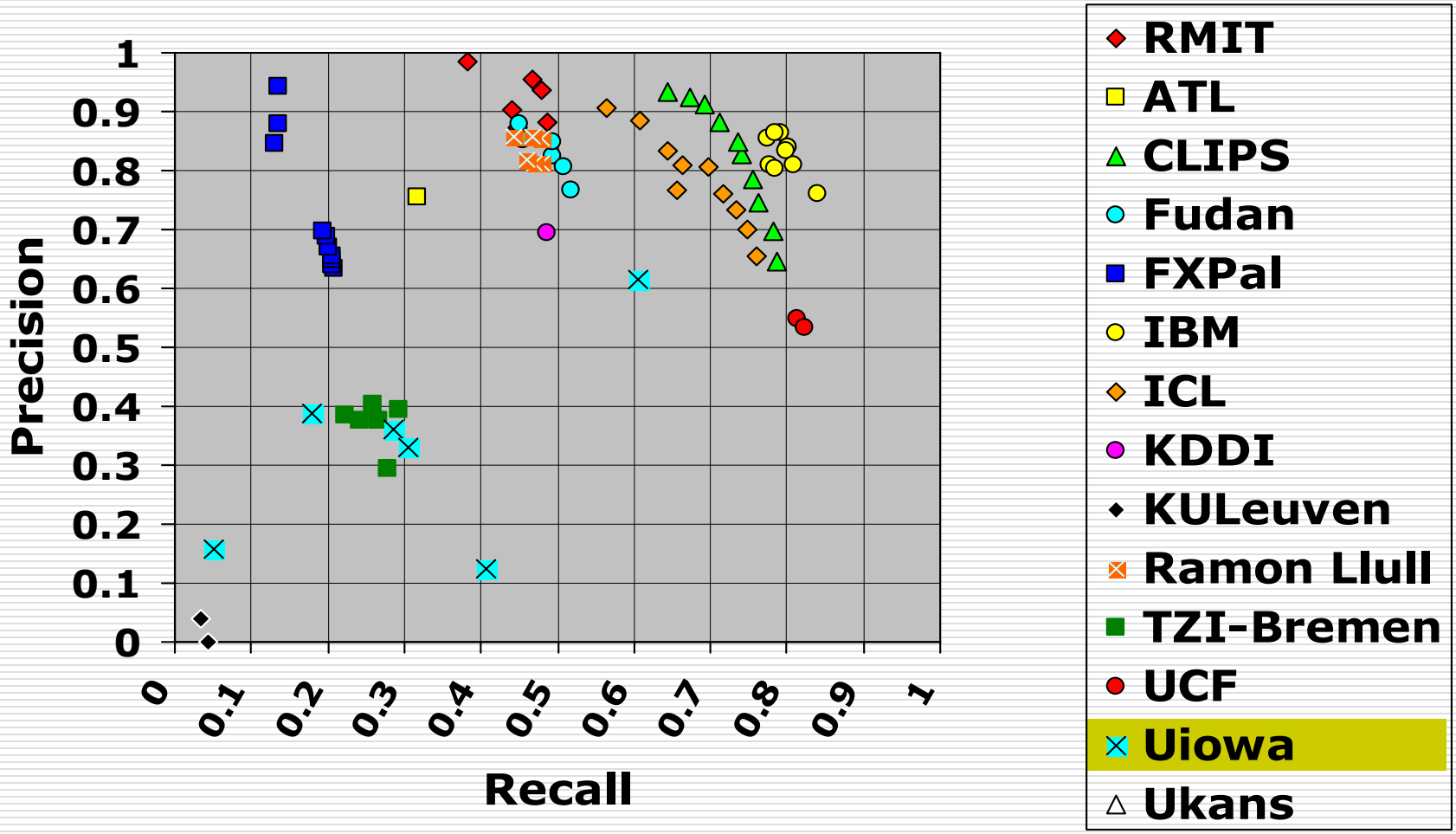
Presentation to follow

Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

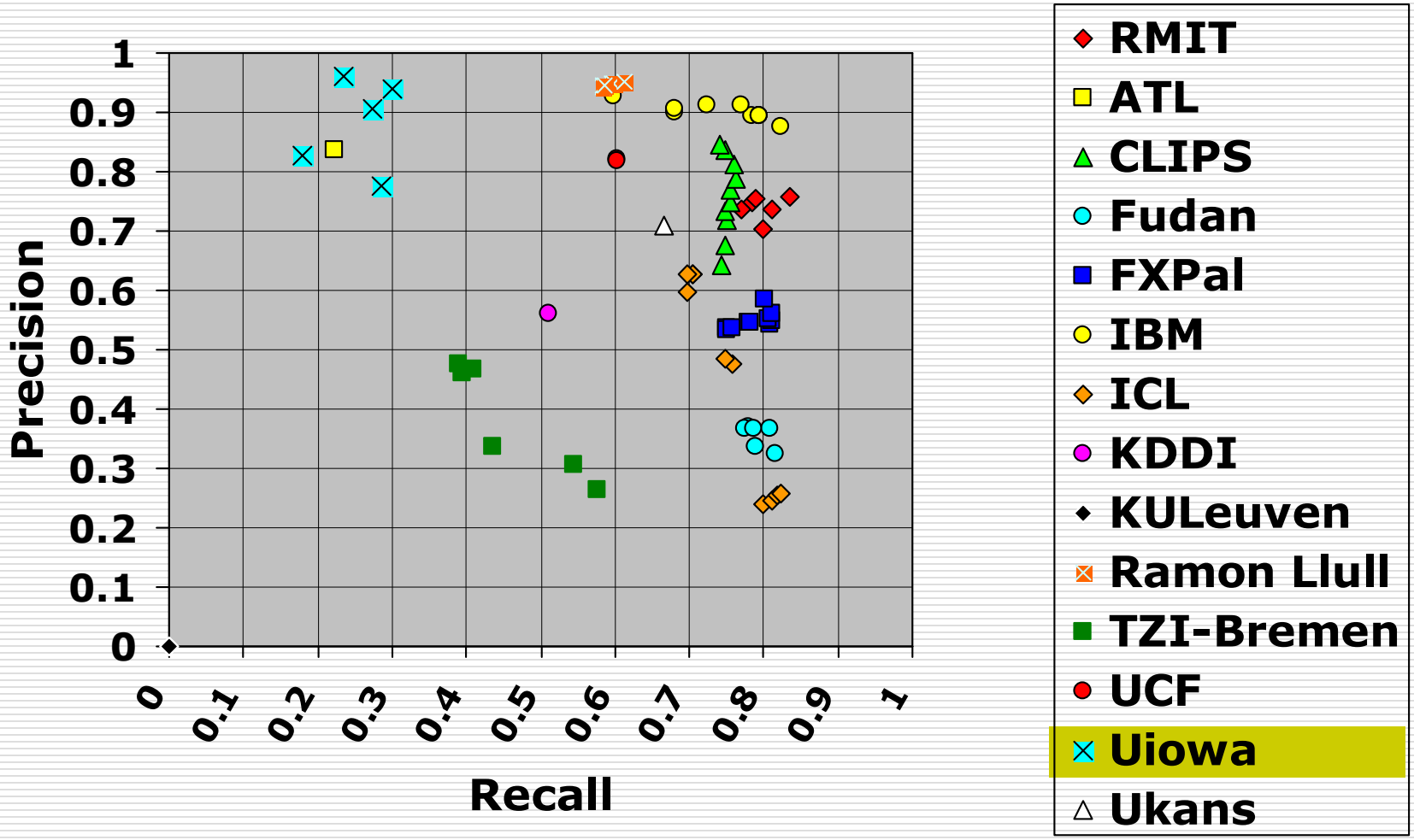
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



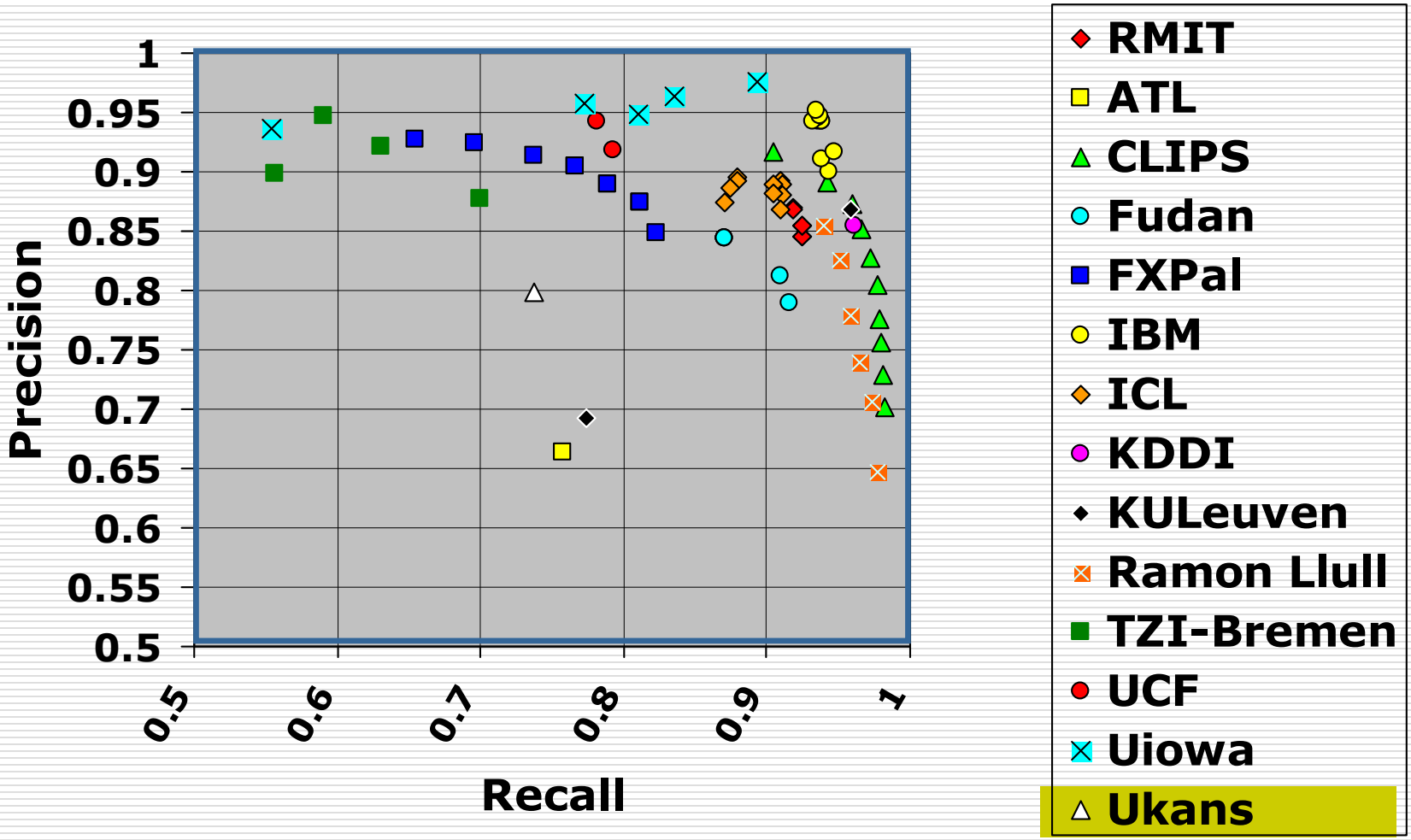
24 Participating Groups

University of Kansas:

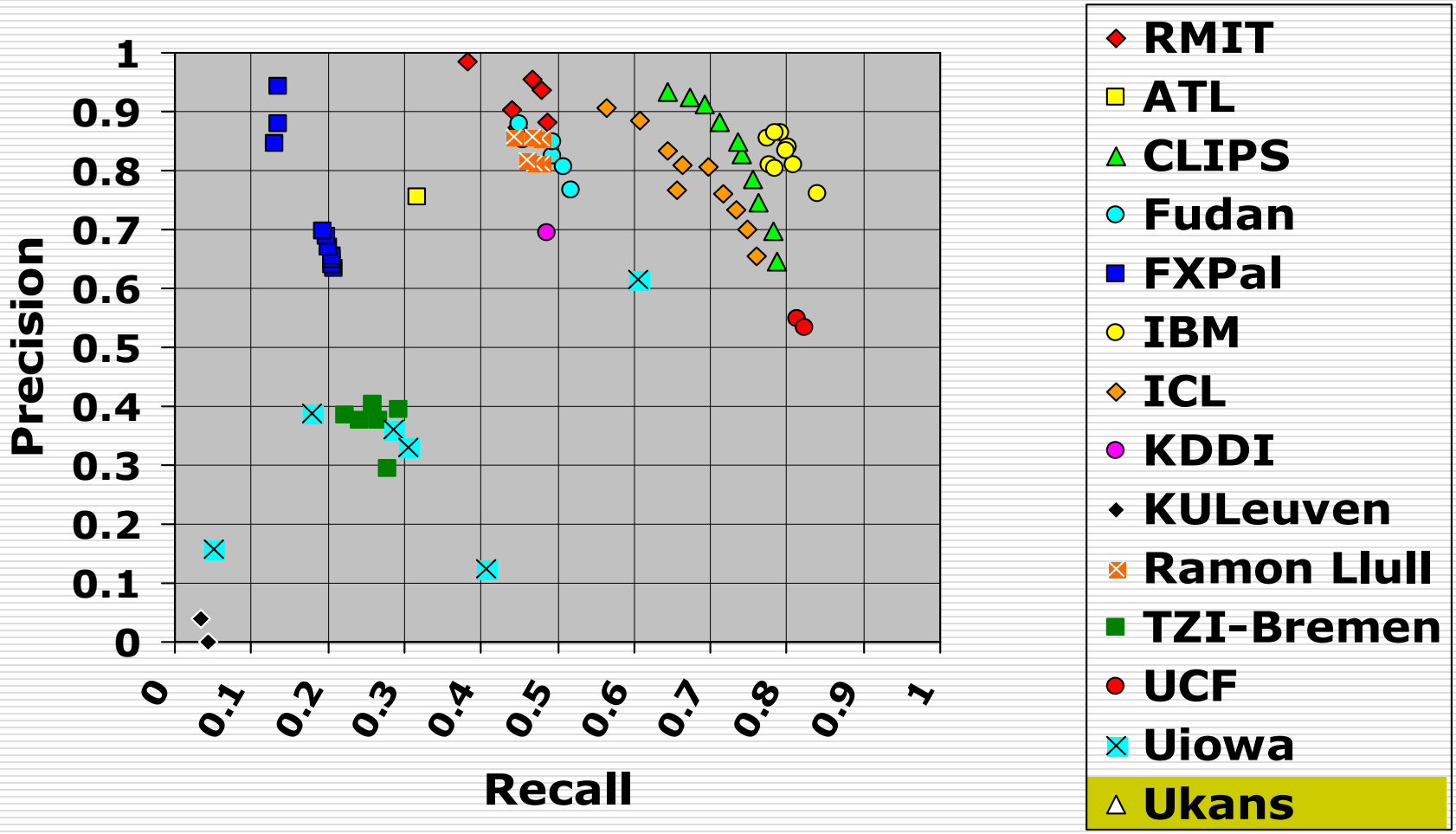
No details available at this time

Indiana University (US)				X
Institut Eurecom (FR)			X	
KDDI (JP)	X	X		
KU Leuven (BE)	X			
Mediamill/U Amsterdam (NL)				X
National Univ. Singapore (Sing.)		X		X
Ramon Llull Univ. (ES)	X			
RMIT University (Aus)	X			
StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

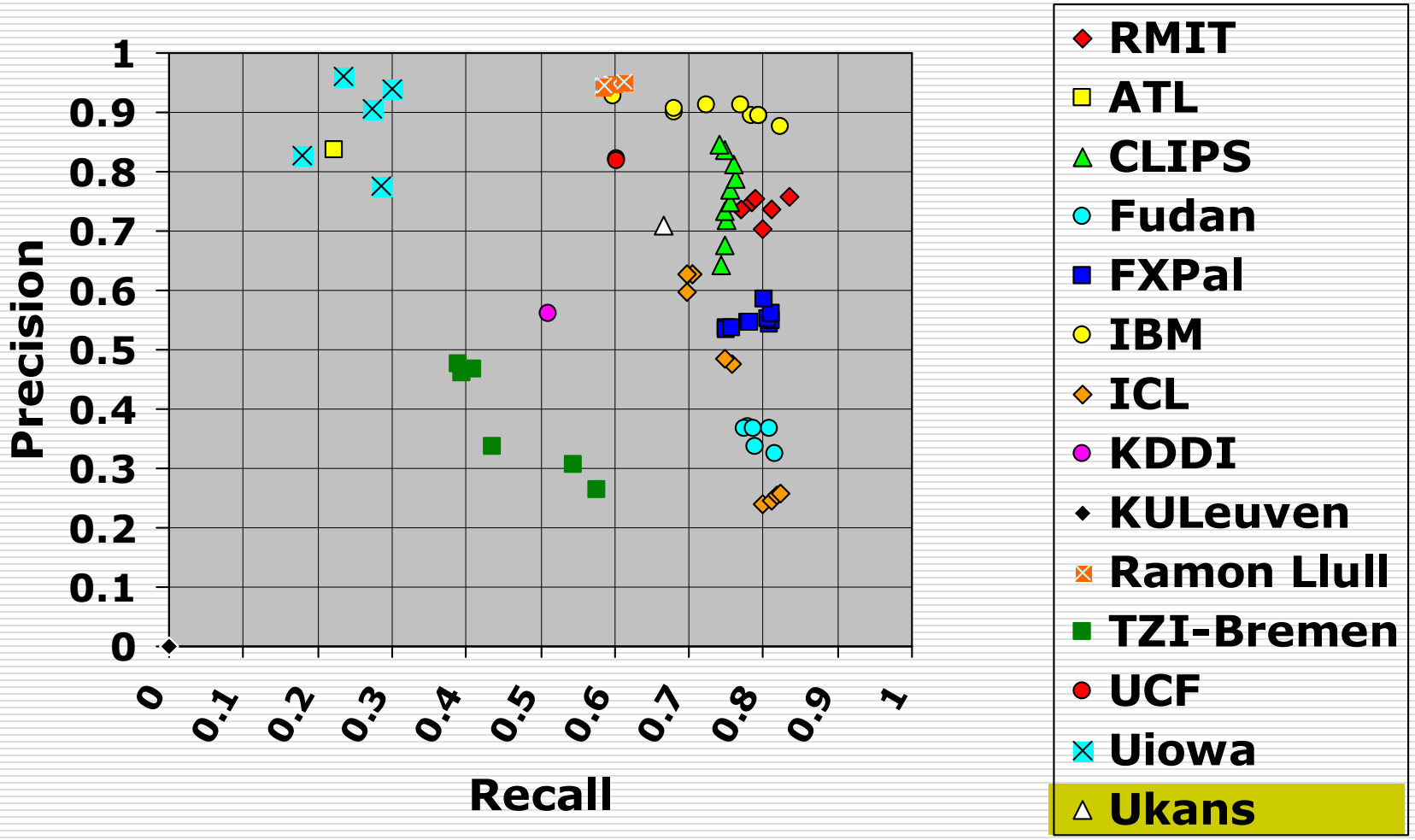
Recall and precision for cuts (zoomed)



Gradual Transitions



Frame-recall & -precision for GTs



Observations

- Most techniques are based on frame-frame comparisons, some with sliding windows;
- Comparisons are based on colour and on luminance, mostly;
- Some use adaptive thresholding, some don’t;
- Most operate on decoded video stream;
- Some have special treatment of motion during GTs, of flashes, of camera wipes;
- Performances are getting better;

Story segmentation and news typing

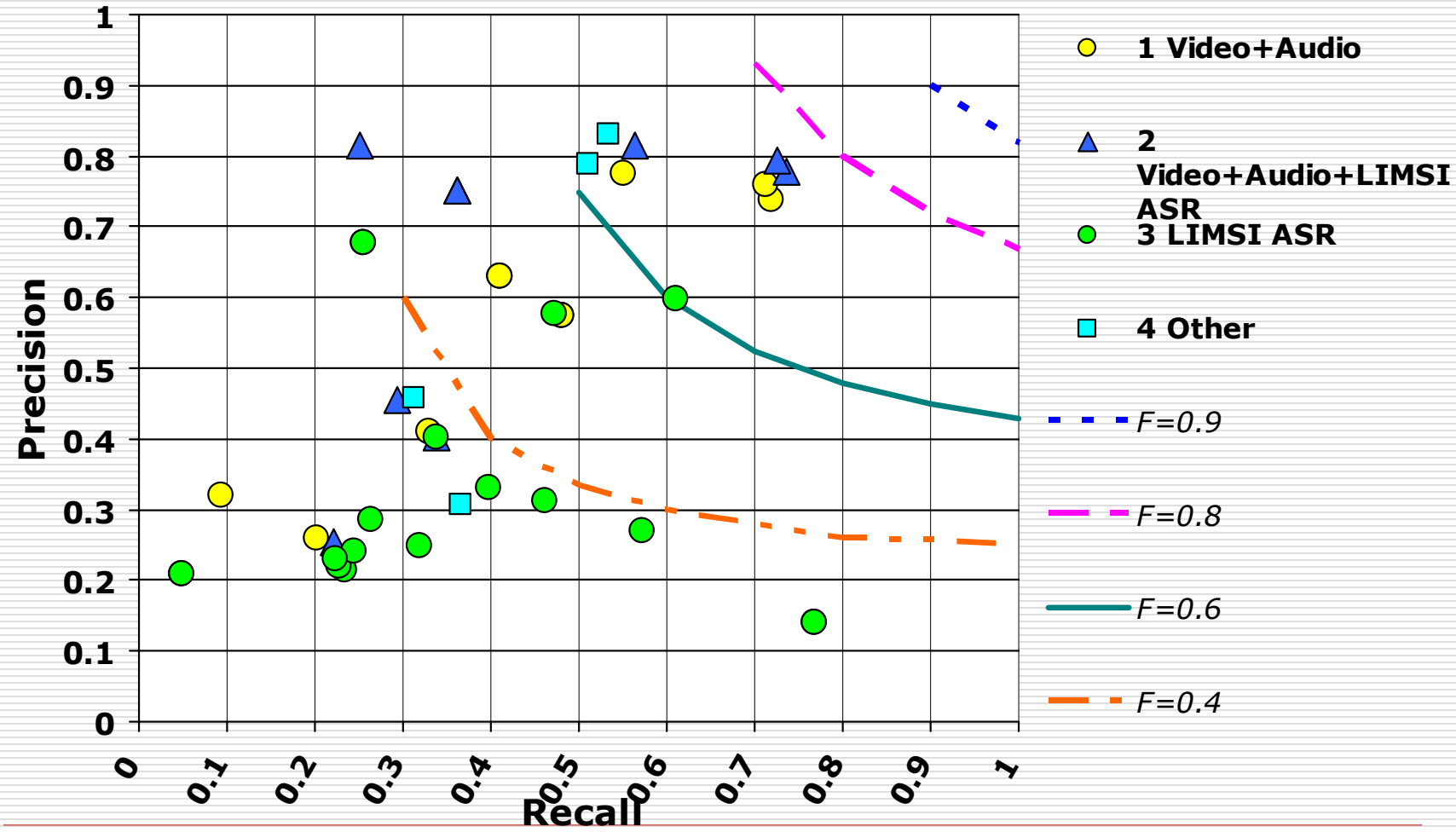
- Identify the individual news items in a news show
- New task in TRECVID, has been studied in ASR/IR community (TDT)
- Hope to show the gain of using video features

1. Segmentation task
   - Identify story boundaries in CNN and ABC news shows
   - Ground truth based on TDT 2 annotations
   - Evaluation based on precision & recall, boundaries have to be within +/- 5 seconds interval around ground truth boundaries
2. News classification task
   - Annotate stories as either news or non-news
   - Evaluation based on percentage of correctly identified news story footage

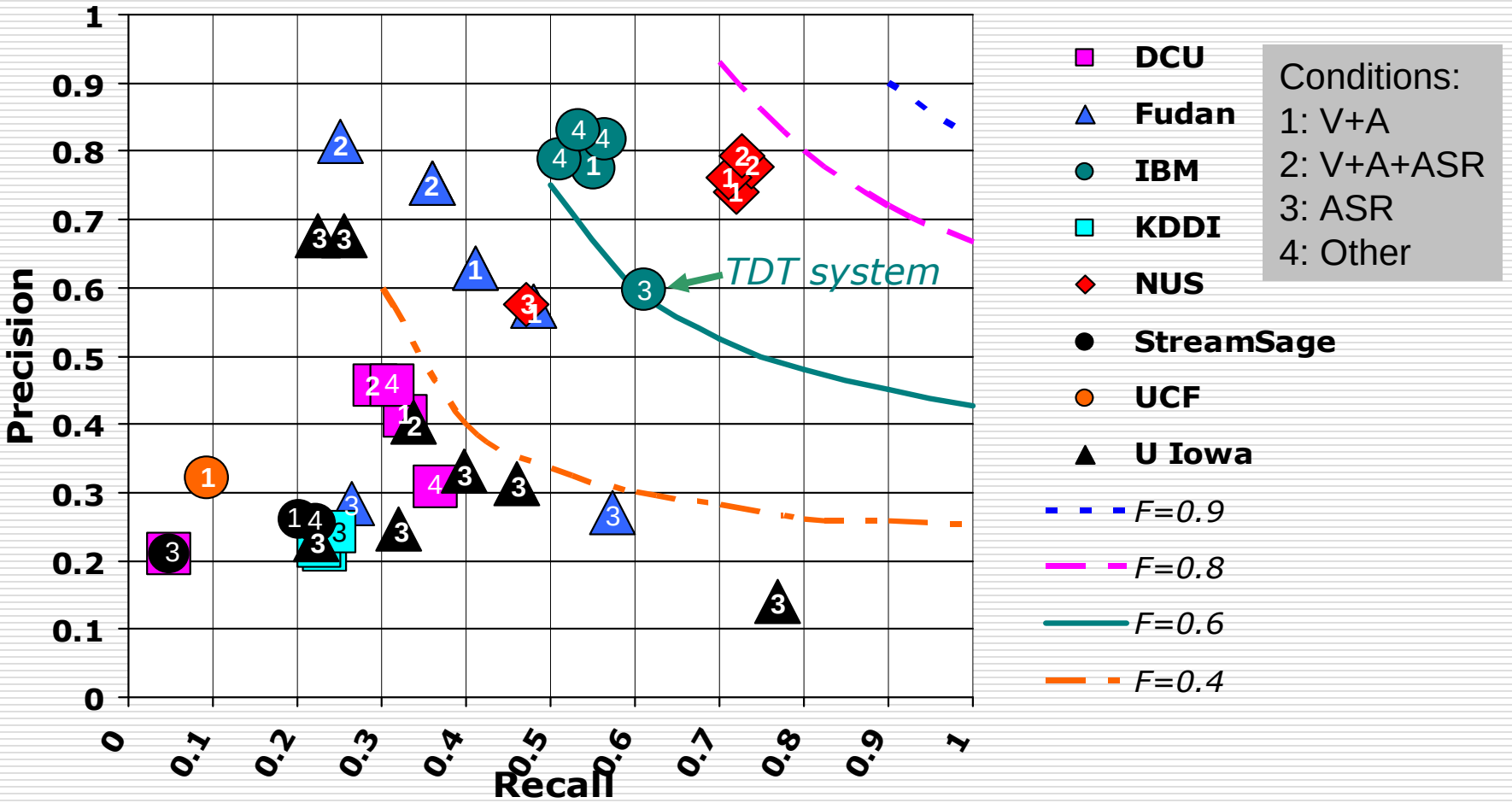
8 Participating Groups

Dublin City University (Irl)
Fudan Univ. (China)
IBM Research (US)
KDDI (JP)
National Univ. Singapore (Sing.)
StreamSage (US)
Univ. of Central Florida (US)
Univ. of Iowa (US)

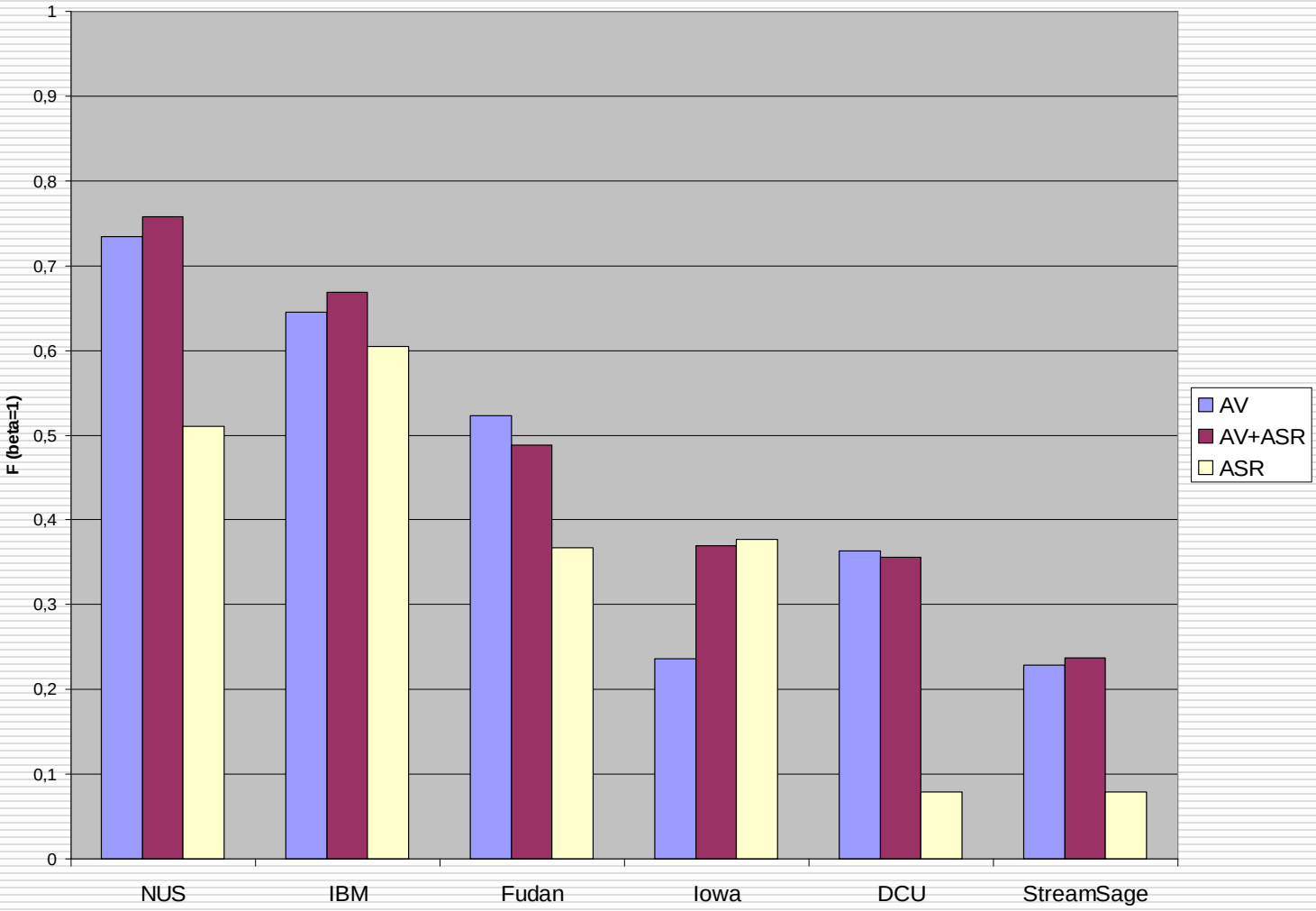
Story segmentation: recall and precision by condition



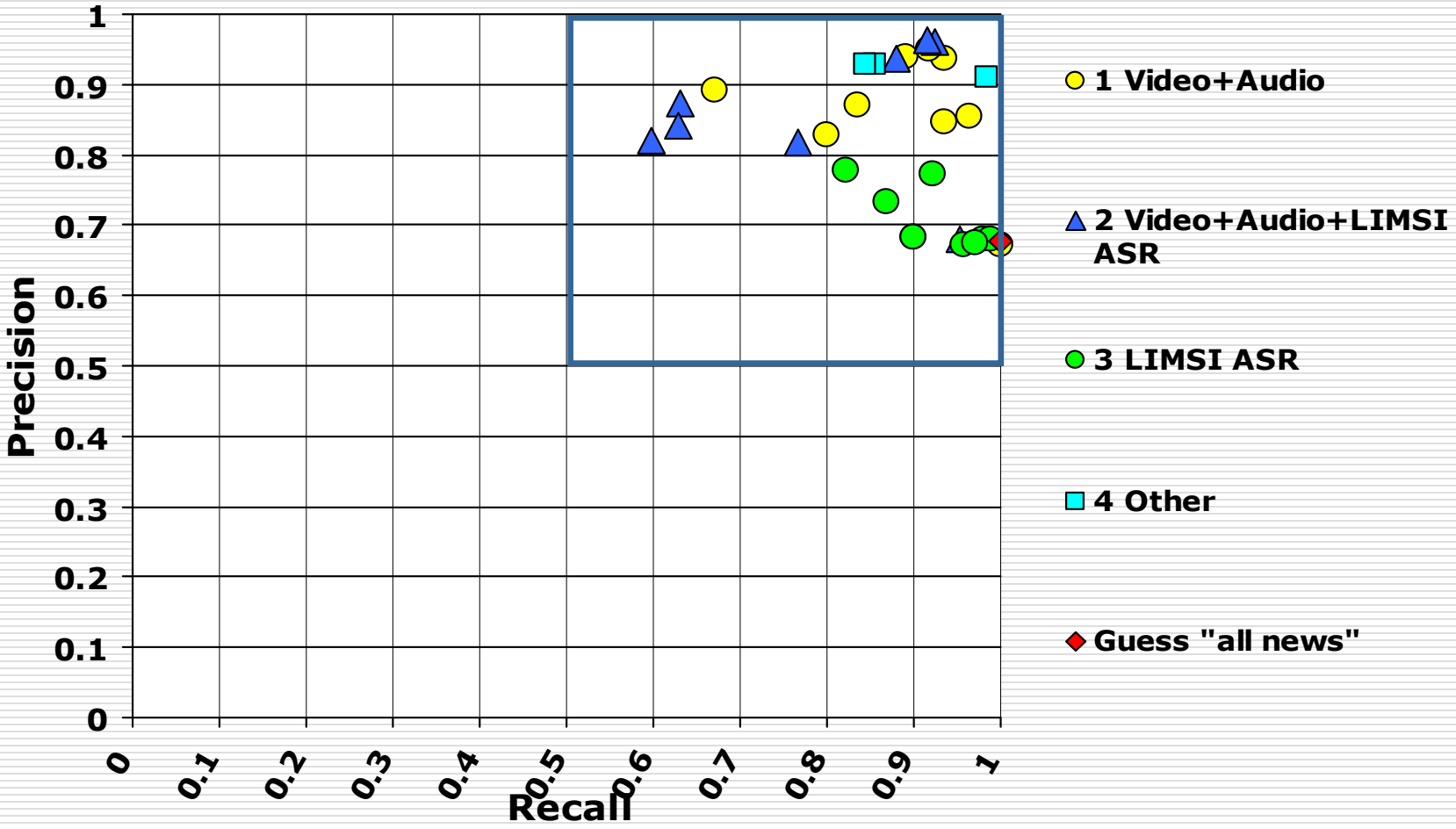
Story segmentation: recall and precision by system and condition (1-4)



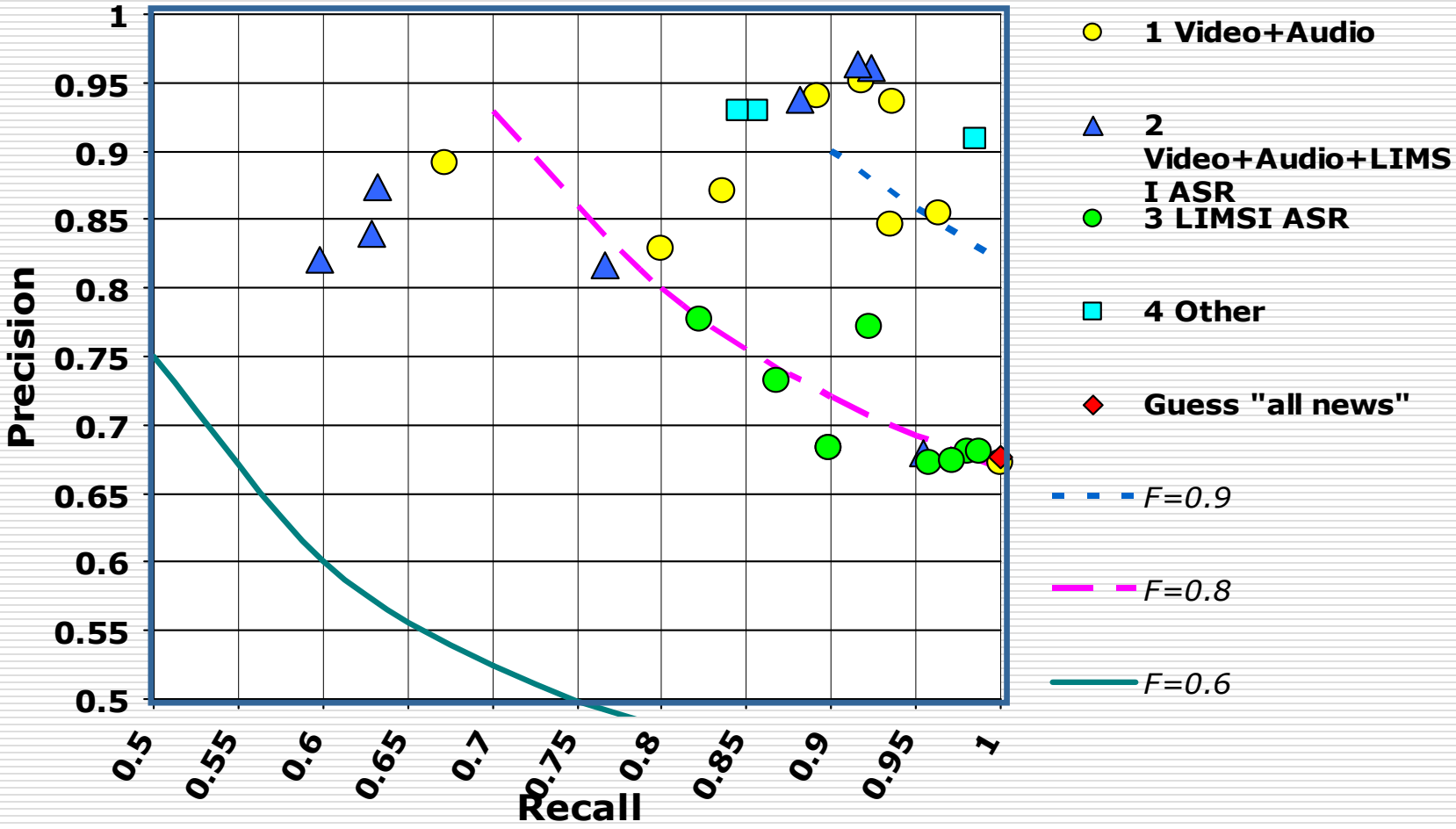
Segmentation, within system (F)



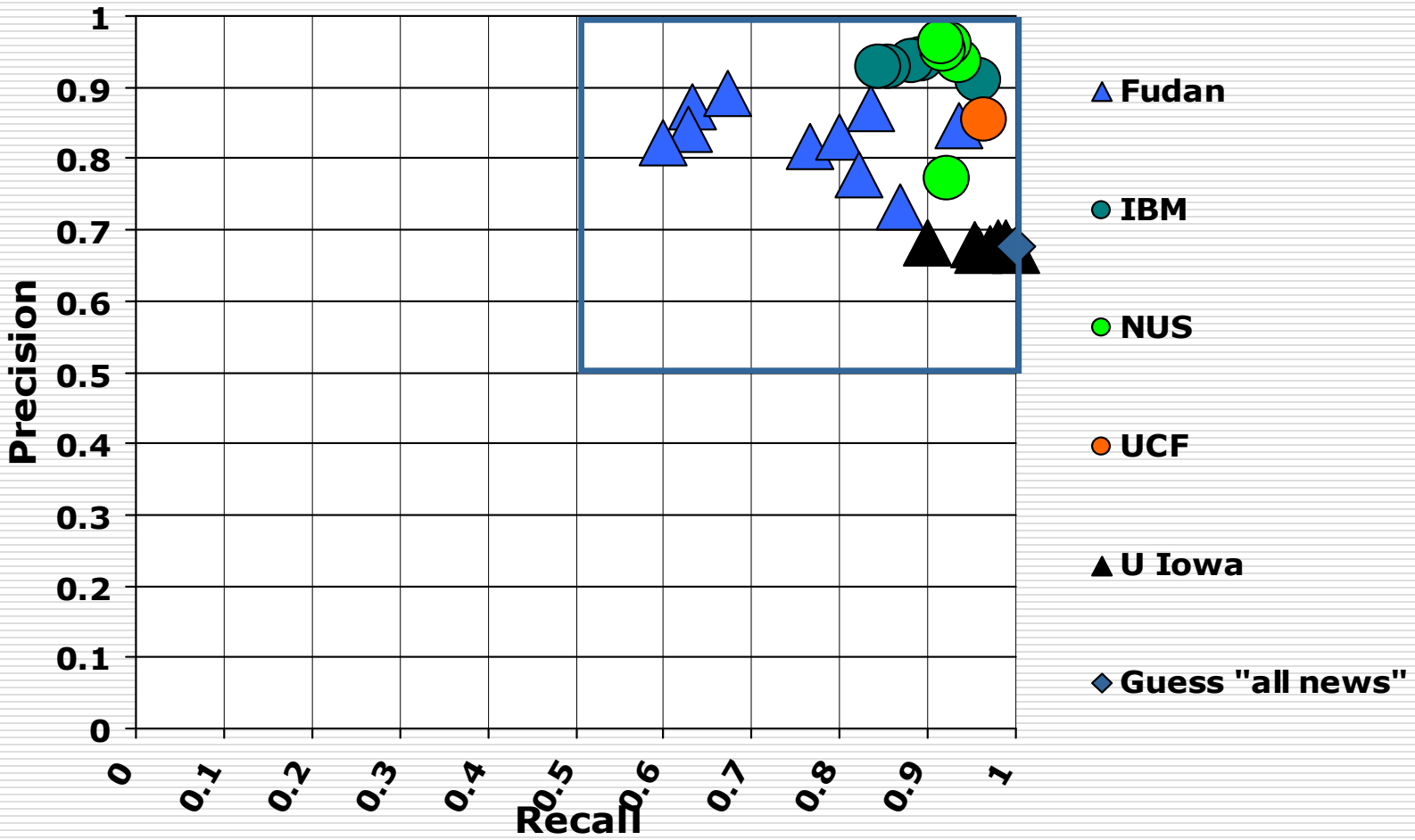
Story classsification: news recall and precision by condition



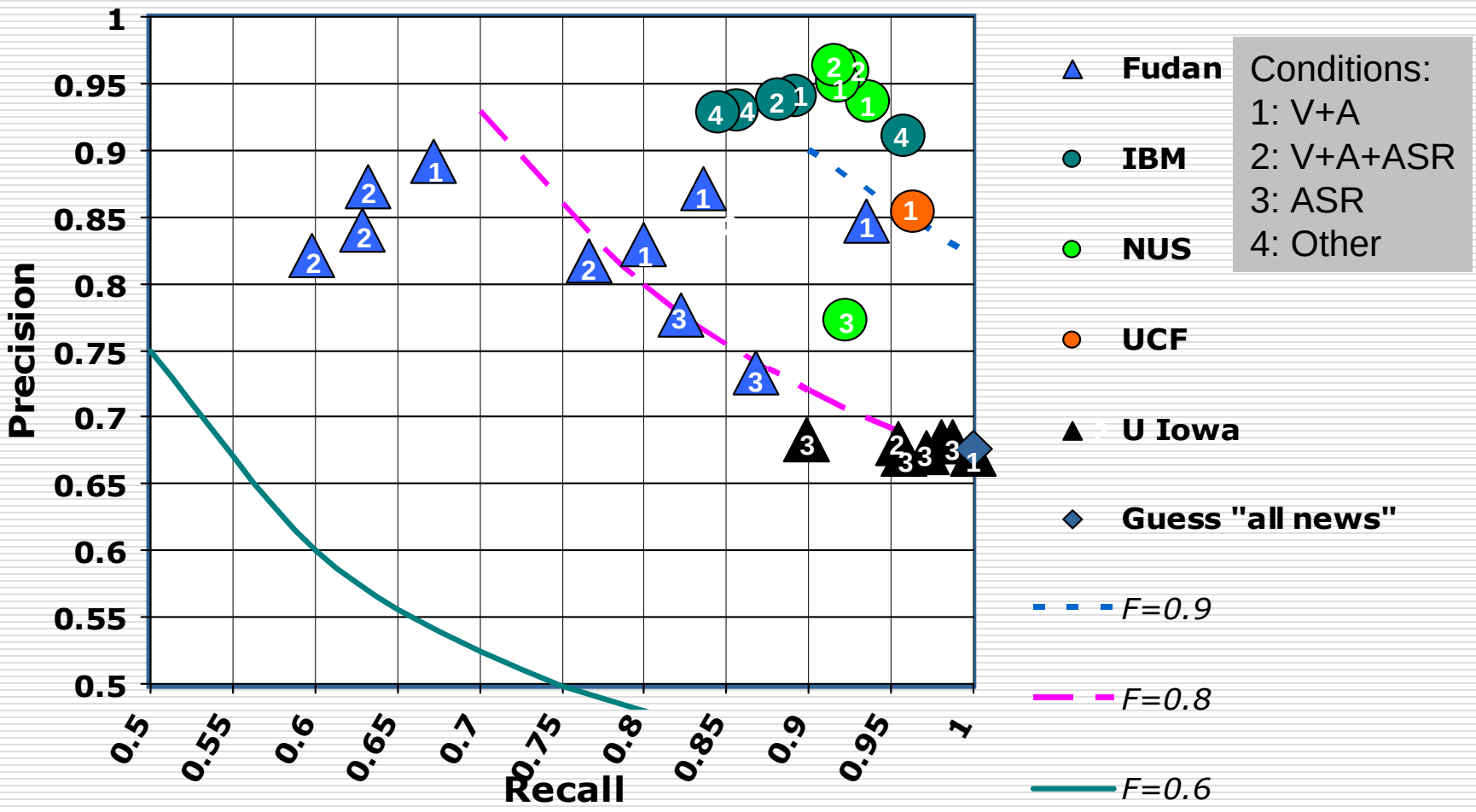
Story classsification: news recall and precision by condition - zoomed



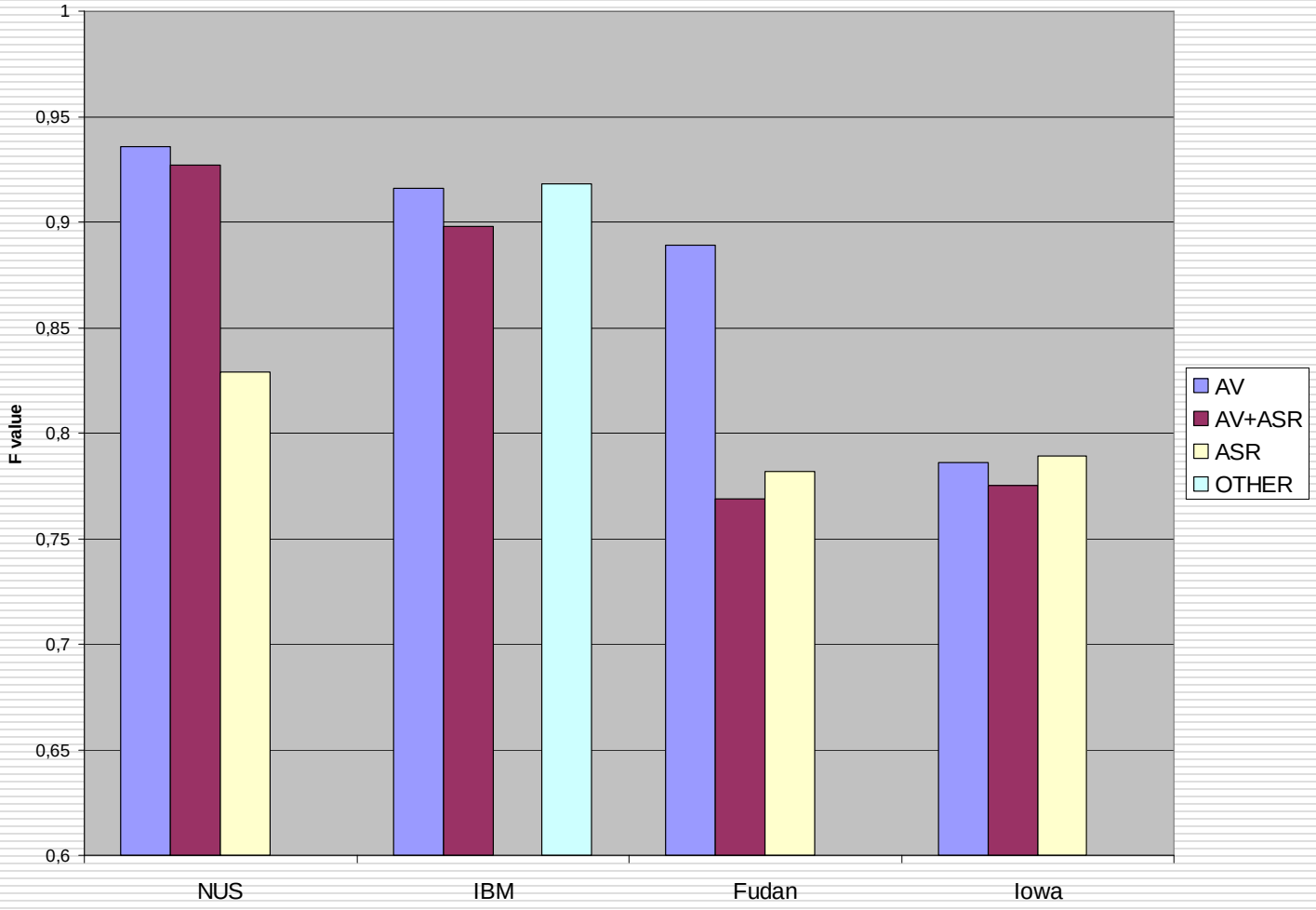
Story classifcation: news recall and precision by system



Story classifcation: news recall and precision by system and condition (1-4) zoomed



Classification, within system (F)



Group headlines

Du
Fu
IB
KD
Na
St
Un
Un

Fudan University

Segmentation

- •Anchor detection based on clustering and heuristics
- •Commercial detection based on ?
- •ASR segmentation using a variant of Text-tiling
- •Rule based and Maxent classifiers

News classification

- •GMM/Maxent using music, commercial and speech proportion as features

Group headlines

Du
Fu
IB
KD
Na
St
Un
Un

KDDI

Segmentation

1. All shots are classified as ANCHOR, REPORT or COMMERCIAL, using audio & motion intensity, color Ł SVM. Subsequently rule based segmentation.
2. Direct classification of boundaries, using the features of two shots before and after the boundary candidate. SVM

Classification

SVM for NEWS-NEWS, NEWS-MISC and MISC NEWS

Group headlines

Du
Fu
IB
KD
Na
St
Ur
Ur

StreamSage (/ DCU)

ASR only segmentation runs

Three methods:

1. lexical chaining to define topically coherent segments
2. Variant of text-tiling
3. Use methods 1 and 2 for compiling a list of cue-phrases that announce topic introduction or closure

Group headlines

University of Central Florida

Combined Segmentation and Classification:

1. Story boundaries are marked by blank frames
2. Long story Ł news, short story Ł non-news
3. Merge adjacent non-news stories

Conclusion: story length is a strong feature for news classification

Group headlines

Dublin City University

IBM Research

National University Singapore

University of Iowa

presentations follow….

Observations

- Video provides strong clues for story segmentation and even more for classification, best runs are either type 1 or 2
- AV runs generally have a higher precision
- Combination of AV and ASR gives a small gain for segmentation
- Most approaches are generic

- Are the combination methods optimal?
- Are the ASR segmentation runs state of the art?

FE Task definition

- Goal: Build benchmark for detection methods of high-level features
- Secondary goal: feature-indexing can help search and navigation
- New: common feature annotation
  - Helps (a.o.) to standardize training resources across sites
  - Category A: sites work with just the common development data and common annotations
  - Category B: sites work with just the common development data and any annotation set
  - Category C: other

FE evaluation

- Each feature is assumed to be binary: absent or present for each shot
- Find shots that contain a certain feature, rank them according to confidence measure, submit the top 2000
- Submissions are pooled
- Evaluate performance quality by measuring the *average precision* of each feature detection method

10 Participating Groups

Accenture Technology Laboratories (US)
Carnegie Mellon Univ. (US)
CLIPS-IMAG (FR)
CWI Amsterdam / Univ. of Twente (NL)
Fudan Univ. (China)
IBM Research (US)
Imperial College London (UK)
Institut Eurecom (FR)
Univ. of Central Florida (US)
Univ. Oulu/VTT (FI)

17 Features

11. Indoors
12. News subject face – not a news show person
13. People – at least three humans
14. Building – walled structure with roof
15. Road
16. Vegetation – living vegetation in its natural env.
17. Animal
18. Female speech – woman speaking (visible, audible)
19. Car/truck/bus – exterior of ..

17 Features

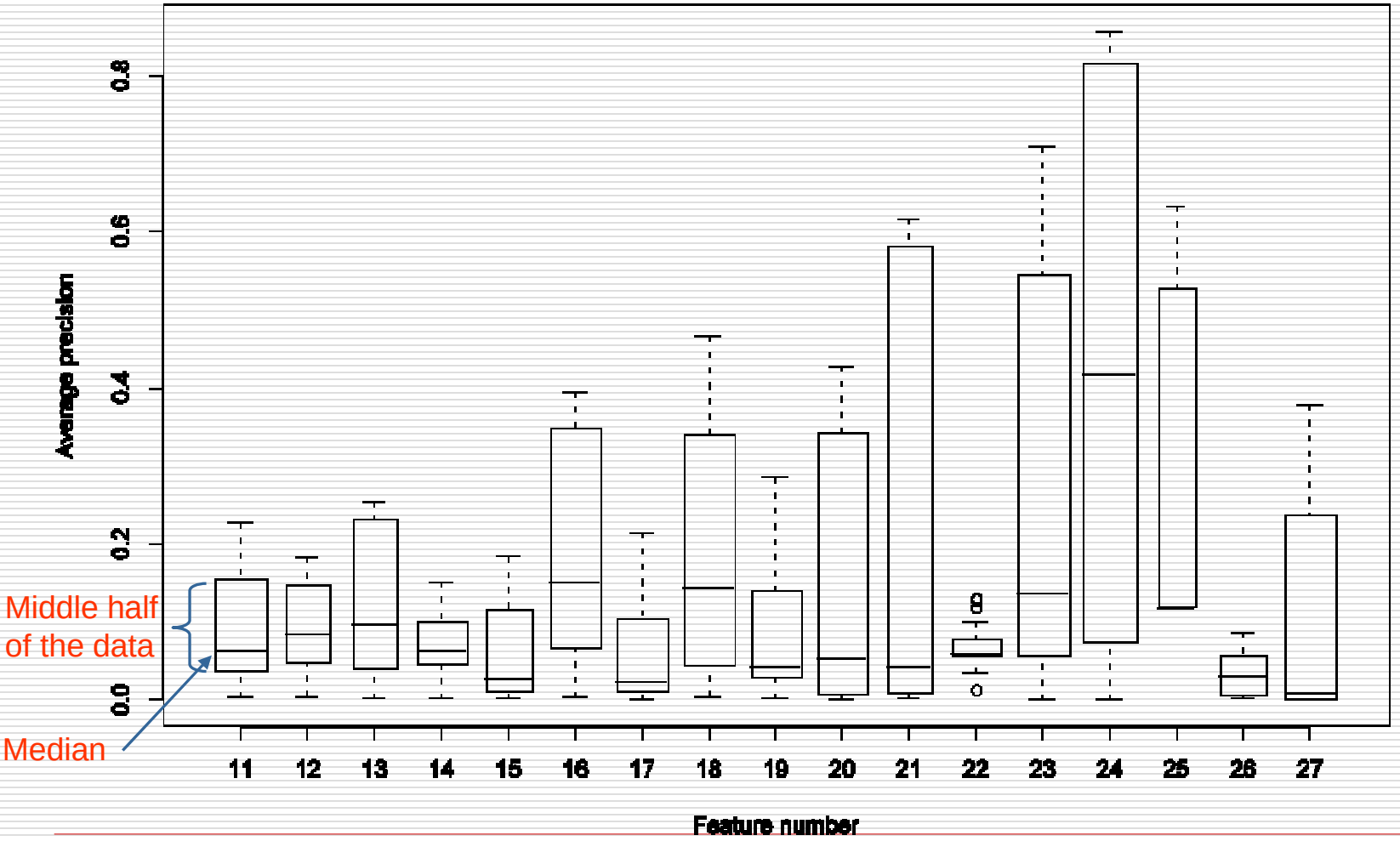
20. Aircraft
21. News subject monologue – uninterrupted
22. Non-studio setting
23. Sporting event
24. Weather news
25. Zoom in
26. Physical violence – between people / objects
27. Madeleine Albright – visible

Who worked on which features

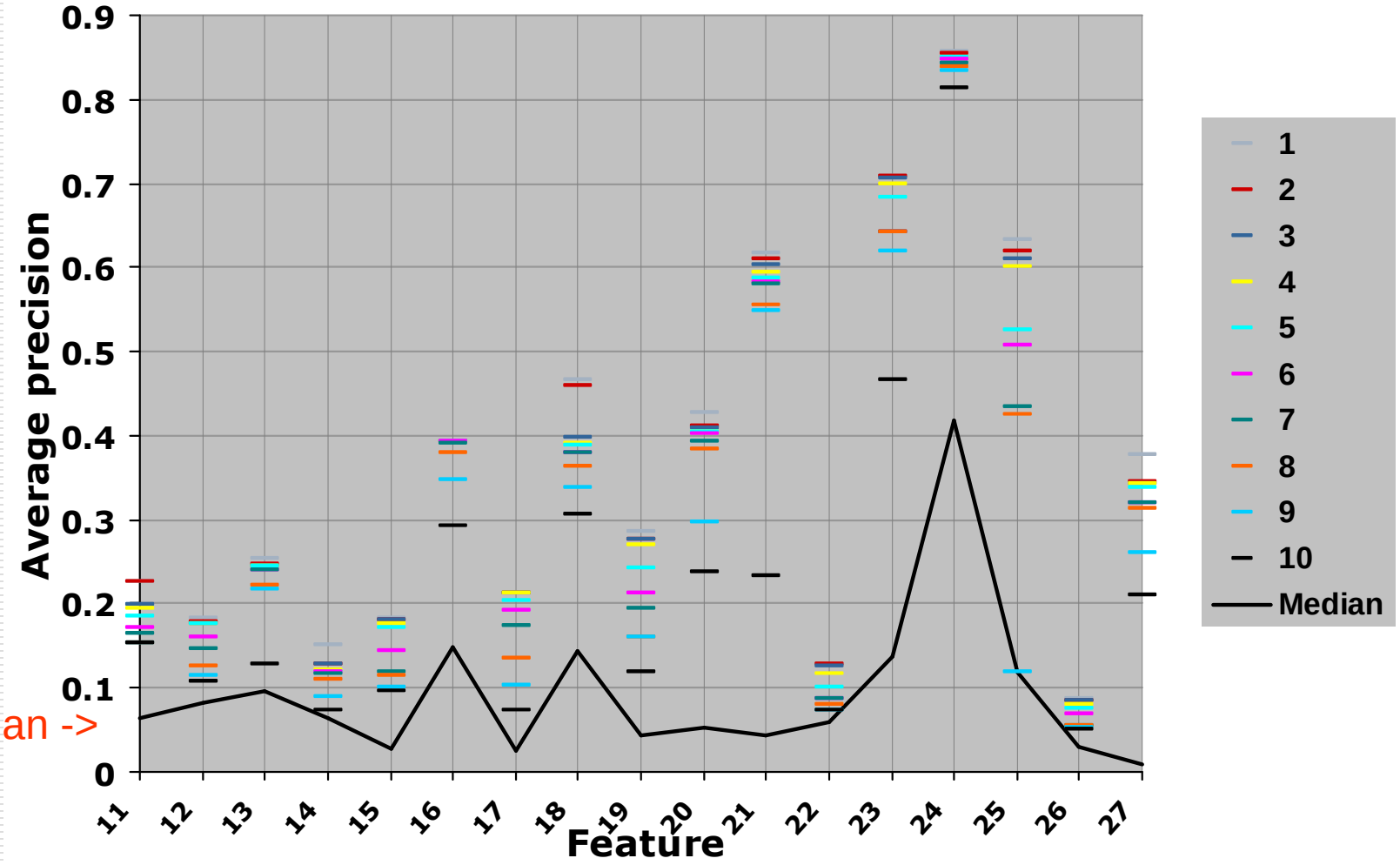
Groups

	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Accenture Technology Laboratories (US)			X					X						X			
Carnegie Mellon Univ. (US)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
CLIPS-IMAG (FR)																	X
CWI Amsterdam / Univ. of Twente (NL)	X	X	X	X	X	X	X		X	X		X	X	X		X	X
Fudan Univ. (China)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
IBM Research (US)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Imperial College London (UK)						X											
Institut Eurecom (FR)	X	X	X	X	X	X	X	X	X	X		X	X	X		X	X
Univ. of Central Florida (US)												X		X			
Univ. Oulu/VTT (FI)	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	
	6	6	6	6	6	7	6	6	6	6	4	7	6	8	3	6	6

AvgP by feature (all runs)

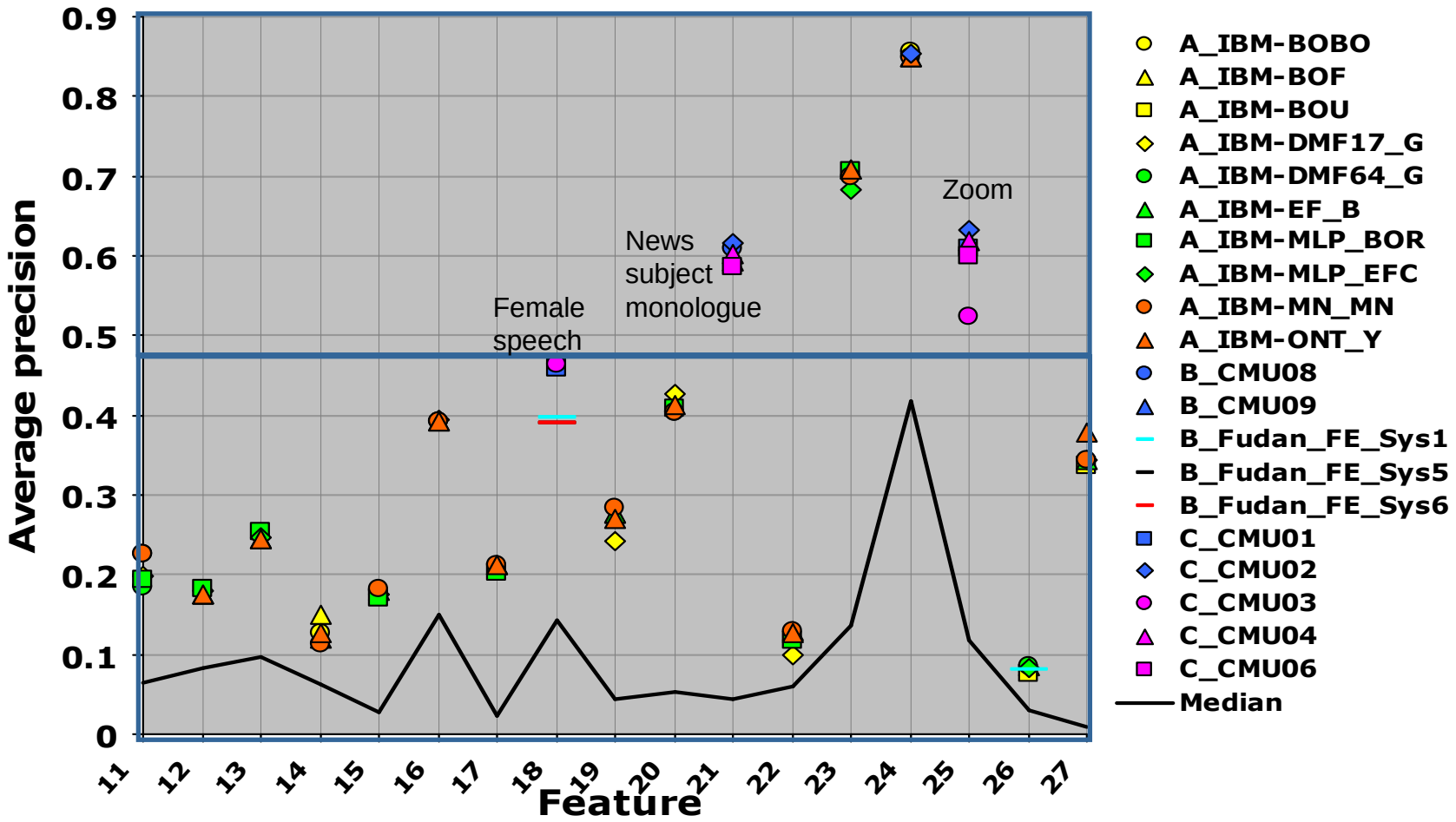


AvgP by feature (top 10 runs)

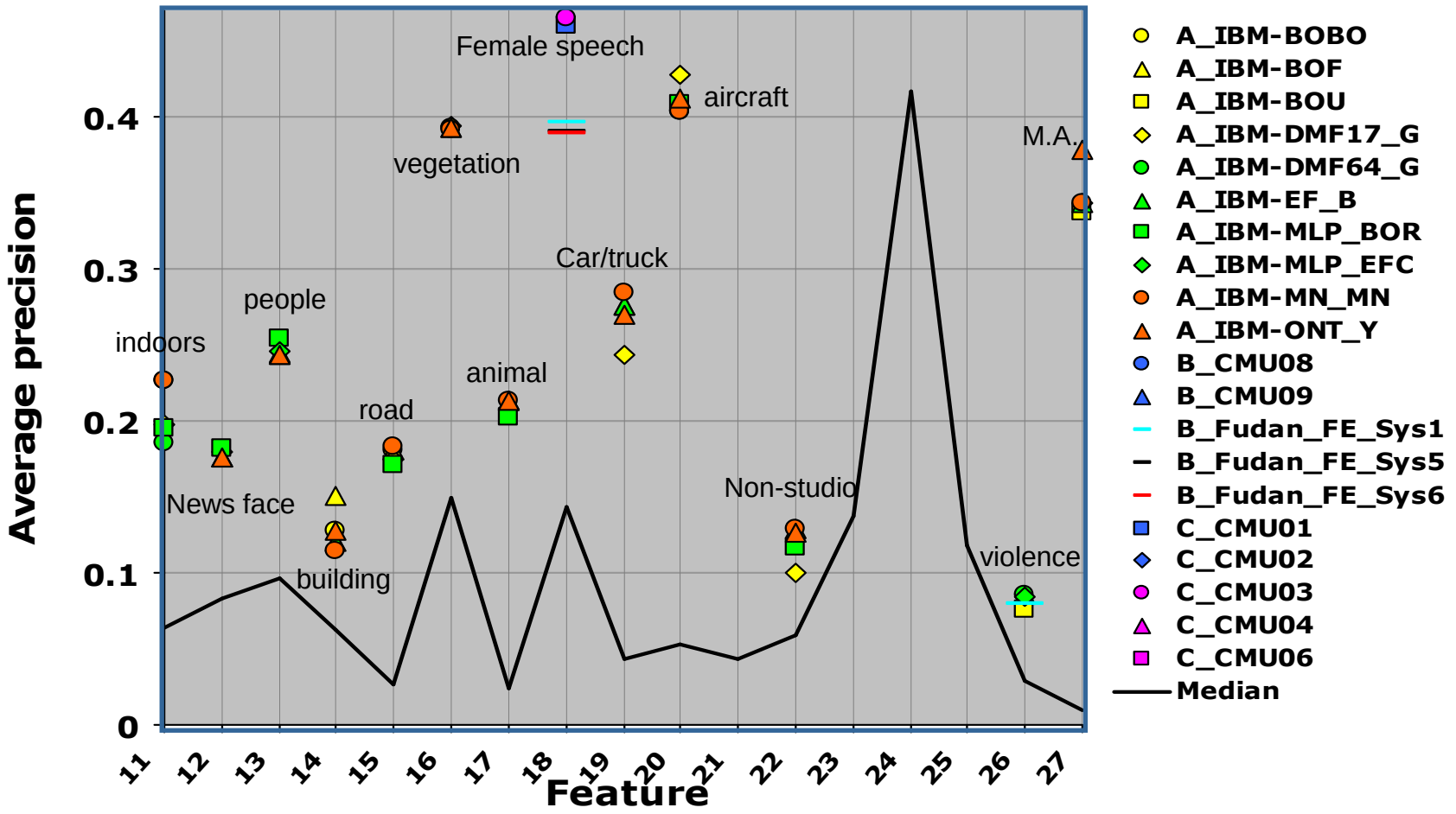


Median ->

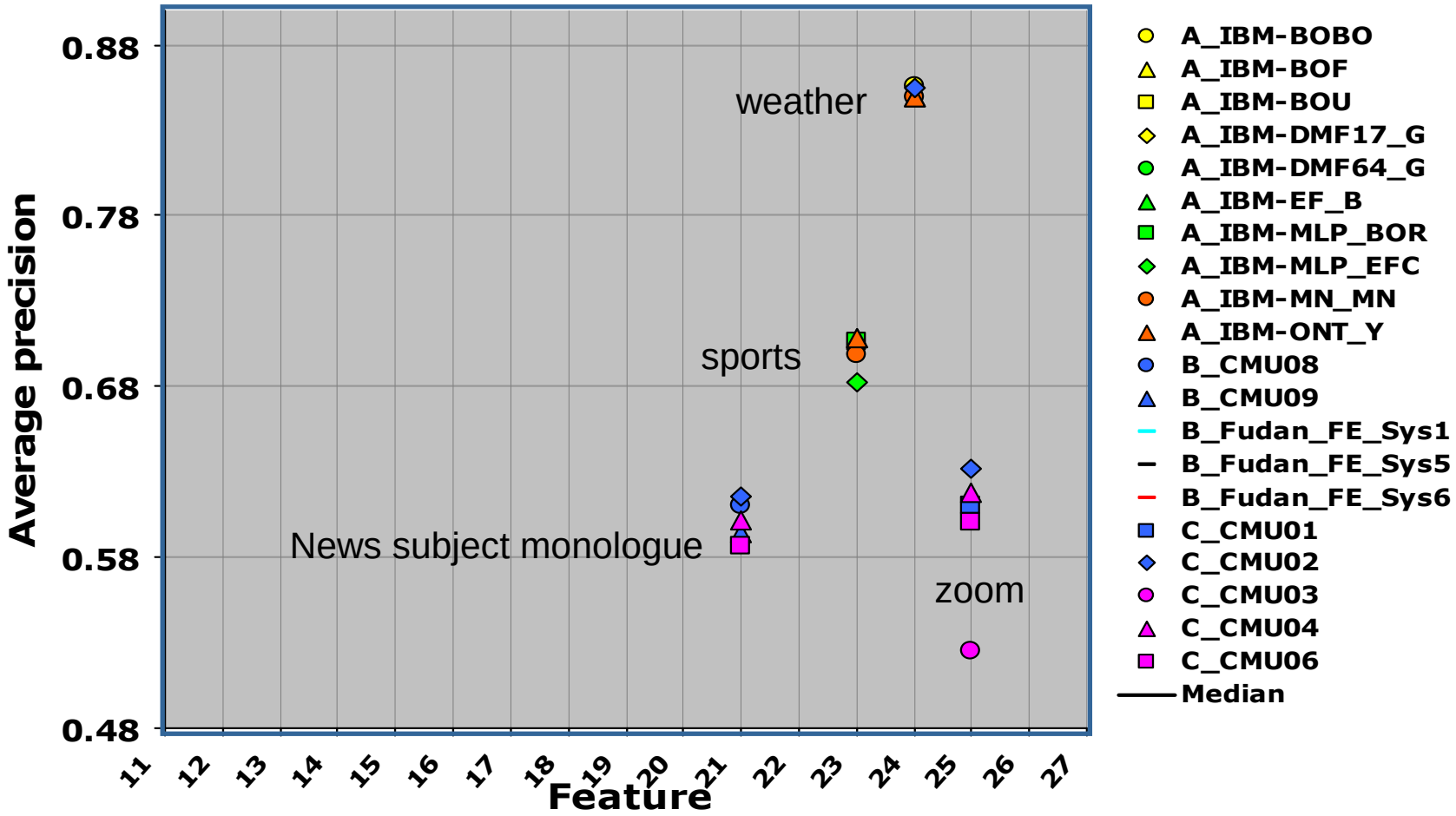
AvgP by feature (top 5 runs by per feature)



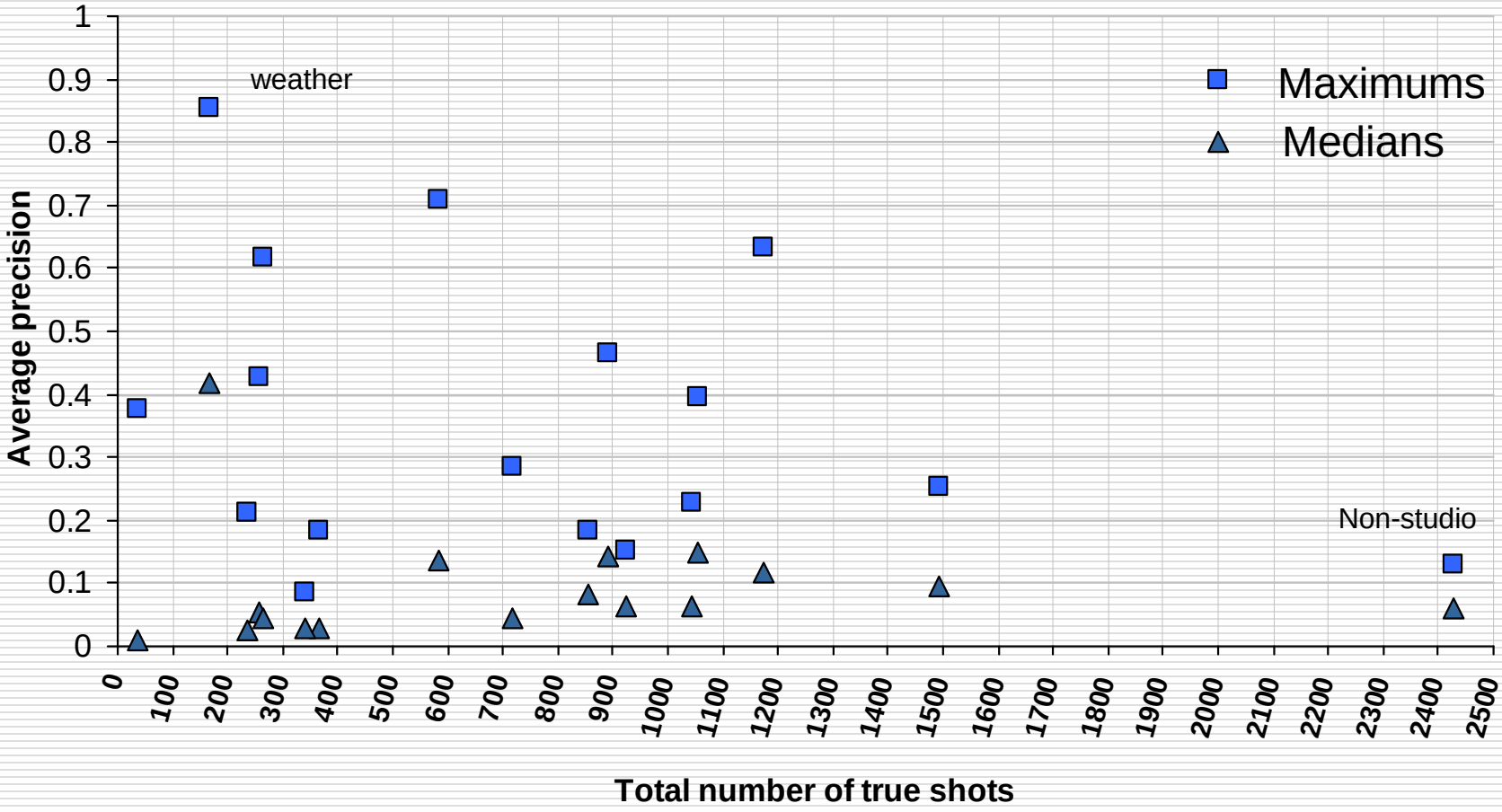
AvgP by feature (top 5 runs by per feature) zoomed: Hard features



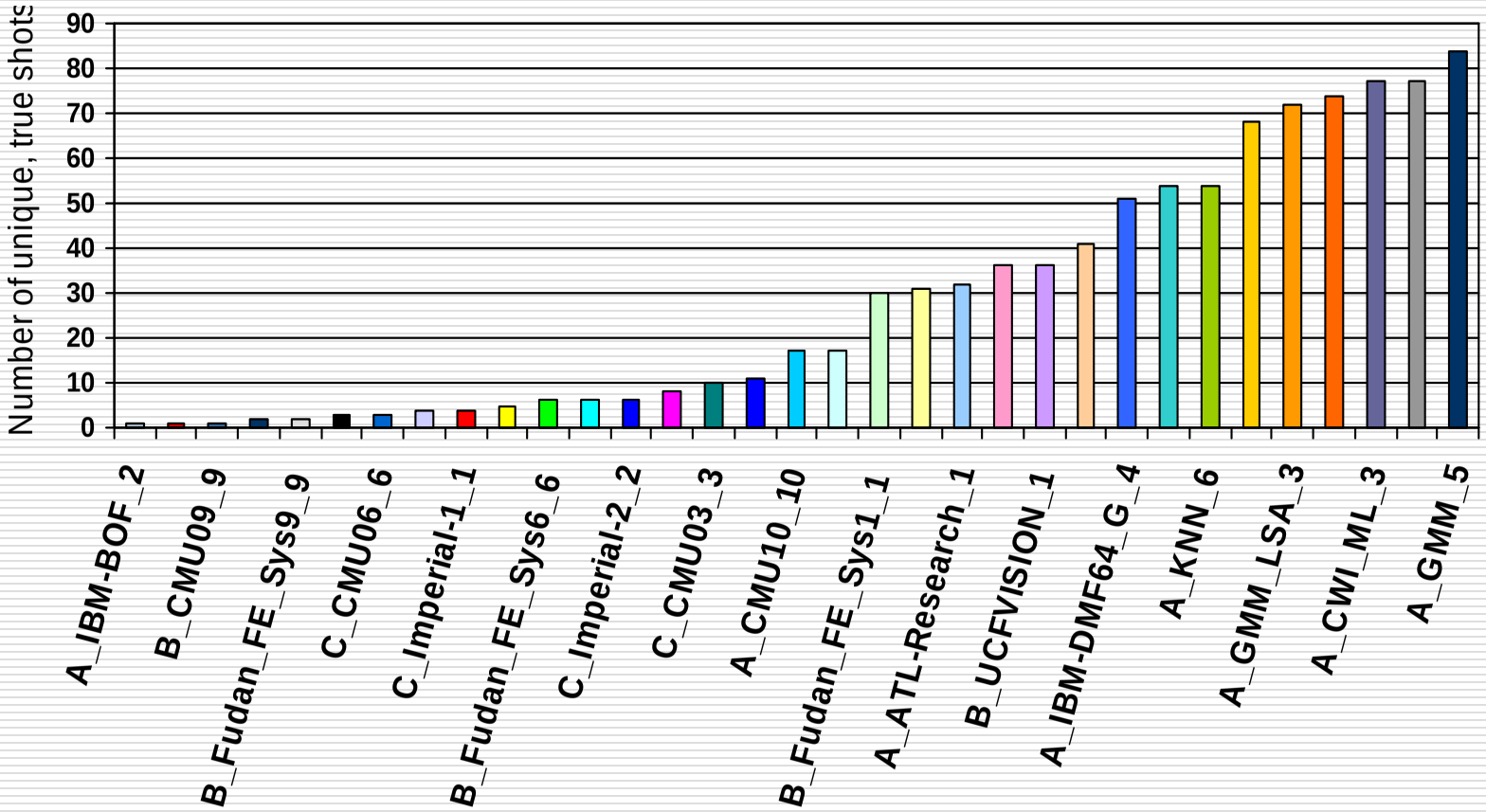
AvgP by feature (top 5 runs per feature) zoomed: Easy features



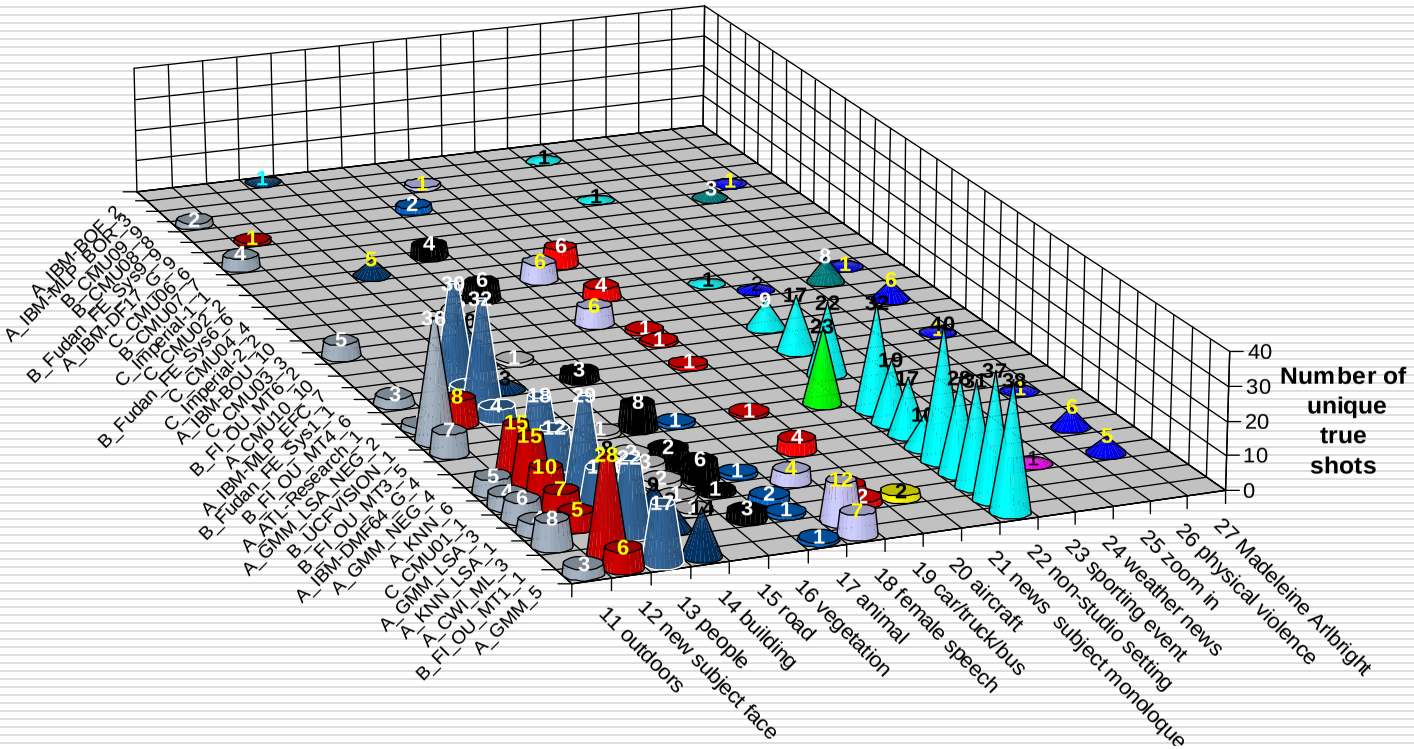
Avg. precision
vs total number true for each feature



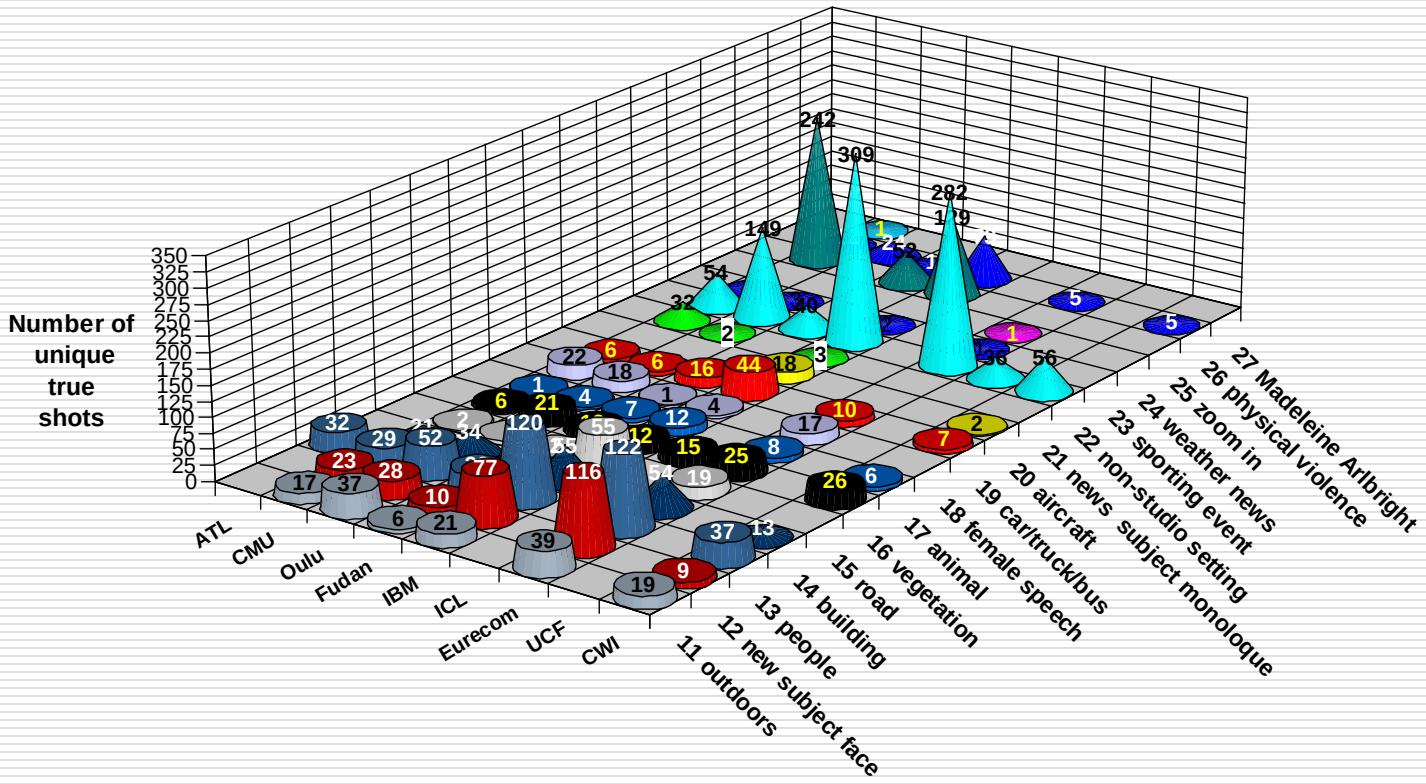
33 of 60 runs contributed one or more unique, true shots



True shots contributed uniquely by run for a feature



True shots contributed uniquely for a feature by a participating group



Group headlines

Accenture Technology Laboratories:

People

- Skin tone detection, count faces

Weather

- 200<length<1000 + color distribution + position of overlay text

Female Speech

- Audio based gender detection + face + moving lips

Group headlines

Carnegie Mellon University:

All features

Presentation follows

IBM Research (US)

Imperial College London (UK)

Institut Eurecom (FR)

Univ. of Central Florida (US)

Univ. Oulu/VTT (FI)

Group headlines

CLIPS-IMAG:

1 feature: M.A.

How would a blind person locate a shot containing Madeline Albright

- Speaker detection (acoustic model)
- M.A. is probably mentioned in one of the preceding shots

Group headlines

A
C
C
C
F
I
I
I
U
U

CWI Amsterdam / University of Twente:

14 features

Working hypothesis: Feature extraction == query by sample

Generative probabilistic retrieval model (same as used for search task), divide frame in pixel blocks

Take a sample of the annotated frames, rank the keyframes based on the likelihood that they generate the query sample

Group headlines

Fudan University: all features

Scene features: grid, color histogram, edge direction, texture, KNN, AdaBoost

Vegetation, Weather: texture+color, SVM, GMM, MaxEnt

Objects:

- Car: Schneiderman
- Animal: vegetation with KNN
- Aircraft: detect context of aircraft

Audio: female speech : 12-MFCC, Pitch, 10-LPC

Group headlines

IBM Research:

All features

Presentation follows

Group headlines

Imperial College London:

Feature 16: Vegetation

Based on grass detector using a colour feature and KNN

Group headlines

Institut Eurecom:

Apply LSI

15 features

Keyframes are segmented into regions

Regions are clustered using K-means

Cluster X frame matrix is reduced by LSI

Use new feature space for GMM and KNN detectors

Group headlines

A
C
C
C
F
I
I
I
U
U

University of Central Florida

2 features

Weather news

- Color histogram similarity

Non-studio setting

- Taken as: all non anchor shots

Group headlines

University of Oulu / VTT:

Extracted 15 features using:

- Motion
- Temporal color correlogram
- Edge gradients
- Several low level audio features (used for outdoors, vehicle noise, sport, monologue
- Feature fusion based on Borda count voting

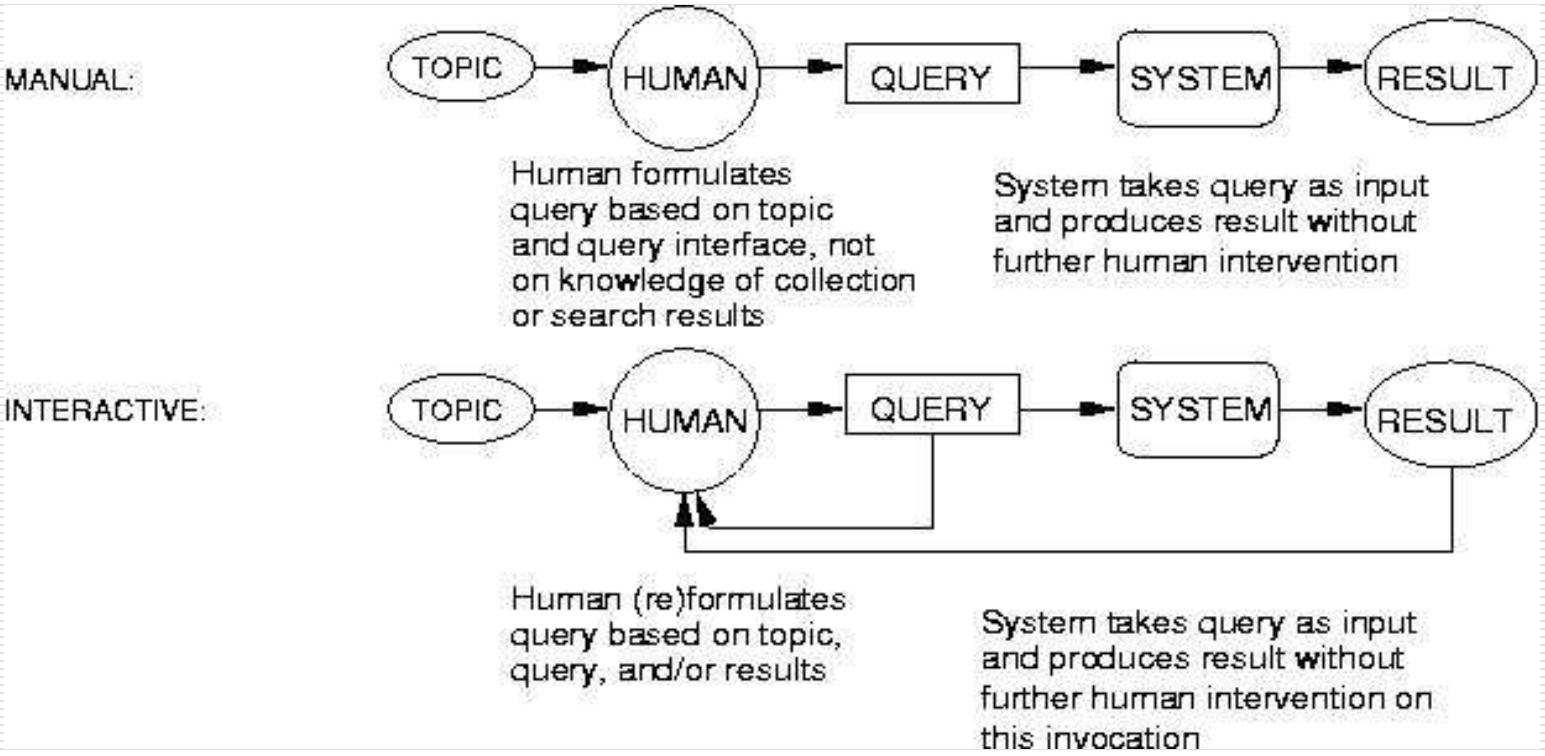
Observations

- Some feature detectors had quite good results
- Are features well chosen for search ?
- Is detection quality good enough?
- Which combination methods work well? Which don't?

TRECVID2003: Search Task

- Search, summarisation, linking, etc. are the ultimate operations on digital video and SBD, features, segmentation, are all enablers for this;
- TRECVID search is an extension of its text-only analogue where systems, including a human in the loop, are presented with a topic and are to return up to 1,000 shots which meet the need;
- Note the unit of retrieval is the *shot*, not the news story;
- Two search modes … manual and interactive, and we're not yet able for full automatic;

Search Types: Interactive and Manual



Search Types: Interactive and Manual

- Topics are MM and the interactions between text, image, video, audio, are complex and understanding how exemplars represent information need, is not really understood;
- This task **really** benefitted from the ASR donated by Jean-Luc Gauvain of LIMSI which is (anecdotally) very accurate;
- One baseline run based on ASR-only was required of every manual system;

Topics

- We can’t achieve the ideal of topics from real users searching our dataset;
- NIST created topics based on a number of basic search types: generic/specific and person/thing/event where there are multiple relevant shots coming from more than one video;
- Videos were viewed by NIST personnel (sound off), notes taken on content, and candidates emerged and were chosen;

25 Topics [total relevant found]

100. Find shots with aerial views containing both one or more buildings and one or more roads [87]
101. Find shots of a basket being made - the basketball passes down through the hoop and net [104]
102. Find shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at [183]
103. Find shots of Yasser Arafat [33]
104. Find shots of an airplane taking off [44]
105. Find shots of a helicopter in flight or on the ground [52]
106. Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery [31]
107. Find shots of a rocket or missile taking off. Simulations are acceptable [62]
108. Find shots of the Mercedes logo (star) [34]

25 Topics

109. Find shots of one or more tanks [16]
110. Find shots of a person diving into some water [13]
111. Find shots with a locomotive (and attached railroad cars if any) approaching the viewer [13]
112. Find shots showing flames [228]
113. Find more shots with one or more snow-covered mountain peaks or ridges. Some sky must be visible behind them. [62]
114. Find shots of Osama Bin Laden [26]
115. Find shots of one or more roads with lots of vehicles [106]
116. Find shots of the Sphinx [12]
117. Find shots of one or more groups of people, a crowd, walking in an urban environment (for example with streets, traffic, and/or buildings) [665]

25 Topics

118. Find shots of Congressman Mark Souder [6]
119. Find shots of Morgan Freeman [18]
120. Find shots of a graphic of Dow Jones Industrial Average showing a rise for one day. The number of points risen that day must be visible. (Manual only) [47]
121. Find shots of a mug or cup of coffee. [95]
122. Find shots of one or more cats. At least part of both ears, both eyes, and the mouth must be visible. The body can be in any position. [122]
123. Find shots of Pope John Paul II [45]
124. Find shots of the front of the White House in the daytime with the fountain running [10]

Evaluation

- Groups allowed to submit up to 10 runs and 37 interactive and 38 manual runs were submitted from 11 groups;
- All submissions were pooled and judged by NIST assessors to variable depths depending on “hit rate” of finding relevant shots;
- Evaluation was trec_eval;

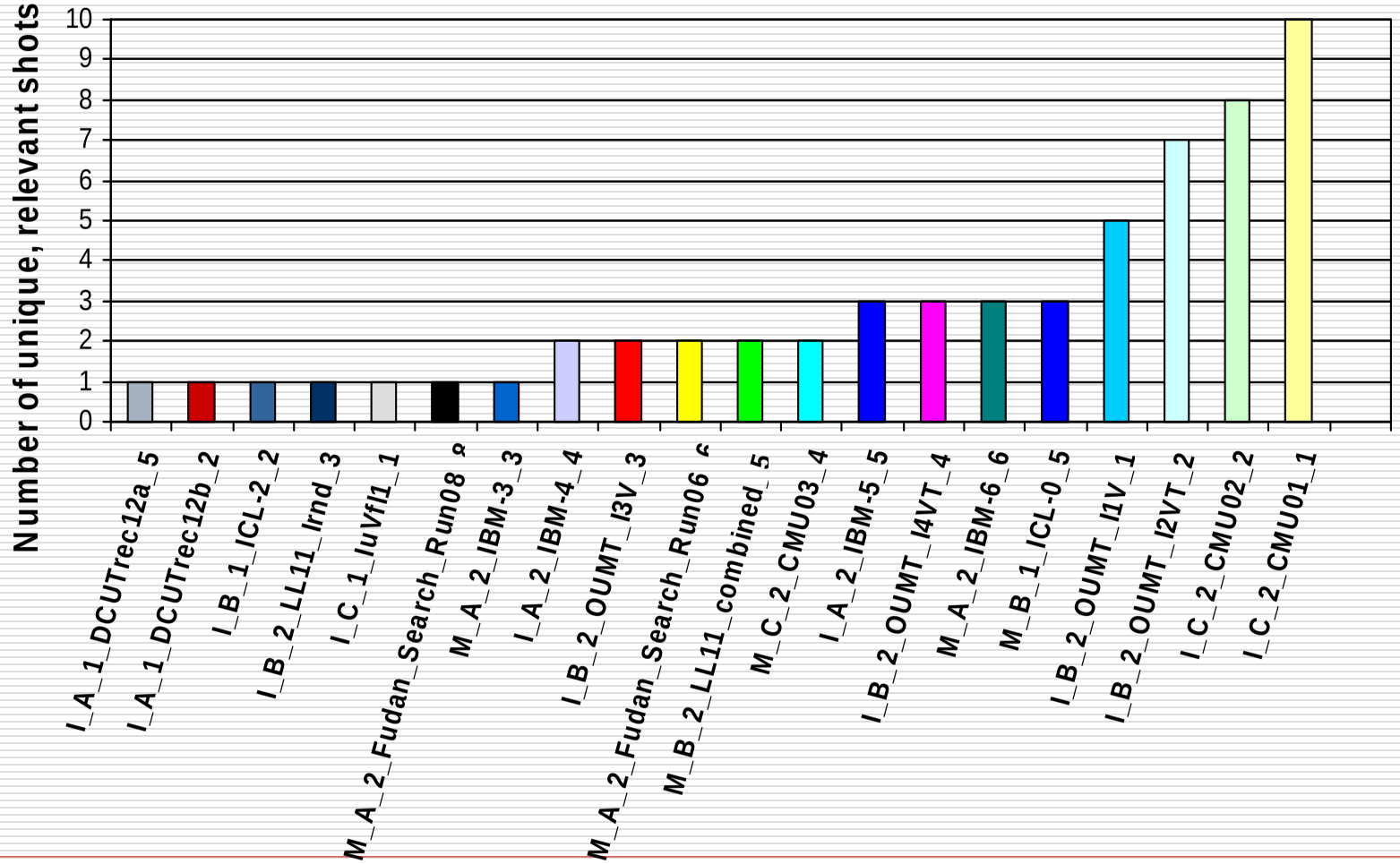
Results

- Absolute performance figures must be taken in their context, so don't believe the numbers … read the papers !
- We tried to level the field by standardising on time spent (15 min.) and thought of introducing a reference system at each site, but TRECVID not yet mature enough for that;
- Also, submitted runs do not necessarily correspond to 1 user, but can be aggregates of multiple users, 2+ groups did this;

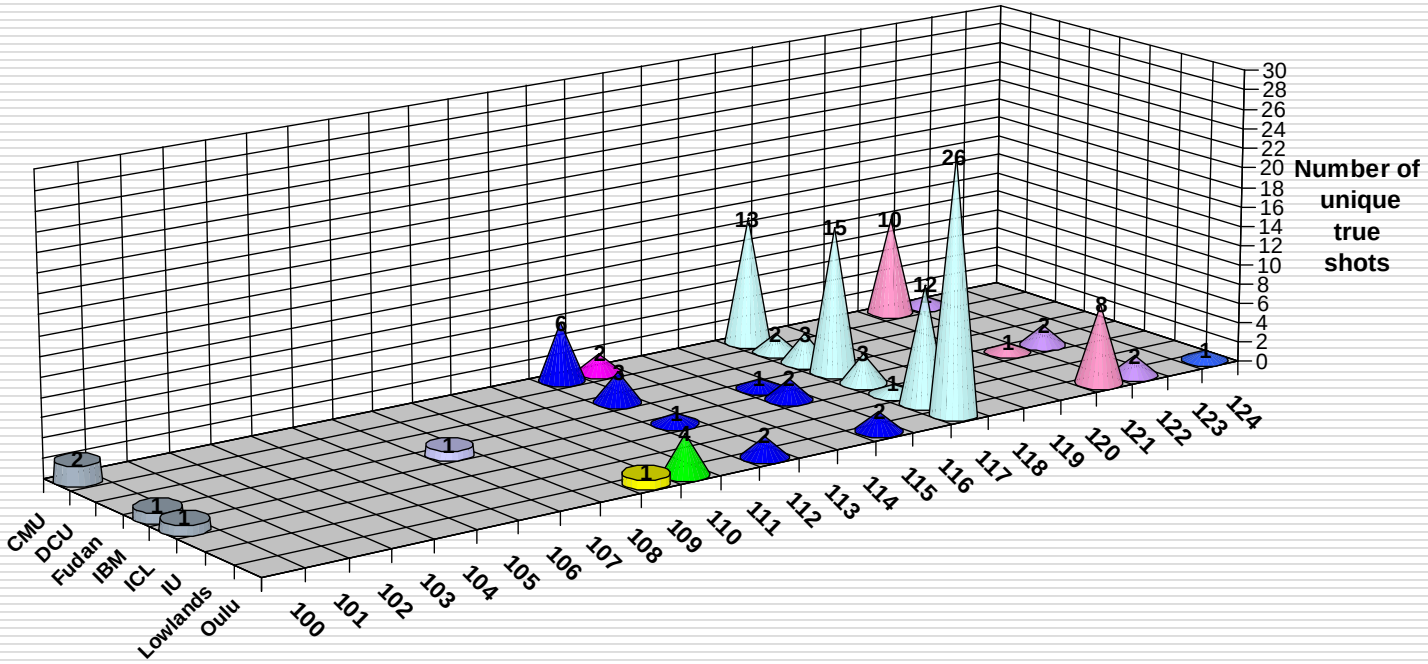
24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	
Carnegie Mellon Univ. (US)			X	X
CLIPS-IMAG (FR)	X		X	
CWI Amsterdam / Univ. of Twente (NL)			X	X
Dublin City University (Irl)		X		X
Fudan Univ. (China)	X	X	X	X
FX-Pal (US)	X			
IBM Research (US)	X	X	X	X
Imperial College London (UK)	X		X	X
Indiana University (US)				X
Institut Eurecom (FR)			X	
KDDI (JP)	X	X		
KU Leuven (BE)	X			
Mediamill/U Amsterdam (NL)				X
National Univ. Singapore (Sing.)		X		X
Ramon Llull Univ. (ES)	X			
RMIT University (Aus)	X			
StreamSage (US)		X		
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

20 of 75 runs contributed 1+ unique, relv. shots

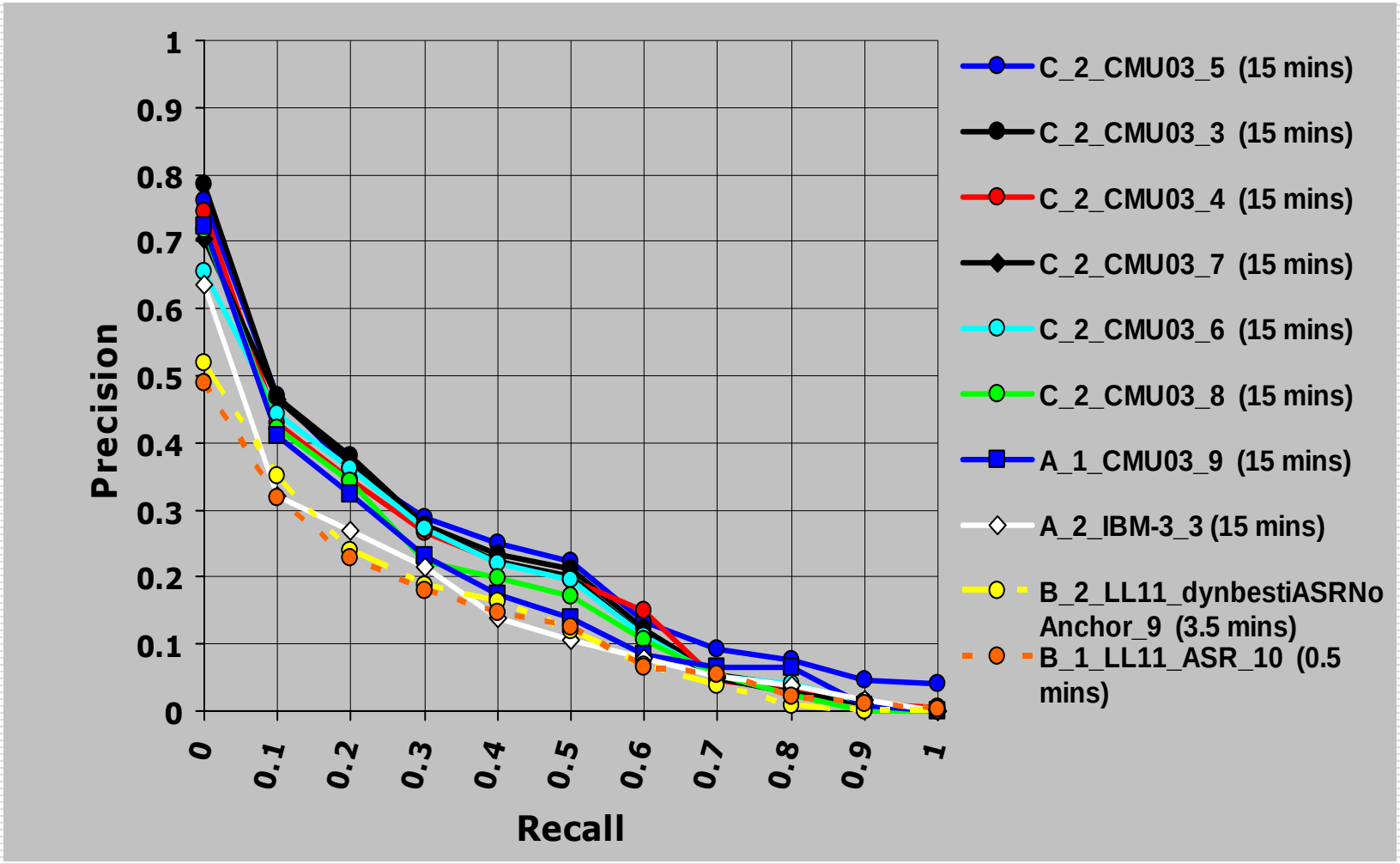


Relevant shots contributed uniquely for a topic by a participating group



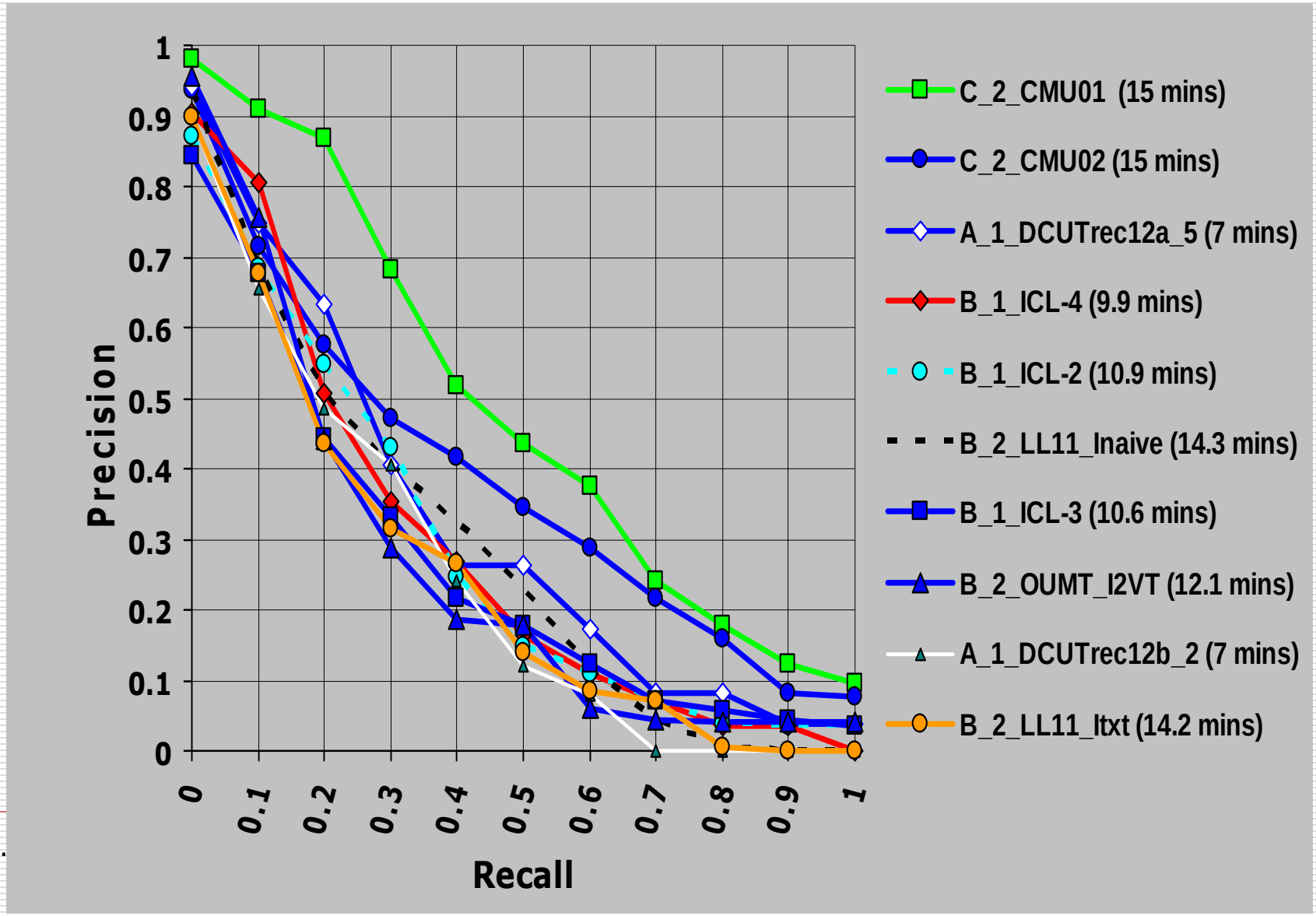
Manual runs - top 10 (of 38)

(with mean human effort / topic)

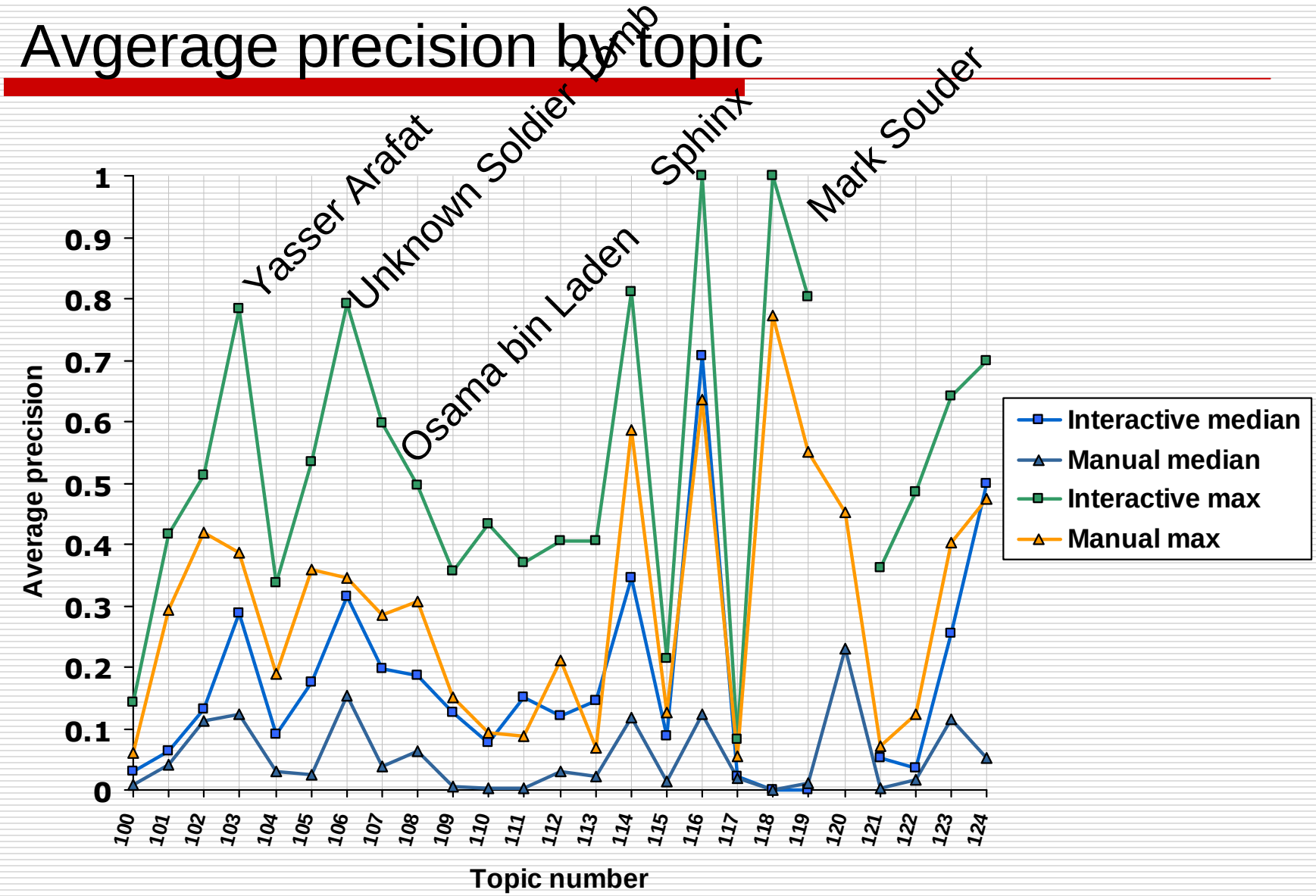


Interactive runs - top 10 (of 36)

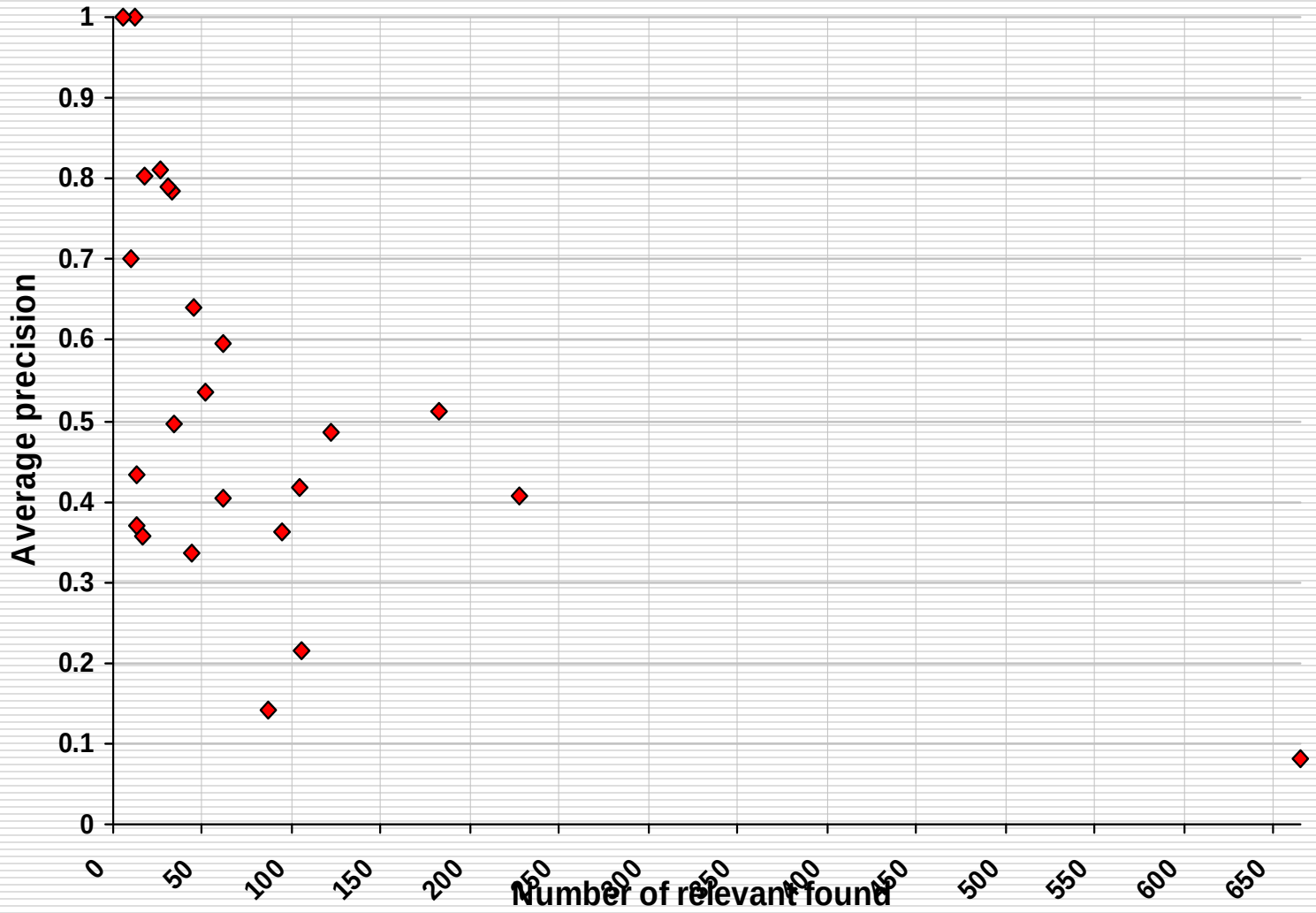
(with mean elapsed time)



Avgerage precision by topic



Average precision (interactive max) vs number relevant shots found



24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

1. Carnegie Mellon University:

Interactive: same system as TV2002 – split topics among 5 individuals, text search across ASR, CC, OCR with storyboarding of keyframes, layout under user control, filtering based on features; another run used improved version with more effective visualisation and browsing;

Manual: multiple retrieval agents across colour, texture, ASR, OCR and some features, combined in different ways, incl. Negative pseudo-RF and "co-retrieval";

Presentation to follow - great results (again);

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

2. Lowlands (CWI & U. Twente):

merging information from multiple modalities:

- run separate Qs for each topic example;
- combine different models of Qs;
- combine sims from system / user judgments;

to build a language model for each shot;

Pre-computing NNs for each keyframe in data;

Interactive better than manual and
combination of text/visual better than text solo

Presentation to follow

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

3. Dublin City University:

Variation of Físchlár in interactive setting with 16 users, 7 mins each, doing 12 topics;

Two system variations were ASR search only and ASR plus query image vs. shot keyframe;

Both had shot-level browsing, user controlled ASR/image search balance, RF allowed by expanding text and/or image;

Aim was to see if users used and benefited from text & image;

Presentation to follow

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	
Univ. Oulu/VTT (FI)			X	X

4. Fudan University:

Manual search using 4 different approaches and then combinations:
- ASR
- colour histogram
- multiple feature (colour hist, edge, coocurrence texture)
- “special search” where user selects most appropriate for topic, from 1. human face recog, 2. general shot features, 3. multiple features, 4. motion (camera and object), 5. colour/texture, 6. colour regions;

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

5. IBM Research:

Examined Spoken Document Retrieval and content based techniques in manual rins

SDR used automatic and phonetic techniques and SDR fusion across multiple match functions, re-ranking shots based on color blobs;

Also did fully automatic multiple example content-based (which is beyond "manual") and fusion of content-based and SDR-based via linear weighting;

Presentation to follow

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

6. Imperial College London:

Used ASR & 11 low-level colour/texture, disregarding image footer likely to contain news ticker;

Features include global colour, colour from frame centre, colour structure descriptors, RGB colour moments,44x27 pixel gray thumbnails, convolution filters, variance, image smoothness and uniformity, ASR;

Retrieval of kNNs, thumbnails on 2D display, RF by user movement of thumbnails, demo ? 2x manual, 4x interactive runs, results good

Presentation to follow

24 Participating Groups

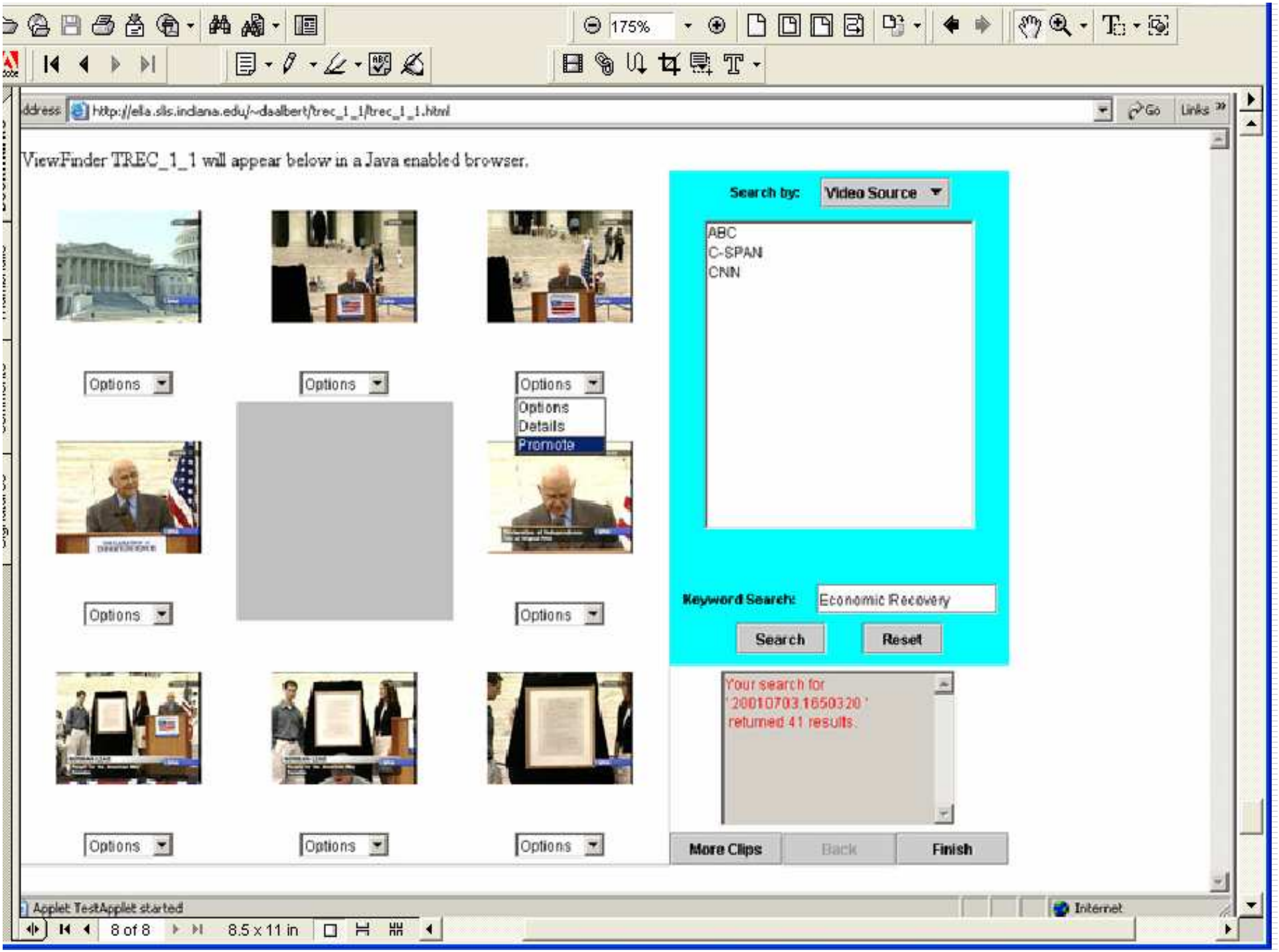
	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	
Univ. of Bremen (D)	X			
Univ. of Central Florida (US)	X	X	X	
Univ. of Iowa (US)	X	X		
Univ. of Kansas (US)	X			
Univ. of North Carolina (US)				X
Univ. Oulu/VTT (FI)			X	X

7. Indiana University:

Used ASR and built a system around interactive text search and query expansion plus video shot browsing;

Interactive search with 1 subject doing all topics, 15 mins max but used only 10 mins;

Future work is to include search based on visual features;


175%
address
http://ella.slis.indiana.edu/~daalbert/trec_1_1/trec_1_1.html
Go
Links
ViewFinder TREC_1_1 will appear below in a Java enabled browser.
Search by:
Video Source
ABC
C-SPAN
CNN
Options
Options
Details
Promote
Keyword Search:
Economic Recovery
Search
Reset
Your search for
'20010703.1650320'
returned 41 results
More Clips
Back
Finish
Applet TestApplet started
Internet
8 of 8
8.5 x 11 in

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

8. MediaMill/University of Amsterdam:

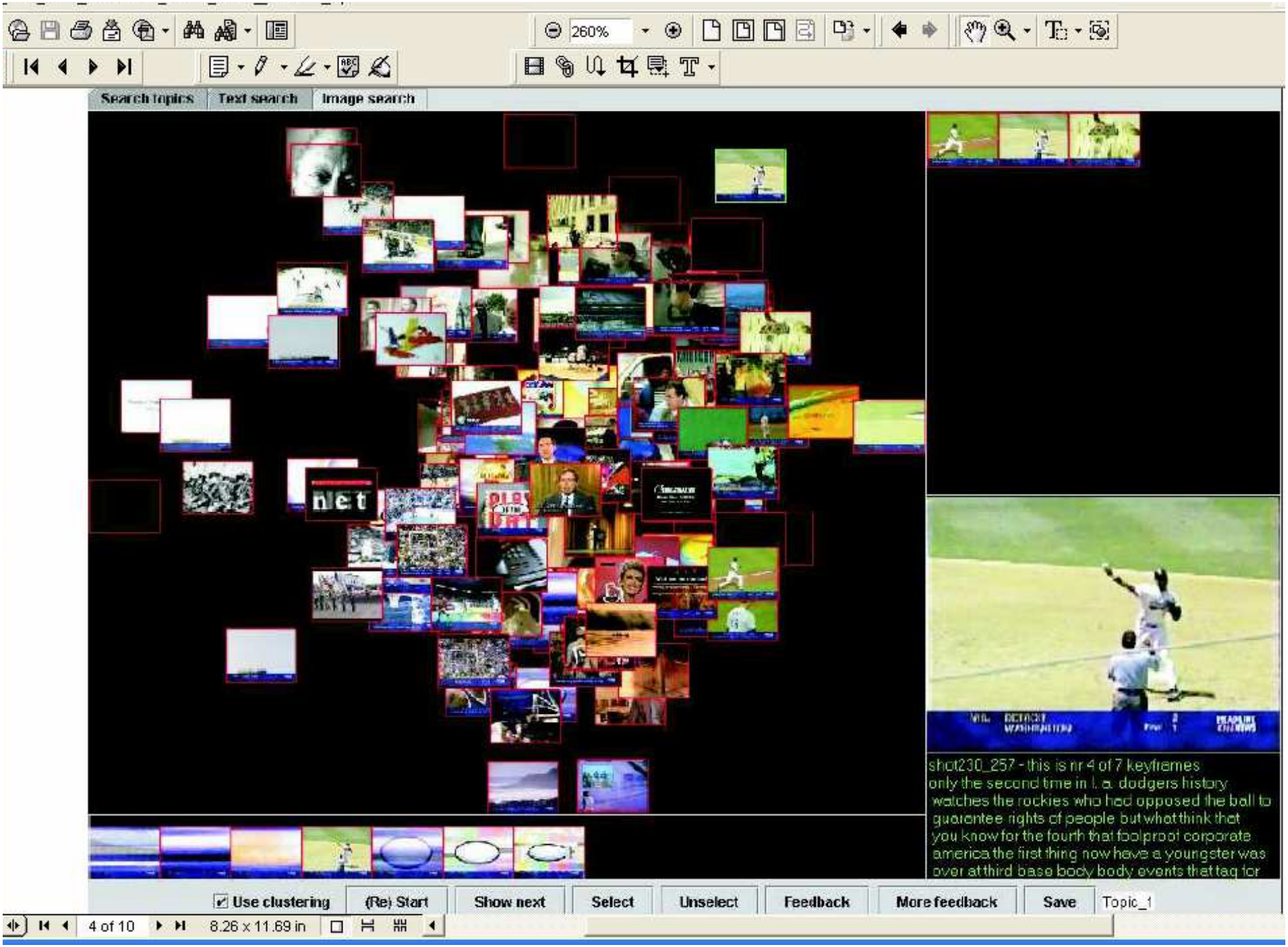
Interactive search with 22 groups of 2 users (in pairs?), using a combination of:
- CMU donated features
- derived "concepts" from LSI over ASR
- keywords from ASR

to yield an active set of 2,000 shots then a snazzy shot browser to select examples;

Only 1 of 11 complete runs submitted.

Used 1 system so no local variant to compare against, and selectively combined sets of users' outputs per topic to generate submission;

"Best" (per topic) objectively selected by submitting the result where the most shots were selected by the users


260%
Search topics
Text search
Image search
net
shot230_257 - this is nr 4 of 7 keyframes
only the second time in l. a. dodgers history
watches the rockies who had opposed the ball to
guarantee rights of people but what think that
you know for the fourth that foolproof corporate
america the first thing now have a youngster was
over at third base body body events that tag for
Use clustering
(Re) Start
Show next
Select
Unselect
Feedback
More feedback
Save
Topic_1
4 of 10
8.26 x 11.69 in

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

9. National University of Singapore:

1. News story retrieval based on ASR and using WordNet and web to expand the original query, POS tagging of query;
2. Filter shots from story based on shot features;
3. Use image & video matching to re-rank remaining shots;

In interactive runs user views top 100 shots and marks relevant ones

Results show marked impact of manual vs. interactive, I.e. user RF;

Presentation to follow

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

10. University of North Carolina (1):

Compare ASR-only, features-only, ASR+features, in interactive search task;

Features: aggregated results of 10 groups from 17 features used in extraction task; ASR was LIMSI, combination was 2xASR;

36 searchers, each doing 12 topics over systems in 15 mins per topic;

Shot browser had annotated storyboard of keyframe + ASR, lots of pre- and post-questionnaire analysis

THE OPEN VIDEO PROJECT

a shared digital video repository

Description of topic 119 :

Find shots of Morgan Freeman

Example images:

Example videos:

Sorry, there are no video examples for this topic.

Enter your search

Search transcripts for:

Submit search

http://152.2.81.55/Csearch.php - http://www.sunvod.com

File Edit View Favorites Tools Help

Back Search Favorites Media

Links Address http://152.2.81.55/Csearch.php Go

Description of topic 112 :

Find shots showing flames

Example images:

Example videos:

Enter your search

Search transcripts for:

Select features :

- ☐ Person in the news
- ☐ People (3 or more)
- ☐ News subject speaking
- ☐ Female speaking
- ☐ Animal (non-human)
- ☐ Vegetation/plants
- ☐ Outdoors
- ☐ Non-studio indoor setting
- ☐ Building
- ☐ Road
- ☐ Car/truck/ bus
- ☐ Aircraft
- ☐ Weather report
- ☐ Physical violence
- ☐ Sporting event
- ☐ Camera Zoom-in

Submit search

Internet

http://152.2.81.55/Fsearch.php - http://www.sunvod.com

File Edit View Favorites Tools Help

Back Search Favorites Media

Links Address http://152.2.81.55/Fsearch.php Go

THE OPEN VIDEO PROJECT

a shared digital video repository

Description of topic 118 :

Find shots of Congressman Mark Souder

Example images:

Example videos:

Sorry, there are no video examples for this topic.

Enter your search

Select features :

- ☐ Person in the news
- ☐ People (3 or more)
- ☐ News subject speaking
- ☐ Female speaking
- ☐ Animal (non-human)
- ☐ Vegetation/plants
- ☐ Outdoors
- ☐ Non-studio indoor setting
- ☐ Building
- ☐ Road
- ☐ Car/truck/ bus
- ☐ Aircraft
- ☐ Weather report
- ☐ Physical violence
- ☐ Sporting event
- ☐ Camera Zoom-in

Submit search

Internet

http://152.2.81.55/CResult1.php?page=1&search_words=pope John Paul&com_features=NewsSubjectFace - http://www.sunvod.com

File Edit View Favorites Tools Help

Back Search Favorites Media

Links Address http://152.2.81.55/CResult1.php?page=1&search_words=pope%20John%20Paul&com_features=NewsSubjectFace Go

Result page: **1** 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 Next

Vertical view

Submit your selection

 vatican says **john paul** is ... says **john paul** is now ... longest serving **pope** this century ... has surpassed **pope** pious the ...	 **pope john paul** ... **pope john paul** went ... **pope john paul** went straight ... say the **pope** avoided the ...	 **pope john paul** ... **pope john paul** the ... **pope john paul** the other ...	 other news **pope john paul** ... news **pope john paul** the ... **pope john paul** the second ...	 crisis develops **pope john paul** ... develops **pope john paul** toda... toda...
 car with **pope john paul** ... with **pope john paul** ... **pope john paul** ...	 eighteen people **pope john paul** ... people **pope john paul** the ... **pope john paul** the second ...	 counties and **pope john paul** ... and **pope john paul** beatified ... **pope john paul** beatified and ...	 just ahead **pope john paul** ... ahead **pope john paul** views ... **pope john paul** views the ...	 accord and **pope john paul** ... and **pope john paul** is ... **pope john paul** is in ...
				

http://152.2.81.55/CResult2.php?page=1&shot_name=shot191_150&videoID=191&search_words=pope John Paul&com_features=NewsSubjectFa Internet

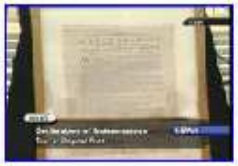
You've selected 0 shot

Search Results for '*actor freeman*':*23* shots found

Result page: **1**

Horizontal view

Submit your selection

	the first printing of the document originally signed by the continental congress in philadelphia on july fourth seventeen seventy six within at tuesday's news conference at the jefferson memorial in washington pennsylvania governor tom ridge **actor** director rob reiner and **actor** morgan **freeman** it's a half hour
	by a landslide so they the most powerful nation can can elect a movie **actor** as a person that don't see a nation way the printer business more come together that ago a movie **actor** as the president
	coming up next former **actor**
	

24 Participating Groups

	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

10. University of North Carolina (2):

Results … no statistical difference in Precision, but statistical difference in recall where features-only was less than the other two … poor feature recognition accuracy ?

Large variability in time taken per search, avg 4 to 6 minutes;

Much evaluation of user's perception and satisfaction;

Some helpful pointers on future assessment of interactive search;

24 Participating Groups

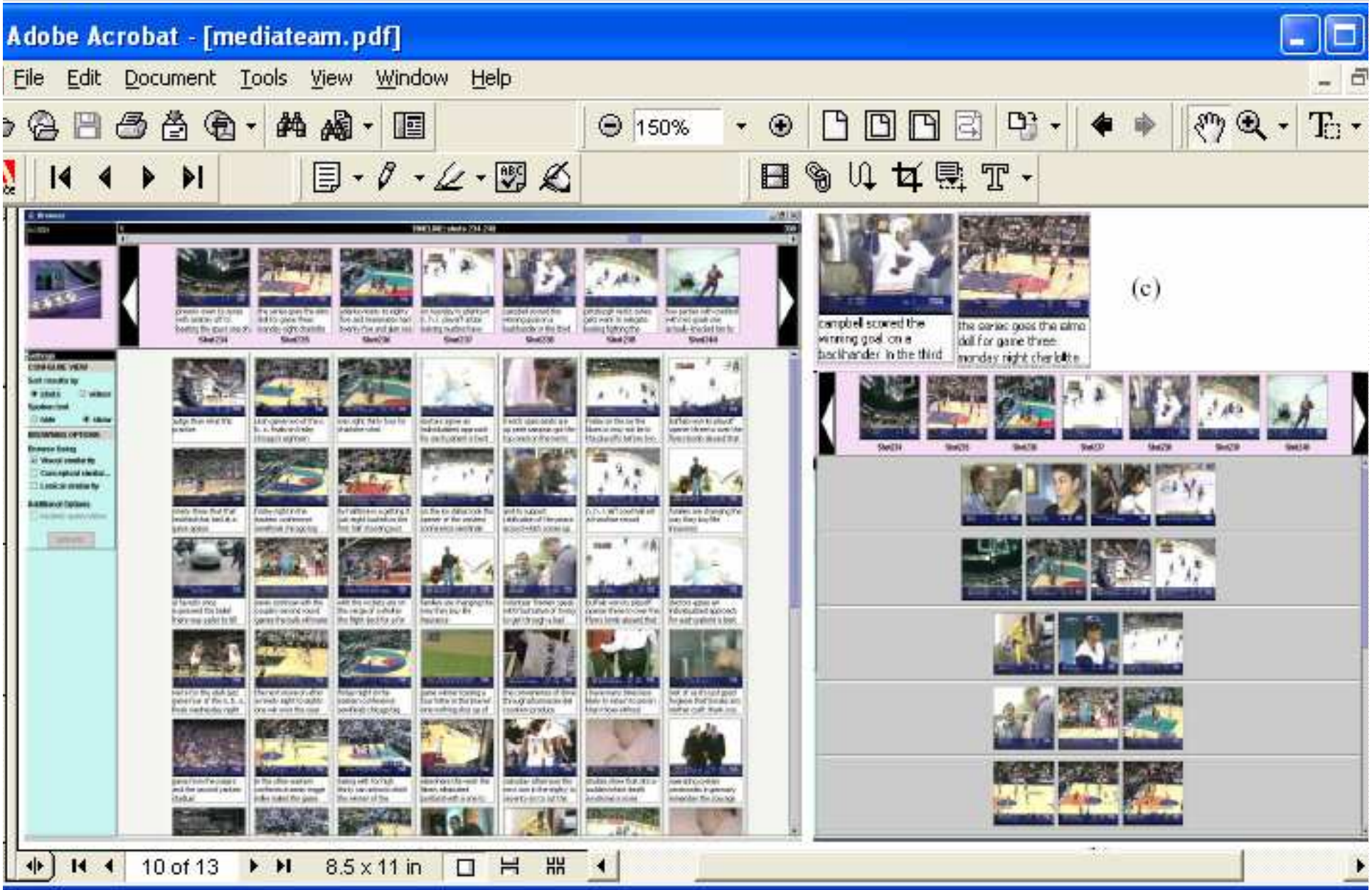
	Shots	Stories	Features	Search
Accenture Technology Laboratories (US)	X		X	

11. University of Oulu/VTT:

VIRE has interactive cluster/temporal shot browsing and shot similarity based on visual (colour, edge structure, motion), conceptual (15x features from feature set) and lexical (from ASR) similarity;

Manual runs .. Pre-select combinations of features and images from topic;

Interactive runs … 8 people, 2 systems, 9.5 mins per topic, (a) browse by visual features only and (b) browse by visual features plus ASR … result indicates no significant difference;


Adobe Acrobat - [mediateam.pdf]
File Edit Document Tools View Window Help
150%
campbell scored the winning goal on a backhander in the third
the senior goes the extra mile for game three monday night charlotte
(c)
10 of 13
8.5 x 11 in

Observations

- Lots of variation, interesting shot browsing interfaces, mixture of interactive & manual;
- Approximately as much use of donated features as TV2002;
- A lot more participation, more runs, better at the upper end … quite respectable curves !
- Nearly a dozen groups can now complete the search task and the demos are impressive;

Plans

- Make notebook papers, presentations, and feedback on plans for TRECVID 2004 available on the website in December
- Make final papers available on the website by mid March 2004
- Plans: probably complete 2 yr plan
  - Add 80 hours of new test data from same news sources
  - Repeat 2003 tasks with some improvements
- More information as it develops at:

  www-nlpir.nist.gov/projects/trecvid