Rich representations for rich semantics: tsinghua@hfe.tv06

Dong Wang, Xiaobing Liu, Linjie Luo, Xiao Zhang, Zhen Xiang, Wanli Peng
Jianmin Li, Fuzong Lin, Bo Zhang

Outline

- Rich representations for rich semantics
- System design and implementation
- Benchmark results
- Future directions

Background

- Video indexing and retrieval is still in its childhood
  - lack of concrete basic indexing unit in video
- Current research trend in TRECVID shows strong favor of the generic visual indexing
  - a cornerstone for video retrieval
- However, generic visual indexing for multimedia archives is far from satisfying
  - low accuracy
  - lack of robustness
  - Non-scalability

Past experience

- no best single feature fits for all concepts
- no best single feature fits for each concept
- Why?
  - Fast changing style and rich semantics
  - ~1000 concepts is challenging

Neuroscience facts

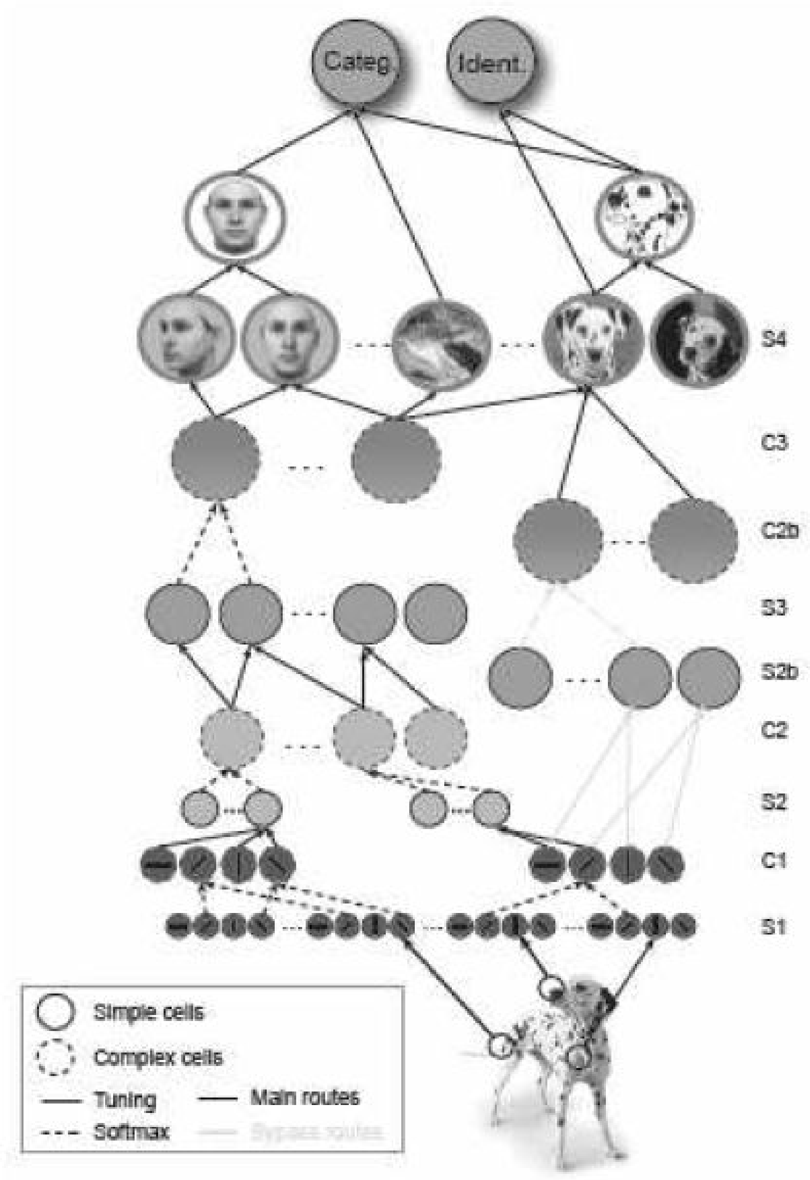
- The human vision system is a real miracle in its generalization ability, computational efficiency and elegancy for recognizing innumerable objects with ease
- human can do *ultra-rapid visual categorization*
  - detect object in complex scenes < 150 ms
  - For both natural and artificial categories
  - Without color information
  - No faster even with training
  - Robust to location/rotation/scale/viewpoint
  - No attention mechanism involved
  - Monkey can also perform this quicker but slightly less accurate

a few accepted properties of the ventral stream for vision [M. Riesenhuber and T. Poggio]

1. A hierarchical build-up of invariances first to position and scale and then to viewpoint and more complex transformations requiring the interpolation between several different object views;
2. in parallel, an increasing size of the receptive fields;
3. an increasing complexity of the optimal stimuli for the neurons;
4. a basic feedforward processing of information (for "immediate" recognition tasks);
5. plasticity and learning probably at all stages and certainly at the level of IT(infero-temporal cortex);
6. learning specific to an individual object is not required for scale and position invariance (over a restricted range).

"Standard Model" of biological object recognition

- Massively parallel model
  - – Only a few levels (5?)
  - – Lowest level has simple feature
- templates with no invariance
  - – Increasing levels add position &
- scale invariance as well as
- increasing feature complexity
- Alternating layers:
  - – S (Simple) layers: template matching
  - – C (Complex) layers: perform MAX over some locations and scales



After [M. Riesenhuber and T. Poggio]

Attempts in computer vision to simulate these principles

- several attempts [Serra05, Mutch06]
- key new aspect
  - a task-specific rich representation for each category of ~6000 *kinds of* features
  - a very simple 'HMAX' fusion model
- Encouraging results
  - 30 examples / category for 101 category to achieve 56% acc. [Mutch06]

rich representation for rich semantics

- Video retrieval demands powerful tool for indexing the rich semantics exhibited in the video content
- Generic and robust approach is required to index the content through detecting a large number of concepts (~1000)
- No best single feature fits for all concepts, and no best single feature fits for every concept either.

- A large number of neurons with increasing complex of the optimal stimuli is found in the human visual system
- Rich representation plus a simple fusion algorithm accounts for recent success in computer vision systems simulating the human visual system
- Do concept-specific complex features are required for robustly detecting the target concept?

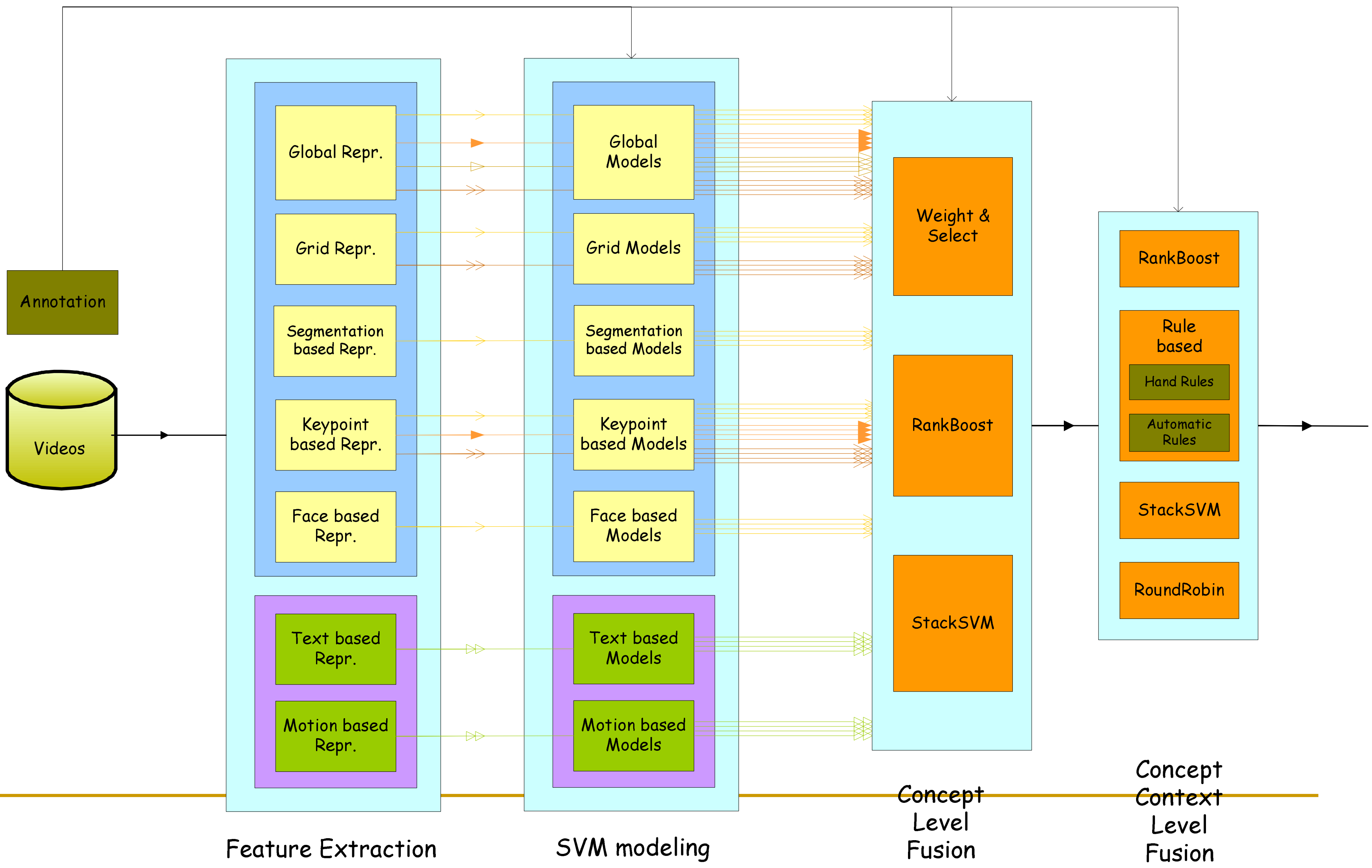
rich representation for rich semantics

- enable our concept detection system by adding richer representations
- hierarchical visual representations (with text and motion based representations altogether)
- many (not so many yet) feature extractors
- a bundle of diversified classifiers for each feature
- simple weight and select fusion is possible at such condition

Outline

- Rich representations for rich semantics
- System design and implementation
- Benchmark results
- Future directions

Concept detection system



Rich Representations

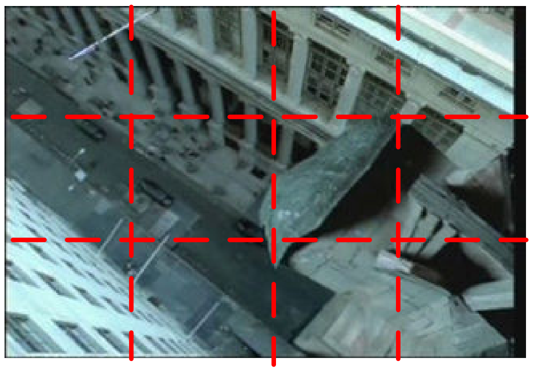
- **The Global Representation**
- **The Grid Representation**
- **The Segmentation based Representation**
- **The Keypoint based Representation**
- **The Face based representation**
- **The Text based Representation**
- **The Motion based Representation**

Rich features

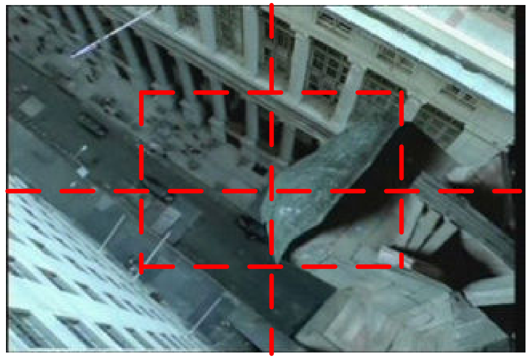
- **The Global Representation**
  - Color Auto-Correlograms (64 dim & 166 dim respectively)
  - Co-occurrence Texture (48 dim)
  - Color Coherence Vector (72 dim)
  - Color Histogram (HSV space, 36 dim)
  - Color Moment (LUV space, 9 dim)
  - Edge Histogram (72 dim)
  - Wavelet Texture (20 dim)

Rich features

- **The Grid Representation**
  - Color Moment (9 dim) * (4x3 grid)
  - Haar Wavelet Moment (10 dim) * (4x3 grid)



- The Edge Histogram (64 dim) * (4 corner + 1 center)

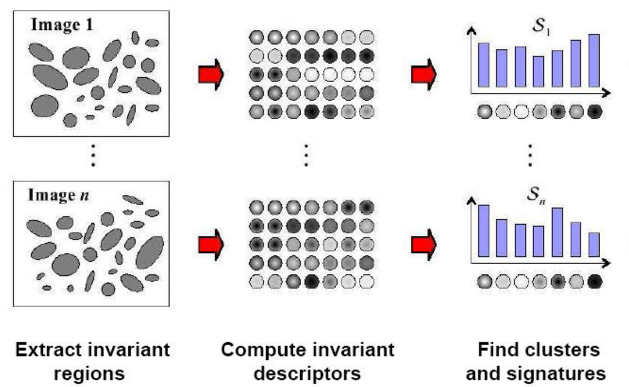


Rich features

- **The Segmentation based Representation**
  - standard JSEG segmentation
  - Keyframes are segmented to at most ten regions.
  - Only Color Moment (LUV space, 9 dim) since region is usually homogenous

Rich features

- **The Keypoint based Representation**
  - SIFT [Lowe04] and SURF [Bay06]
  - Codebook form K-mean in concept-(in)dependent style
  - Result in six kinds of histogram features
    - Codebook_500 ==> histogram_500
    - Codebook_20 ==> histogram_20 * 4x3grid layout
    - Codebook_50 ==> histogram_50 * 3x2grid layout
    - Concept_codebook ==> histogram_100
    - 39*concept_codebook_10 ==> histogram_390
    - 39*concept_codebook_10 + codebook_200 ==> histogram_590



Rich features

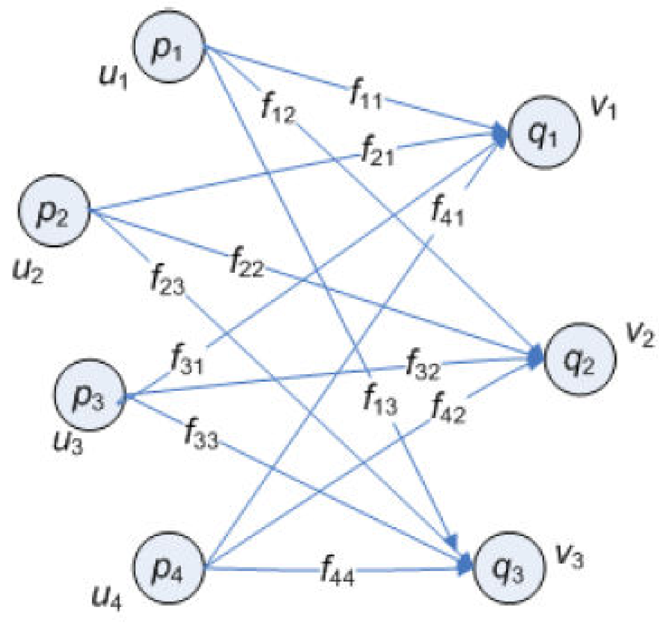
- **The Face based representation**
  - Produced with a state-of-the-art multi-view face detector [Huang05]
  - a human-oriented segmentation (human body and background)
  - to capture the invariance of different kinds of roles, e.g. government leaders or military

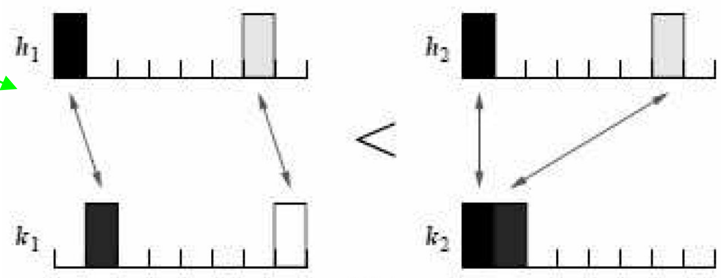
Rich features

- **The Text based Representation**
  - Simple TF-IDF features are extracted
- **The Motion based Representation**
  - Our Low Level Feature extraction algorithm [Thu_notebook05] and motion activity from [Peker01]

Modeling

- SVM classifier is appreciated, follow this respectable tradition
- Different kernels for different features
  - RBF/EMD/χ^2
- RankBoost to produce a bundle of diversified classifiers
- 110 dim model vector for each keyframe for each concept
  - from the 22 features used in 7 representations with 5 model score for each feature





Concept Level Fusion

- simple Weight and Select
- RankBoost again
- StackSVM

Concept Context Level Fusion

- RankBoost
- StackSVM
- Rule based
  - Automatic generated Rule
  - Hand generated Rules
- Roundrobin

Computational issues

- Some partial estimation of the running time of training sums up to 600 days for one computer!
- Fortunately the parallel computing paradigm, which is a natural choice for the uncorrelated concept detection task, ends in less than 10 days.

Outline

- Rich representations for rich semantics
- System design and implementation
- Benchmark results
- Future directions

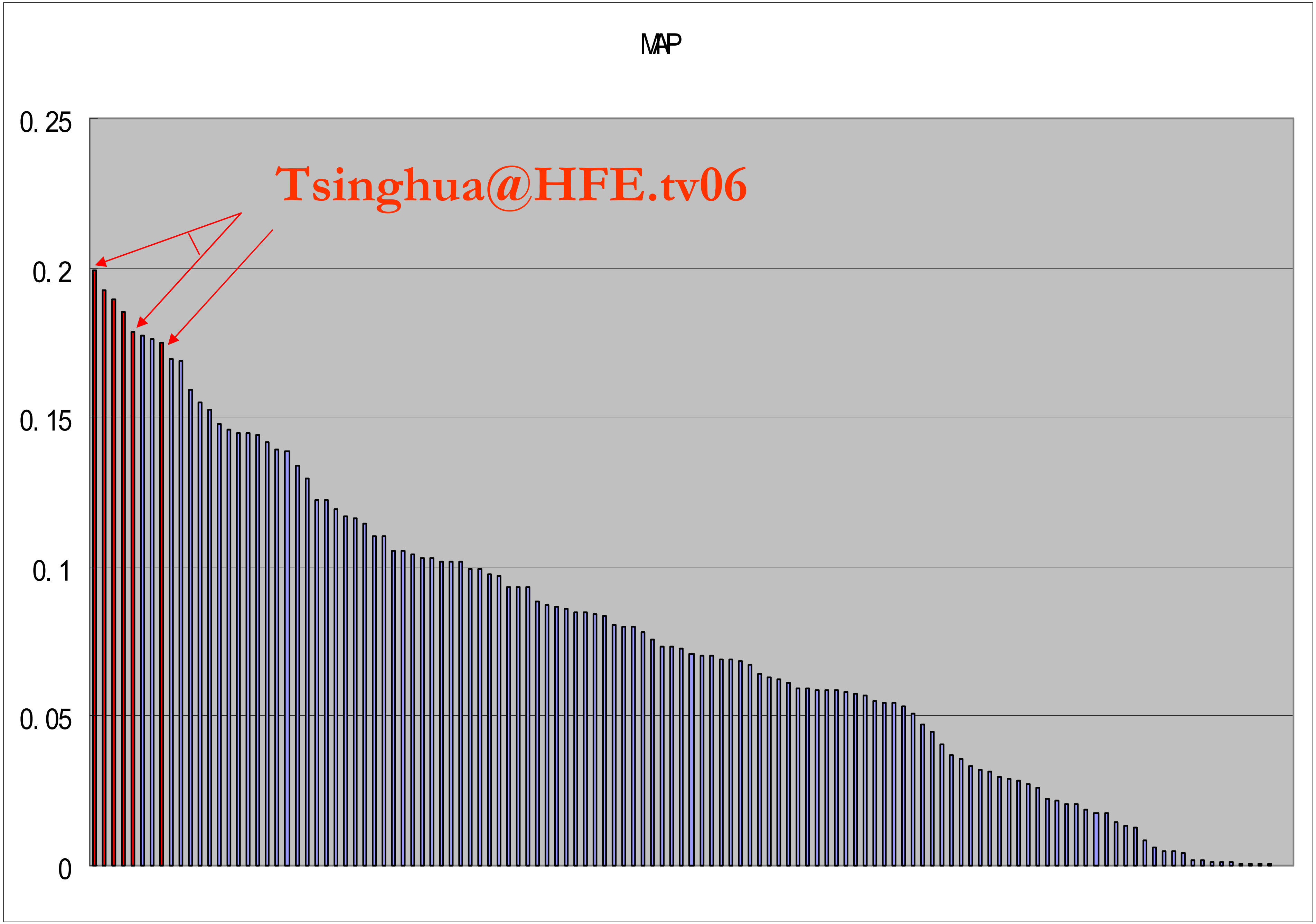
How to evaluate such systems?

- TRECVID benchmark try to provide
  - a common basis for comparison/evaluation for video retrieval which is reliable and on a large scale
- 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
  - Invite participation and see what happens…

Submission and Result

HFE Runs	MAP	Description
B_hua_1 (Tai)	0.189	RankBoost which takes baseline shot score plus Mediamill 404 dim features as weak rankings
B_hua_2 (Hua)	0.175	RankBoost which takes 110 dim baseline keyframe score as weak rankings, 50 top rankings selected
B_hua_3 (Huang)	0.199	Roundrobin which combines all five other runs on the shot rank basis[1]
B_hua_4 (HengNorth)	0.179	RankBoost which takes 110 dim baseline keyframe score as weak rankings, 200 top rankings selected
B_hua_5 (HengSouth)	0.185	Automatic or manual rule for setting concept context with 39 dim concept vector for each shot
A_hua_6 (Song)	0.192	Baseline, select and weight top 50 SVM classifiers out of 110 trained on 22 features respectively

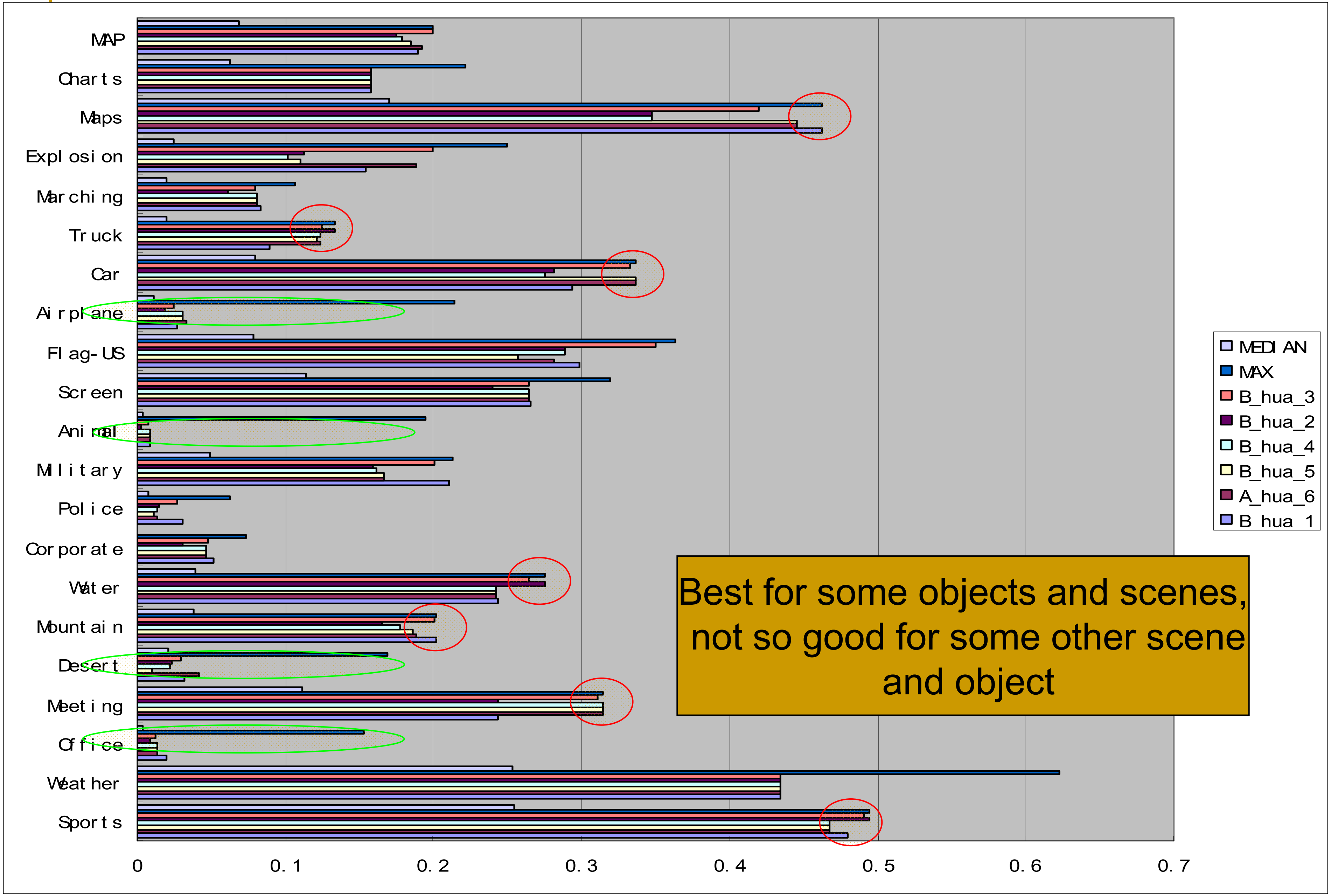
[1] After a bug-fix, the roundrobin run turns out to be the best run among both our submitted runs and all submitted runs.


MAP
Tsinghua@HFE.tv06
0. 25
0. 2
0. 15
0. 1
0. 05
0

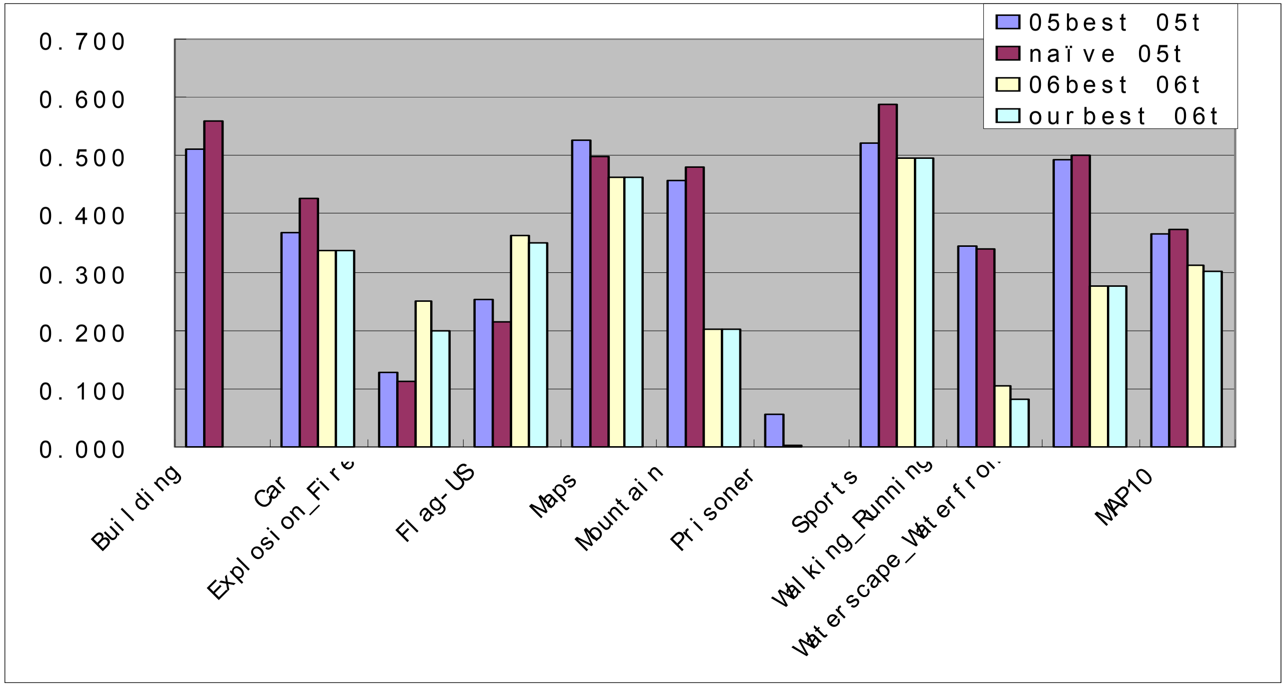
Concept Detection Lessons

- 1. The RankBoost algorithm builds a diversified bundle of classifiers for each feature and alleviates the burden of the fusion process. In other words, what to fuse is more important than how to fuse.
- 2. The Weight and Select fusion algorithm outperform the other ones since over-fitting occurs quite often in fusion, especially when there is strong mismatch of concept occurrence and broadcasting style between the training and testing data.
- 3. The pooling strategy used does not give the Roundrobin method significantly higher MAP (+3.5%) though it is looked more than other runs.

Performance for concepts

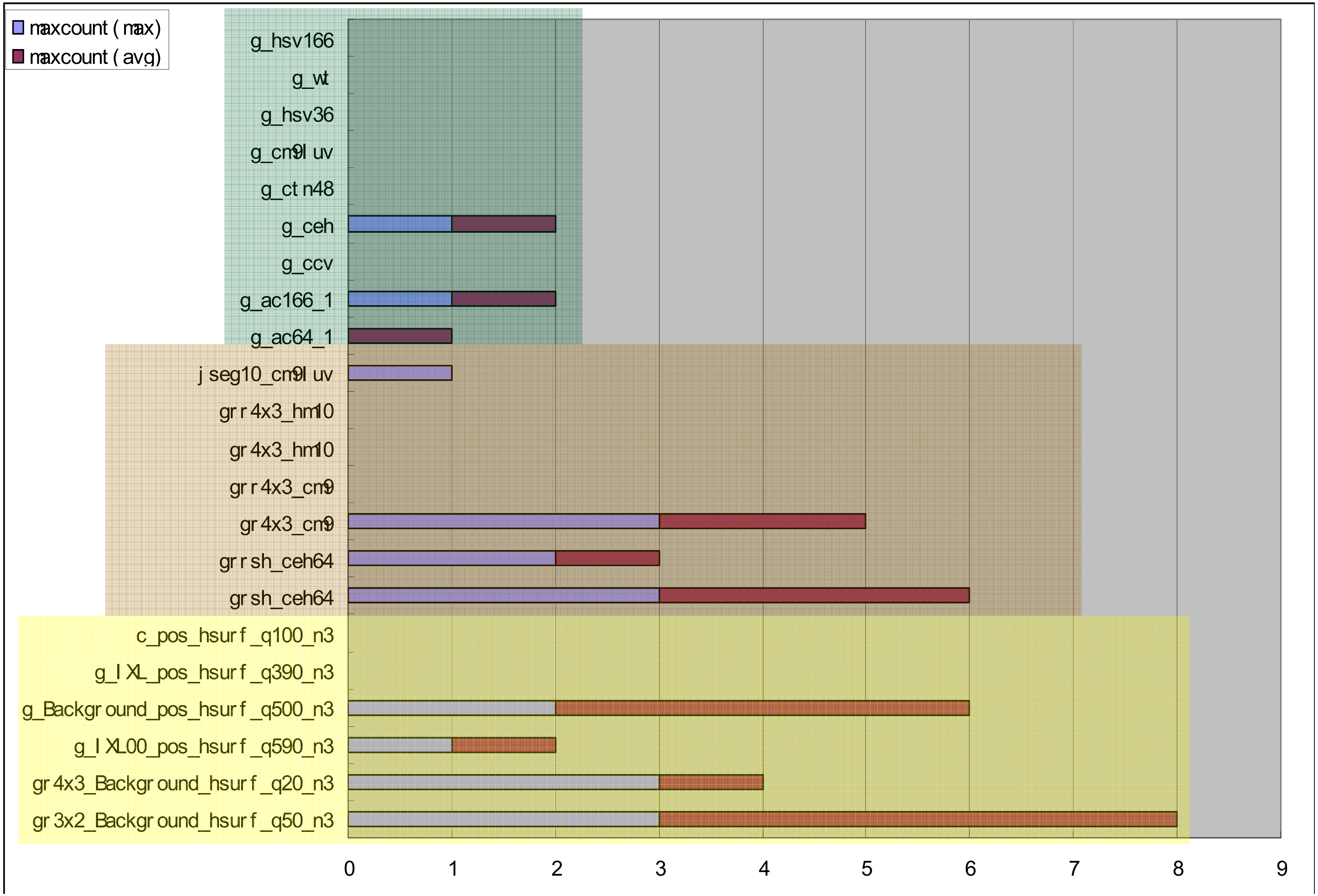


Evaluating our system on both tv05 and tv06

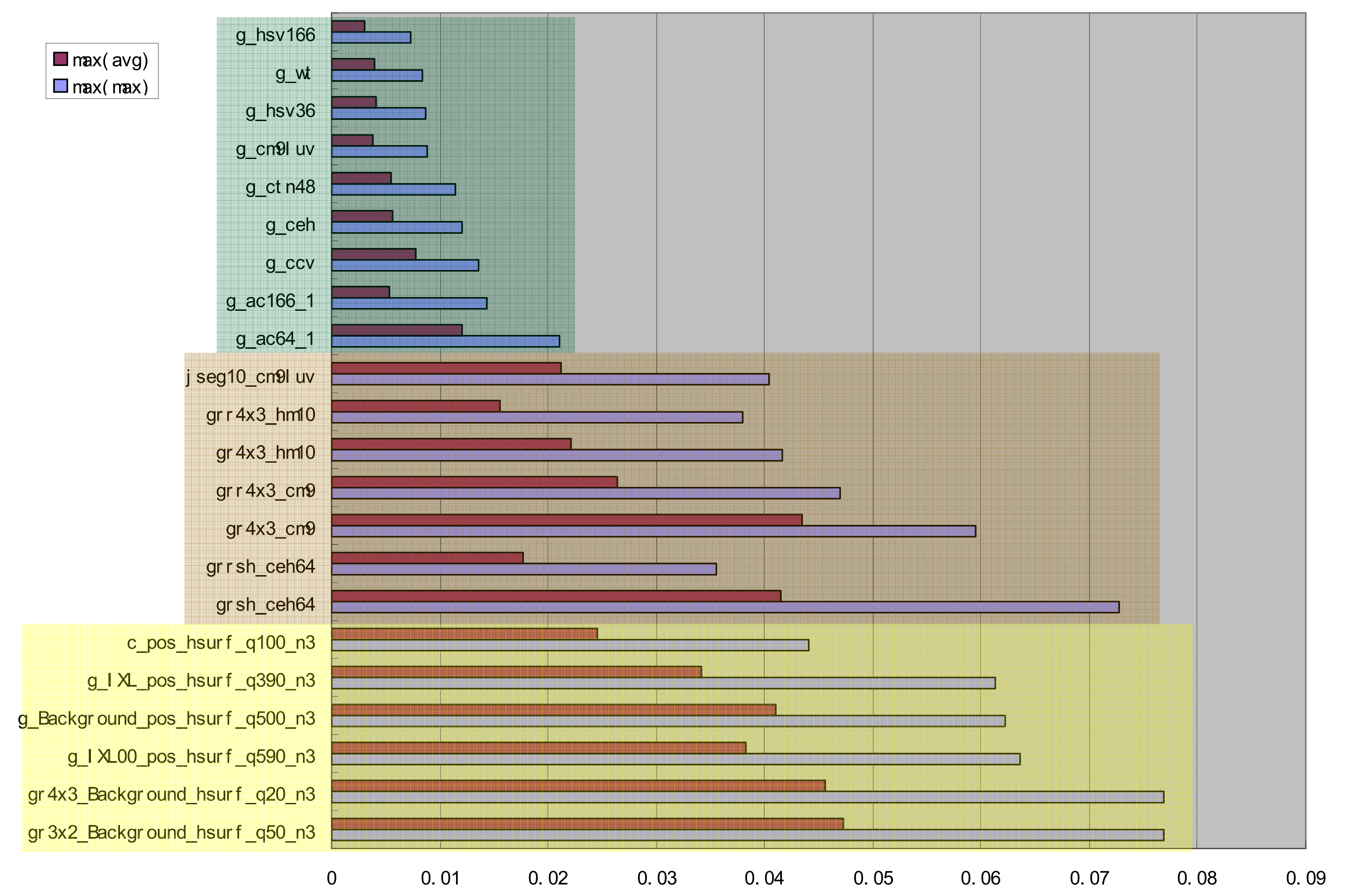


- We are slightly better than best result for last year
- Significant performance drop from last year (-24%!) maybe due to the large time gap, a real challenge for style analysis
- Performance on Explosion and Flag have increased
- Sport Visual only is better than Multilingual last year

Performance decomposition-maxcount



Performance decomposition-maxAP



Performance for representation and features

- No best feature for all concept
- No best feature for every concept either
- Grid based layout outperforms segmentation since it captures the spatial layout
- But global features can not be neglected
- SIFT/SURF features generalize well for scene and events
- χ^2 kernel is better than EMD for the grid layout

Outline

- Rich representations for rich semantics
- System design and implementation
- Benchmark results
- Future directions

Further work

- Neuroscientifically sound vs. biologically motivated
  - Learning more concept relevant features
  - Incorporate feature processing hierarchy
- Incorporate more spatial and temporal cues
- Refine the face/body segmentation features for social roles
- Better fusion strategy and method
- Re-annotate the 05t for retrain StackSVM and RankBoost?

Acknowledgment

- For trecvid 2006 benchmark
  - Prof. AI Haizhou for face detection
  - Computation Platform from NLIST
  - Intel China Research Center (ICRC)
  - D. Lowe for SIFT binary
  - H. Bay for SURF binary
  - C.-J. Lin for LIBSVM
  - MediaMill for donating their detection results
  - LSCOM workshop for annotation

Thanks!

Q/A☺

wdong01@mails.tsinghua.edu.cn