Facing the Challenge

A Divide and Conquer Approach to Concept Detection

C.G.M. Snoek, J.C. van Gemert, Th. Gevers, B. Huurnink, D.C. Koelma,
K.E.A. van de Sande, A.W.M. Smeulders, C.J. Veenman, M. Worring

Intelligent Systems Lab Amsterdam,
University of Amsterdam, The Netherlands

NIST TRECVID benchmark

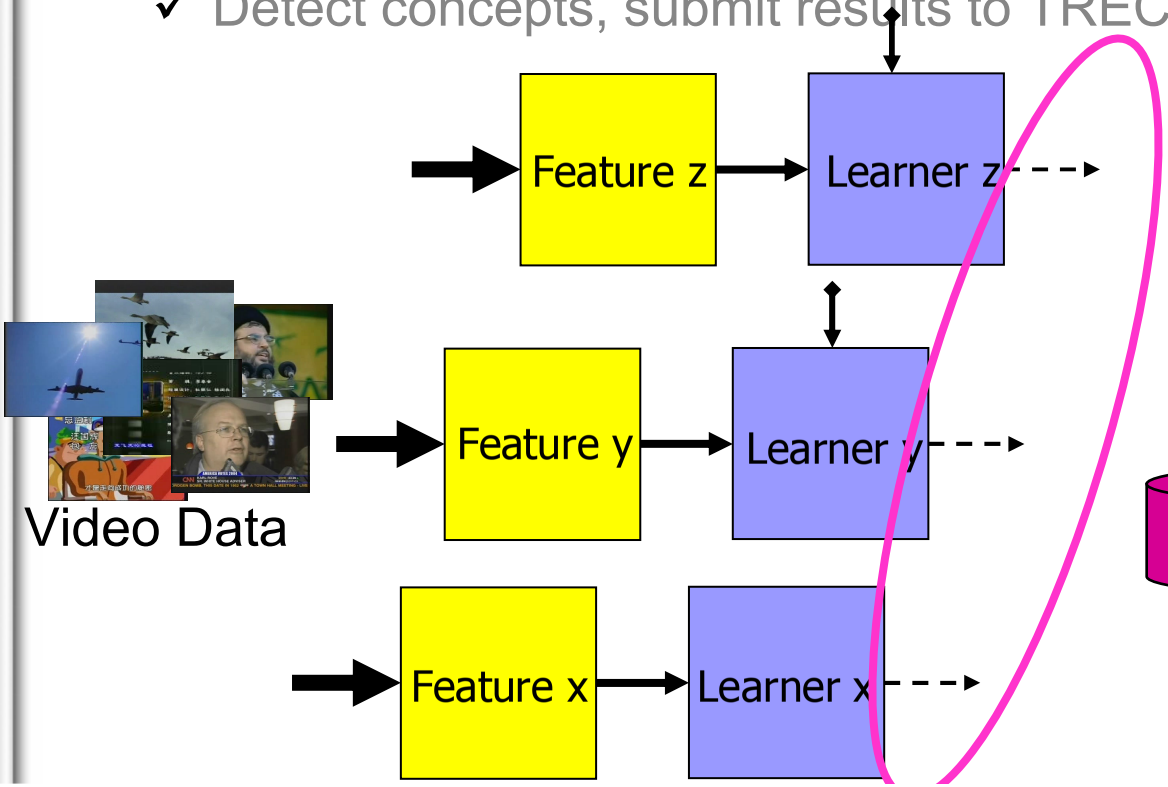
- Introduction
- Challenge
- Visual-only
- Results
- Lessons

Benchmark objectives

- ✓ Promote progress in video retrieval research
- ✓ Provide common dataset (shots, annotations, key frames)
- ✓ Use open, metrics-based evaluation

High-level feature task

- ✓ Detect concepts, submit results to TRECVID

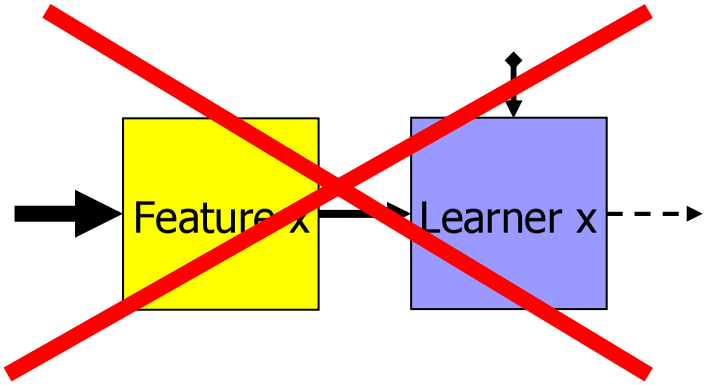


NIST

Ground truth

Benchmark limitations

- Focus is on the final result
  - ✓ TRECVID judges relative merit of indexing methods
  - ✓ Ignores repeatability of intermediate analysis steps
- Systems are becoming more complex
  - ✓ Typically combining several features and learning methods

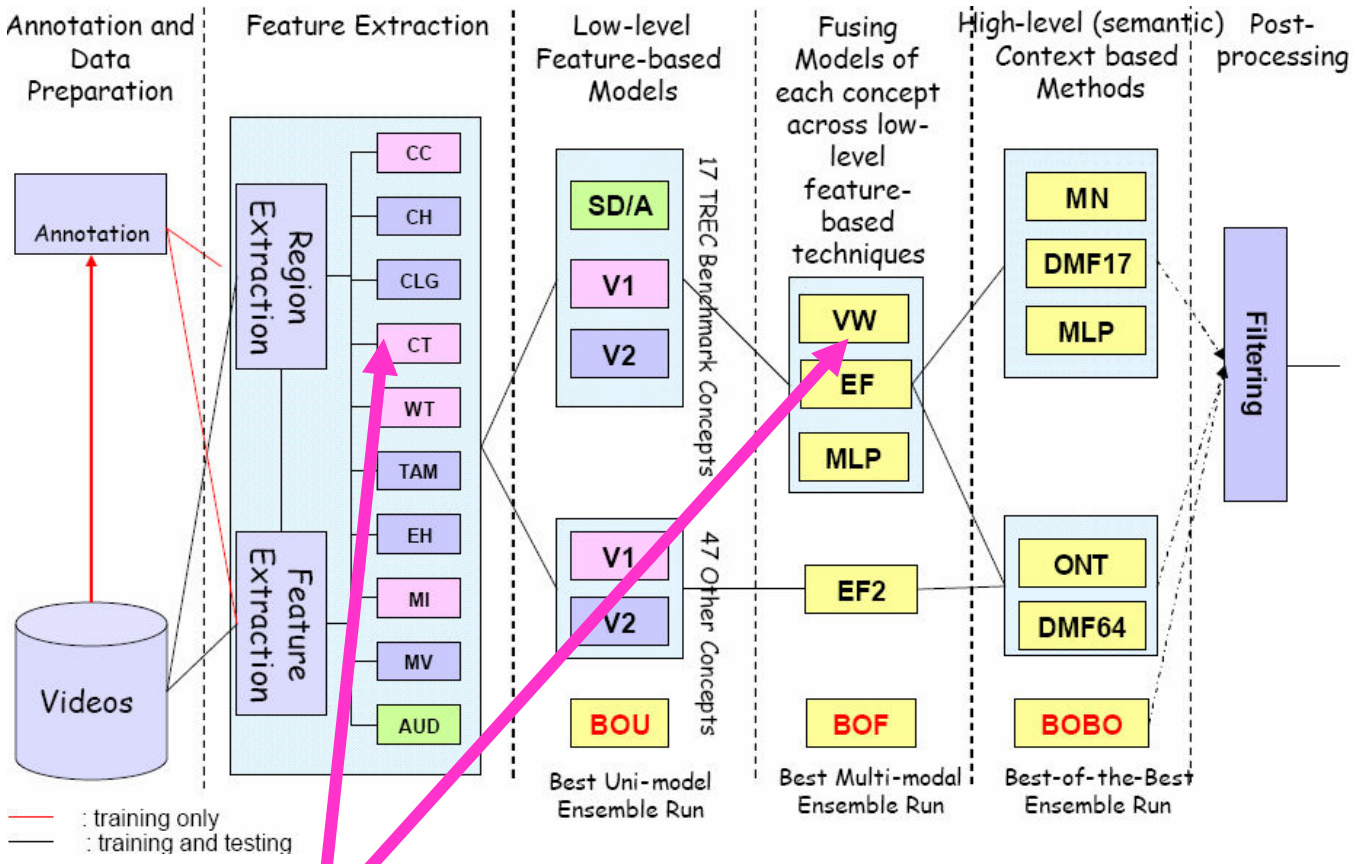


- Share video data, not the annotations and features
  - ✓ Component-based optimization and comparison impossible

IBM
Semantic video indexing

- Introduction
- Challenge
- Visual-only
- Results
- Lessons

Analysis Pipeline



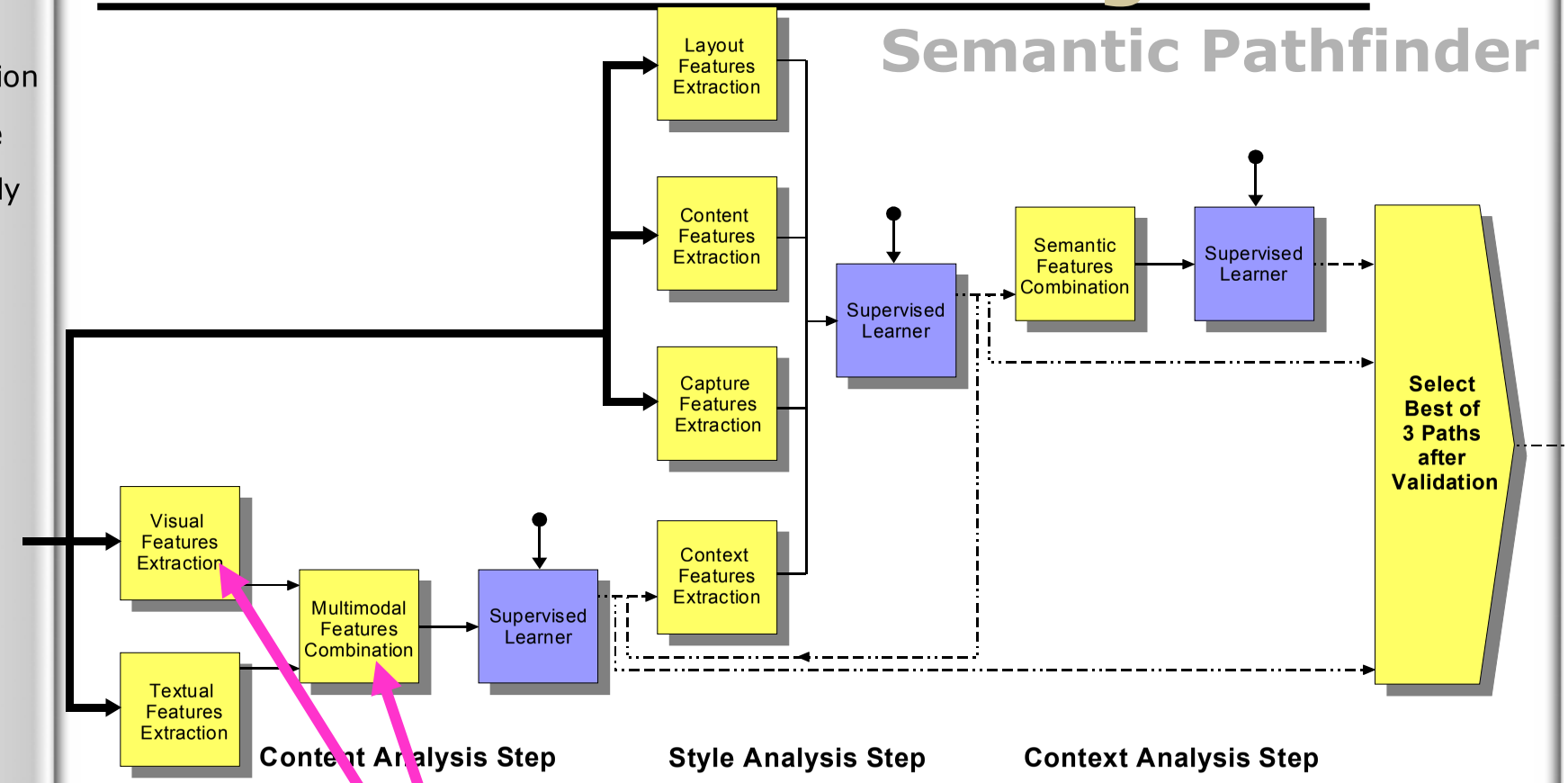
What is contribution of these components?

MediaMill

Semantic video indexing

- Introduction
- Challenge
- Visual-only
- Results
- Lessons

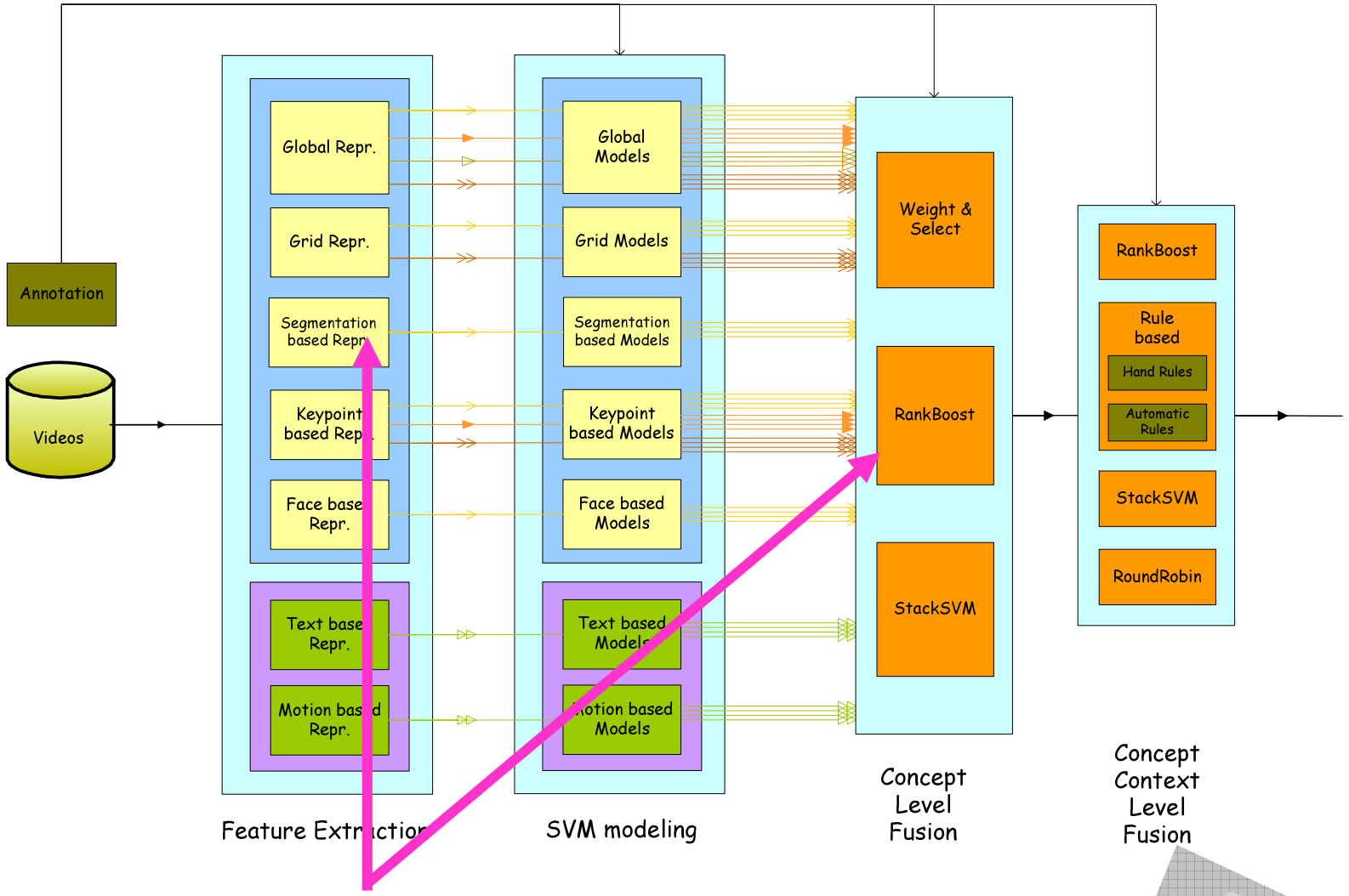
Semantic Pathfinder



What is contribution of these components?

Tsinghua
Semantic video indexing

- Introduction
- Challenge
- Visual-only
- Results
- Lessons

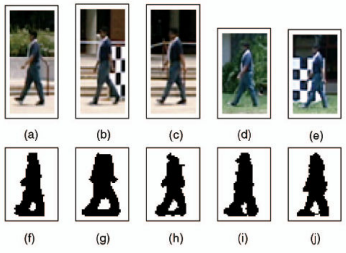


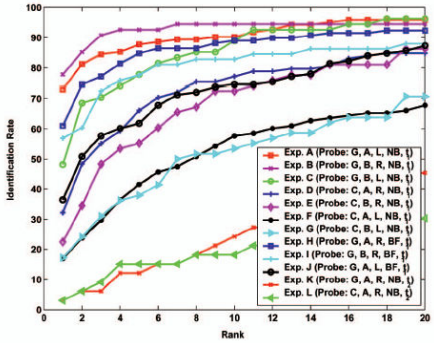
What is contribution of these components?

Lessons from computer vision

➢ Gait-based identification of humans

- ✓ Standardize common components





- ✓ What is needed?
  - ❖ annotated data set
  - ❖ baseline implementation
  - ❖ baseline results

} Challenge Problem

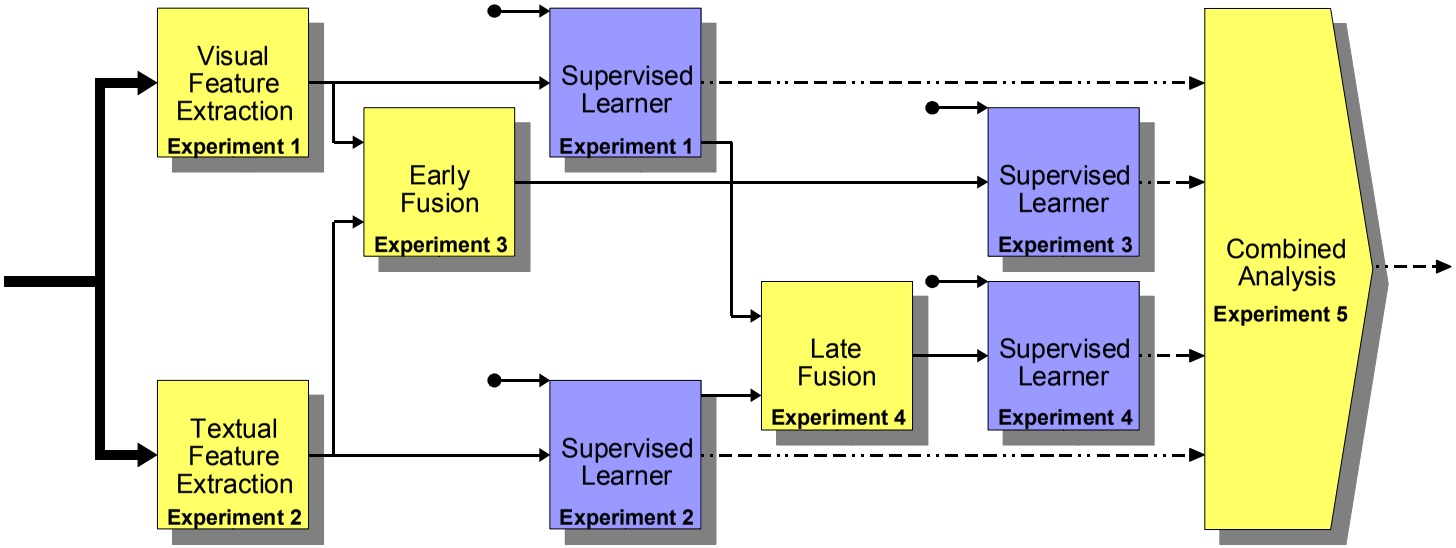
➢ We propose concept detection Challenge Problem

- ✓ Evaluates semantic video indexing methods
- ✓ Allows for component-based optimization
- ✓ Offers reference during feature development

The MediaMill Challenge

- Introduction
- Challenge
- Visual-only
- Results
- Lessons

- Given a large video data set, we provide:
  - ✓ A lexicon of **101** annotated concepts
  - ✓ 5 pre-cooked multimedia analysis experiments
  - ✓ Baseline detector algorithms
  - ✓ Baseline performance

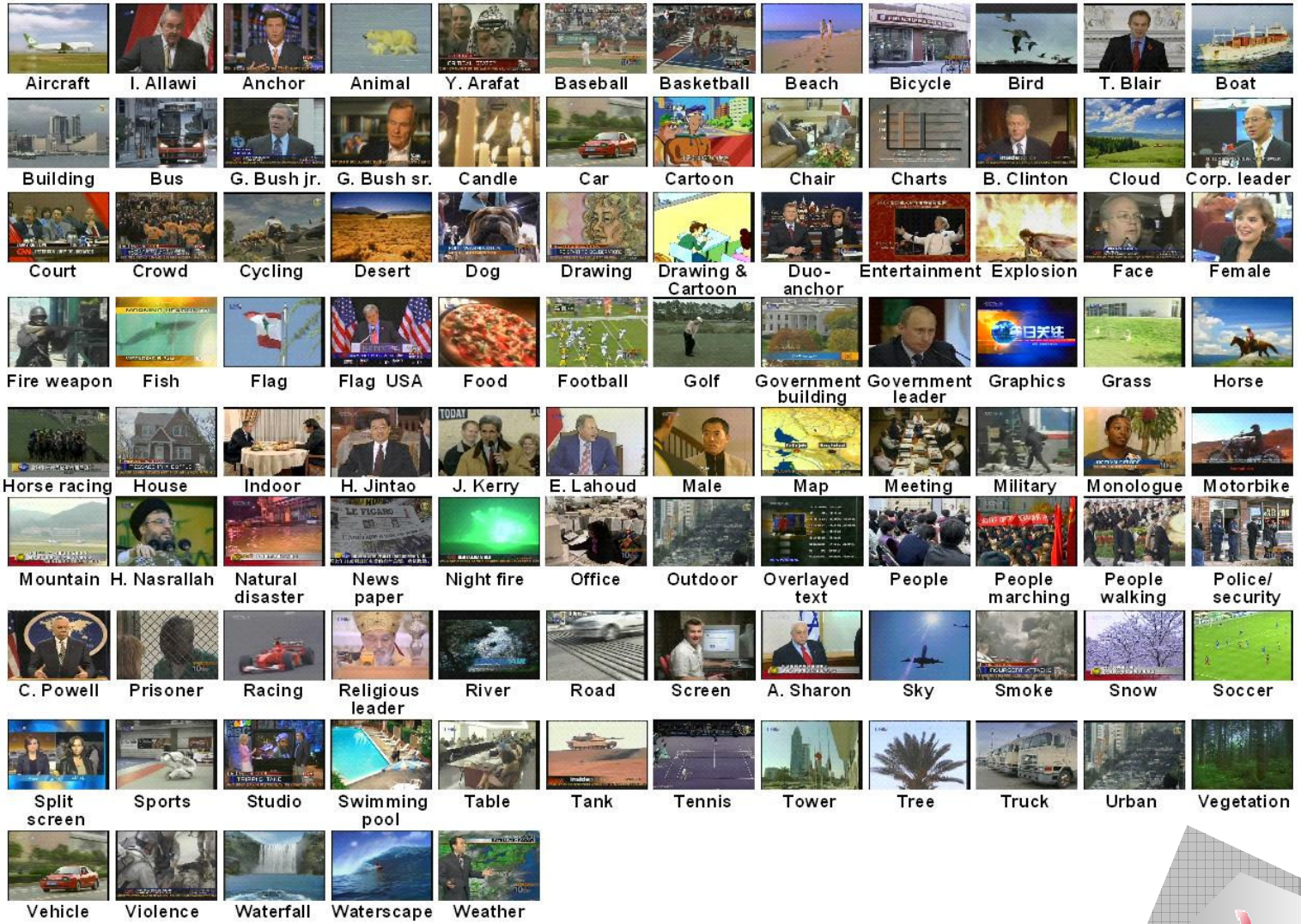


- Can the multimedia research community improve?

MediaMill

Annotated 101 concept lexicon

- Introduction
- Challenge
- Visual-only
- Results
- Lessons



Baseline results

- Introduction
- Challenge
- Visual-only
- Results
- Lessons

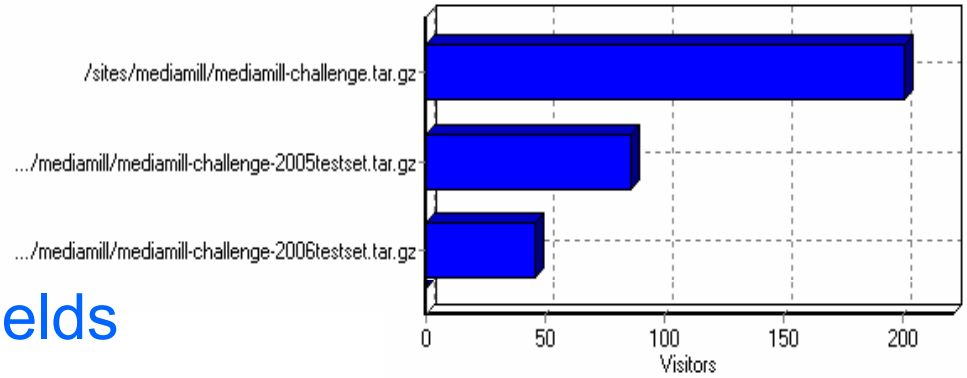
➢ Reference for comparison

	Concept	Ground Truth Train (%)	Ground Truth Test (%)	Challenge Experiments 1	2	3	4	5
1	People	77.67	75.87	0.831	0.817	0.890	0.840	0.840
2	Face	64.15	62.37	0.895	0.737	0.892	0.890	0.890
3	Overlayed text	36.33	34.30	0.669	0.533	0.642	0.666	0.669
4	Outdoor	32.68	38.33	0.688	0.579	0.709	0.691	0.691
5	Entertainment	19.64	12.55	0.166	0.179	0.257	0.146	0.179
6	Indoor	19.59	21.20	0.593	0.460	0.592	0.606	0.592
7	Studio	13.66	14.20	0.636	0.490	0.664	0.651	0.664
8	People walking	13.61	16.83	0.353	0.294	0.338	0.296	0.338
9	Urban	11.78	8.80	0.222	0.178	0.195	0.201	0.195
10	Crowd	11.48	16.12	0.480	0.288	0.490	0.440	0.490
11	Sky	10.77	11.38	0.478	0.218	0.496	0.463	0.496
12	Government leader	9.35	7.87	0.213	0.213	0.222	0.236	0.213
13	Violence	8.07	9.75	0.317	0.301	0.334	0.237	0.334
14	Road	7.76	6.60	0.195	0.138	0.212	0.188	0.195
15	Vehicle	7.61	8.53	0.221	0.167	0.271	0.190	0.271
16	Building	6.86	11.16	0.316	0.154	0.233	0.291	0.154
17	Male	5.71	2.38	0.086	0.034	0.068	0.069	0.086
18	Anchor	5.09	4.85	0.631	0.201	0.620	0.618	0.631
19	Car	4.87	5.93	0.252	0.118	0.246	0.215	0.252
20	Meeting	4.53	4.86	0.257	0.158	0.211	0.257	0.257
21	Female	4.38	2.11	0.086	0.020	0.061	0.068	0.086
22	Military	4.14	6.58	0.217	0.206	0.235	0.203	0.235
23	Vegetation	3.87	4.64	0.183	0.051	0.161	0.150	0.183
24	Sports	3.76	2.61	0.304	0.267	0.231	0.308	0.304
25	Monologue	3.10	2.33	0.094	0.051	0.074	0.081	0.094
26	Graphics	2.89	3.48	0.365	0.275	0.379	0.367	0.365
27	Corporate leader	2.57	1.30	0.016	0.020	0.014	0.018	0.020
28	Waterscape	2.31	1.89	0.150	0.079	0.134	0.142	0.134
29	People marching	1.93	4.13	0.228	0.087	0.267	0.109	0.267
30	Soccer	1.67	0.29	0.503	0.000	0.079	0.372	0.503
31	Mountain	1.64	1.01	0.141	0.022	0.092	0.157	0.141
32	G. Bush jr.	1.61	0.54	0.062	0.065	0.040	0.060	0.062
33	Office	1.56	1.75	0.077	0.024	0.045	0.037	0.024
34	Screen	1.53	1.90	0.101	0.063	0.058	0.121	0.063
35	Flag	1.26	1.12	0.189	0.029	0.120	0.166	0.189
36	Truck	1.16	1.02	0.038	0.019	0.042	0.038	0.019
37	Map	1.16	1.21	0.476	0.220	0.313	0.407	0.476
38	Smoke	1.13	2.14	0.250	0.103	0.366	0.149	0.250
39	Animal	1.00	0.91	0.209	0.204	0.199	0.239	0.199
40	Weather	0.99	1.25	0.405	0.730	0.701	0.566	0.730
41	Aircraft	0.99	0.94	0.073	0.033	0.115	0.030	0.115
42	Police/security	0.92	0.77	0.012	0.053	0.082	0.017	0.053
43	Flag USA	0.92	0.94	0.227	0.036	0.157	0.184	0.227
44	Grass	0.90	0.59	0.064	0.004	0.028	0.054	0.064
45	Cloud	0.87	1.54	0.117	0.042	0.078	0.129	0.117
46	Split screen	0.86	0.60	0.630	0.100	0.321	0.566	0.630
47	Desert	0.81	1.44	0.103	0.032	0.093	0.052	0.032
48	Natural disaster	0.81	0.93	0.055	0.091	0.139	0.084	0.091
49	Boat	0.78	0.54	0.096	0.109	0.083	0.020	0.109
50	Tree	0.78	0.84	0.124	0.011	0.063	0.087	0.124
51	Charts	0.76	0.51	0.327	0.301	0.254	0.355	0.327
52	Table	0.75	0.52	0.073	0.006	0.037	0.060	0.073
53	Tower	0.75	0.63	0.057	0.009	0.023	0.033	0.057
54	Basketball	0.69	0.34	0.382	0.219	0.179	0.239	0.382
55	Y. Arafat	0.62	0.88	0.026	0.072	0.034	0.013	0.072
56	Chair	0.60	0.58	0.486	0.101	0.261	0.467	0.486
57	Explosion	0.53	1.04	0.098	0.038	0.078	0.046	0.038
58	Food	0.50	0.83	0.287	0.085	0.188	0.170	0.287
59	Bus	0.43	0.64	0.013	0.007	0.009	0.005	0.007
60	Snow	0.41	0.53	0.085	0.018	0.045	0.004	0.085
61	Fire weapon	0.35	0.52	0.121	0.013	0.060	0.047	0.121
62	Tennis	0.34	0.56	0.448	0.195	0.299	0.397	0.397
63	Prisoner	0.33	0.22	0.047	0.027	0.051	0.004	0.051
64	News paper	0.31	0.27	0.375	0.000	0.121	0.384	0.375
65	E. Lahoud	0.30	0.15	0.289	0.080	0.115	0.196	0.289
66	J. Kerry	0.29	0.01	0.000	0.012	0.002	0.001	0.000
67	House	0.29	0.36	0.023	0.004	0.007	0.009	0.004
68	Government building	0.27	0.19	0.011	0.038	0.079	0.002	0.002
69	Religious leader	0.27	0.23	0.043	0.026	0.035	0.041	0.026
70	Fish	0.27	0.12	0.489	0.068	0.408	0.312	0.489
71	Duo-anchor	0.26	0.18	0.634	0.022	0.108	0.287	0.634
72	Golf	0.25	0.31	0.091	0.007	0.042	0.143	0.091
73	I. Allawi	0.21	0.02	0.000	0.030	0.002	0.000	0.000
74	Bicycle	0.20	0.04	0.006	0.454	0.223	0.733	0.454
75	Court	0.20	0.30	0.093	0.041	0.030	0.052	0.041
76	G. Bush sr.	0.20	0.01	0.000	0.000	0.000	0.000	0.000
77	Football	0.20	0.39	0.048	0.016	0.020	0.043	0.016
78	Cycling	0.18	0.03	0.042	0.950	0.888	0.608	0.950
79	Bird	0.18	0.23	0.724	0.577	0.761	0.743	0.724
80	Drawing & Cartoon	0.17	0.38	0.265	0.207	0.181	0.191	0.207
81	Horse	0.16	0.02	0.000	0.000	0.000	0.000	0.000
82	Dog	0.14	0.38	0.225	0.012	0.103	0.019	0.012
83	Night fire	0.14	0.05	0.526	0.001	0.249	0.000	0.526
84	Horse racing	0.12	0.02	0.000	0.000	0.000	0.000	0.000
85	River	0.10	0.09	0.310	0.710	0.654	0.098	0.710
86	Racing	0.09	0.12	0.029	0.176	0.175	0.004	0.176
87	Candle	0.08	0.10	0.011	0.057	0.080	0.001	0.057
88	Cartoon	0.08	0.21	0.259	0.671	0.278	0.285	0.671
89	Drawing	0.08	0.17	0.293	0.011	0.044	0.026	0.293
90	Tank	0.08	0.08	0.008	0.003	0.011	0.004	0.008
91	Swimming pool	0.08	0.10	0.003	0.001	0.001	0.002	0.001
92	Beach	0.08	0.06	0.027	0.001	0.065	0.007	0.065
93	Waterfall	0.07	0.08	0.381	0.011	0.415	0.001	0.011
94	Motorbike	0.05	0.16	0.006	0.029	0.007	0.005	0.006
95	T. Blair	0.05	0.26	0.005	0.031	0.015	0.048	0.048
96	B. Clinton	0.05	0.21	0.004	0.010	0.189	0.002	0.189
97	H. Nasrallah	0.05	0.19	0.006	0.068	0.004	0.001	0.006
98	C. Powell	0.05	0.47	0.010	0.022	0.085	0.008	0.010
99	A. Sharon	0.04	0.19	0.050	0.019	0.035	0.001	0.050
100	H. Jintao	0.03	1.03	0.030	0.023	0.044	0.018	0.023
101	Baseball	0.01	0.41	0.003	0.066	0.003	0.011	0.003
	Mean			0.216	0.147	0.201	0.191	0.237

Using the Challenge

➢ Challenge donation to TRECVID community

- ✓ 80+ participants downloaded the data
- ✓ Some even used it for their TRECVID submission



➢ Used in other fields

- ✓ Concept detectors used at INEX benchmark
- ✓ Referenced as standard dataset on LIBSVM website

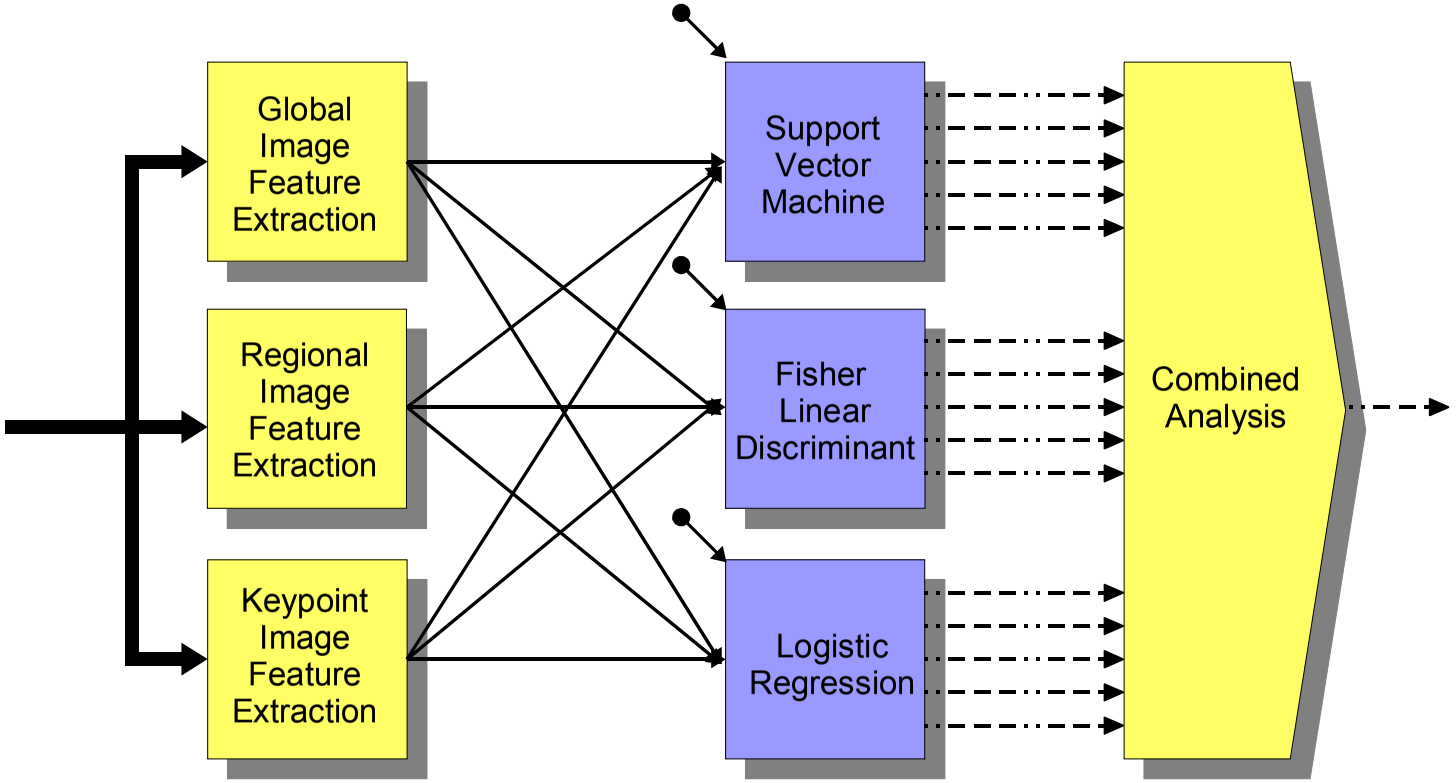
➢ Usage in education

- ✓ M.Sc. Student lab course
- ✓ Lowers threshold to enter field

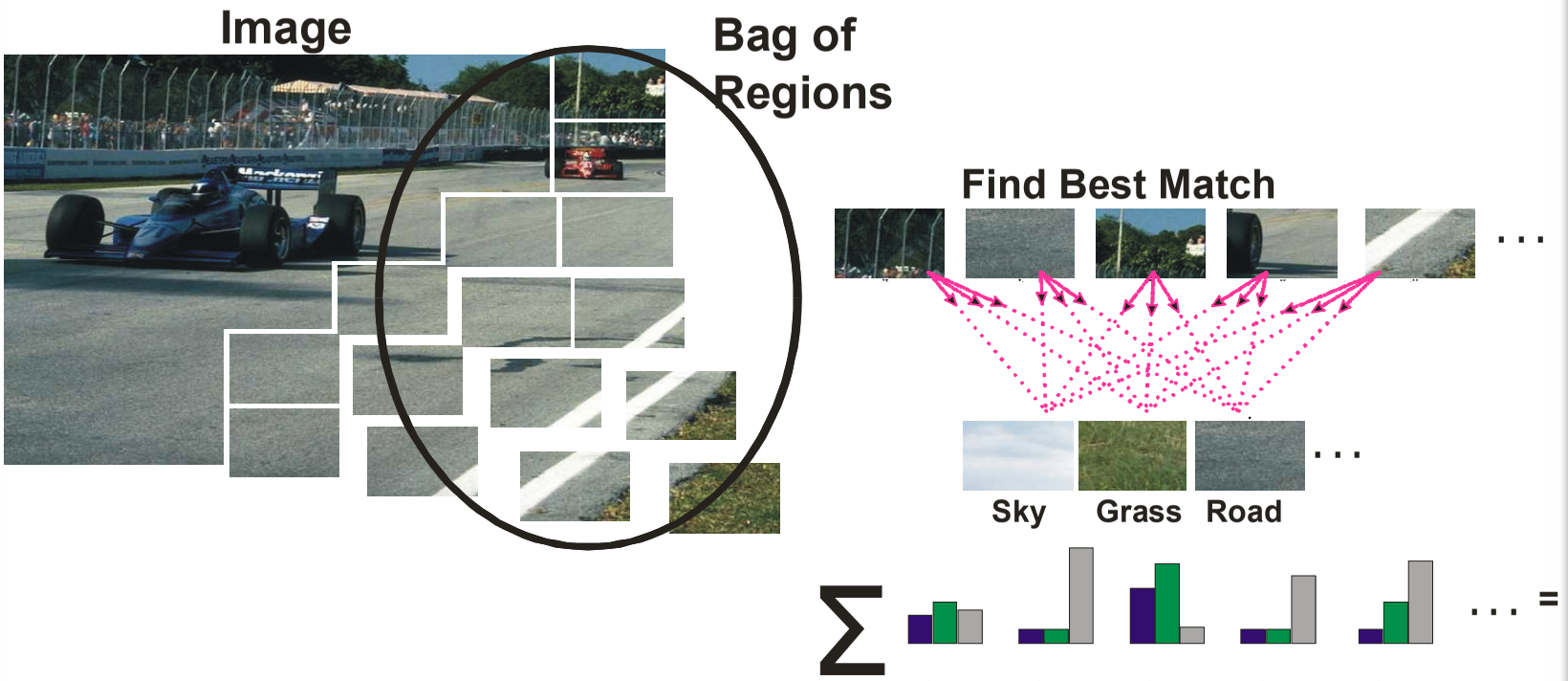
MediaMill TRECVID 2006 approach

- Introduction
- Challenge
- Visual-only
- Results
- Lessons

- Emphasize visual-only analysis
  - ✓ Viz.: experiment 1 of the MediaMill Challenge



Decompose image into proto-concepts

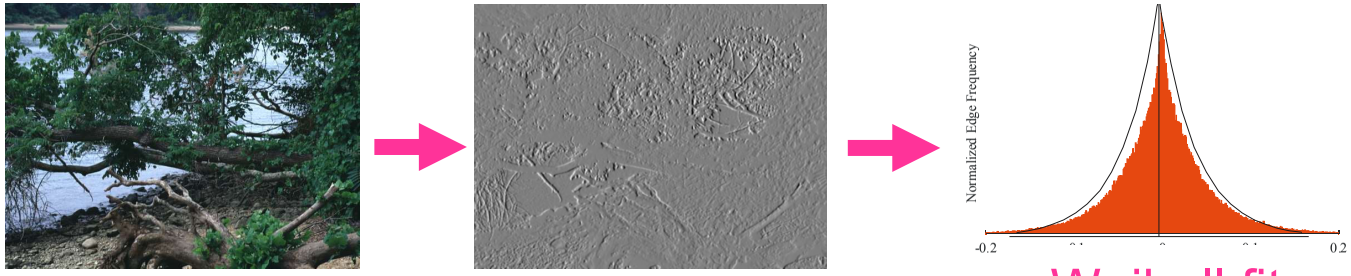


- **Problem**
  - ✓ Assumes one proto-concept per region is sufficient
- **Solution**
  - ✓ Use similarity distribution

Global image features

- Color Invariant features
  - ✓ Robust to Shadows
- Histogram of Edges
  - ✓ Resembles Weibull distribution

Original → x-edges → x-edges histogram



Weibull fit

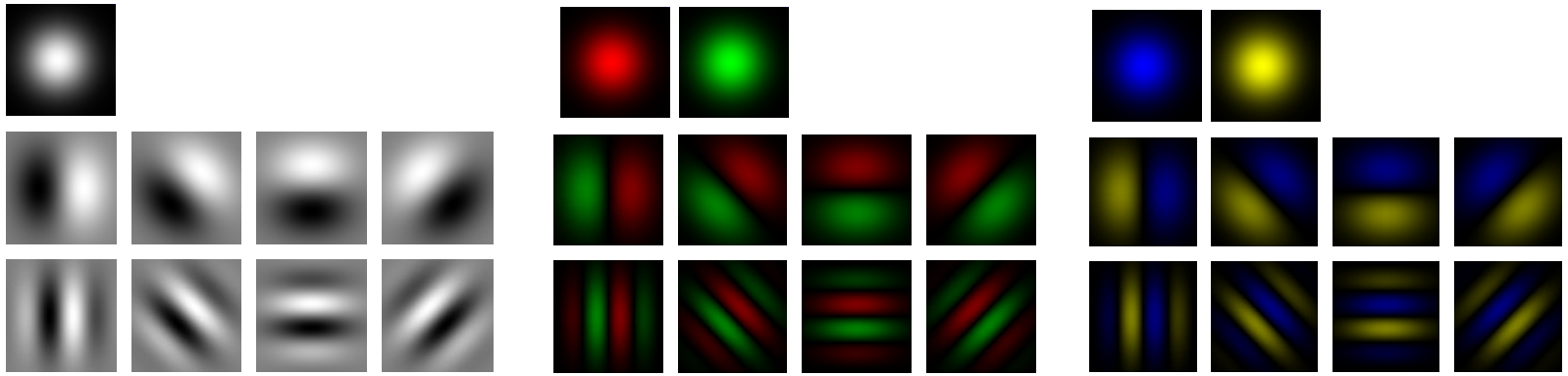
- Weibull depends on two parameters only
  - ✓ Substantial data reduction in feature vector
- Proto-concept similarity measure
  - ✓ Based on Weibull parameters

Regional image features

- We also use Weibull fit on region level

+

- Histogram of Gabor filter responses
  - ✓ zero order = color histogram
  - ✓ higher order = texture



- Proto-concept similarity measure
  - ✓ Histogram intersection

Keypoint image features

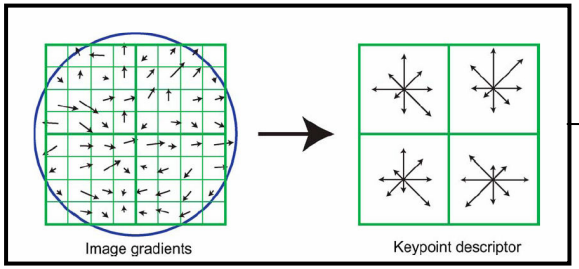
- Introduction
- Challenge
- Visual-only
- Results

Harris-Laplace

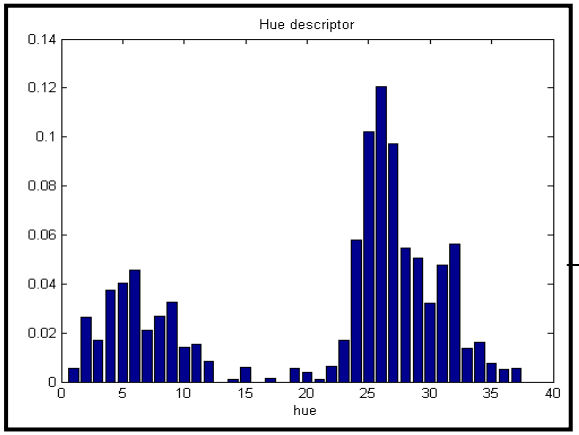
Boosted ColorHarris-Laplace

Detector

Descriptor



SIFT



Hue descriptor

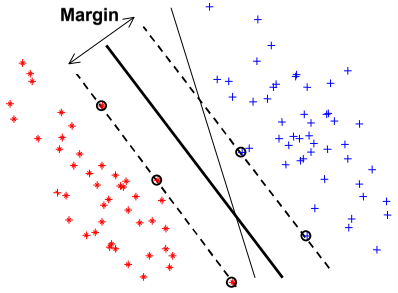
- Clustering in descriptor space
  - ✓ yields ‘proto-concept’ codebook
  - ✓ 10 codebook elements per concept, totalling ~400
  - ✓ dissimilarity codebook/descriptors yields vector

- Introduction
- Challenge
- Visual-only
- Results
- Lessons

Supervised learners

Support vector machine

- ✓ Maximizes margin between two classes
- ✓ Problematic when data not balanced
- ✓ Solution: expensive parameter tuning
  - ❖ We experiment with several settings
  - ❖ We use 3 supercomputer clusters



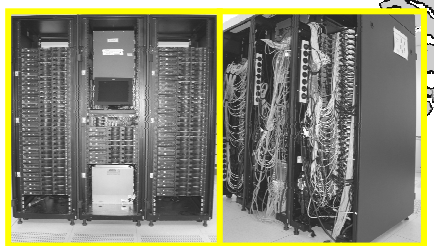
$$p(\omega|\vec{s}) = \frac{1}{1+\exp(\alpha\gamma(\vec{s})+\beta)}$$

Problem: SVMs are accurate but demanding

- ✓ Convergence too slow for large feature vectors

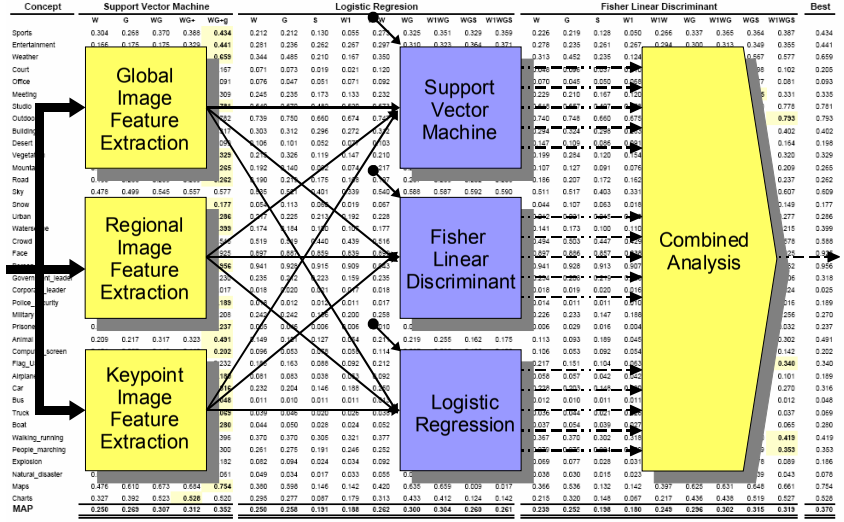
Solution: try less demanding classifiers

- ✓ Logistic regression
- ✓ Fisher linear discriminant



Visual-only analysis results on Challenge

- Extensive and repeatable experiments by
  - ✓ modular distribution of our efforts
  - ✓ emphasizing features, classifiers, and settings
- 'Best-of' conquers the baseline by more than 50% MAP
  - ✓ Overview in notebook

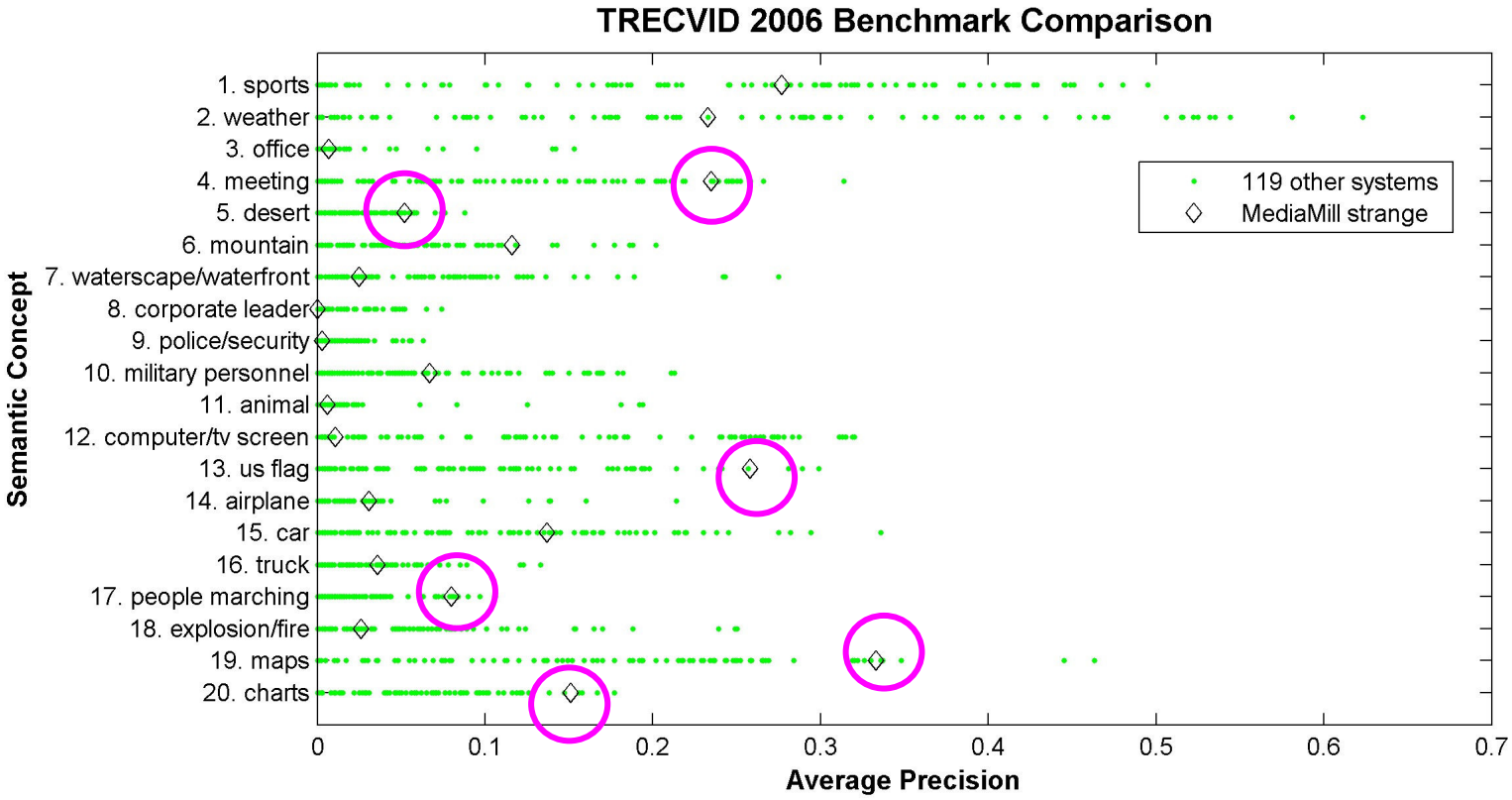


- Lessons:
  - ✓ Regional image features + SVM (incl. γ) most effective
  - ✓ Logistic/Fisher allows for classification using long vectors
    - ❖ Surprisingly good performance
    - ❖ Sometimes outperforming SVM

Best-of visual-only analysis

run: strange

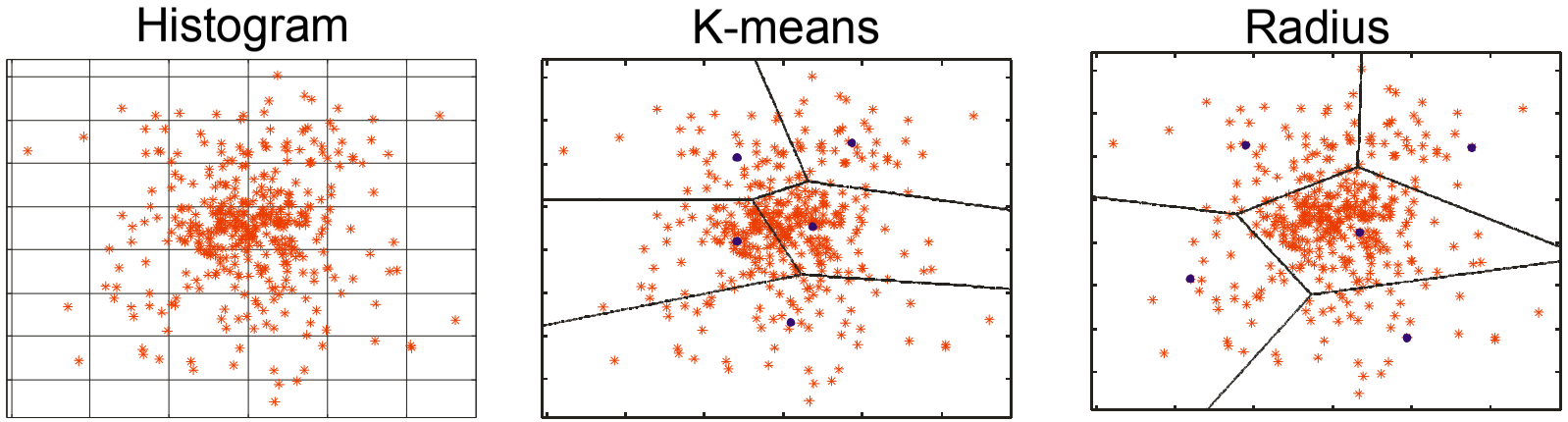
- MediaMill TRECVID2006 baseline run
  - ✓ OK for some concepts, but not spectacular



- All submitted runs should improve upon baseline
  - ✓ Three will be highlighted

Proto-concept clustering

- Problem: pre-defined codebook is restrictive
  - ✓ Use data-driven codebook instead
  - ✓ Tested for regional Weibull and Gabor features only
- Data-driven prototypes: radius clustering



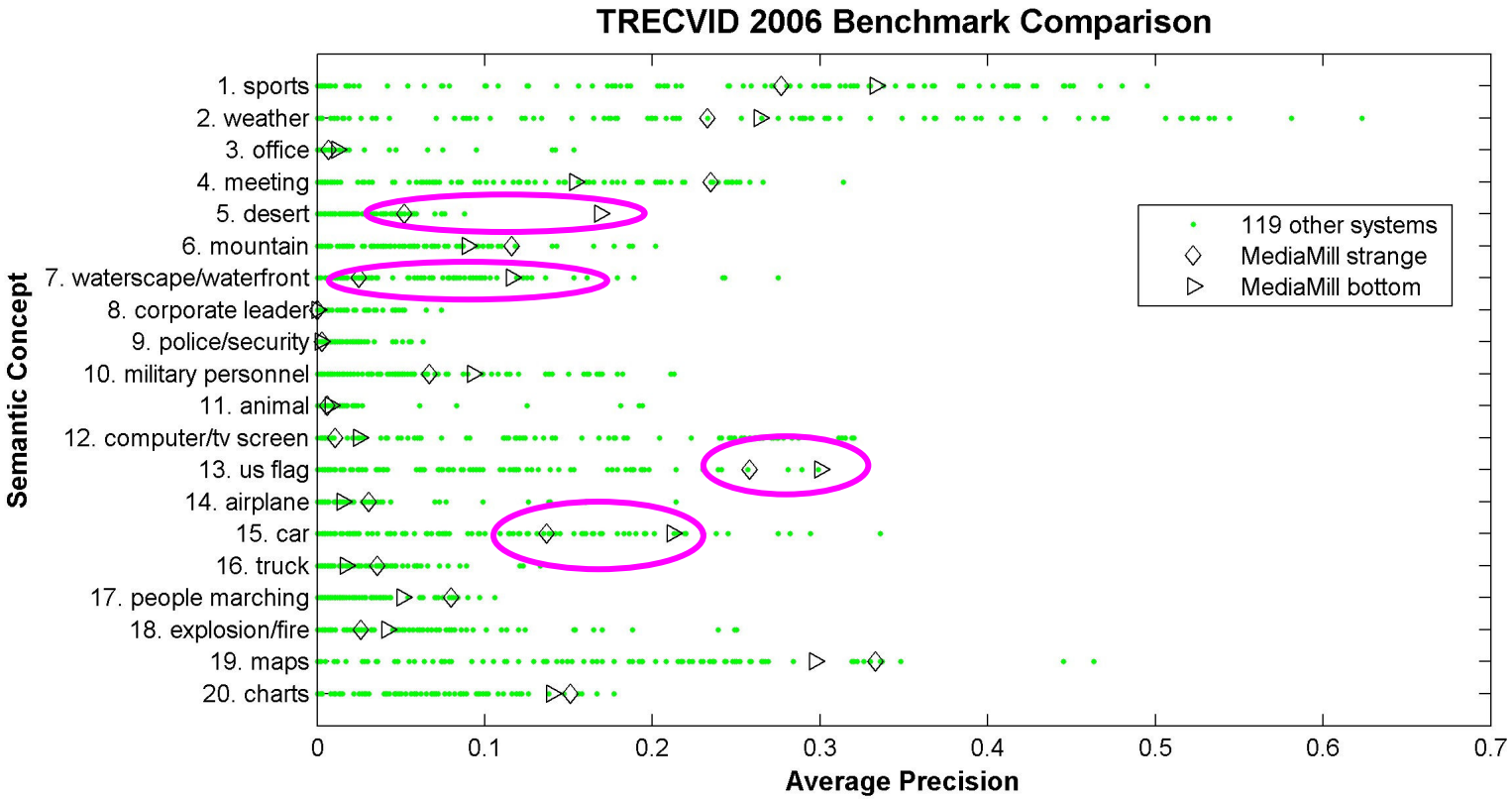
- Investigated in parallel to other experiments
  - ✓ No fusion used

- Introduction
- Challenge
- Visual-only
- Results
- Lessons

Proto-concept clustering

run: bottom

- Outperforms 'best-of visual' for half of the concepts
  - ✓ Desert = overall best



Coloring keypoints

on Challenge

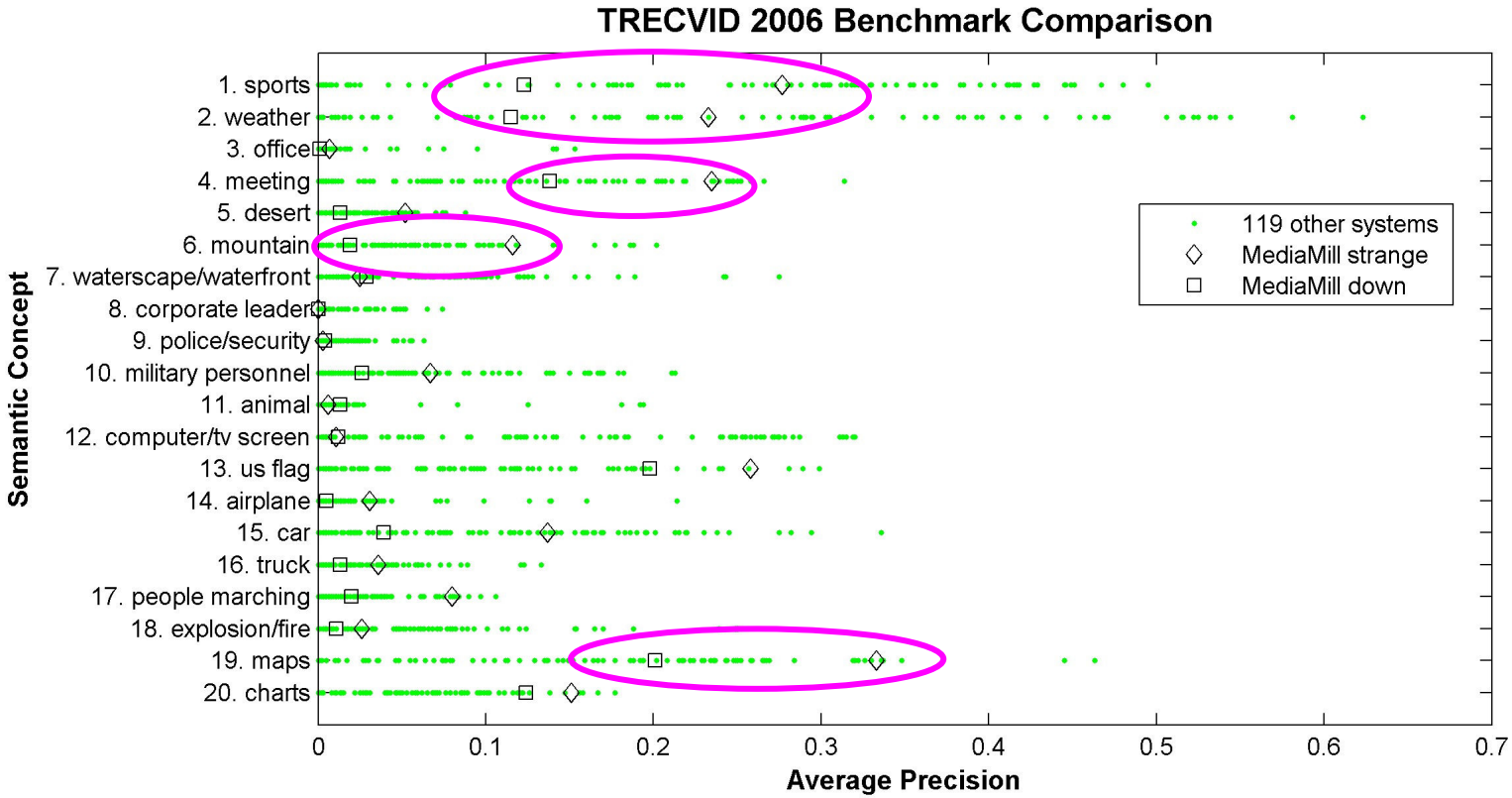
➢ Extending SIFT with color pays off

Detector	Descriptor	MAP (Challenge-39)	Change
Harris-Laplace	SIFT	0,198	-
Harris-Laplace	Hue+SIFT	0,231	+16%
Boosted ColorHarris-Laplace	SIFT	0,170	-14%
Boosted ColorHarris-Laplace	Hue	0,169	-15%
Harris-Laplace *and* Boosted ColorHarris-Laplace	SIFT	0,212	+7%
Late fusion		0,270	+36%

Coloring keypoints

run: down

- Modest performance
  - ✓ Unstable for images with only few interest regions



- A keypoint-only approach seems a bad idea

Late fusion of visual-only analysis

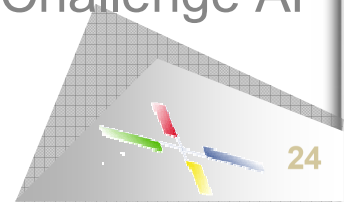
- All experiments yield for each shot a confidence

$$p(\omega_j|\vec{x}_i)$$

- Fuse concept detection scores using geometric mean
  - ✓ Requires no additional learn set
  - ✓ Favors single high confidence over many low confidences

$$\exp\left[\frac{1}{n}\sum_{k=1}^{n}\ln p_k(\omega_j|\vec{x}_i)\right]$$

- Pathfinder Procedure
  - ✓ Per concept
  - ✓ Include keypoint run as 1 experiment
  - ✓ Determine 9 additional experiments based on Challenge AP
  - ✓ Compute optimal path of experiments based on Challenge AP



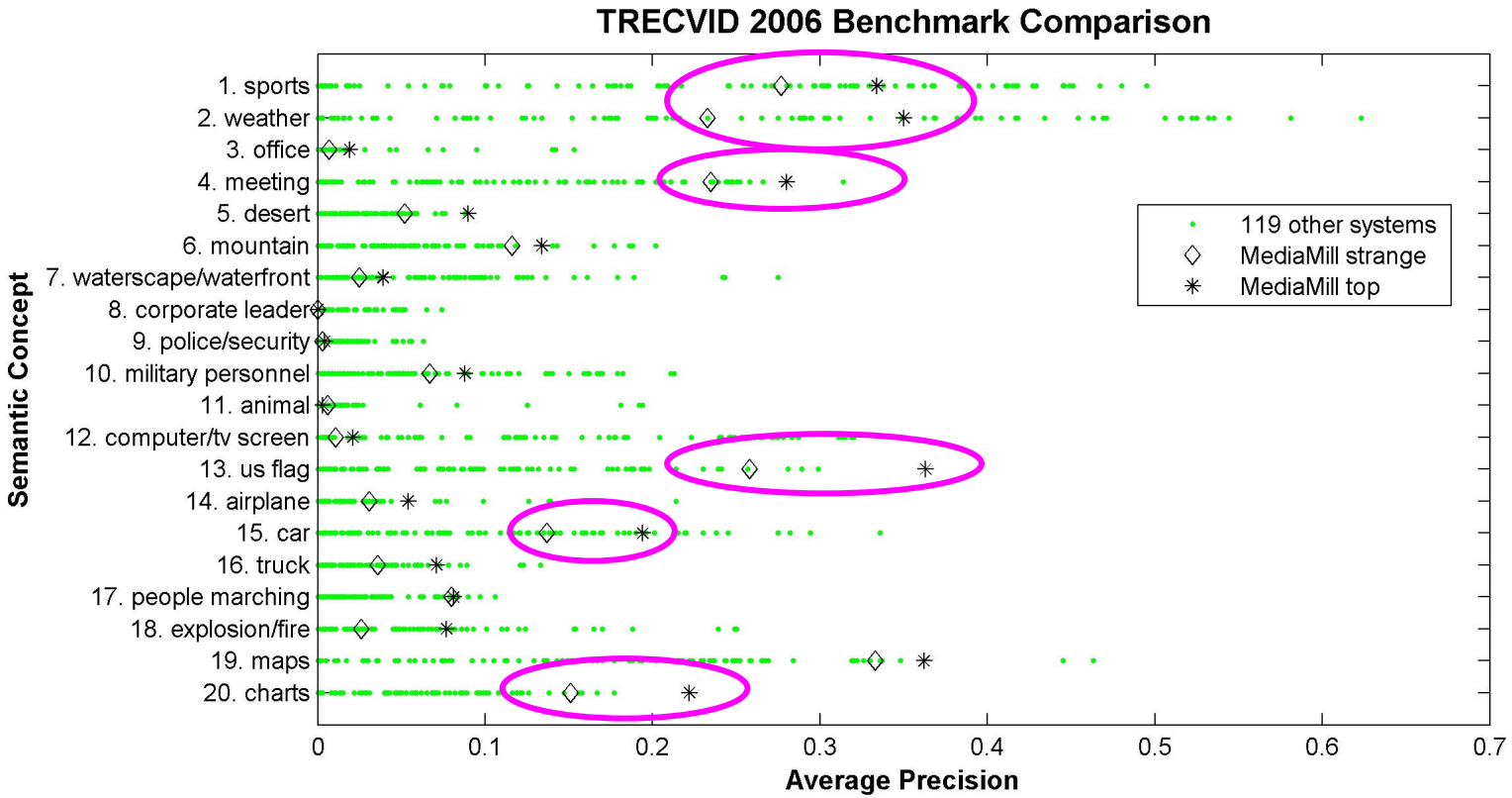
- Introduction
- Challenge
- Visual-only
- Results
- Lessons

Late fusion of visual-only analysis

run: top

➢ Fusion always improves over best single approach

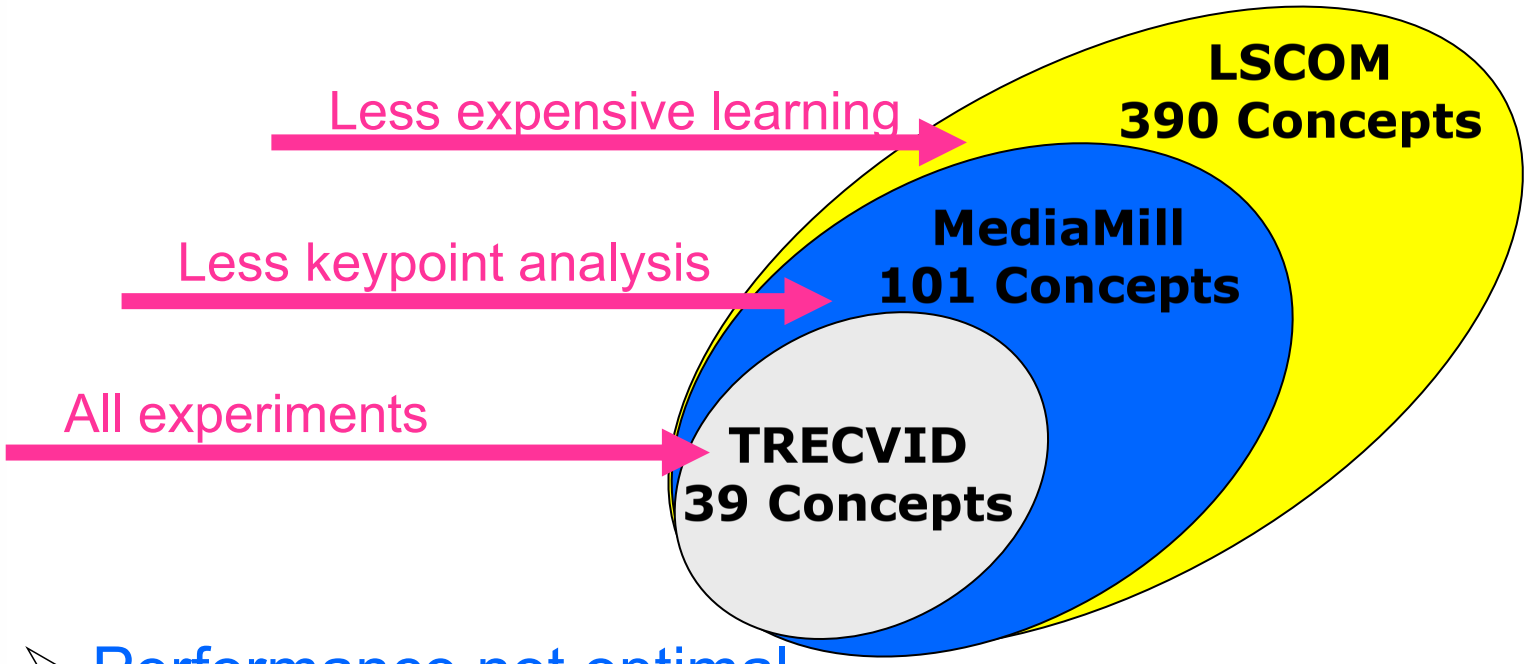
✓ Overall best for US flag and Charts



Scaling-up to 491 concept detectors

➢ Graceful degradation

✓ Not all experiments were available for all concepts



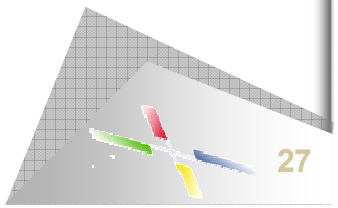
➢ Performance not optimal

✓ Detectors might be useful for semantic video retrieval

Lessons learned

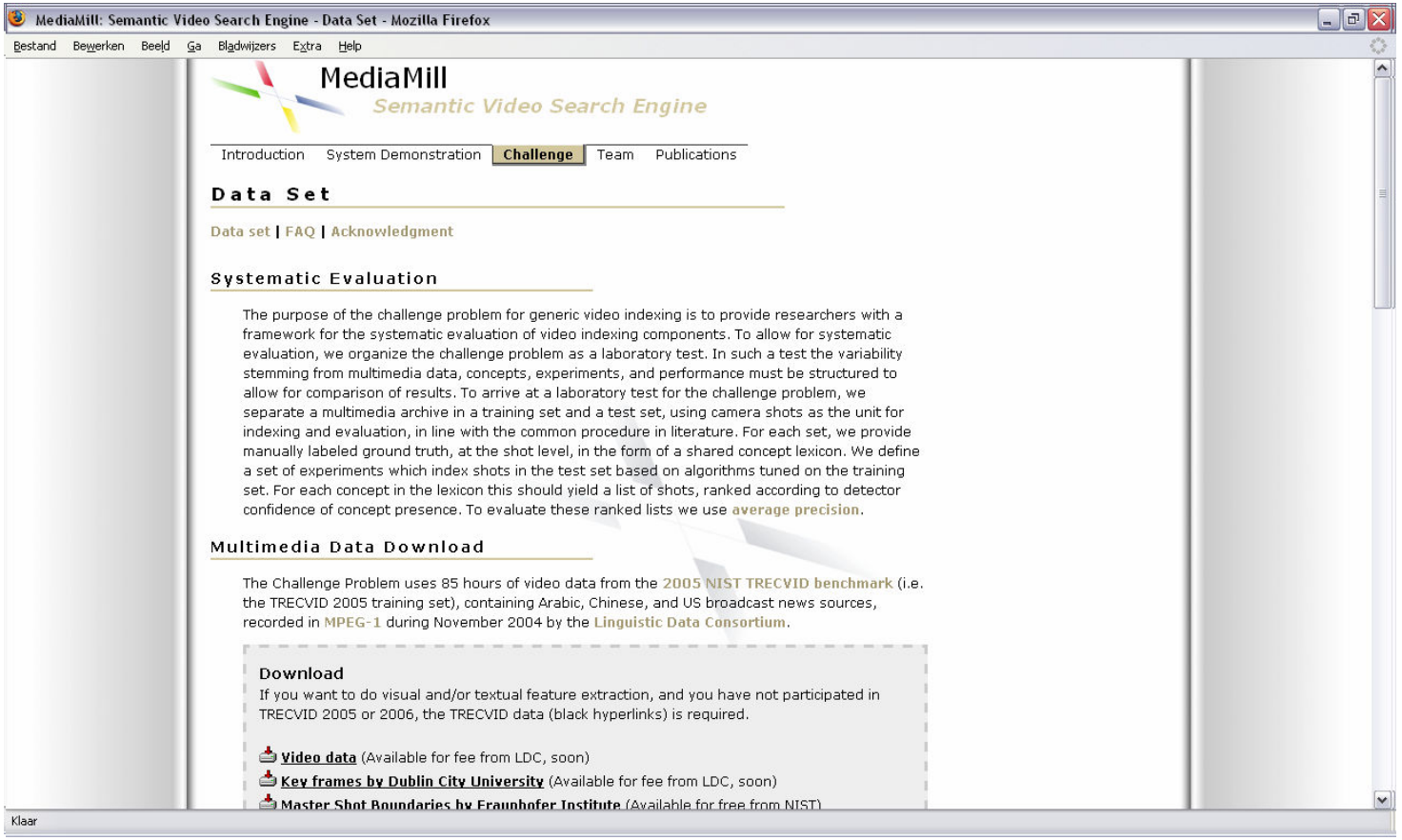
- The MediaMill Challenge allows to
  - ✓ Gain insight in intermediate video analysis steps
  - ✓ Foster repeatability of experiments
  - ✓ Optimize video analysis systems on a component level
  - ✓ Compare and improve upon baseline

- For visual-only analysis we learned that
  - ✓ A combination of various techniques pays off
  - ✓ Regional image features seem most effective
  - ✓ Data-driven clustering is more effective than fixed codebook
  - ✓ Keypoint methods unstable for images with few interest points
  - ✓ High-dimensional feature vectors can be handled effectively by relatively simple classifiers like Fishers linear discriminant
  - ✓ Fusion using geometric mean is cheap and effective
  - ✓ Scaling-up to 1,000+ detectors is a matter of annotated examples



- Introduction
- Challenge
- Visual-only
- Results
- Lessons

Download location:
www.mediamill.nl/challenge/



Check it out and beat the baseline!