IBM Research & Columbia University Multimedia Event Detection System

Speaker: Paul Natsev <natsev@us.ibm.com>
IBM T. J. Watson Research Center

On Behalf Of:

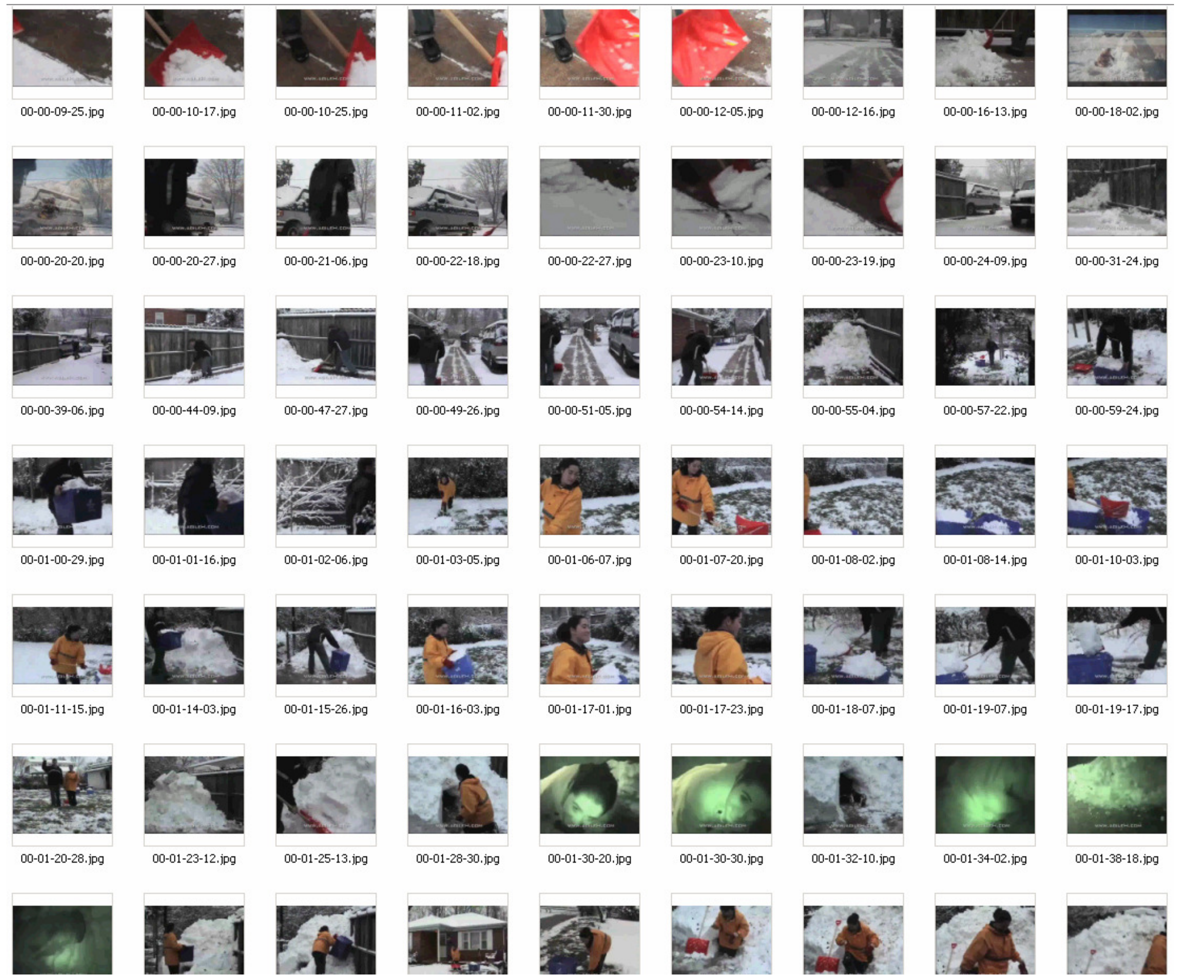
IBM Research: Matthew Hill, Gang Hua, John R. Smith, Lexing Xie
IBM Interns: Bert Huang, Michele Merler, Hua Ouyang, Mingyuan Zhou
Columbia Univ.: Shih-Fu Chang, Dan Ellis, Yu-Gang Jiang

TRECVID-2010 Workshop
Gaithersburg, MD
Nov. 15-17, 2010

Multimedia Event Detection (MED) Task Overview

- **Judge Y/N for each target event given a YouTube-style video**
- **Challenging dataset**
  - 1700+ diverse videos
  - A few shots vs long and varied
  - Only 50 examples/event

Category	#Videos	#Keyframes
Assembling shelter	48	2,123
Making cake	48	3,119
Batting in run	50	347
Random	1,577	49,247



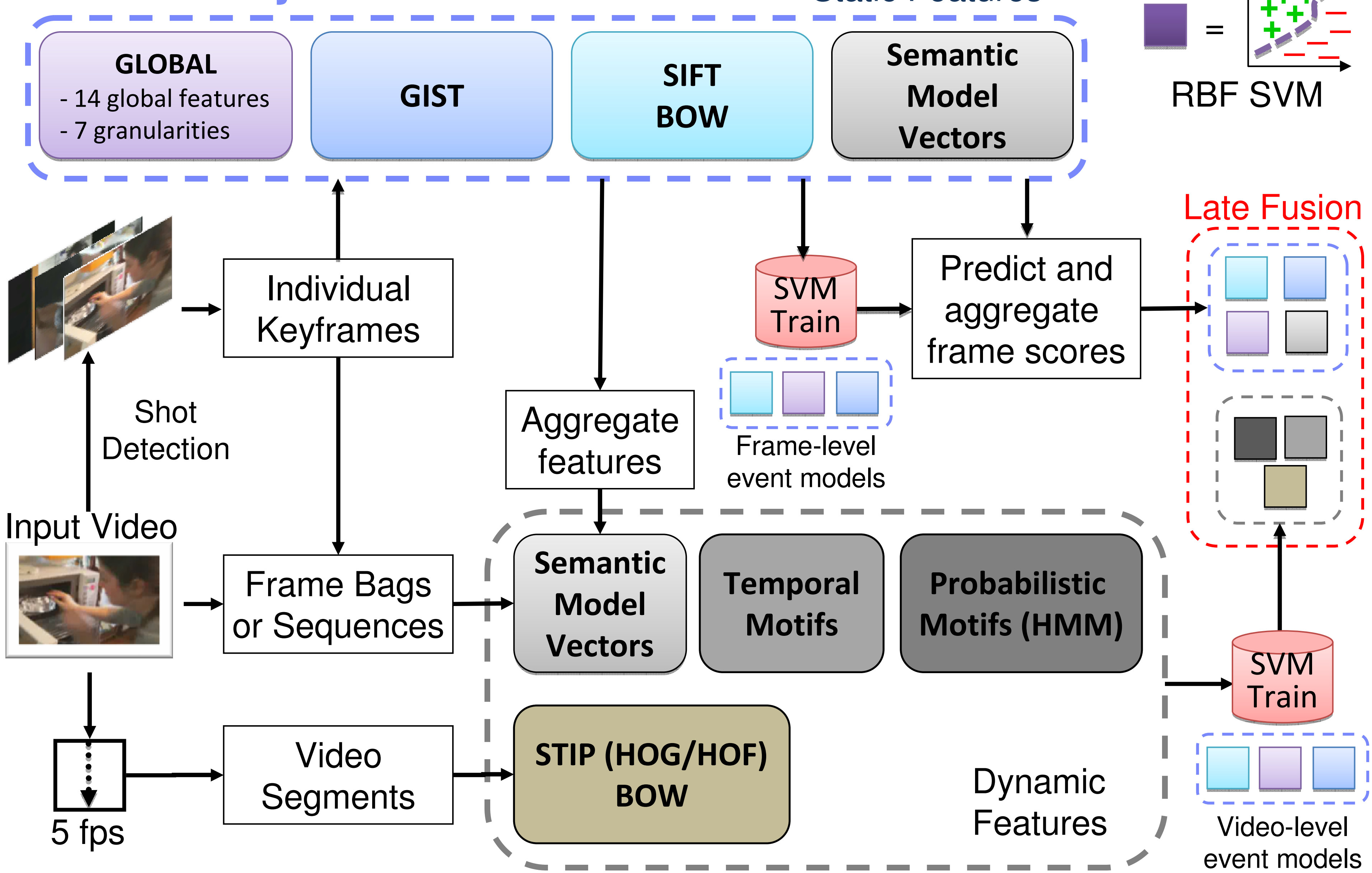
Key Questions

- **Do cross-domain concept classifiers help for complex event detection?**
- **Answer: YES! Our best performing feature…**

- **How do static features/models compare to dynamic ones?**
- **Answer: Surprisingly similarly…**

- **Can we move beyond bag-of-X representations to sequence-of-X?**
- **Answer: Exploratory temporal motif features show promise, 2nd best feature…**

IBM MED System Architecture



Static and Dynamic Features

- **Static Features:**
  - Break down video into keyframes
  - Extract 98 global image features
  - GIST features
  - Dense SIFT descriptors (BOW, 1K codebook)
  - Semantic model vectors (272 semantic concept classifiers)
- **Dynamic features**
  - Transcode videos to 5 frames per second
  - Extract Space-Time Interest Points [Laptev et al.]
  - Build dynamic visual words from HOG and HOF descriptors (BOW, 1K codebook, 1x2 temporal pyramid)
  - Temporal motifs (co-occurring sequences or bags of features)
  - Probabilistic motifs (Hierarchical HMM-based)

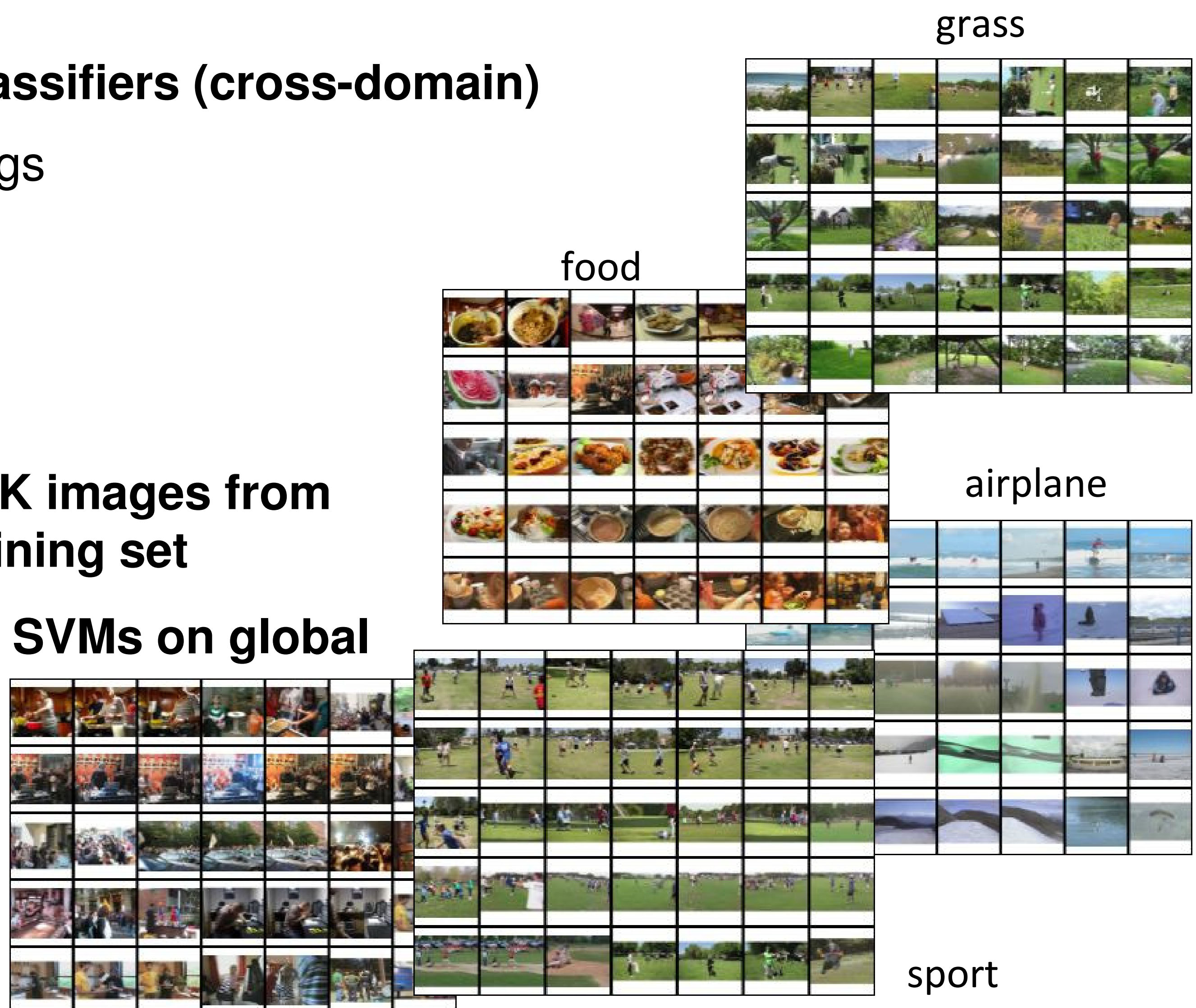
Breakdown of features and event modeling approaches

Event Models \ Features	Static features	Dynamic features
Frame-level models	98 Global features GIST SIFT BoW Semantic Model Vector	–
Video-level models	Semantic Model Vector SIFT BoW (Columbia*)	STIP HOF + Temporal Pyramid Temporal motifs Probabilistic motifs (HMM-based) STIP HOG + HOF (Columbia*) Audio BoW (Columbia*)

* For details on Columbia features/runs, see Columbia notebook paper and presentation

Single Best Performing Feature – Semantic Model Vector

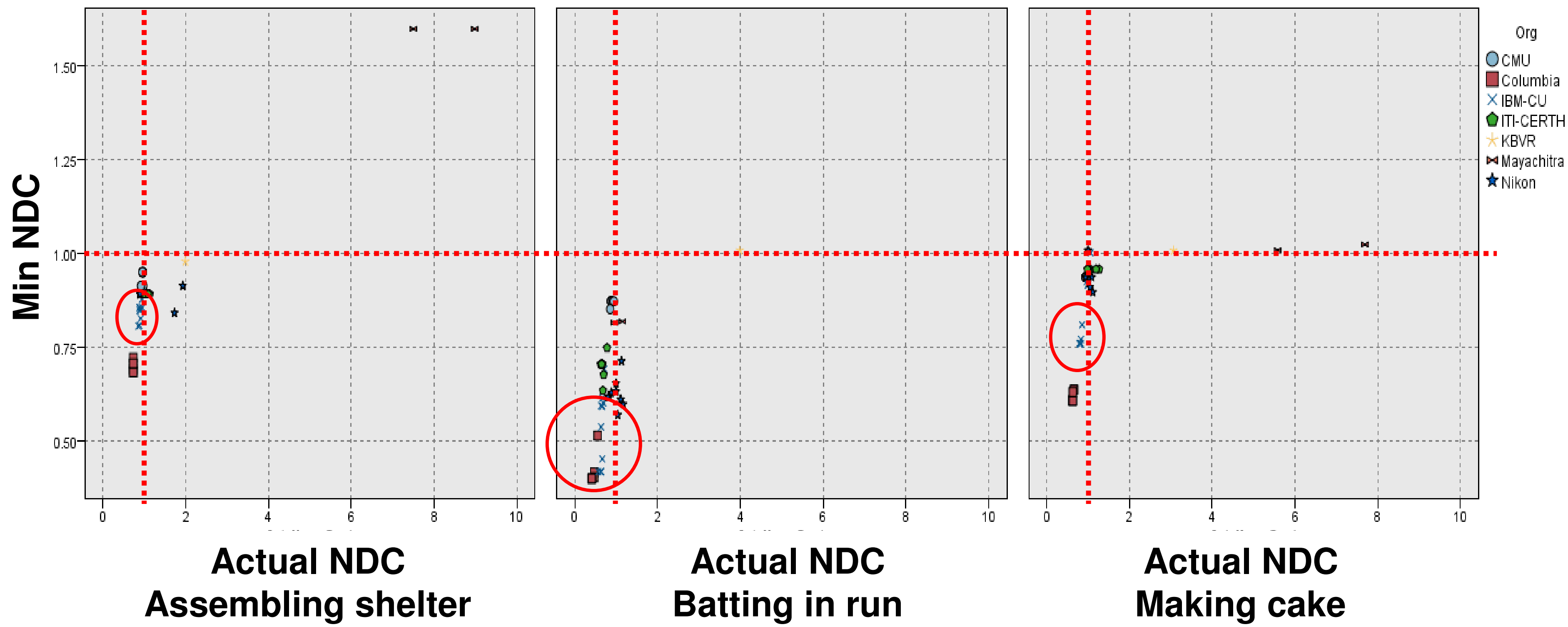
- **272 semantic classifiers (cross-domain)**
  - Scenes/settings
  - Objects
  - People
  - Activities
- **Trained on ~600K images from independent training set**
- **Using ensemble SVMs on global image features**



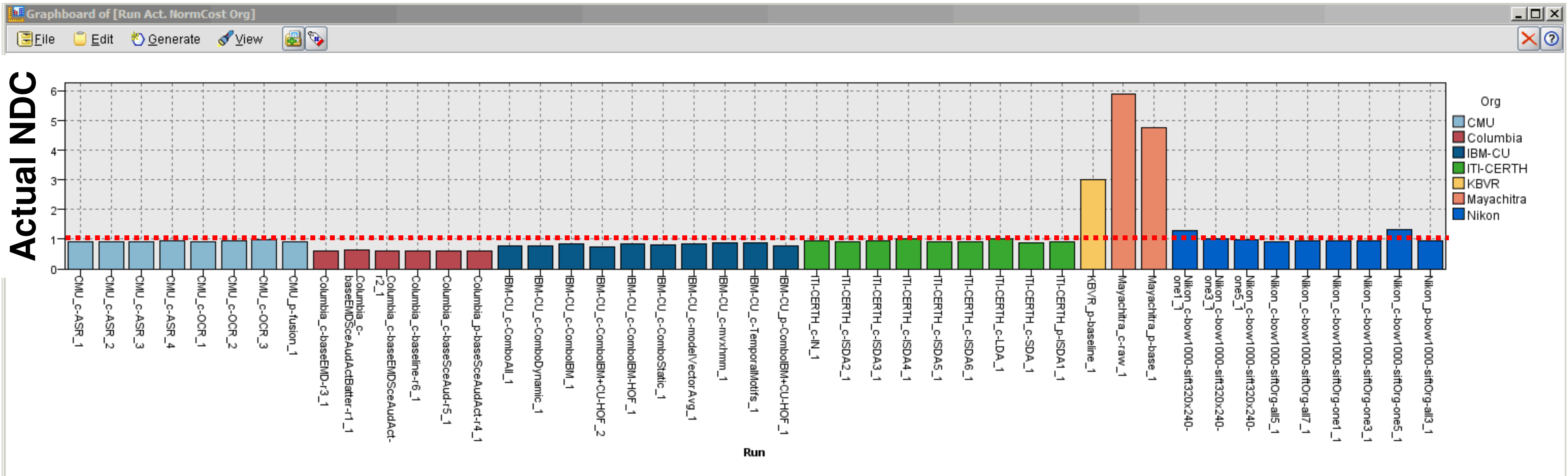
Other Notable Features

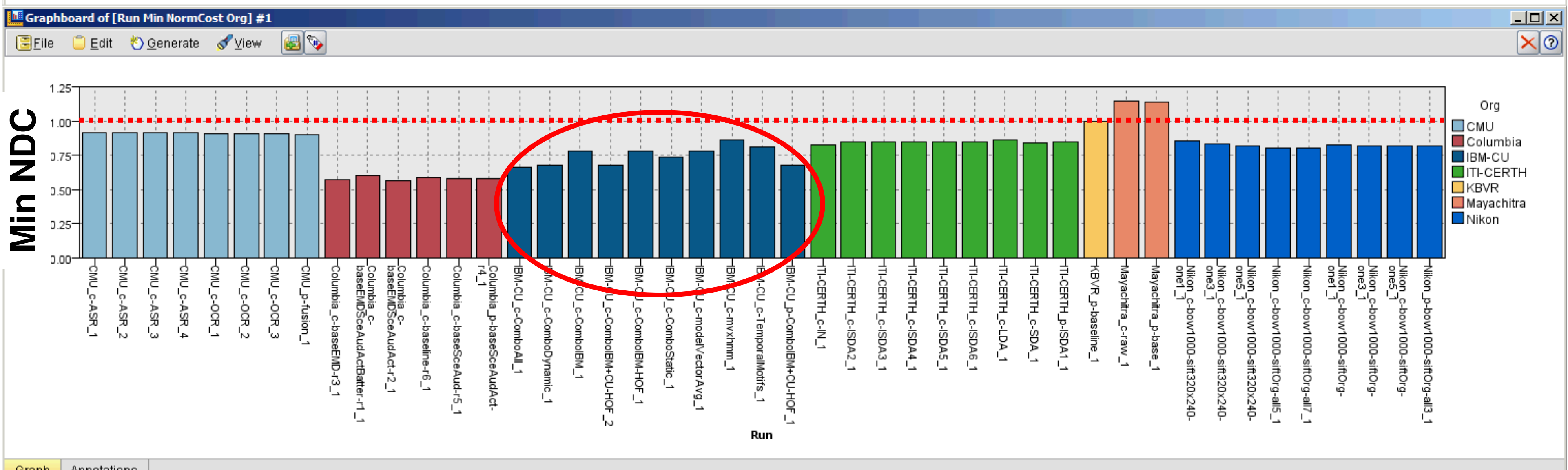
- **Bag-of-visual words**
  - IBM: dense SIFT, 1000-D visual word codebook, soft assignment
  - Columbia: SIFT with DoG and Hessian detectors, 500-d codebooks, spatial pyramid (frame + 4 quadrants), 5000-D total feature length
- **Bag-of-audio-words**
  - Columbia: MFCCs for every 32ms, 4000-d audio word codebook
- **Spatio-Temporal Interest Points (STIP) [Laptev et al.]**
  - Histogram of Gradients (HOG) and Histogram of Flow (HOF)
  - IBM: 1000-D codebook + temporal pyramid, HOF only
  - Columbia: 4000-D codebook, concatenated HOG+HOF
- **Temporal motifs**
  - Mine sequential frequent item-sets from training data
  - Use the presence/absence of item-sets as features
- **Probabilistic motifs**
  - Learn a group of HMMs on feature partitions
  - Use the state histogram of HMMs as features

Results – Normalized Detection Cost (NDC) Per Event



Results – Aggregated NDC Over All Events

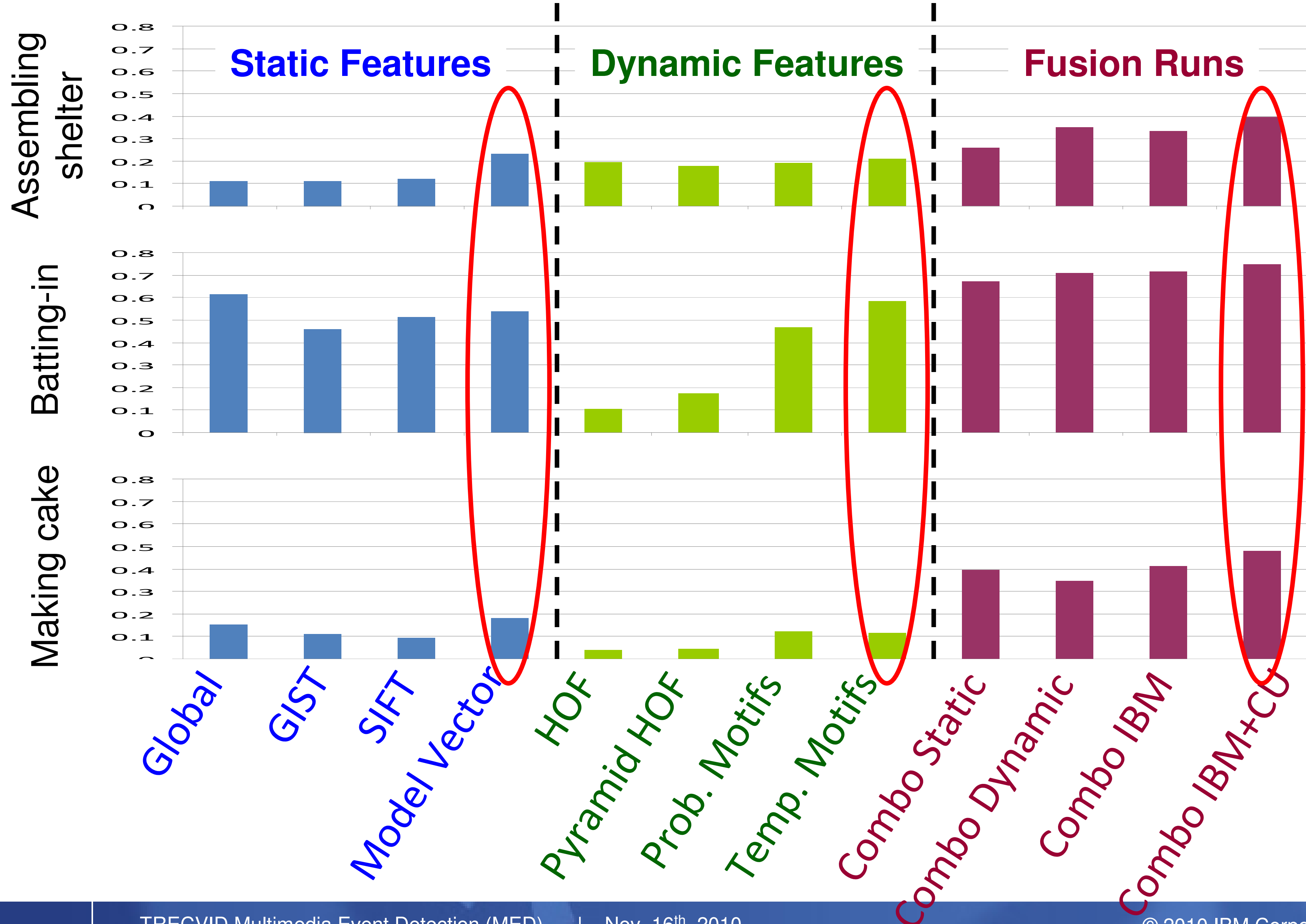




Results – Mean Average Precision Over All Events

Run	Mean AP (submitted)	Mean AP (*with bug fix)
Global	0.10	0.29*
GIST	0.08	0.23*
SIFT BoW	0.08	0.24*
Semantic Model Vector	0.32	0.32
Combo Static Features	0.39	0.44*
HoF	0.11	0.11
HoF Temporal Pyramid	0.13	0.13
Temporal Motifs	0.30	0.30
Probabilistic Motifs	0.26	0.26
Combo Dynamic Features	0.47	0.47
Combo IBM Runs	0.34	0.49*
Columbia Audio BoW	0.37	0.37
Columbia STIP BoW	0.45	0.45
Columbia SIFT BoW	0.47	0.47
Combo IBM + CU Runs	0.49	0.54*

Per-Event Performance Breakdown of Constituent Runs



Per-Event Observations

- **Assembling shelter & making cake events**
  - Not clear they are very temporal in nature
  - Static features perform on par with, or better than, dynamic features
  - Semantic model vectors outperform everything else
  - Fusion runs dramatically improve upon all constituent runs (over 2x better)

- **Batting-in event**
  - Most homogeneous event, highest performance of the 3 events
  - Sequence features (motifs) outperform other dynamic features
  - Fusion runs modestly improve upon all constituent runs (over 25% better)

- **Fusion with Columbia runs brings an extra 10% improvement → 0.54 MAP**

Summary

- **Semantic Model Vector is our single best-performing feature**
  - The cross-domain semantic concept classifiers are **very** useful
- **New temporal motif representation (sequence-of-X) shows promise**
  - Our second-best feature overall
- **Dynamic and static features perform comparably, surprisingly…**
  - Not all complex events are truly dynamic in nature
  - Still, fusion of dynamic and static features performs best (2x gains)
- **Columbia features/runs bring in complementary info (e.g., audio)**
  - Lead to overall MAP of 0.54 with only 50 training examples per event

- **Comments for the task**
  - If no localization required, AP and NDC give similar rankings
  - So can we use the simpler AP metric? How is cost profile motivated?

Acknowledgments: The Team (in alphabetical order)

- **IBM Research**
  - Matthew Hill
  - Gang Hua
  - Paul Natsev
  - John R. Smith
  - Lexing Xie

- **Summer Interns @ IBM**
  - Bert Huang, Columbia U.
  - Michele Merler, Columbia U.
  - Hua Ouyang, Georgia Tech
  - Mingyuan Zhou, Duke U.

- **Columbia University**
  - Shih-Fu Chang, Dan Ellis, Yu-Gang Jiang