TRECVID-2004 Feature extraction task: Overview

Wessel Kraaij
TNO TPD
&
Joaquim Arlandis
NIST

FE Task definition

- Goal: Build benchmark collection for detection methods of high-level features
- Secondary goal: feature-indexing could help search and navigation
- Feature set was mostly taken from the common annotation effort of TV2003
- Examples of thing/activity/person/location

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 (new strategy)
- Evaluate performance quality by measuring the *average precision* of each feature detection method
- Resource use:
  - Type A: sites work with just the common development data, common annotations and tv2003 truth data for tv2003 topics
  - Type B: sites work with just the common development data and any annotation set
  - Type C: other

10 Features

28. Boats / Ships – any type
29. Madeleine Albright
30. Bill Clinton
31. Trains or railroad cars
32. Beach – water and shore visible
33. Basket score – ball passing down the hoop and the net
34. Airplane taking off
35. People walking or running
36. Physical violence between people and/or objects
37. Road – Any size, paved or not

Who worked on which features

12 participant groups (tv2003:10)	Boats/ships 28	M.Albright 29	B.Clinton 30	Train 31	Beach 32	Basket scr. 33	Airplane takeoff 34	People walk-run 35	Phys.violence 36	Road 37
CLIPS-LSR-LIS (CLIPS+)	X	X	X	X	X	X	X	X	X	X
Carnegie Mellon University (CMU)	X	X	X	X	X	X	X	X	X	X
Eurecom (EURECOM)	X	X	X	X	X	X	X	X	X	X
IBM Research (IBM)	X	X	X	X	X	X	X	X	X	X
IDIAP			X				X			
Imperial College London						X				
National University of Singapore (NUS)	X	X	X	X	X	X	X	X	X	X
University of Iowa (UIOWA)	X		X		X	X				
Tsinghua National Lab for I & T (TSINGHUA)			X		X	X				
Mediamill/University of Amsterdam (UvA)	X	X	X	X	X	X	X	X	X	X
National Institute of Informatics (NII)		X	X							
University of Central Florida (UCF)		X	X		X	X		X		
TOTAL groups by feature	7	8	11	6	9	10	7	7	6	6

Number of runs each training type

Tr-Type	2004	2003
A	45 (54.2%)	22 (36.7%)
B	27 (32.5%)	20 (33.3%)
C	11 (13.3%)	18 (30.0%)
Total runs	83	60

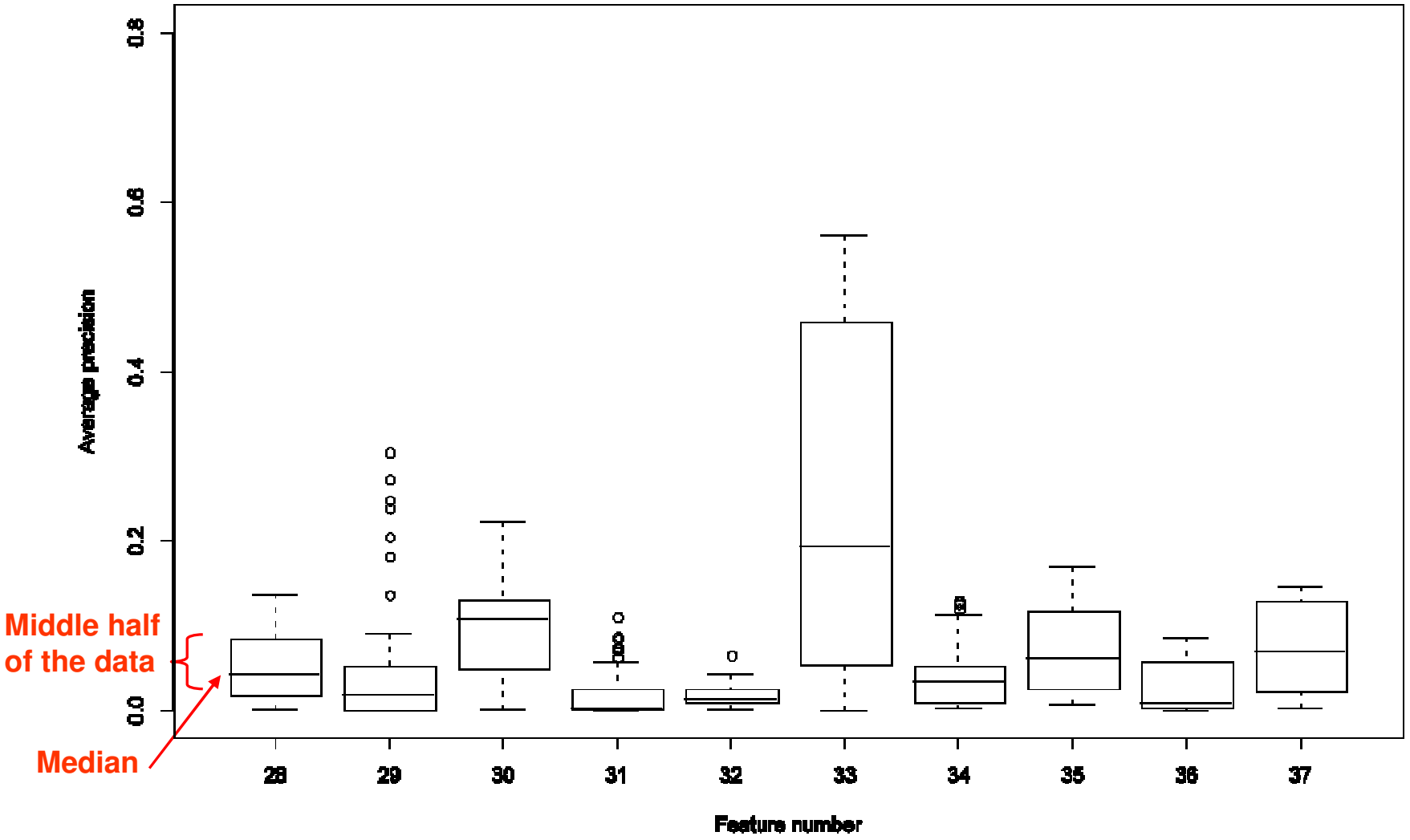
System training type:

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

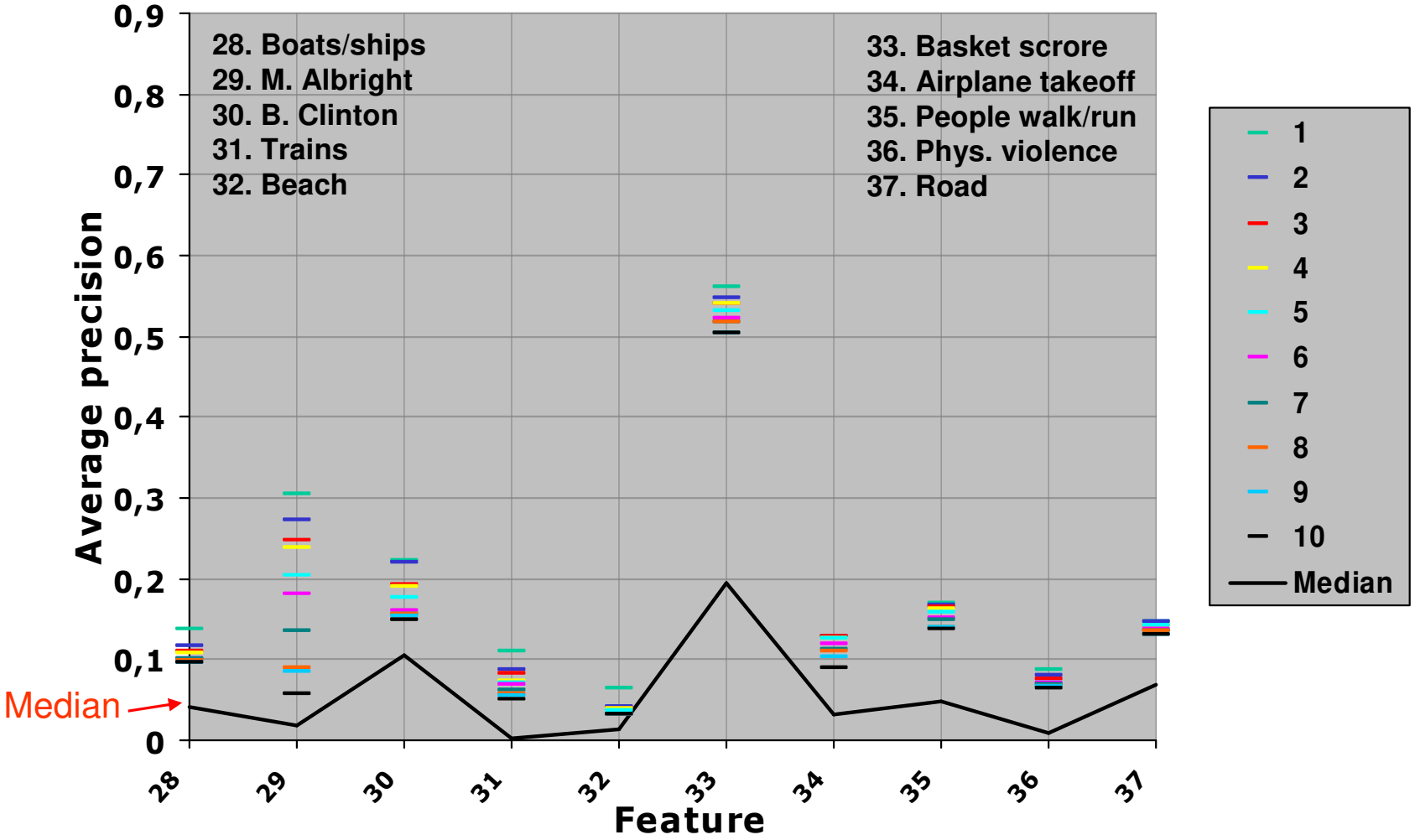
B - Only on common dev. collection but not on (just) the common annotation

C - not of type A or B

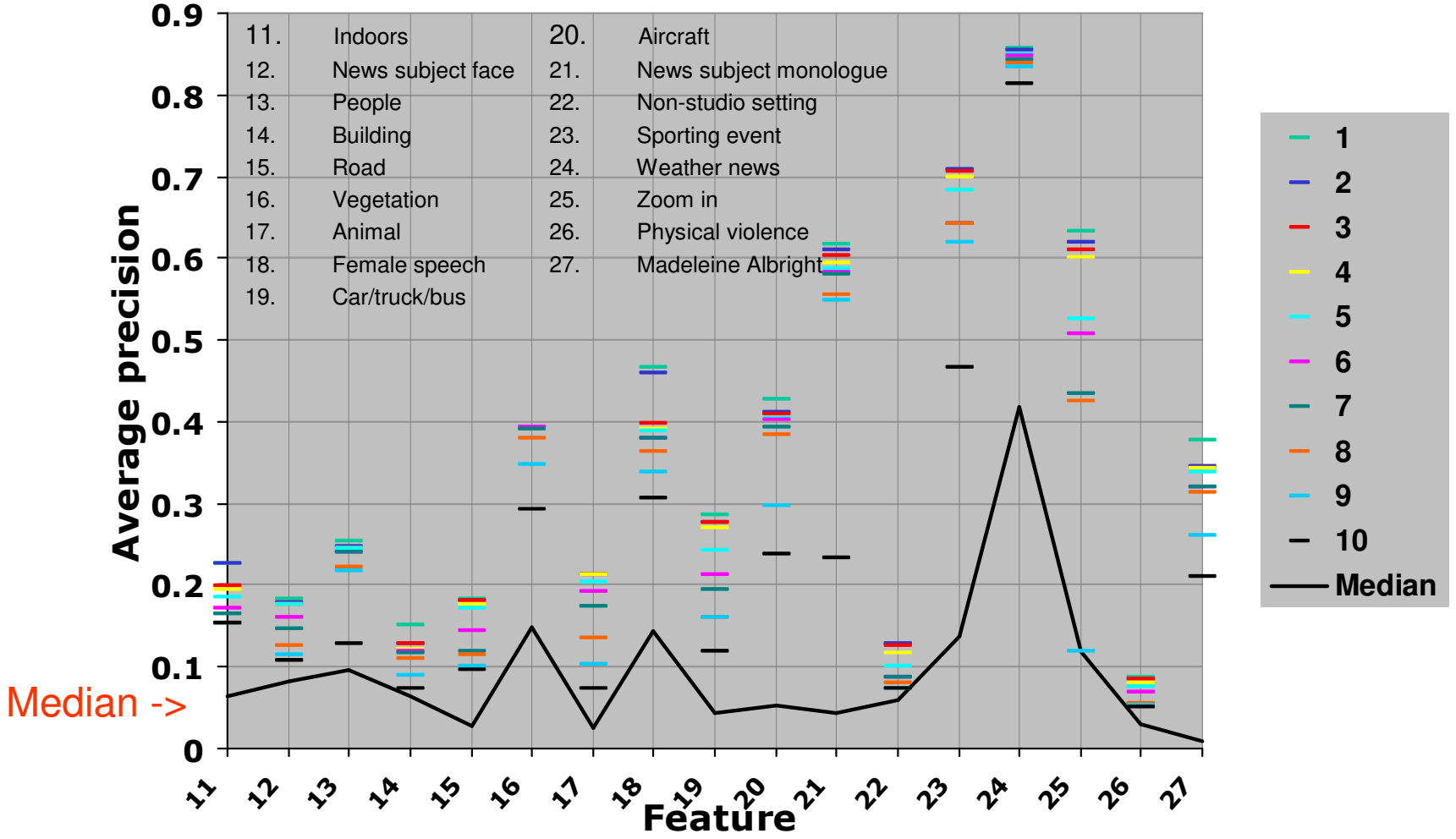
2004: AvgP by feature (all runs)



2004: AvgP by feature (top 10 runs)

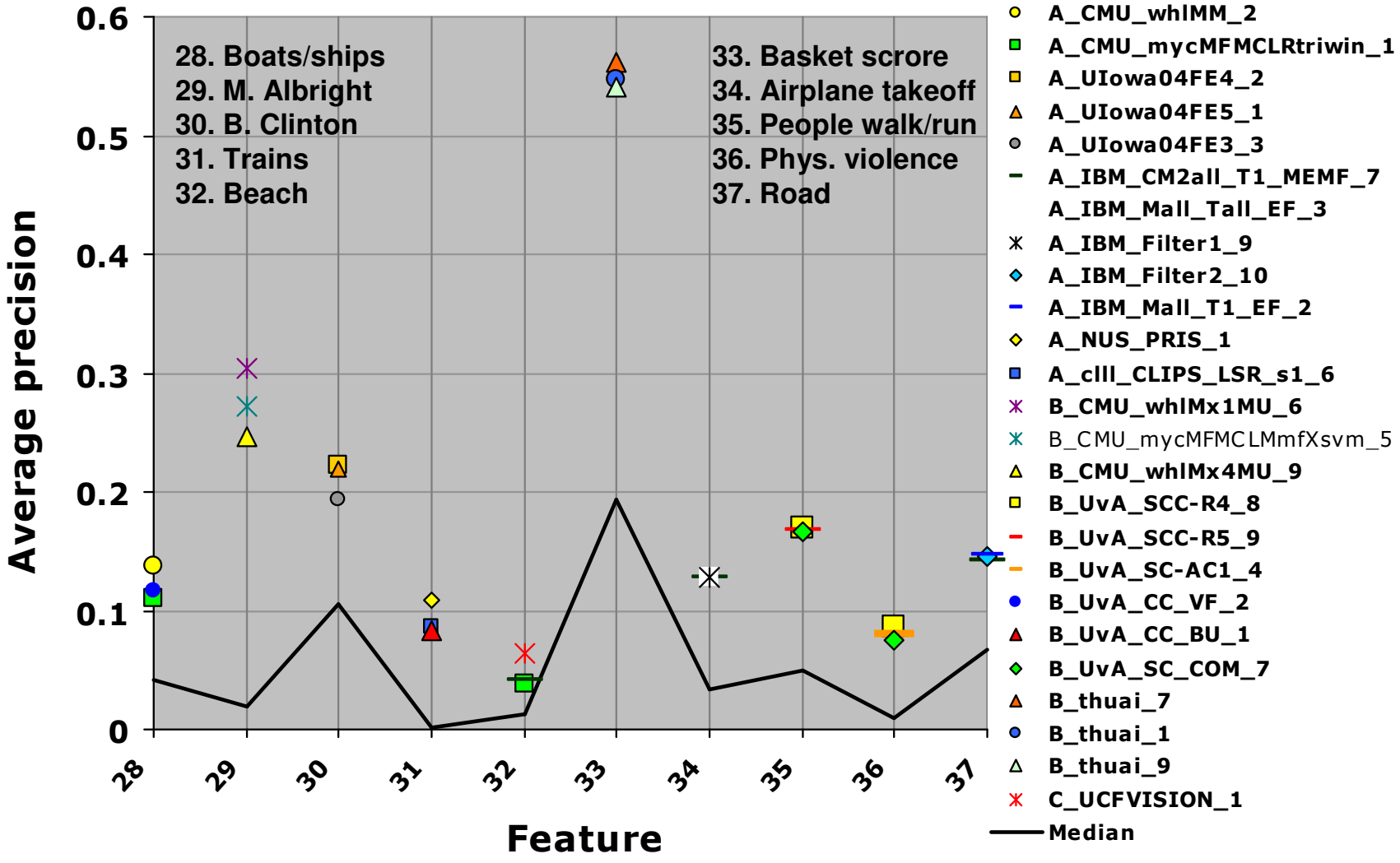


2003: AvgP by feature (top 10 runs)

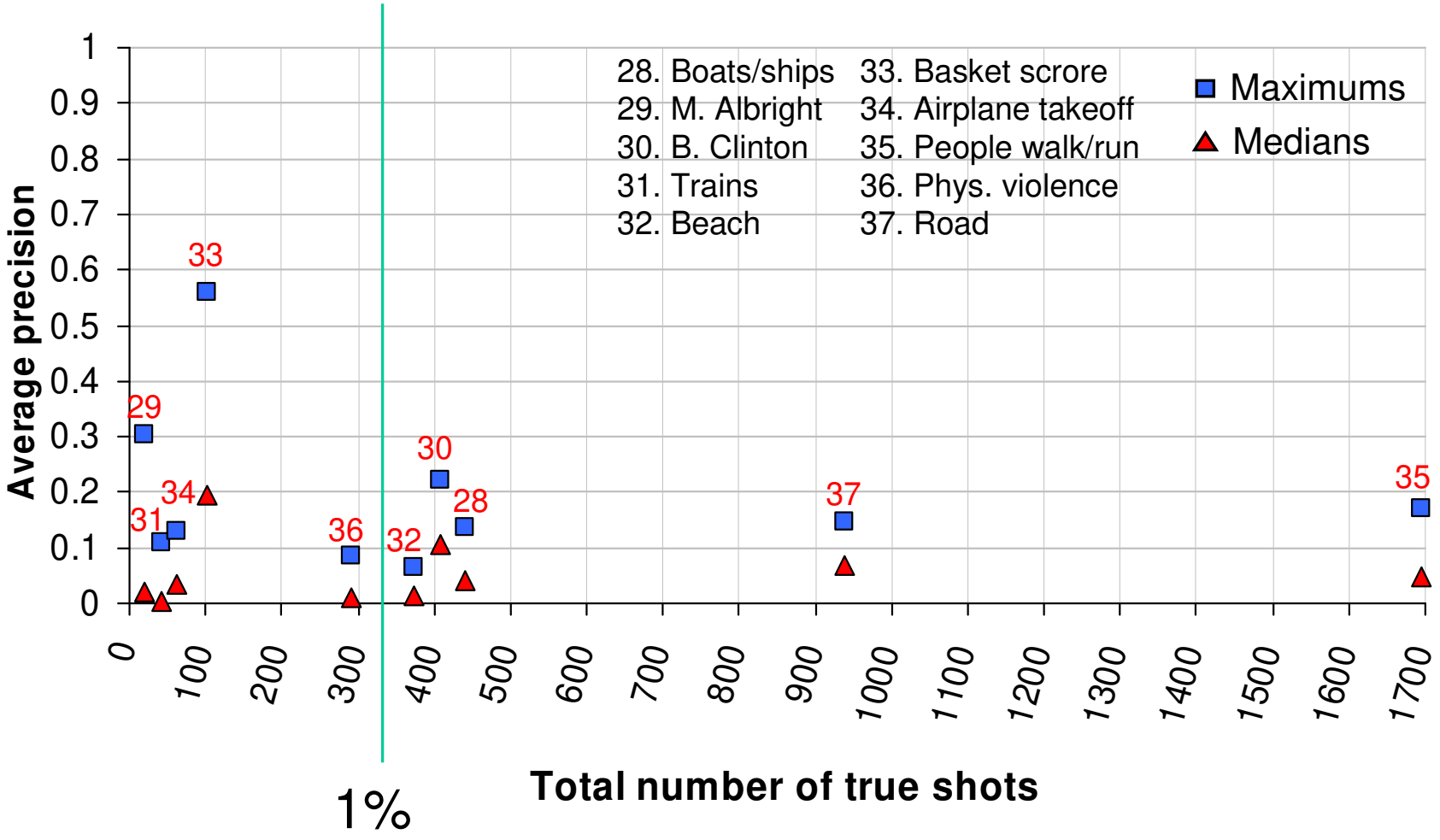


Median ->

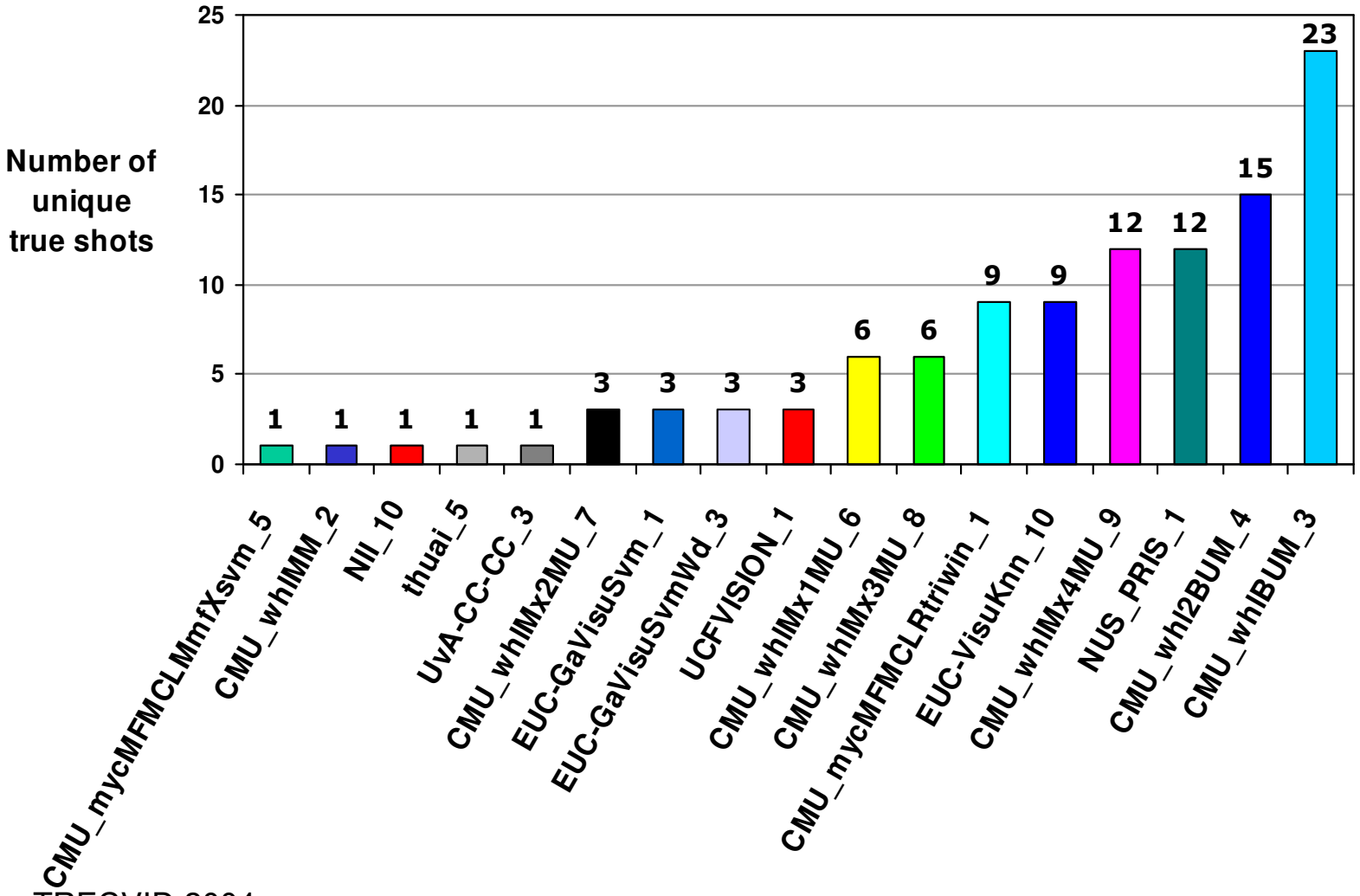
2004: AvgP by feature (top 3 runs by per feature)



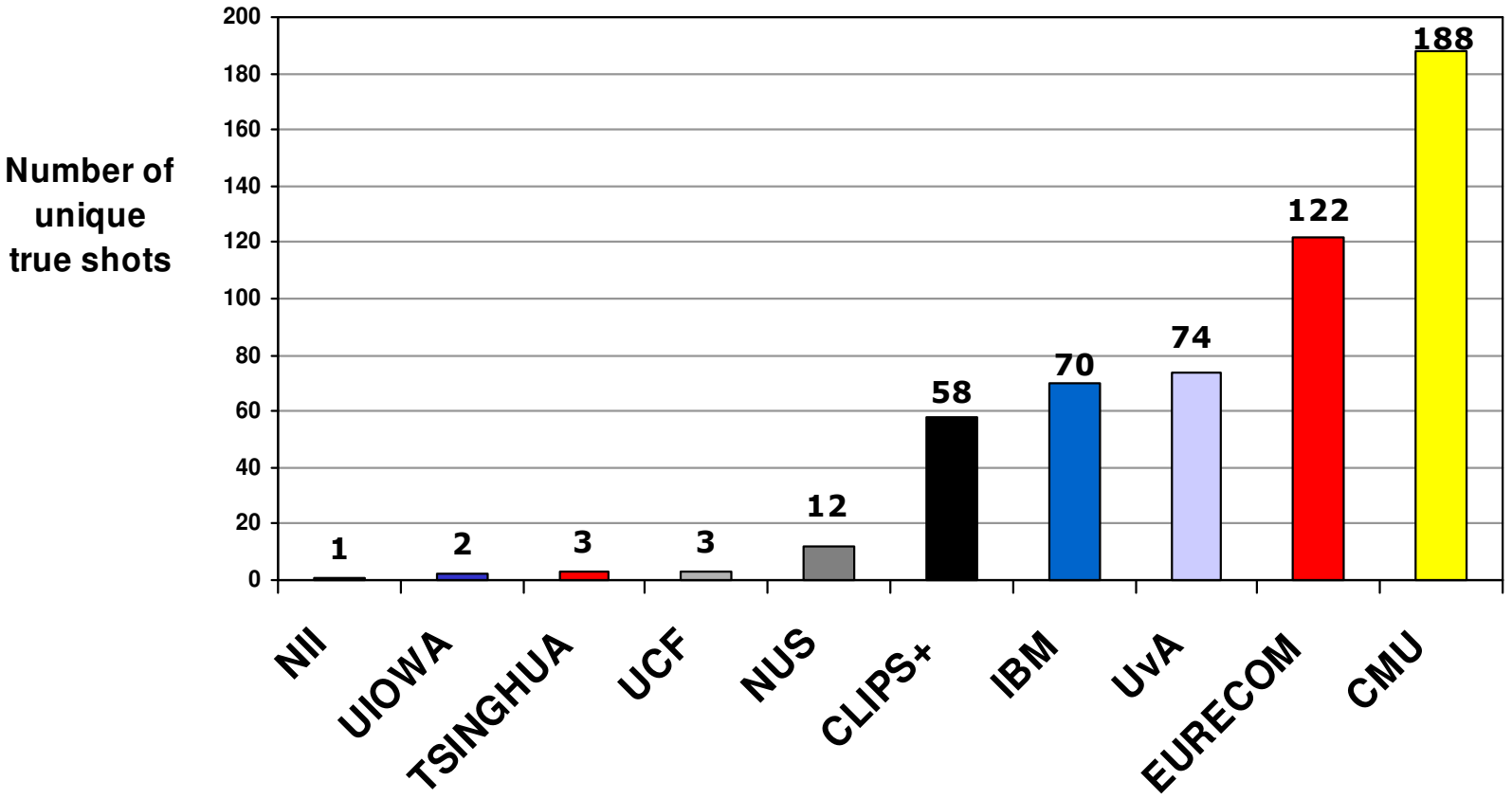
2004: Avg. precision *vs* total number true for each feature



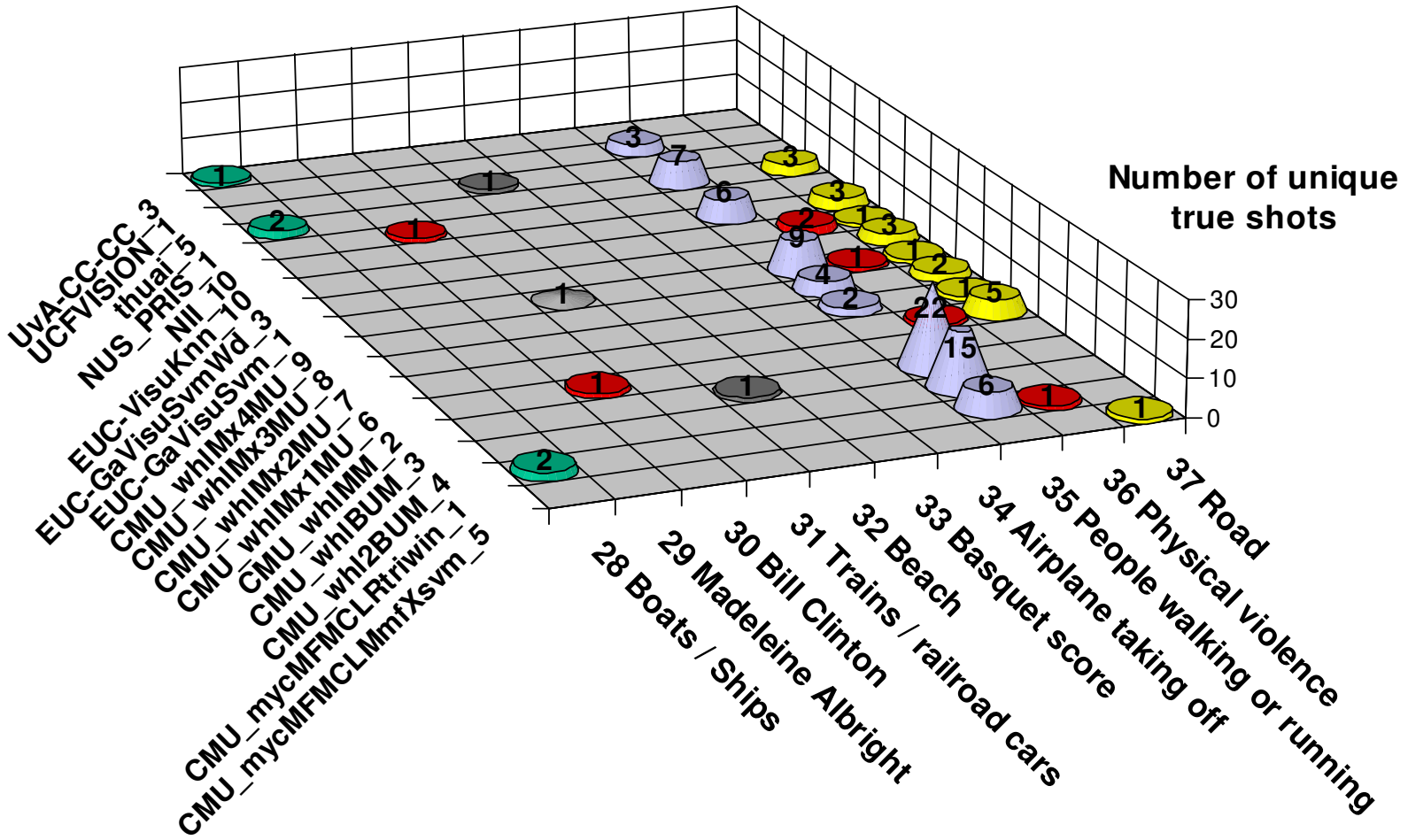
2004: 17 of 83 runs contributed one or more unique, true shots



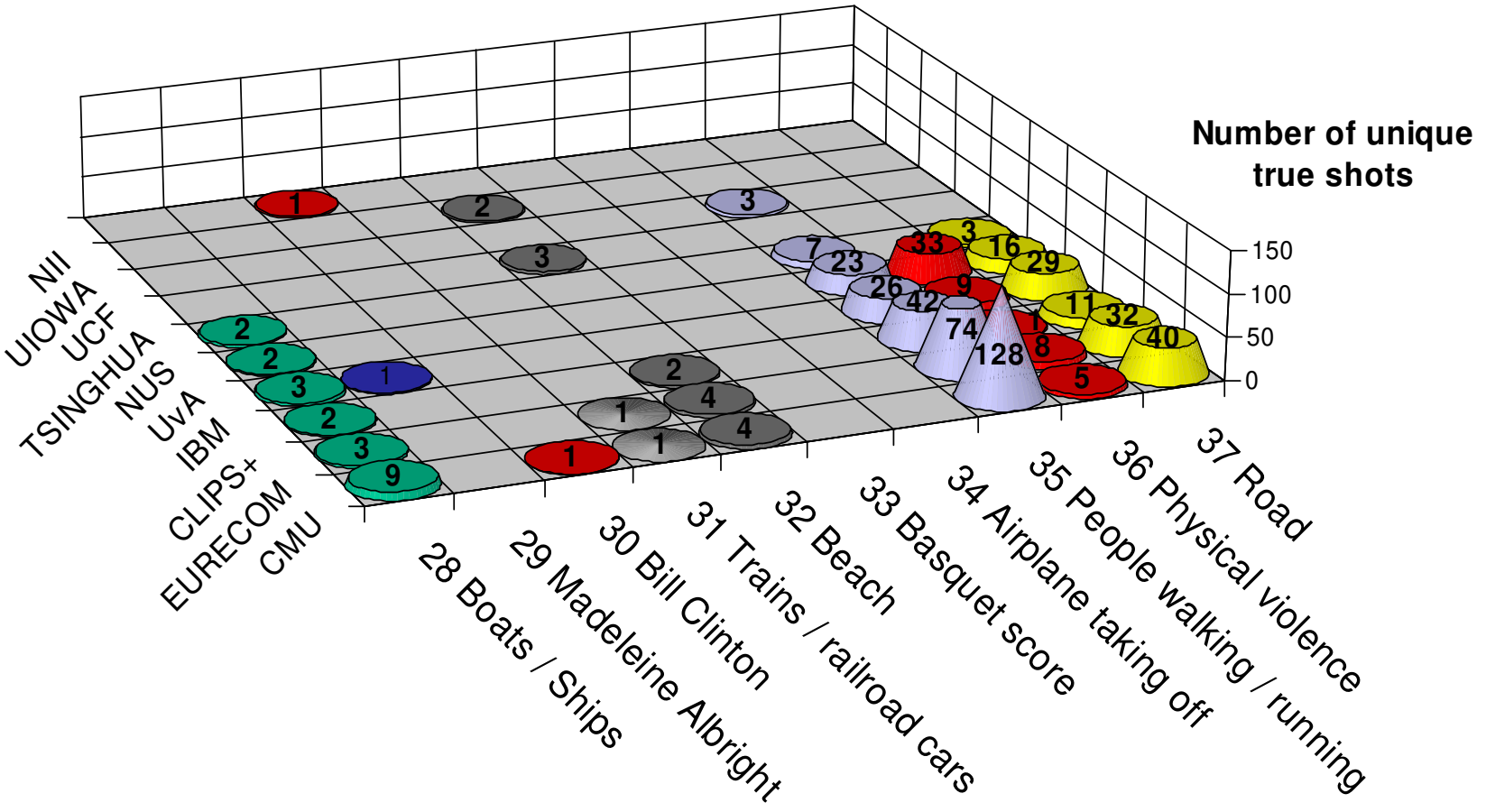
2004: 10 of 12 teams contributed one or more unique, true shots



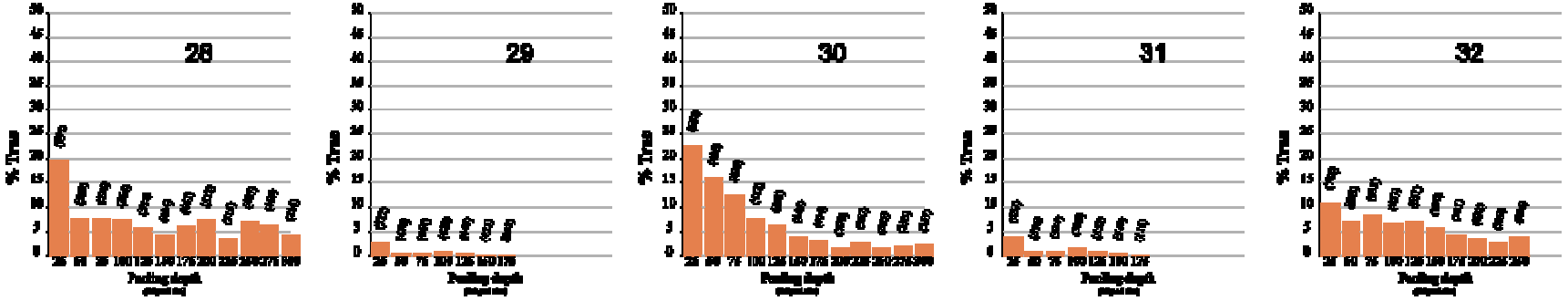
2004: True shots contributed uniquely by run for a feature



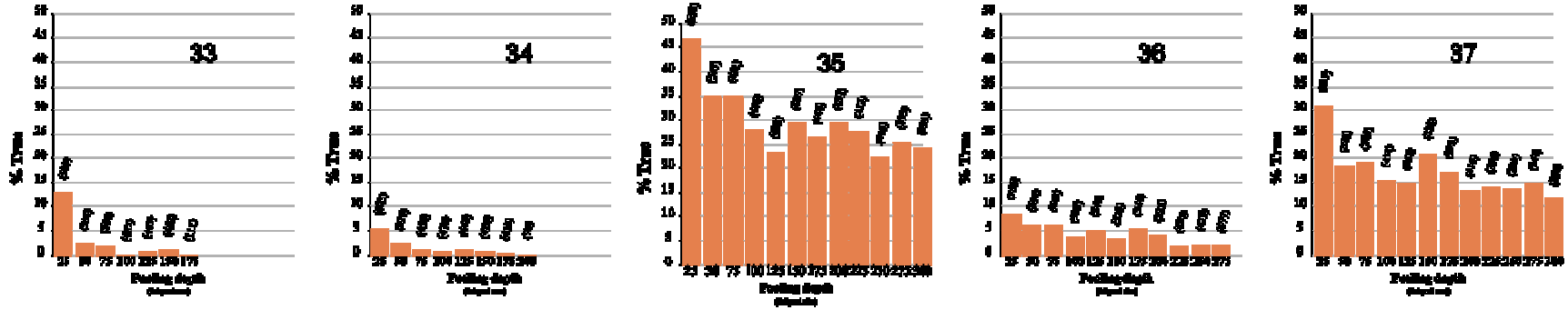
2004: True shots contributed uniquely by team for a feature



2004: Pooling, true shots judged by feature



Percentage of true shots found at deeper ranges (subpools 1-25, 26-50, …)



CLIPS-LSR-LIS: headlines

- *Approach*
  - ASR, AV and combination
  - Fusion by linear combination
- *Features*
  - Lexical context: determine P(text|F)
  - Color histogram (fuzzy quantization)
  - Texture (gabor, PCA)
  - Face detection/recognition
- *Results*
  - ASR run was best CLIPS run
  - Close to maximum type A performance for “Trains”

Carnegie Mellon University: headlines

- *Approach*: No paper available
- *Results*
  - Type A: best results for “boats/ships”,”basket score”
  - Type B: best results for “M.A”,”beach”
- *Presentation follows*

Eurecom: headlines

- *Approach*
  - Three fusion techniques: SVM, KNN, GA
- *Features*
  - Visual: colorhistogram,LSI
  - Text: 2000 word vocabulary frequency vector
  - Motion: camera motion & motion histogram extracted from MPEG motion vectors
- *Results*
  - Type A: best result for “M.A.”

IBM Research: headlines

- *Approach:* Cross feature Ensemble Learning
  - Feature extraction
  - SVM and MaxEnt classification
  - Ensemble fusion or MaxEnt fusion
  - Remove anchors
- *Features*
  - Visual: correlogram, histograms, edges, color moments, wavelet texture, co-occurrence texture
  - Text
- *Results*
  - Type A: best results for “beach”, “airplane takeoff”,”physical violence”,”road”
- *Presentation follows*

IDIAP: headlines

- *Approach*: No paper available
- *Results*
  - 2 type A features:
    - Bill Clinton: well below median
    - Airplane takeoff: slightly above median

Imperial College London: headlines

- *Approach*
  - Nearest neighbour
- *Features:*
  - Global HSV
  - Thumbnail
  - Convolution
  - Word “basketball”
  - Linear combination of inverse normalized distances
- *Results*
  - 1 Type A feature run: “basket score” around median

MediaMill: headlines

- *Approach*
  - Generic approach for all features: “Semantic Value Chain”
  - Lexicon of 32 semantic concepts trained on 2003 data, common annotation effort
  - Video analysis= inverted authoring
  - SVM & Parallel architecture
- *Features*
  - Words associated with concept
  - Visual: colorspace corrected for intensity and object size variation
  - Layout: shot length, overlayed text, silence, voice over,
  - Content: faces, face location, cars, object motion, frequent speaker, overlayed text length, named entity
  - Capture: camera distance, camera work, camera motion
- *Results*
  - Type B: best results for all features except “M.A.”, “Beach”,”Basket score”

National University of Singapore: headlines

- *Approach*
  - Visual auto-concept annotation:
  - Fusion of text and specialized detectors
  - 4X4 blocks, block clustering, SVM training
  - Generic except use of face detector for person X
- *Features*
  - Colorhistogram, texture, edge histogram
  - Select informative keywords, using IDF and MI using Google
- *Results*
  - Type A: best result for “trains”

National Institute of Informatics: headlines

- *Approach*
  - Face detection (Viola/Jones2001)
  - Face alignment
  - Face recognition: positive training example: MA/BC, negative training examples from 1210 individuals in FERET database for training SVM classifier, Gaussian radial basis function kernel
- *Results*
  - Two type C features (MA/BC): B.C. above median

Tsinghua National Lab for I&T: headlines

- *Approach*
  - 1st step: SVM classifier for several features
  - 2nd step: Voting
  - 3rd step: interpolation with text based score
  - Filter known false positives
- *Features*
  - Visual and text features
  - Exploit clusters of positive shots
  - Position of shot in show
- *Results*
  - Three type B features: best results for “basket score”

University of Central Florida: headlines

- *Approach*
  - Specific detectors for 5 features
  - Fusion using SVM or Adaboost
- *Features*
  - Text (for B.C. , M.A. and basketball)
  - Face
  - Colorhistogram
  - Correlogram
  - Skin
  - Motion (people walking)
  - Sports shot
- *Results*
  - Best result for beach (type C), other features above median

University of Iowa: headlines

- *Approach*
  - Combination of general low level features
  - Fusion by sum or **product**
- *Features*
  - Image: colorhist., edgehist, feature specific cropping
  - Text: manual selected trigger words
  - anchor
- *Results*
  - 4 Type A features, best result for “Bill Clinton” (text feature most influential)

Observations

- TV2004 features are more difficult than tv2003
  - Hard features are sometimes easy (basketbal hoop)
- Most features were rare enough for reliable pooling
- Max performance for MA and Road decreased, did not change for Physical Violence
- Many groups chose type A, increasing comparability

- Less specialized detector approaches
- Strong performance of feature neutral systems using a wide range of features
- FE is hard, even if substantial training data is available
- What about conditioning on AV, ASR and combination?