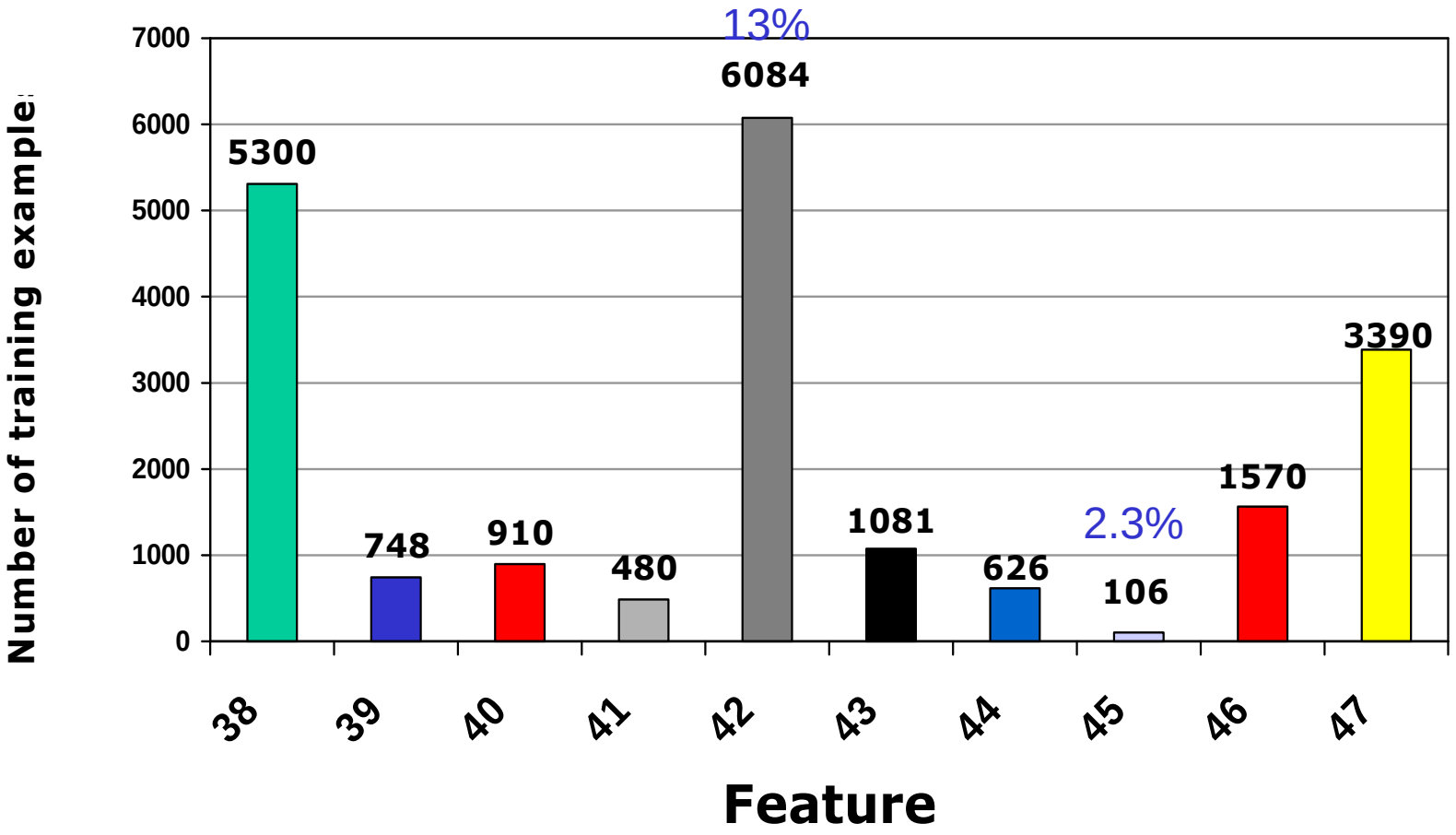
TRECVID-2005 High-Level Feature task: Overview

Wessel Kraaij
TNO
&
Paul Over
NIST

High-level feature task

- Goal: Build benchmark collection for detection methods
- Secondary goal: feature-indexing could help search/browsing
- Feature set selected from feature set used for annotation of development data (LSCOM-lite)
- Examples of thing/activity/person/location
- Collaborative development data annotation effort
  - Tools from CMU and IBM (new tool)
  - 39 features and about 100 annotators
  - multiple annotations of each feature for a given shot
- Range of frequencies in the common development data annotation

True examples in the common training data



High-level feature evaluation

- Each feature assumed to be binary: absent or present for each master reference shot
- Task: Find shots that contain a certain feature, rank them according to confidence measure, submit the top 2000
- NIST pooled submissions to depth 250
- Evaluate performance quality by measuring the *average precision* etc. of each feature detection method

10 Features

38. People walking/running: segment contains video of more than one person walking or running (tv4: 35)
39. Explosion or fire: segment contains video of an explosion or fire
40. Map: segment contains video of a map
41. US flag: segment contains video of a US flag
42. Building exterior: segment contains video of the exterior of a building (tv3: 14)
43. Waterscape/waterfront: segment contains video of a waterscape or waterfront
44. Mountain: segment contains video of a mountain or mountain range with slope(s) visible
45. Prisoner: segment contains video of a captive person, e.g., imprisoned, behind bars, in jail, in handcuffs, etc.
46. Sports: segment contains video of any sport in action (tv3: 23)

Participants (22/42) *(up from 12/33 in 2004)*

Bilkent University	Turkey	-- LL HL SE
Carnegie Mellon University	USA	-- -- HL SE
CLIPS-IMAG, LSR-IMAG, Laboratoire LIS	France	SB –- HL --
Columbia University	USA	-- -- HL SE
Fudan University	China	SB LL HL SE
FX Palo Alto Laboratory	USA	SB –- HL SE
Helsinki University of Technology	Finland	-- -- HL SE
IBM	USA	SB –- HL SE
Imperial College London	UK	SB –- HL SE
Institut Eurecom	France	-- -- HL --
Johns Hopkins University	USA	-- -- HL --
Language Computer Corporation (LCC)	USA	-- -- HL SE
LIP6-Laboratoire d'Informatique de Paris 6	France	-- -- HL --
Lowlands Team (CWI, Twente, U. of Amsterdam)	Netherlands	-- -- HL SE
Mediamill Team (Univ. of Amsterdam)	Netherlands	-- LL HL SE
National ICT Australia	Australia	SB LL HL --
National University of Singapore (NUS)	Singapore	-- -- HL SE
SCHEMA-Univ. Bremen Team	EU	-- -- HL SE
Tsinghua University	China	SB LL HL SE
University of Central Florida / University of Modena	USA,Italy	SB LL HL SE
University of Electro-Communications	Japan	-- -- HL --
University of Washington	USA	-- -- HL --

Who worked on which features

Bilkent University	38
Carnegie Mellon University	38 39 40 41 42 43 44 45 46 47
CLIPS-IMAG, LSR-IMAG, Laboratoire LIS	38 39 40 41 42 43 44 45 46 47
Columbia University	38 39 40 41 42 43 44 45 46 47
Fudan University	38 39 40 41 42 43 44 45 46 47
FX Palo Alto Laboratory	38 39 40 41 42 43 44 45 46 47
Helsinki University of Technology	38 39 40 41 42 43 44 45 46 47
IBM	38 39 40 41 42 43 44 45 46 47
Imperial College London	38 39 40 41 42 43 44 45 46 47
Institut Eurecom	38 39 40 41 42 43 44 45 46 47
Johns Hopkins University	38 39 40 41 42 43 44 45 46 47
Language Computer Corporation (LCC)	38 39 40 41 42 43 44 45 46 47
LIP6-Laboratoire d'Informatique de Paris 6	40
Lowlands Team (CWI, Twente, U. of Amsterdam)	38 39 40 41 42 43 44 45 46 47
Mediamill Team (Univ. of Amsterdam)	38 39 40 41 42 43 44 45 46 47
National ICT Australia	38 39 40 41 42 43 44 45 46 47
National University of Singapore (NUS)	38 39 40 41 42 43 44 45 46 47
SCHEMA-Univ. Bremen Team	40 41 43
Tsinghua University	38 39 40 41 42 43 44 45 46 47
University of Central Florida / Univ. of Modena	39 40 41 42 43 44 45 46 47
University of Electro-Communications	43 44
University of Washington	38 39 40 41 42 43 44 45 46 47

Number of runs each training type

Tr-Type	2005	2004	2003
A	79 (71.8%)	45 (54.2%)	22 (36.7%)
B	24 (21.8%)	27 (32.5%)	20 (33.3%)
C	7 (6.3%)	11 (13.3%)	18 (30.0%)
Total runs	110	83	60

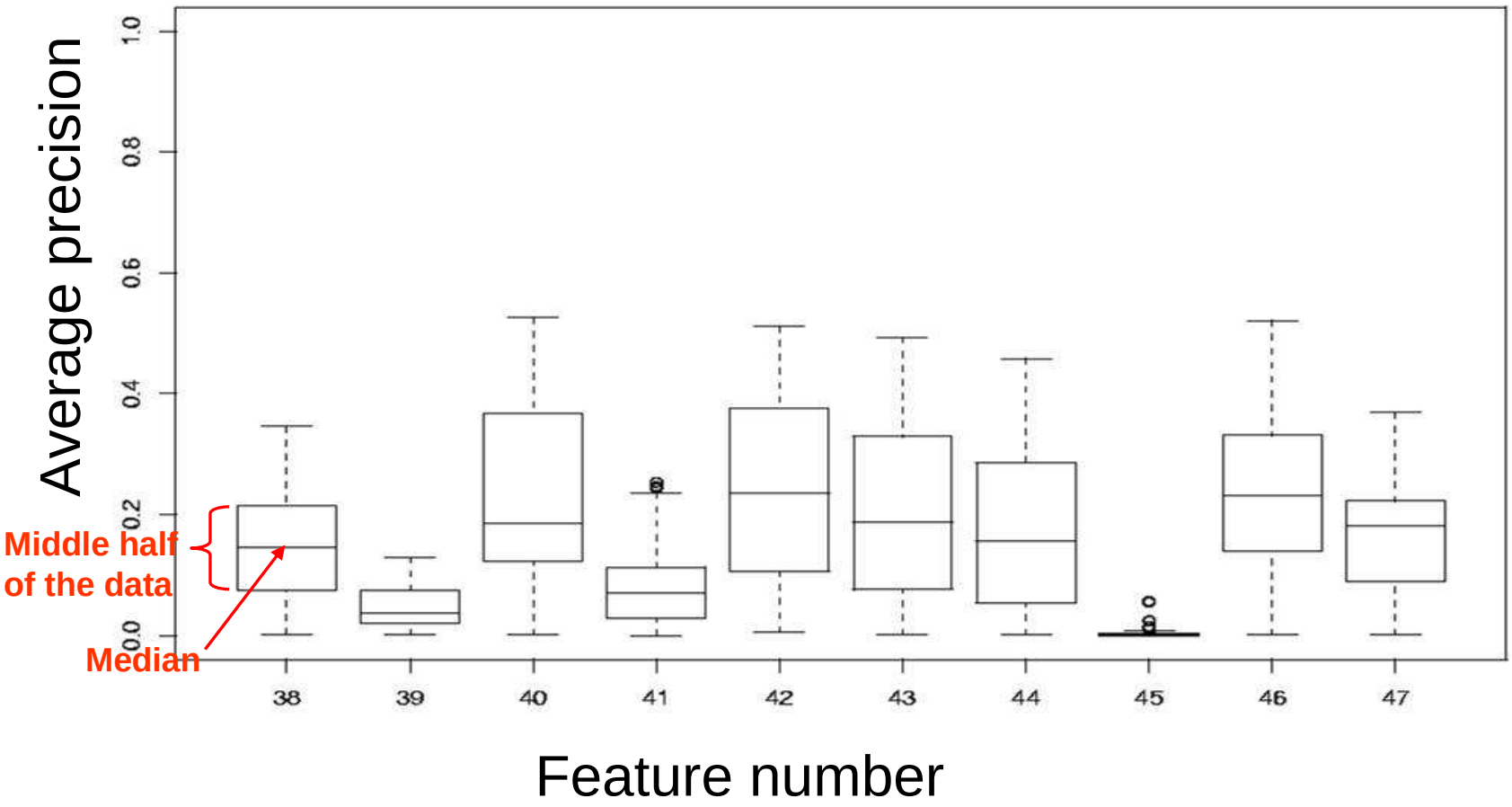
System training type:

A - Only on common dev. collection and the common annotation

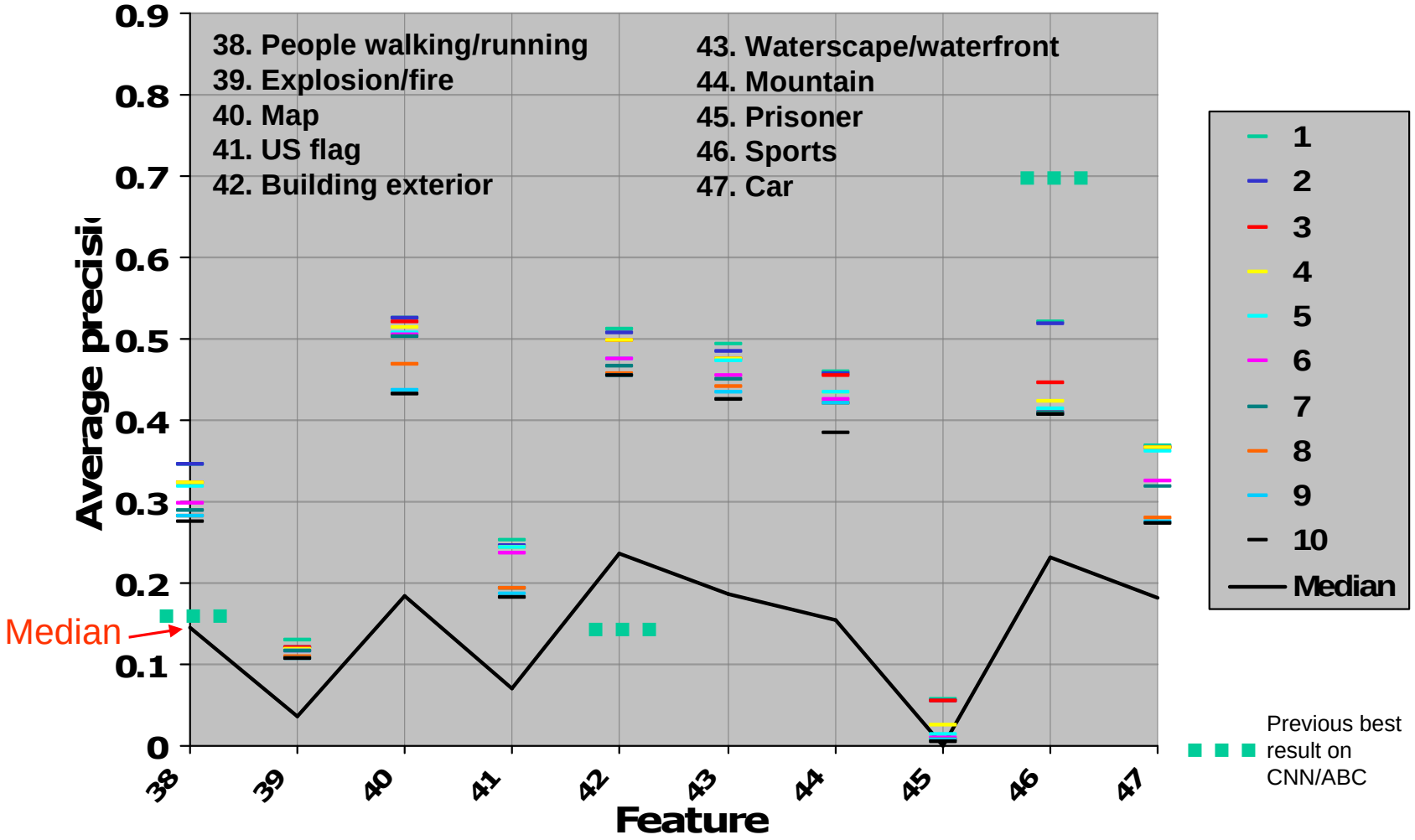
B - Only on common dev. collection but not on (just) the common annotation

C - not of type A or B

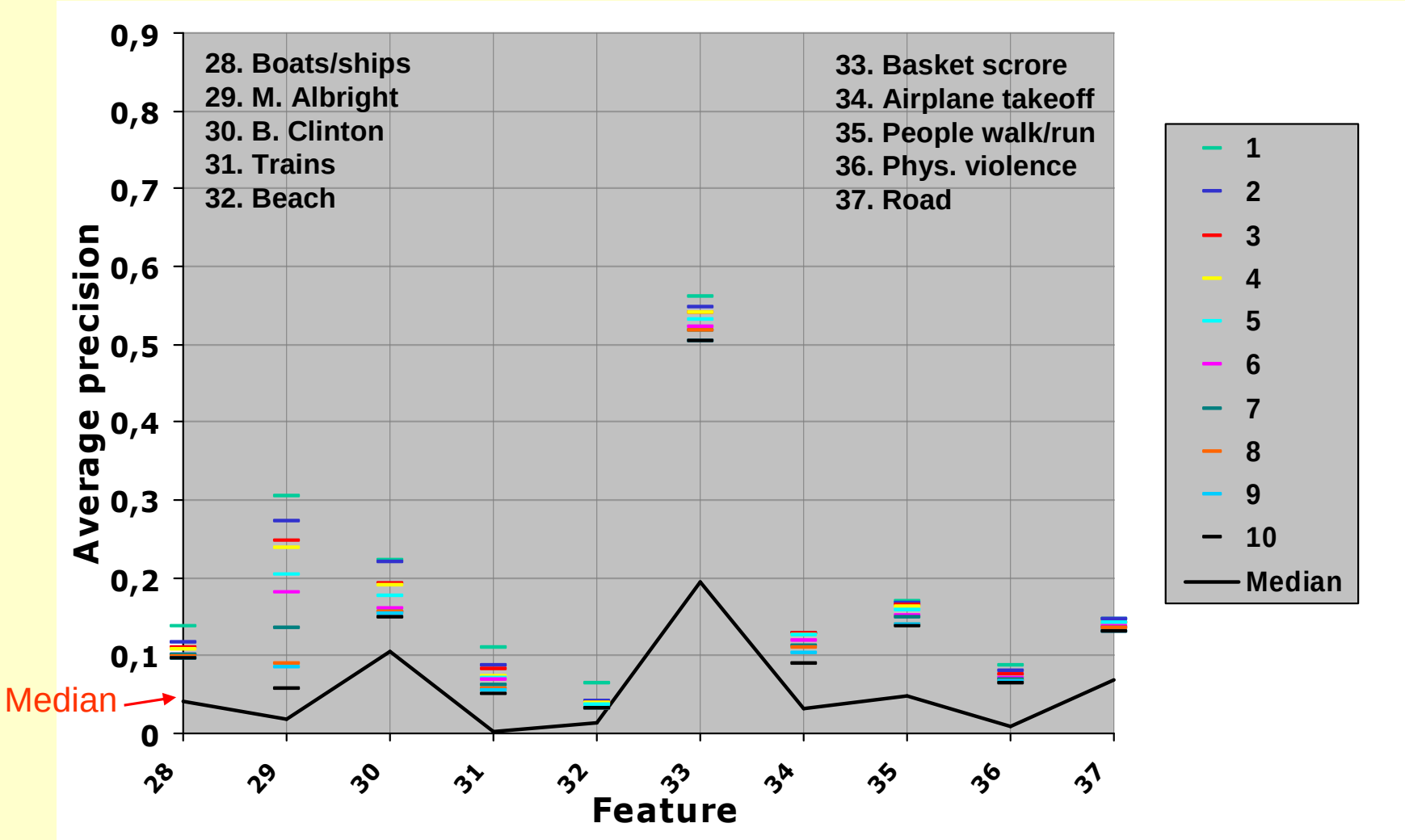
AvgP by feature (all runs)



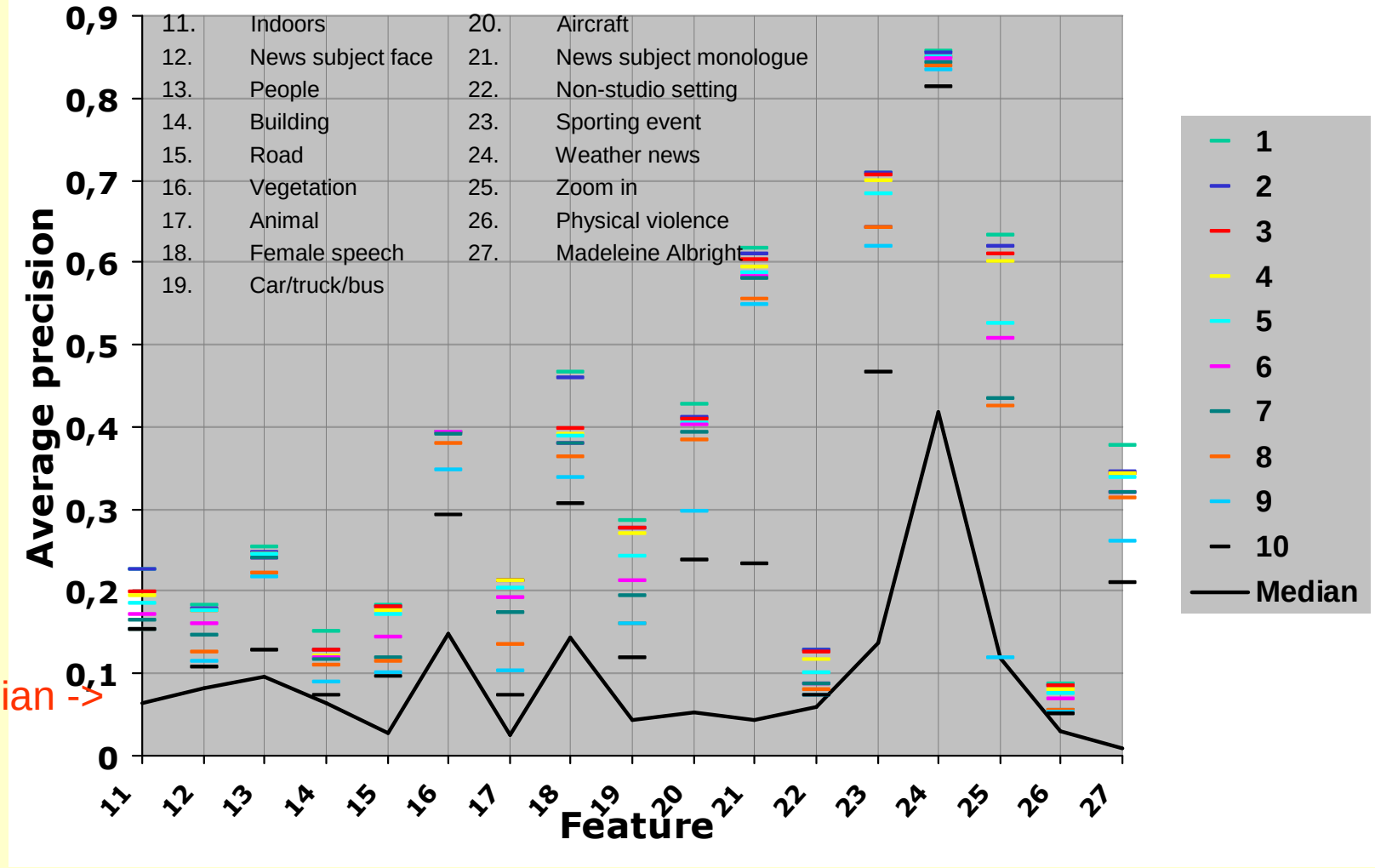
2005: AvgP by feature (top 10 runs)



2004: AvgP by feature (top 10 runs)

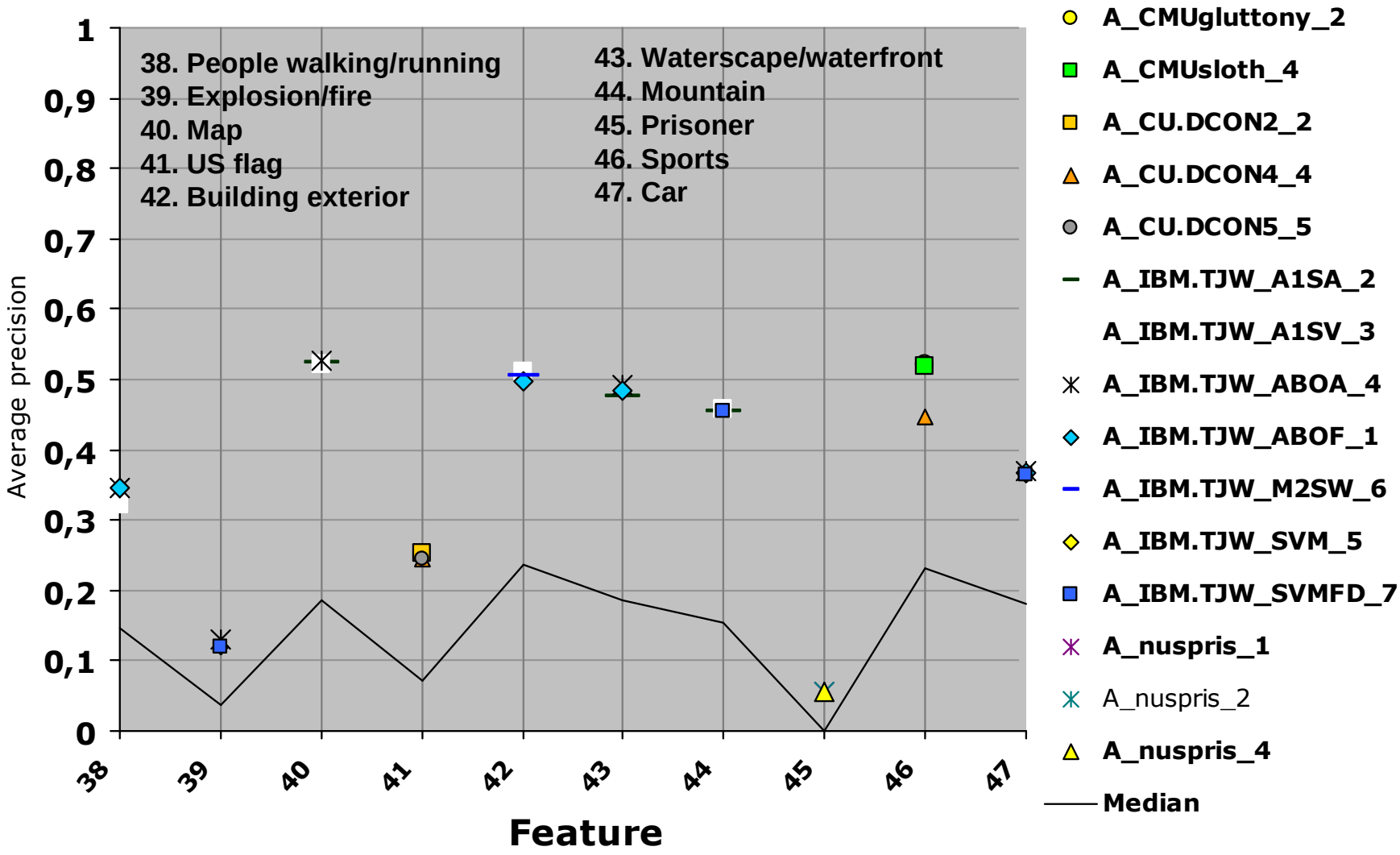


2003: AvgP by feature (top 10 runs)

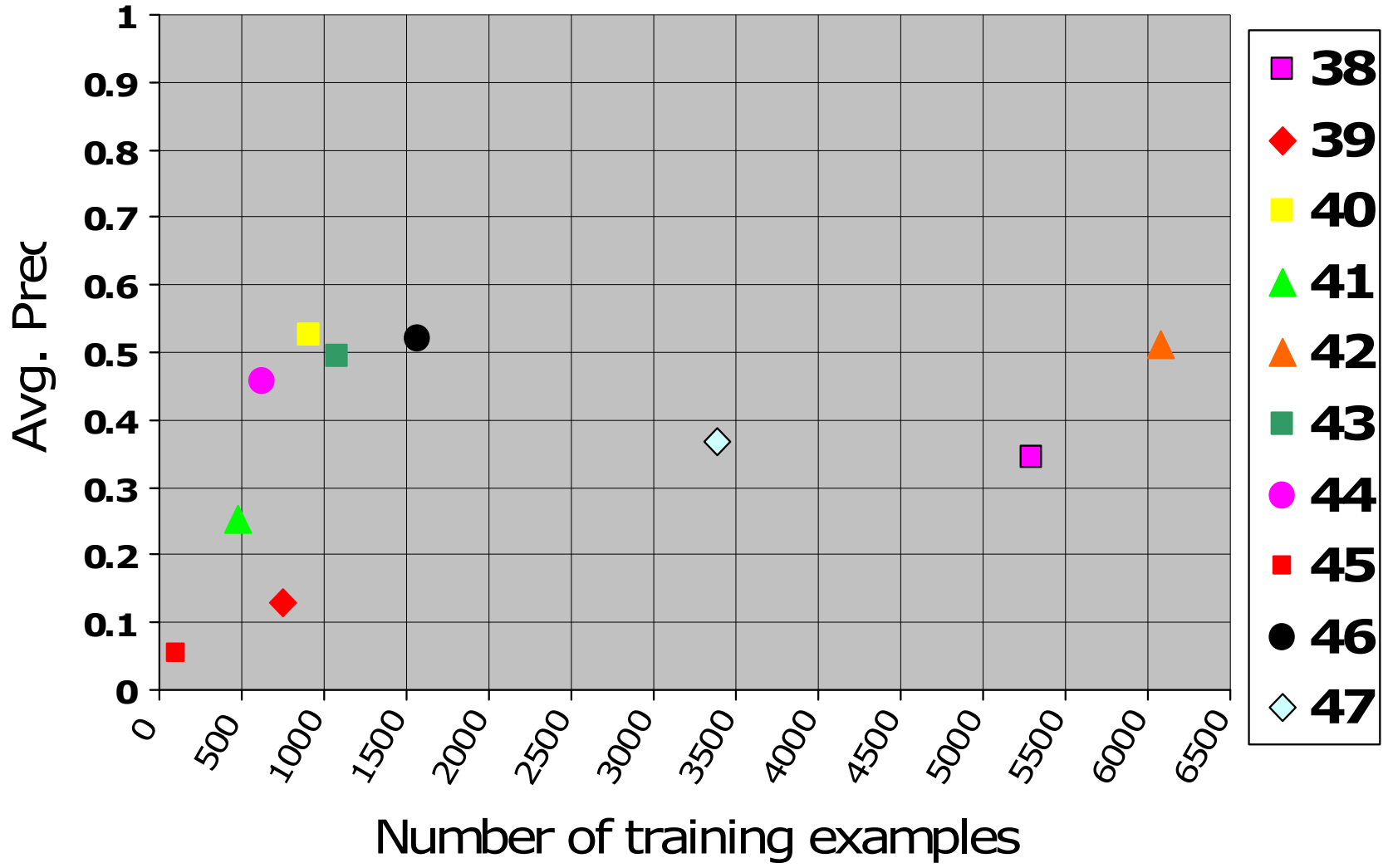


Median ->

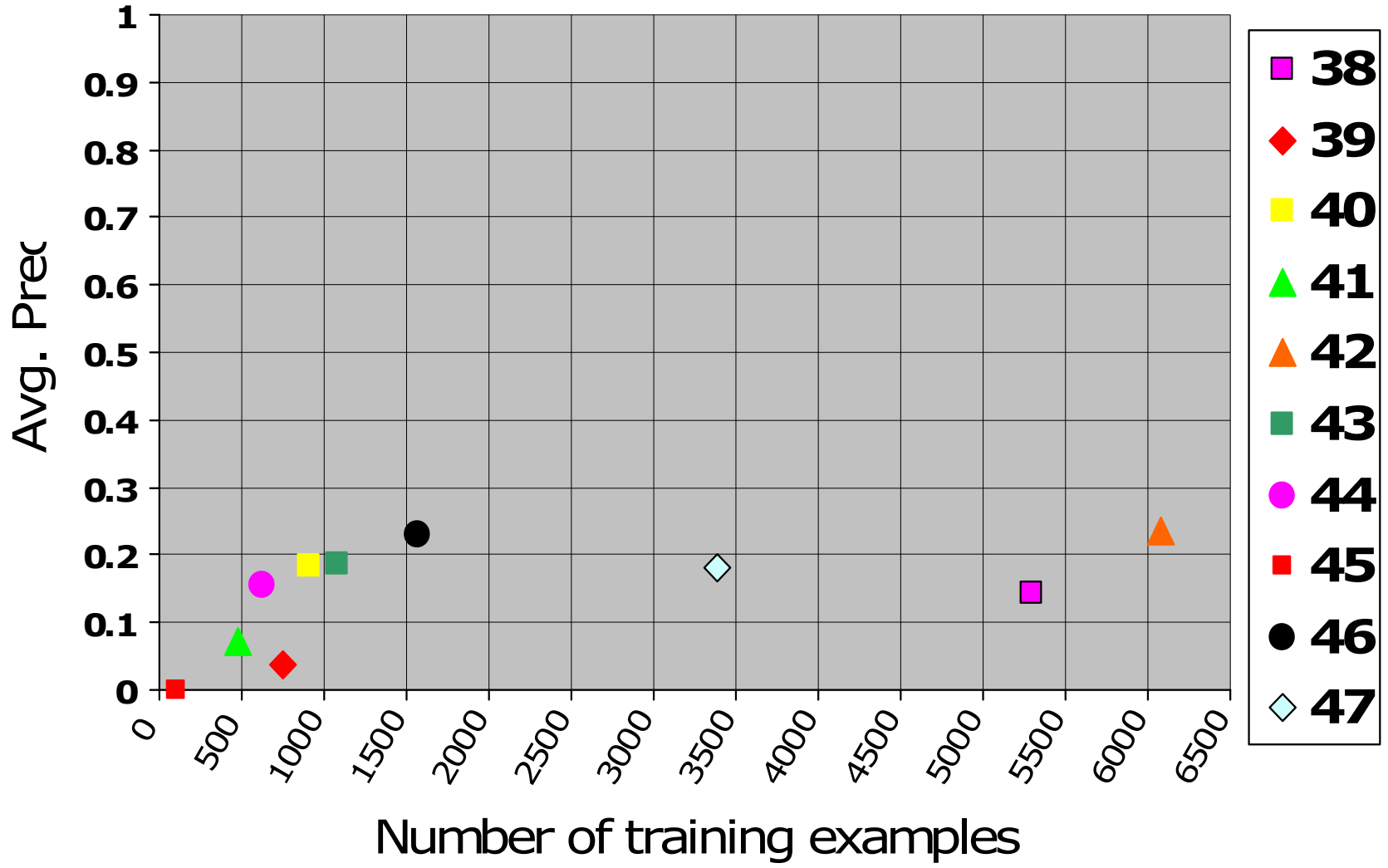
AvgP by feature (top 3 runs by per feature)



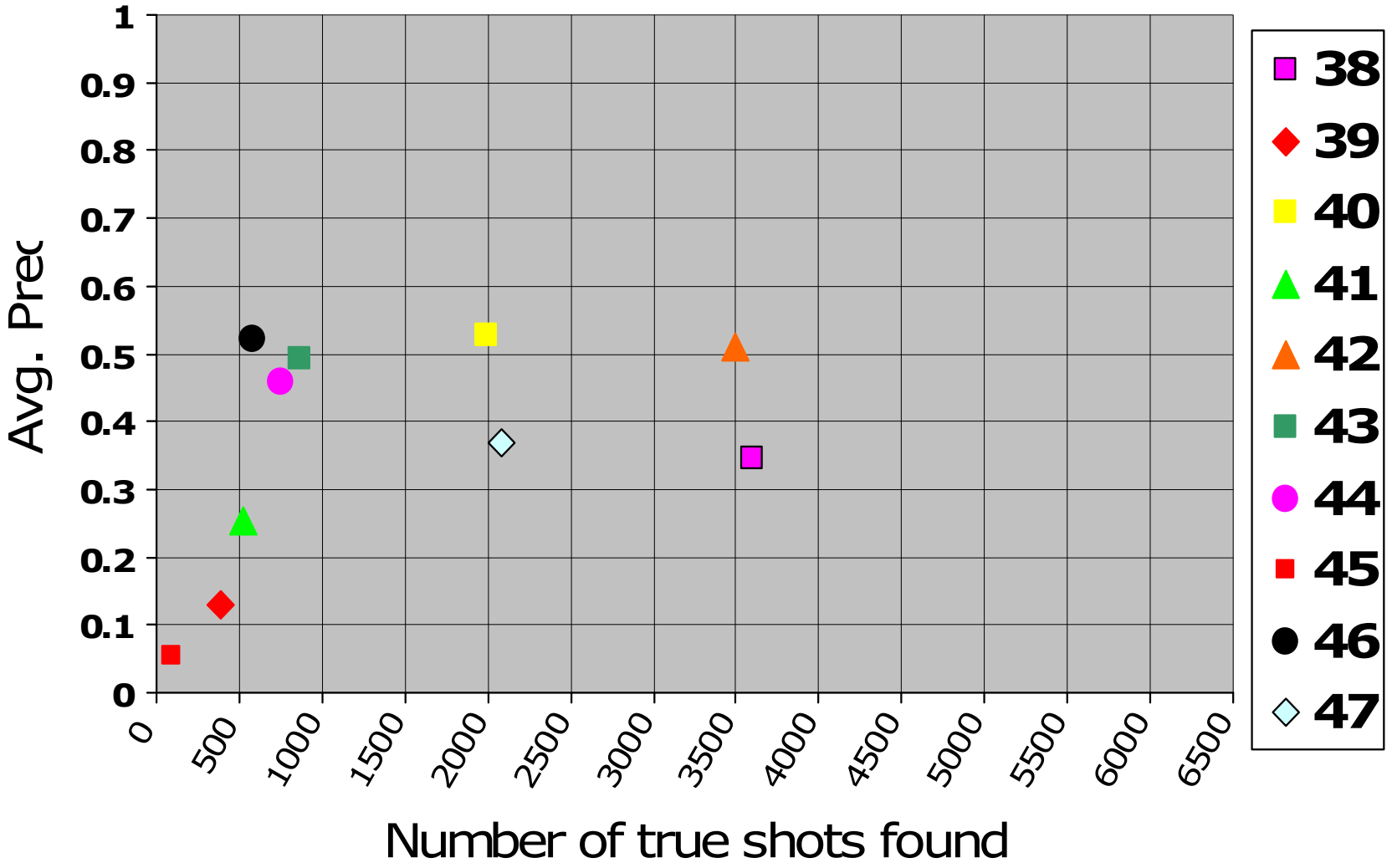
Max AvgP by number of annotated training examples



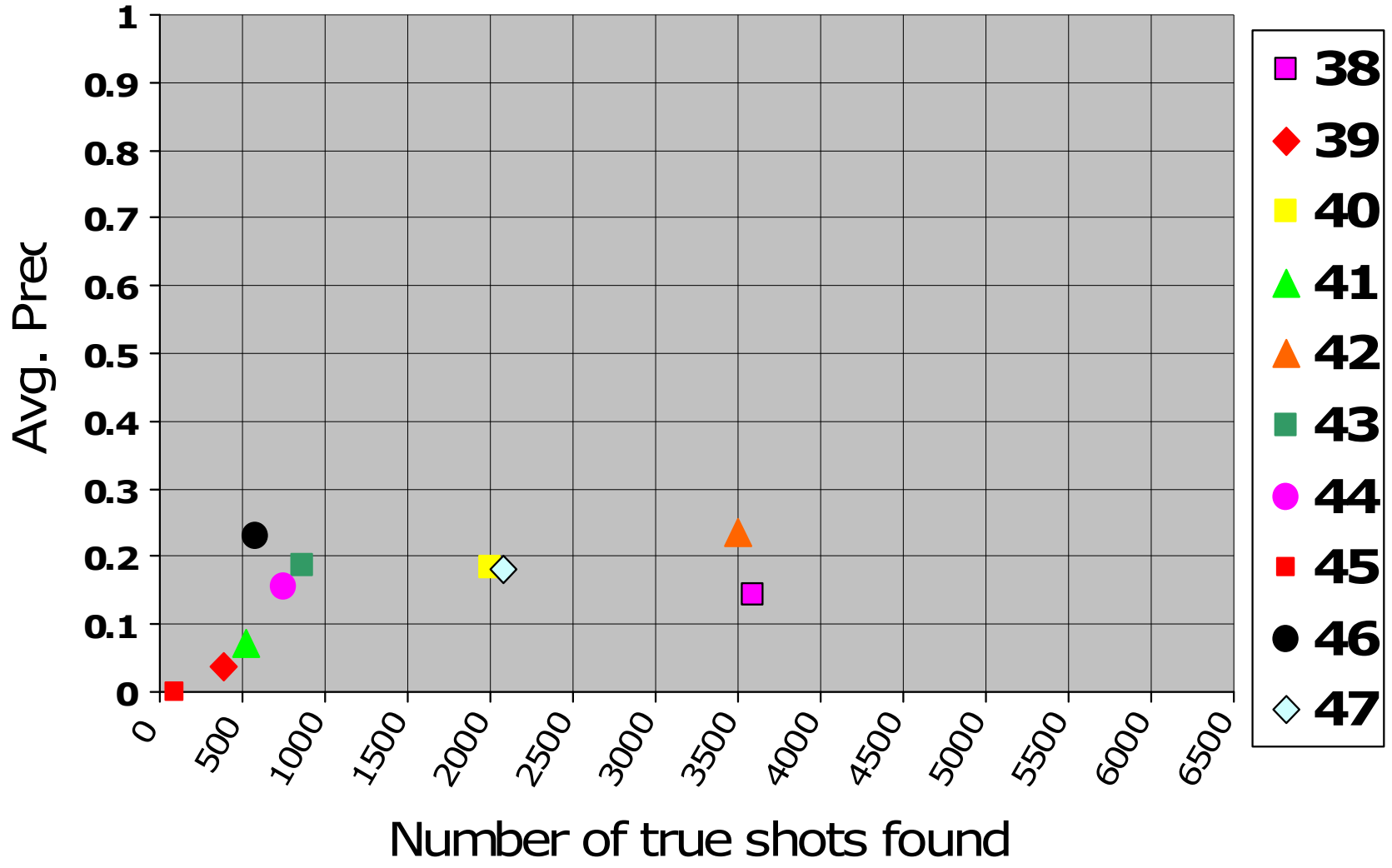
Median AvgP by number of annotated training examples



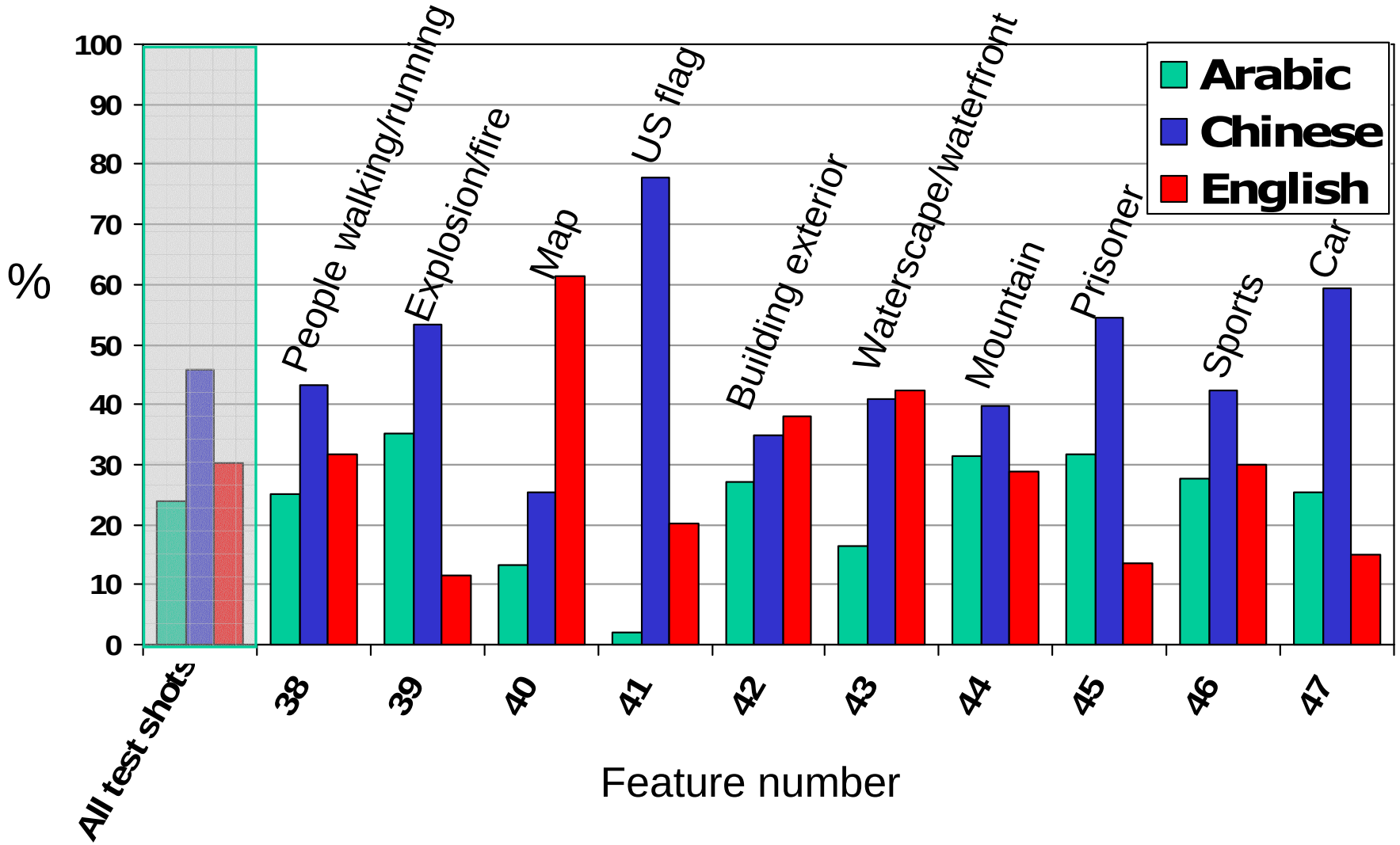
Max AvgP by number true shots found



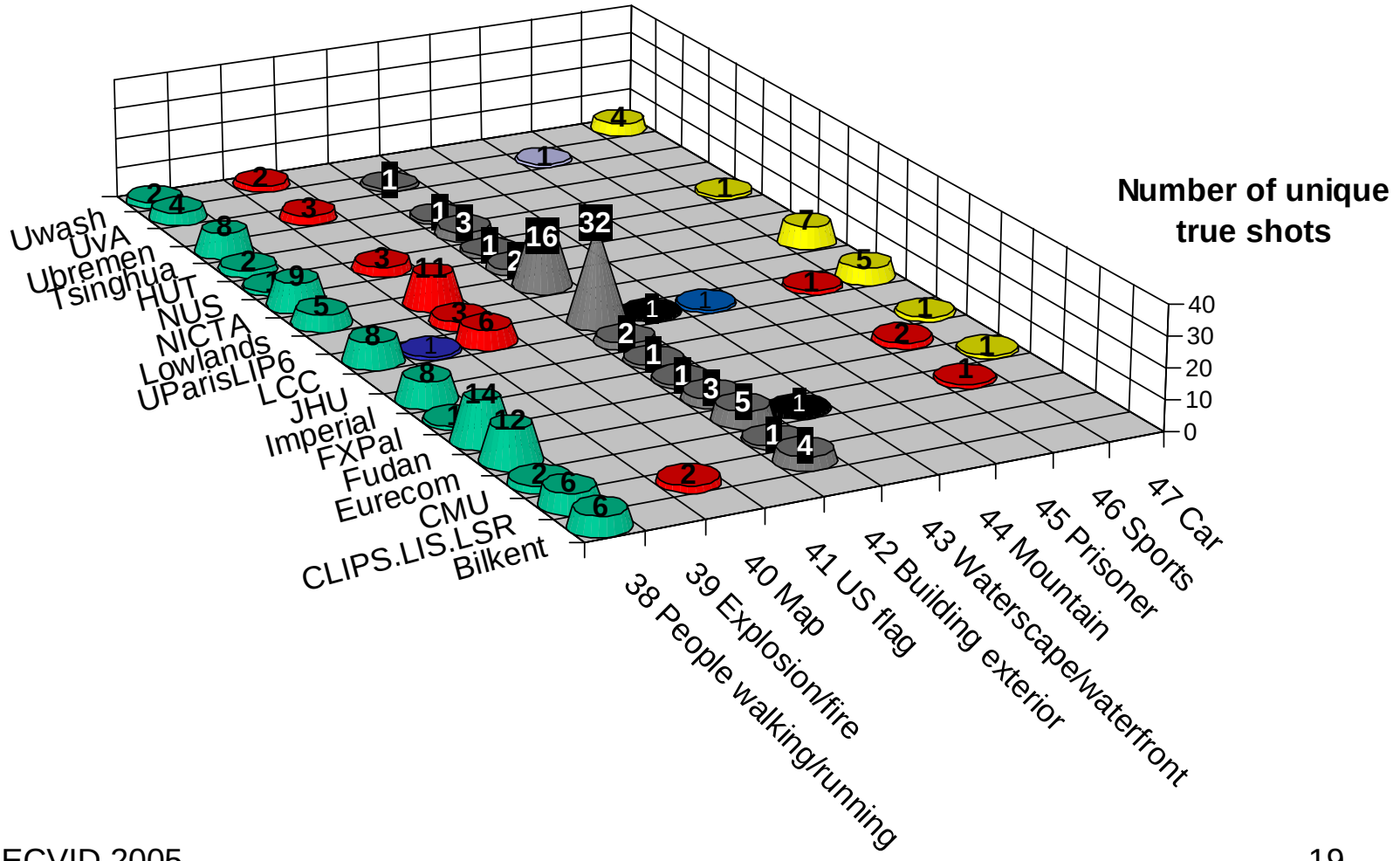
Median AvgP by number true shots found



% of true shots by source language for each feature



True shots contributed uniquely by team for each feature



Observations

- Participation almost doubled over 2004 (12 -> 22)
- Focus on category A runs (increased comparability)
- Scores are generally higher than in 2004 despite
  - new sources
  - errorful text from speech (via MT)
  - What does it mean?
- Did anybody run last year's system on this year's task?

- Features were generally found in all language sources
- Top scores come from fewer systems/groups

To follow: overview of the systems with map > 0.16 (median)

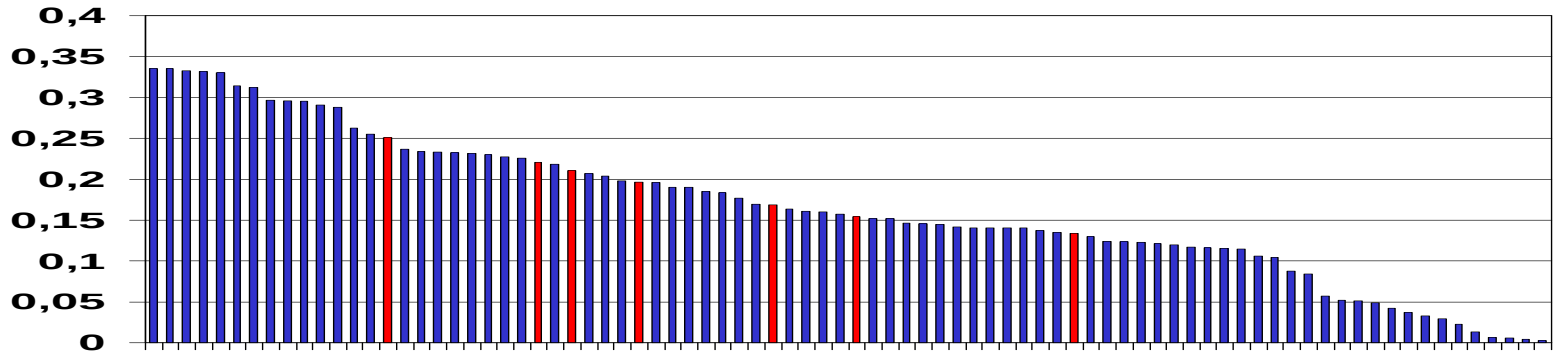
- Only systems that were tested on all 10 features
- Only category A
- Runs were compared on map across 10 features

Overview of approaches

- HLF systems draw from a very wide range of signal processing and machine learning techniques
  - Generic vs feature specific
  - How to do feature selection for visual modalities such as color and texture
  - Visual representation: grid or salient feature clusters
  - Various fusion methods, normalization methods
  - Range of classifiers

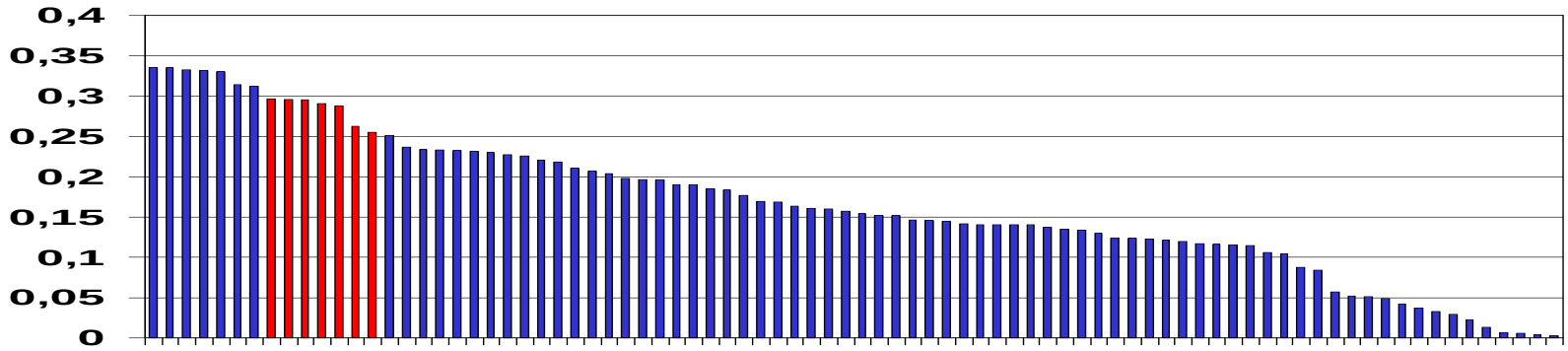
Carnegie Mellon University

- Approach
  - unimodal / multimodal (as in 2004)
  - learn dependencies between semantic features (by using various graphical model representations): inconclusive
  - global fusion < local fusion
  - multilingual > monolingual
  - multiple text sources > single text source
  - Best run: local fusion
- Results:



Columbia University
- presentation follows -

- Approach
  - Parts based object representation (ARG)
  - Captures:
    - topological structure (spatial relationships among parts)
    - Local attributes of parts
  - Model learns the parameter distribution properties due to differences in photometric conditions and geometry
  - Runs vary across classifier combination schemes (fusion/selection)
- Results:
  - Significantly better than global (i.e. grid based) approach
  - Esp. good for visual concepts where topology and local attributes are important (e.g. US flag)
  - Text features play only a marginal role (contrastive experiment)



Fudan University

- Approach:
  - Several runs
    - Specific feature detectors
    - ASR based
    - Fusion of several unimodal SVM classifiers
    - Contrastive experiments with different dimension reduction techniques (PCA, locality preserving projection
- Results:
  - Best run: 0.19

FXPAL

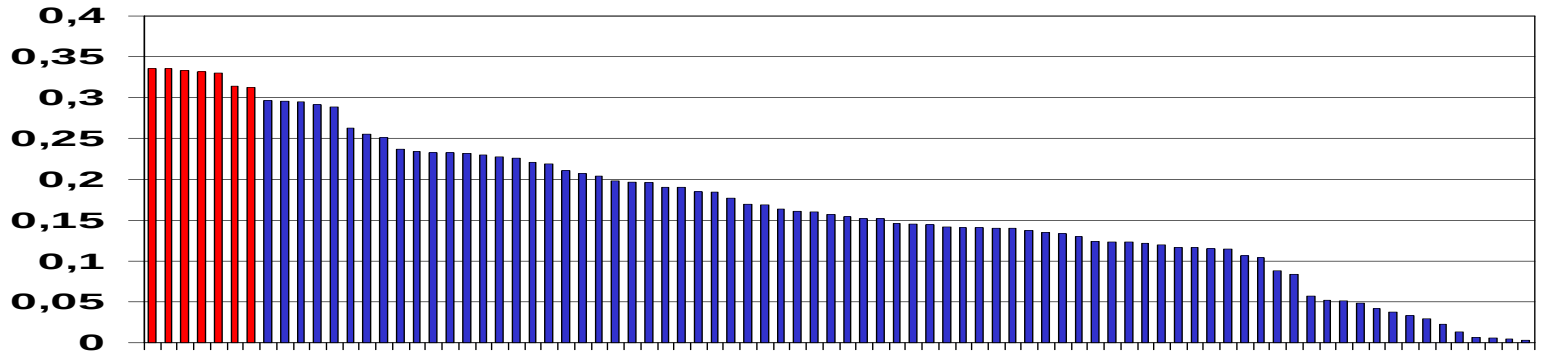
- Approach
  - SVM trained on low level features donated by CMU
  - Classifier combination schemes based on various forms of regression
  - 1st time participation
- Results
  - Best result: map=0.18

Helsinki University of Technology

- Approach:
  - Self Organizing maps trained on multimodal features and LSCOM lite annotations
- Result:
  - 1 run : map 0.2

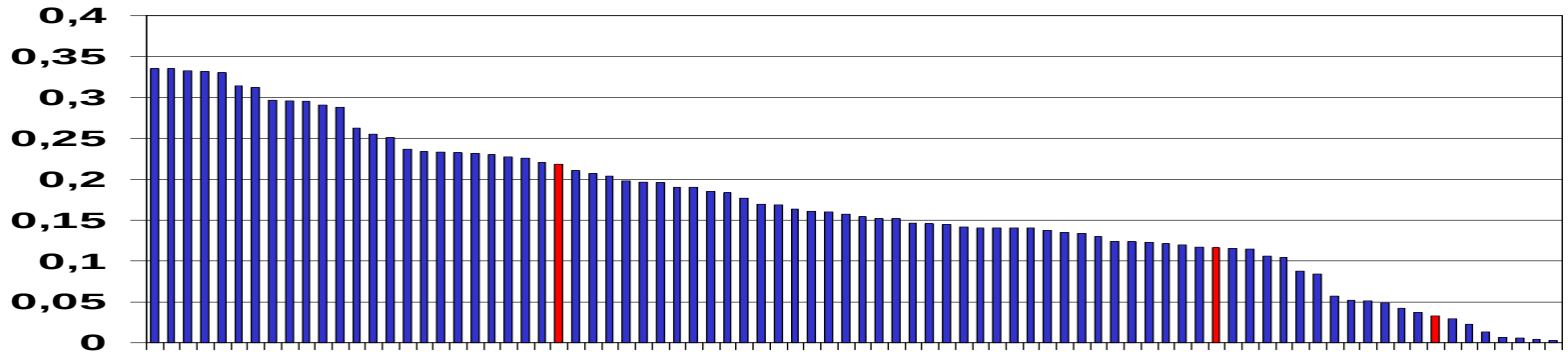
IBM

- Approach
  - Features:
    - Visual: Extensive experiments for selecting best feature type and granularity for individual modalities (color, texture etc.)
    - Motion, Text, LSCOM LITE concepts
    - Features also included meta-information such as time of broadcast, channel etc.
  - SVM > (ME, KNN, GMM)
  - Flat and hierarchical feature fusion
  - Variations in classifier fusion methods
  - Feature specific approaches (selection based on held-out data)
- Results:



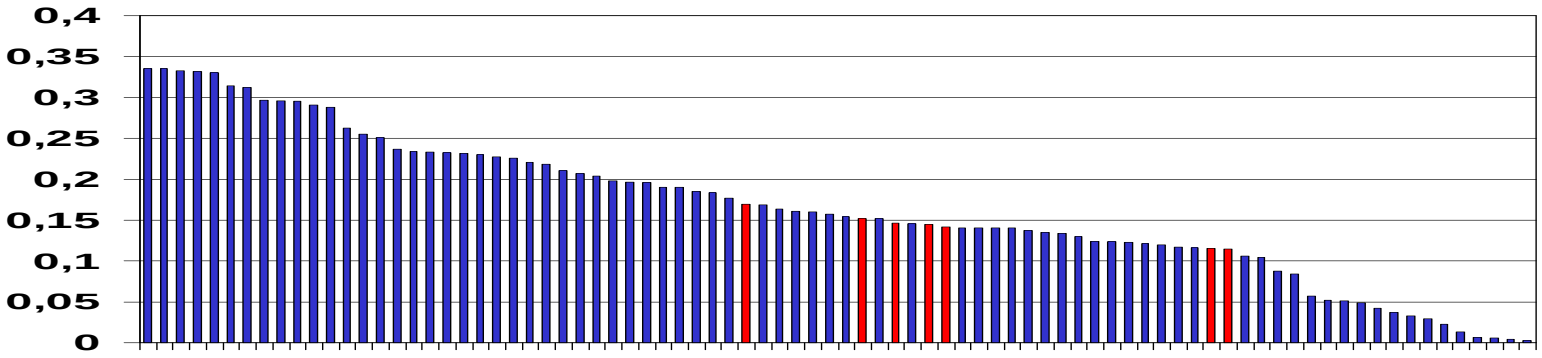
Imperial College London

- Approach
  - 1. “Naïve model”:
    - locate salient clusters in feature space
    - Learn HLF<-> clusters models
  - 2. Nonparametric Density estimation (kernel smoothing)
- Results:
  - Naïve model: performance problems
  - NPDE >> Naïve model



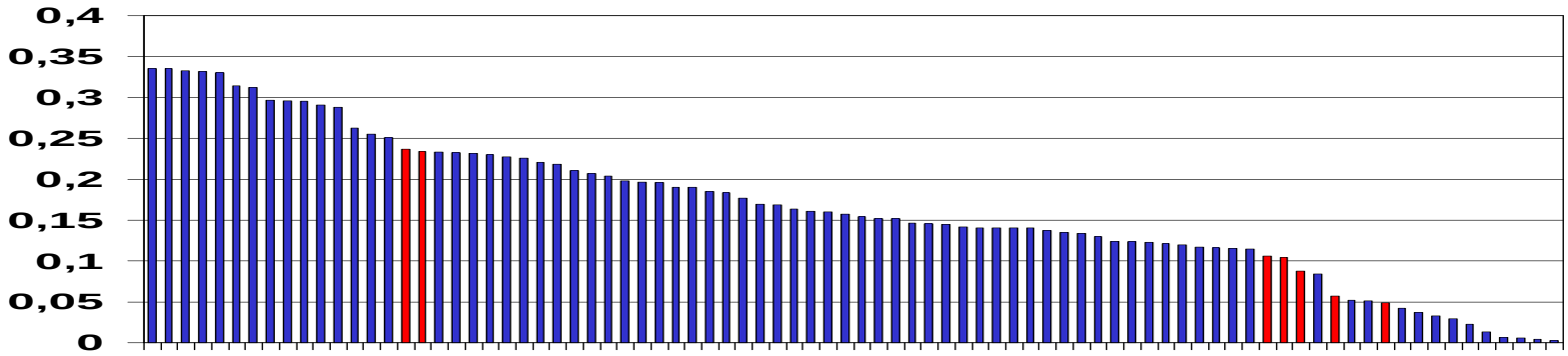
Mediamill team (Univ. of Amsterdam) - presentation follows -

- Approach
  - Authoring metaphor
  - Feature specific combination of content, style and context analysis
  - 101 concept lexicon
- Results:
  - Textual features contribute only a small performance gain



National University of Singapore (NUS)

- Approach
  1. Ranked maximal figure of merit: ASR only, texture only, 2 fused runs
  2. HMM for visual dependency (4X4 grid): ASR only, +visual, +audio,genre,OCR . RankBoost fusion
- Results:
  - 2nd approach >> 1st approach



University of Washington

- Approach: ?
  - (notebook paper not available yet)

- Results: