COLUMBIA UNIVERSITY

Joint Audio-Visual Signatures for Web Video Analysis

Dan Ellis, Shih-Fu Chang

Yu-Gang Jiang, Xiaohong Zeng,
Guangnan Ye, Courtenay Cotton

Department of EE, Columbia University, NY

NGA NARP Symposium, Aug 31, 2011

What are Consumer (Web) Videos?

- Original unedited videos made from consumers
  - Interesting and very diverse contents
  - Very weakly indexed: 3 tags per consumer video vs. 9 tags avg
  - Original audio tracks - good for audio-visual joint analysis

YouTube vimeo flickr ...

- Challenge: Content-based retrieval
  - Find items similar to example(s)

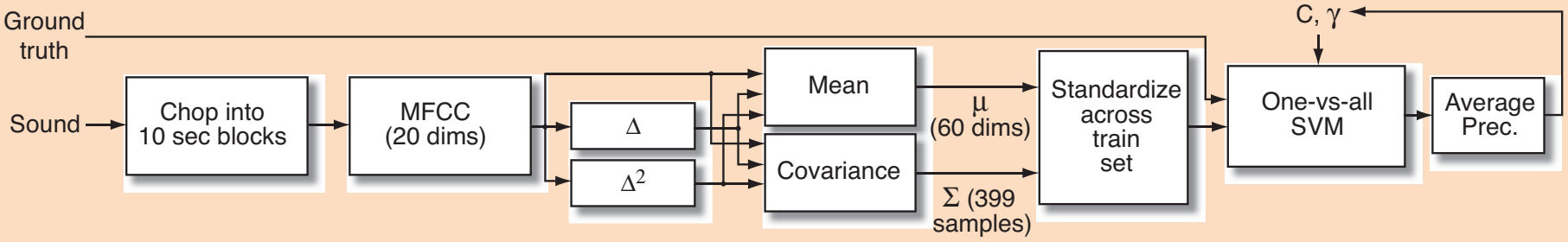
→

Highlights 2010-2011

- Novel audio features for events (transients) and environments (textures)
- Release of Columbia Consumer Video dataset annotated via Amazon Mechanical Turk
- Best result in TRECVID 2010 Multimedia Event Detection evaluation

Event + Environment Soundtrack Features

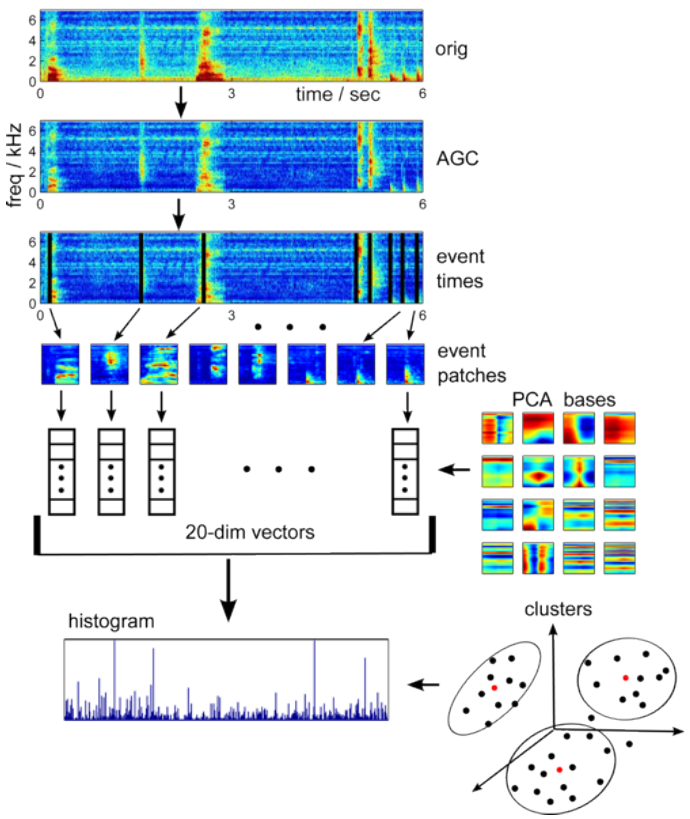
- Conventional Bag-of-MFCC features:

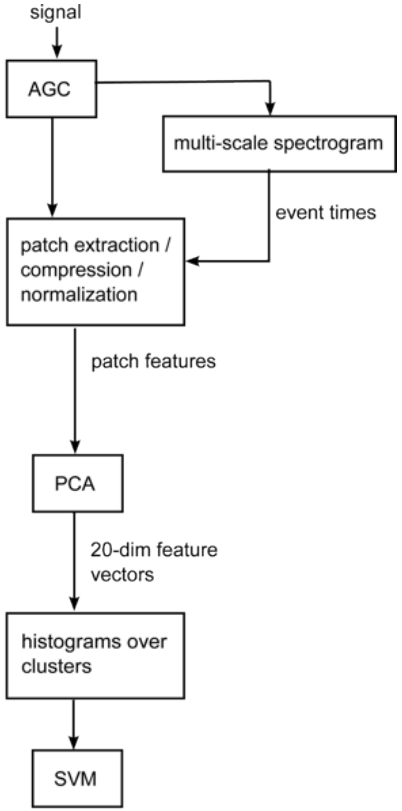


  - everything mixed in together

- Can we differentiate foreground and background?

Foreground: Transient Features

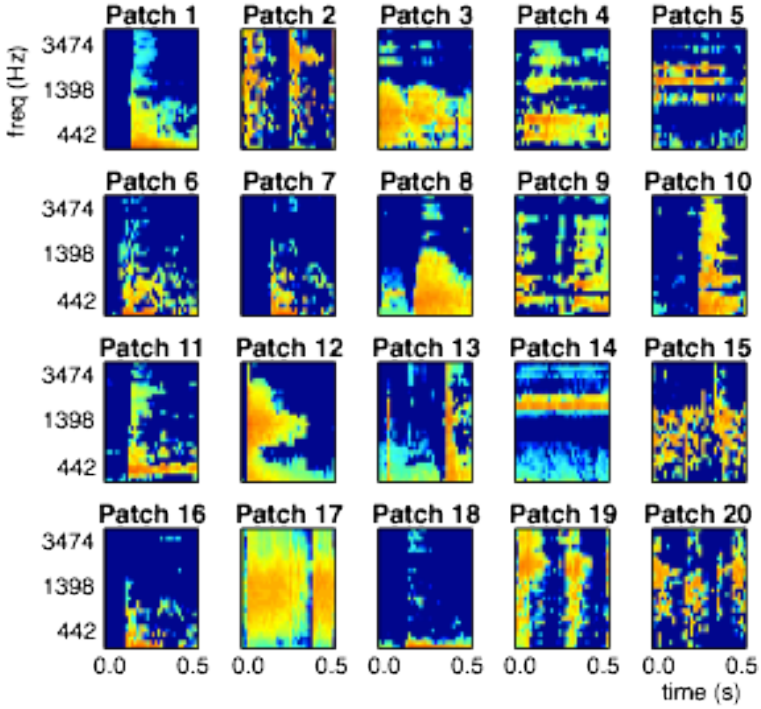


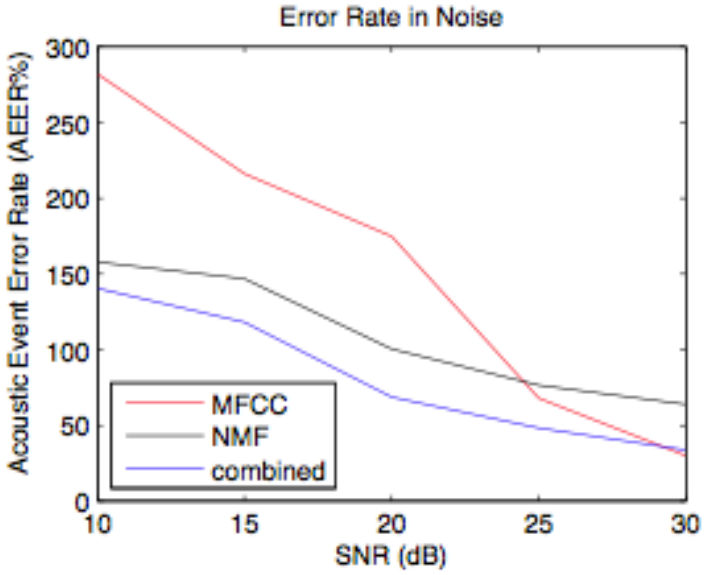


- Transients = foreground events?
- Onset detector finds energy bursts
  - best SNR
- Represent with PCA basis
  - 300 ms x aud freq
- "bag of transients"

NMF Transient Features

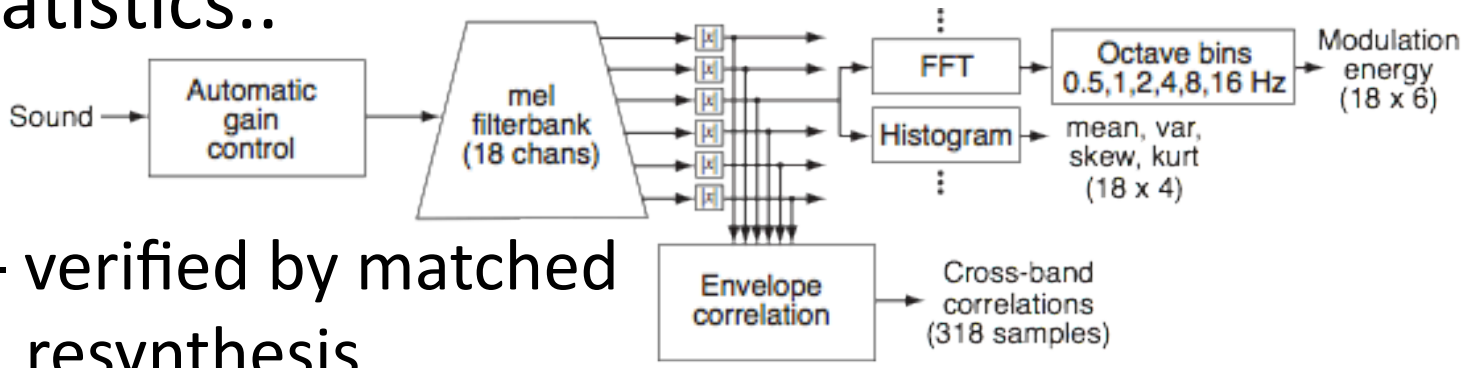
- Learn 20 patches by Nonnegative Matrix Factorization
- Compare to MFCC-HMM



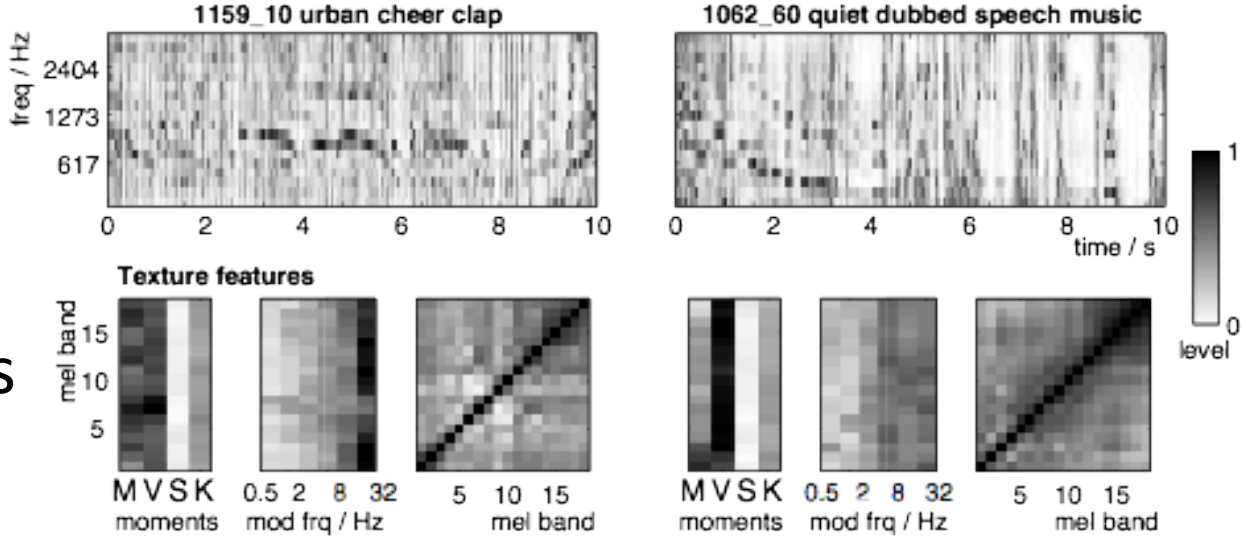


- NMF more noise-robust
  - combines well

Background: Texture features

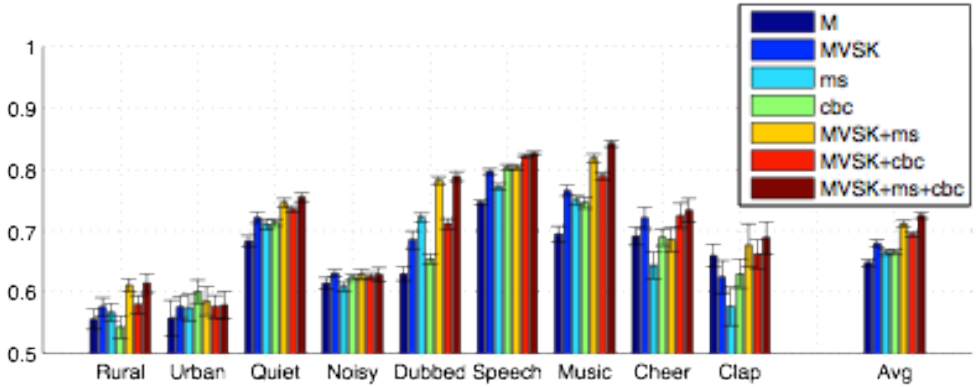
- Characterize sounds by perceptually-sufficient statistics..
-


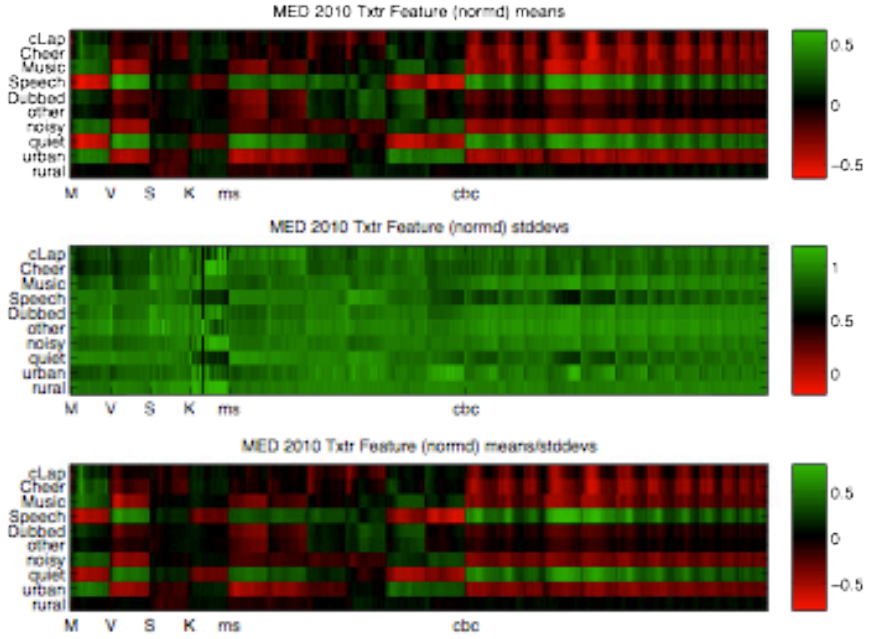
  - verified by matched resynthesis
- Subband distributions & env x-corrs
  - Mahalanobis distance ...



Texture Feature Results

- Test on MED 2010 development data
  - 10 labels



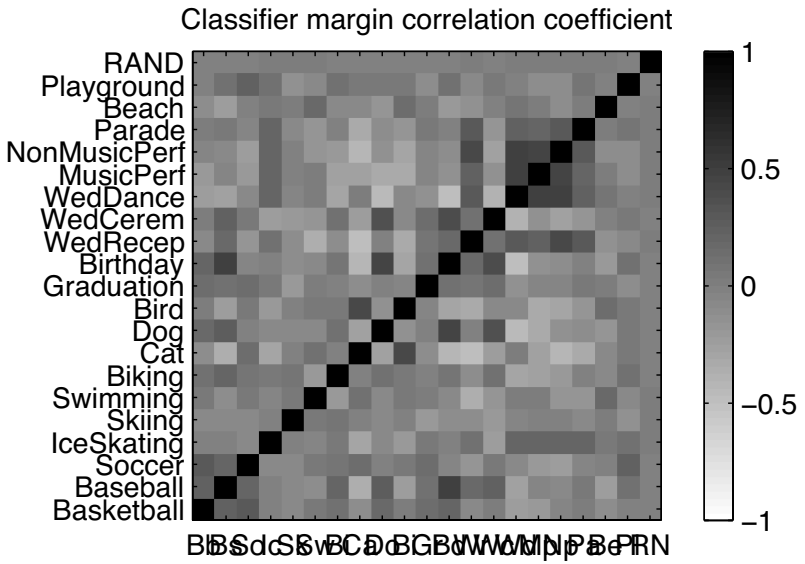


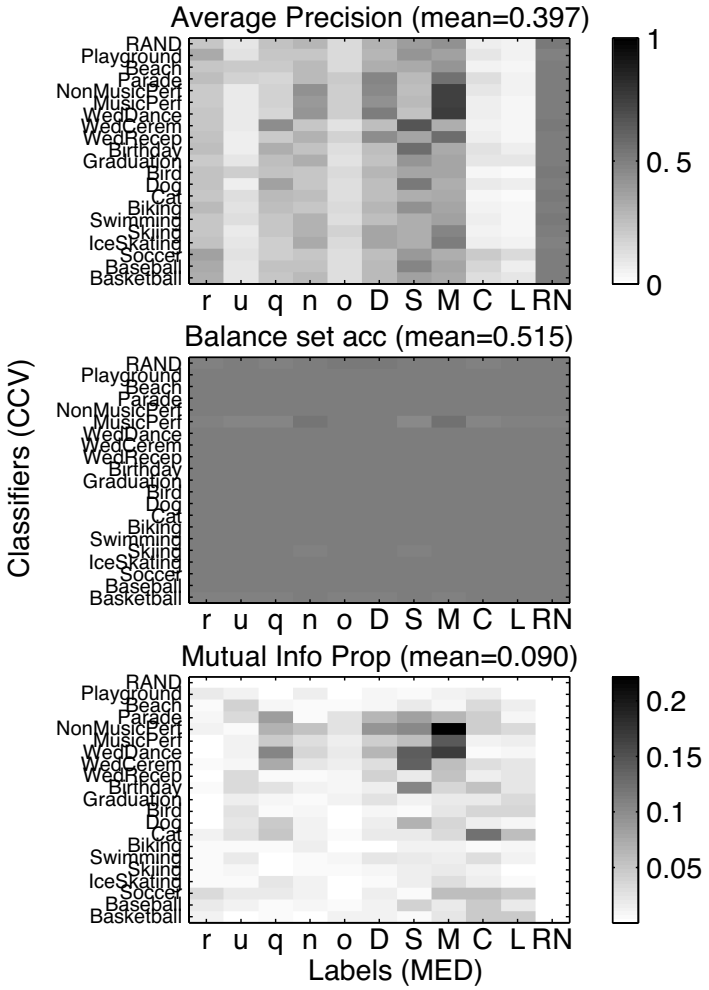
- Contrasts in feature sets
  - correlation of labels
- Perform ~ same as MFCCs
  - combine well

Audio Classifier Evaluation

- Investigating beyond mAP…
  - Accuracy, Mutual Information Proportion, Correlation

$$MIP = \frac{I(\text{classifier}; \text{label})}{H(\text{label})}$$

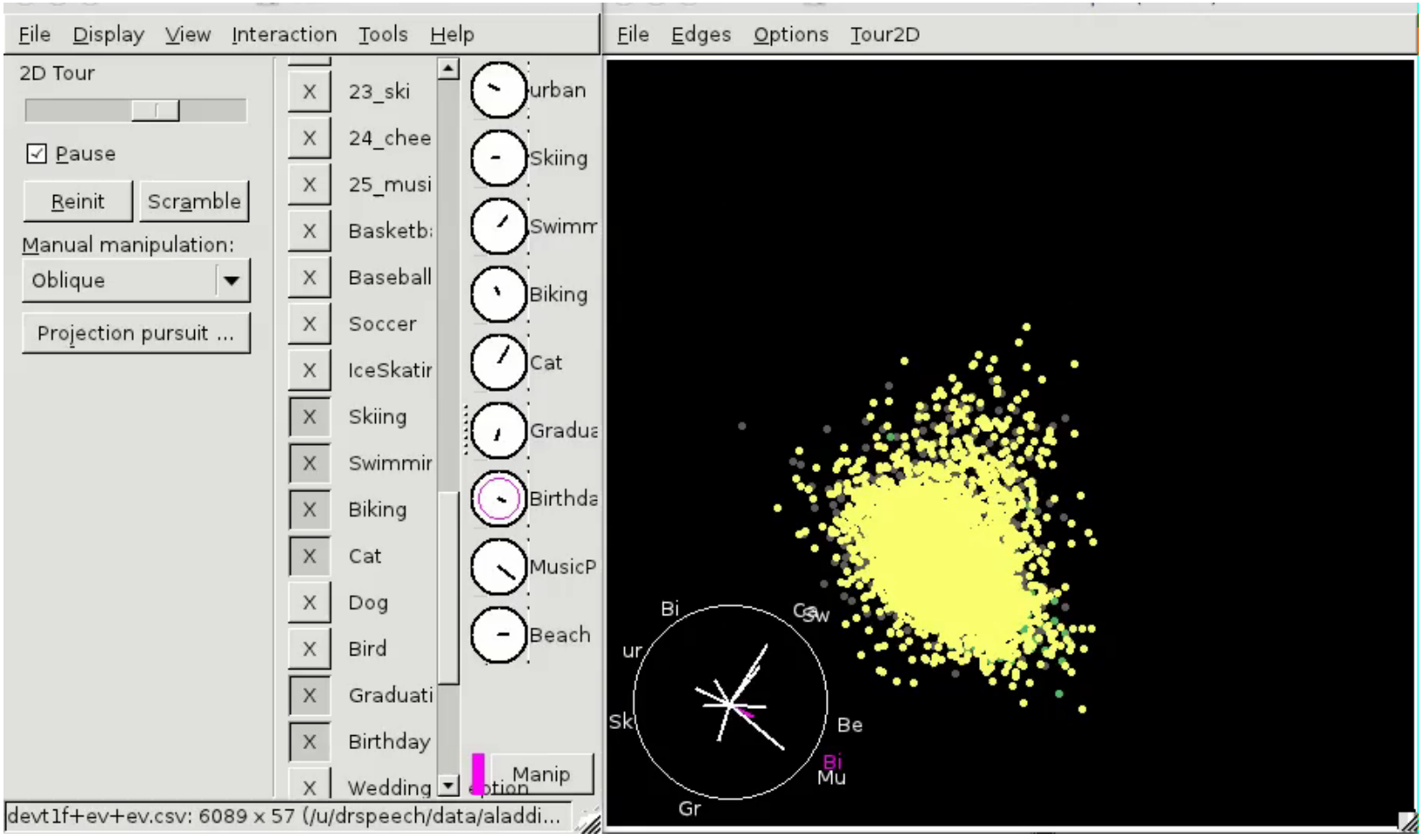




  - Cross-corpus evaluations

Audio Classifier Results Browsing

- Customized version of GGobi links to Movie Player
  - Rapid investigation of high-dimensional data sets
  - Each point is a video, colored by label (e.g. Event)

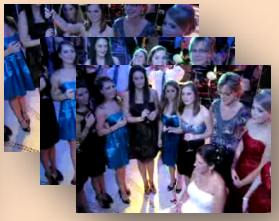


Columbia Consumer Video (CCV) Database


Basketball


Skiing


Dog


Wedding Reception

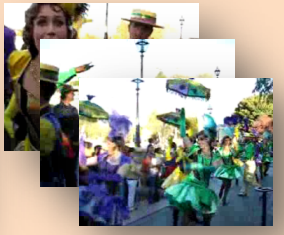

Non-music Performance


Baseball

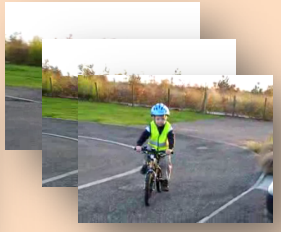

Swimming

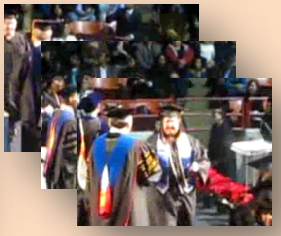

Bird

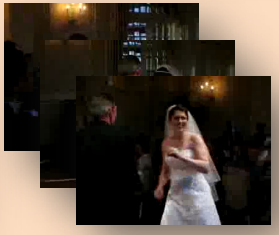

Wedding Ceremony

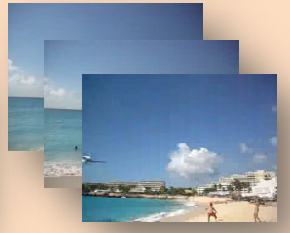

Parade


Soccer

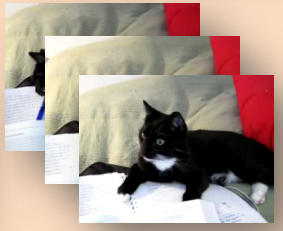

Biking


Graduation

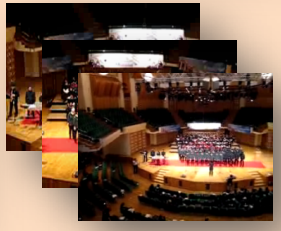

Wedding Dance


Beach


Ice Skating


Cat

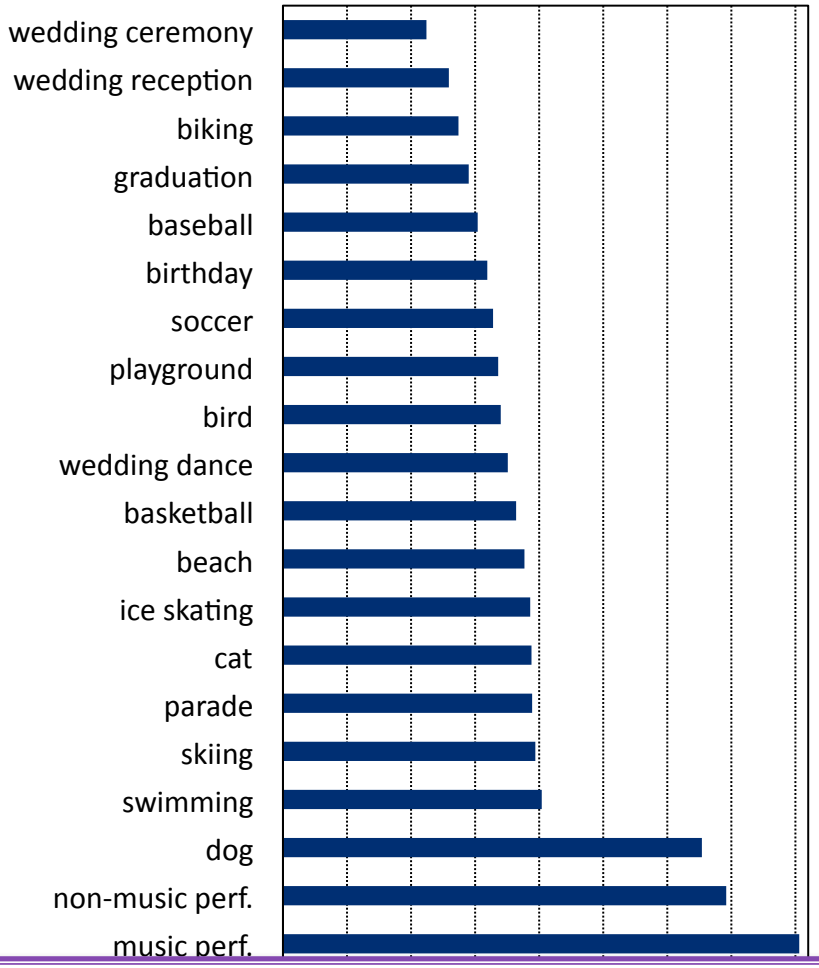

Birthday Celebration


Music Performance


Playground

CCV Snapshot

- # videos: 9,317
  - (210 hrs in total)
- video genre
  - unedited consumer videos
- video source
  - YouTube.com
- average length
  - 80 seconds
- # defined categories
  - 20
- annotation method
  - Amazon Mechanical Turk



The trick of digging out consumer videos from YouTube:
Use default filename prefix of many digital cameras: "**MVI** and parade".

Existing Database?

- Human Action Recognition
  - KTH & Weizmann
    - (constrained environment) 2004-05
  - Hollywood Database
    - (12 categories, movies) 2008
  - UCF Database
    - (50 categories, YouTube Videos) 2010
- Kodak Consumer Video
  - (25 classes, 1300+ videos) 2007
- LabelMe Video
  - (many classes, 1300+ videos) 2009
- TRECVID MED 2010
  - (3 classes, 3400+ videos) 2010

CCV Database

Unconstrained YouTube videos

Higher-level complex events

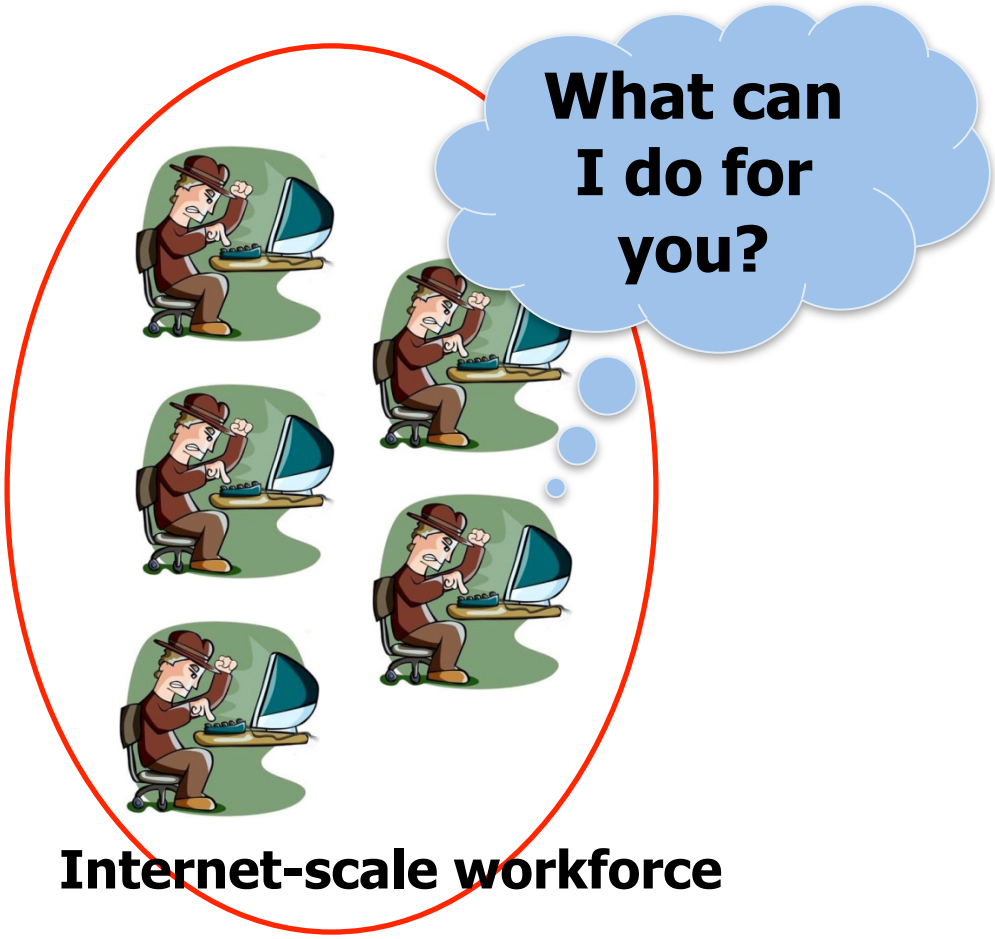
More videos & better defined categories

More videos & larger content variations

More videos & categories

Crowdsourcing: Amazon Mechanical Turk

- A web services API that allows developers to easily integrate human intelligence directly into their processing



Internet-scale workforce

$?.??

financial rewards

MTurk: Annotation Interface

Mark all the categories that appear in any part of the video.

Instructions:

- Watch the entire video as more categories may appear over time.
- Mark all the categories that appear in any part of the video.
- Make sure audio is on.
- If no matching category is found, mark the box in front of "None of the categories matches".
- For categories that appears to be relevant but you're not completely sure, please still mark it.
- Please mouse-over or click on the category names to read detailed definitions.

Replay Continue Playing

Original URL: http://www.youtube.com/watch?v=-0n50a7seNI

Sports	Animal	Celebration	Others
☐ Basketball	☐ Cat	☐ Graduation	☐ Music Performance
☐ Baseball	☐ Dog	☑ Birthday	☐ Non-music Performance
☐ Soccer	☐ Bird	☐ Wedding Reception	☐ Parade
☐ Ice Skating		☐ Wedding Ceremony	☐ Beach
☐ Skiing		☐ Wedding Dance	☐ Playground
☐ Swimming	☐ None of the categories matches.		
☐ Biking	☐ I don't see any video playing.		

Current Time: 10 sec

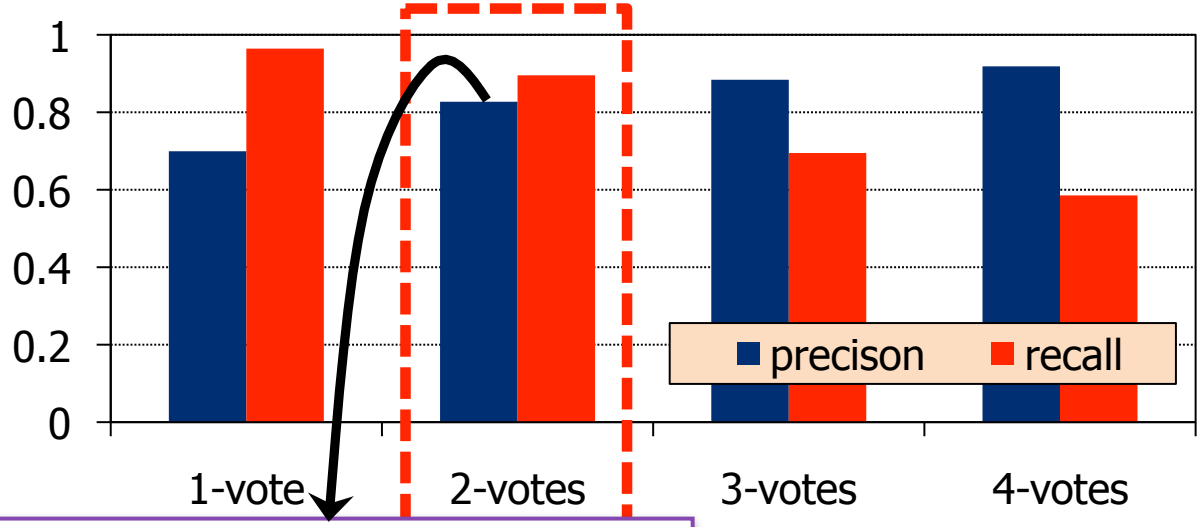
Submit

Reliability of Labels: each video was assigned to four MTurk workers

$ 0.02

Human Recognition Performance

- How to measure human (MTurk workers) recognition accuracy?
  - We manually and carefully labeled 896 videos
    - Golden ground truth!
- Consolidation of the 4 sets of labels



Plus additional manual filtering of 6 positive sample sets: 94% final precision

Download

- **Unique YouTube Video IDs,**
- **Labels,**
- **Training/Test Partition,**
- **Three Audio/Visual Features**

http://www.ee.columbia.edu/dvmm/CCV/

Fill out this ...

TRECVID MED 2010

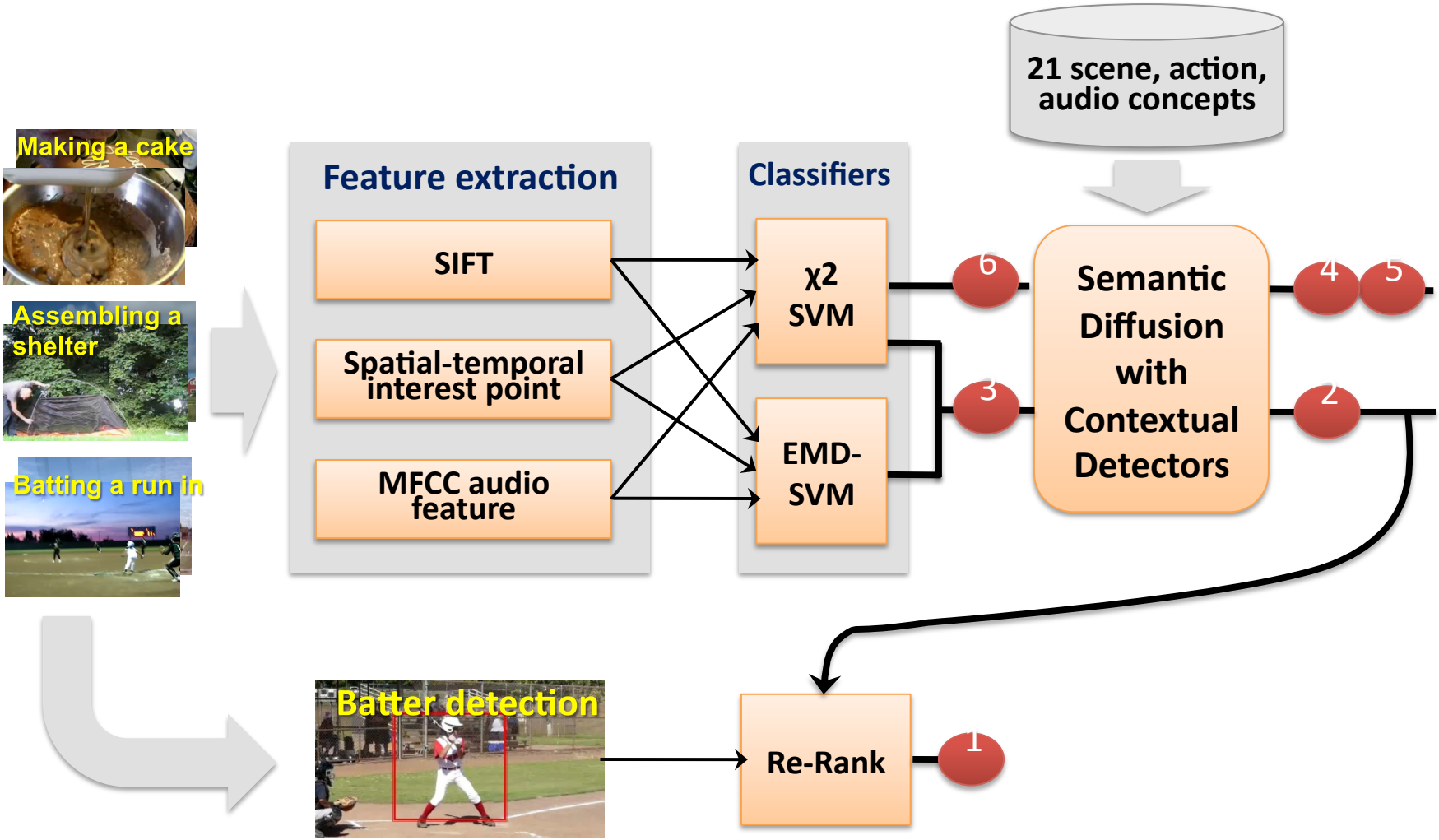
- Find “multimedia events” among 1700 videos
- 3 target event categories:

Making a cake

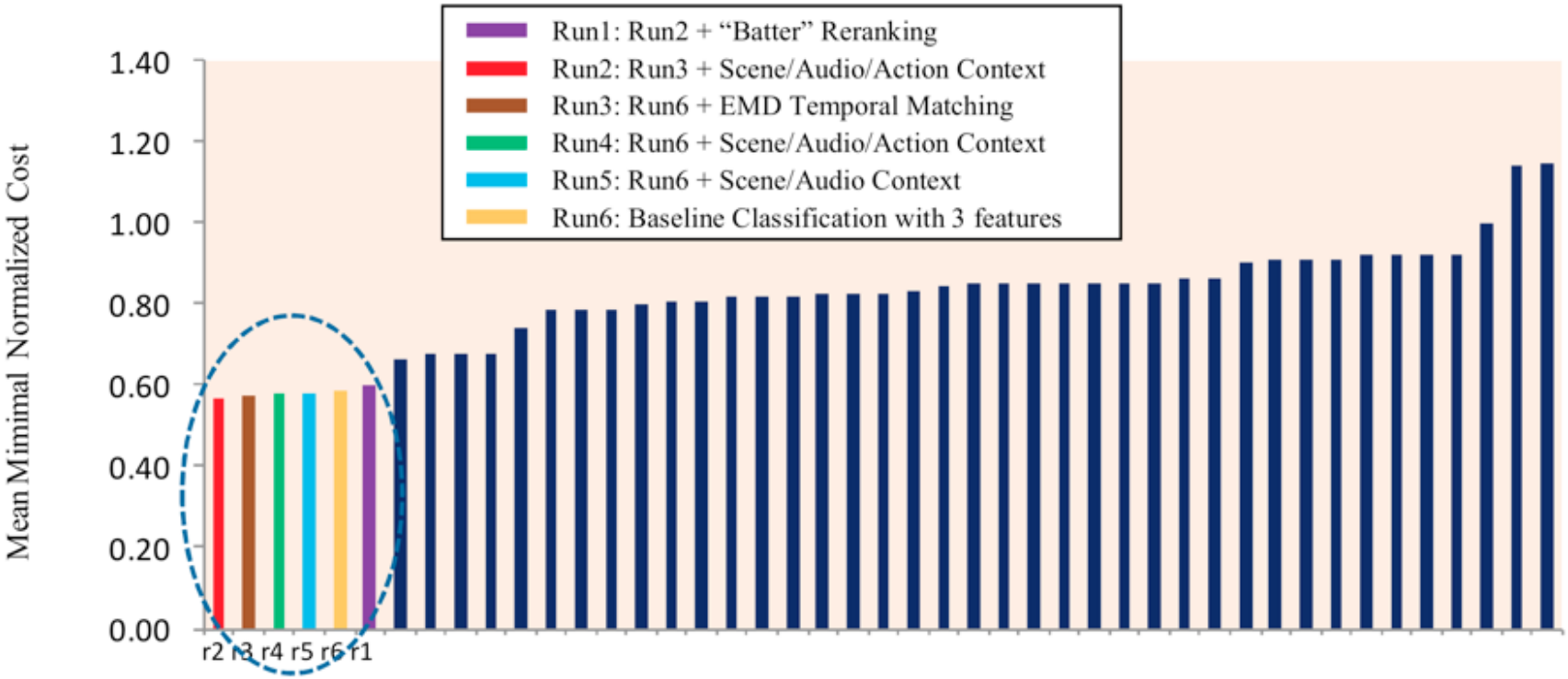
Assembling a shelter

Batting a run in

Overview: 4 major components & 6 runs

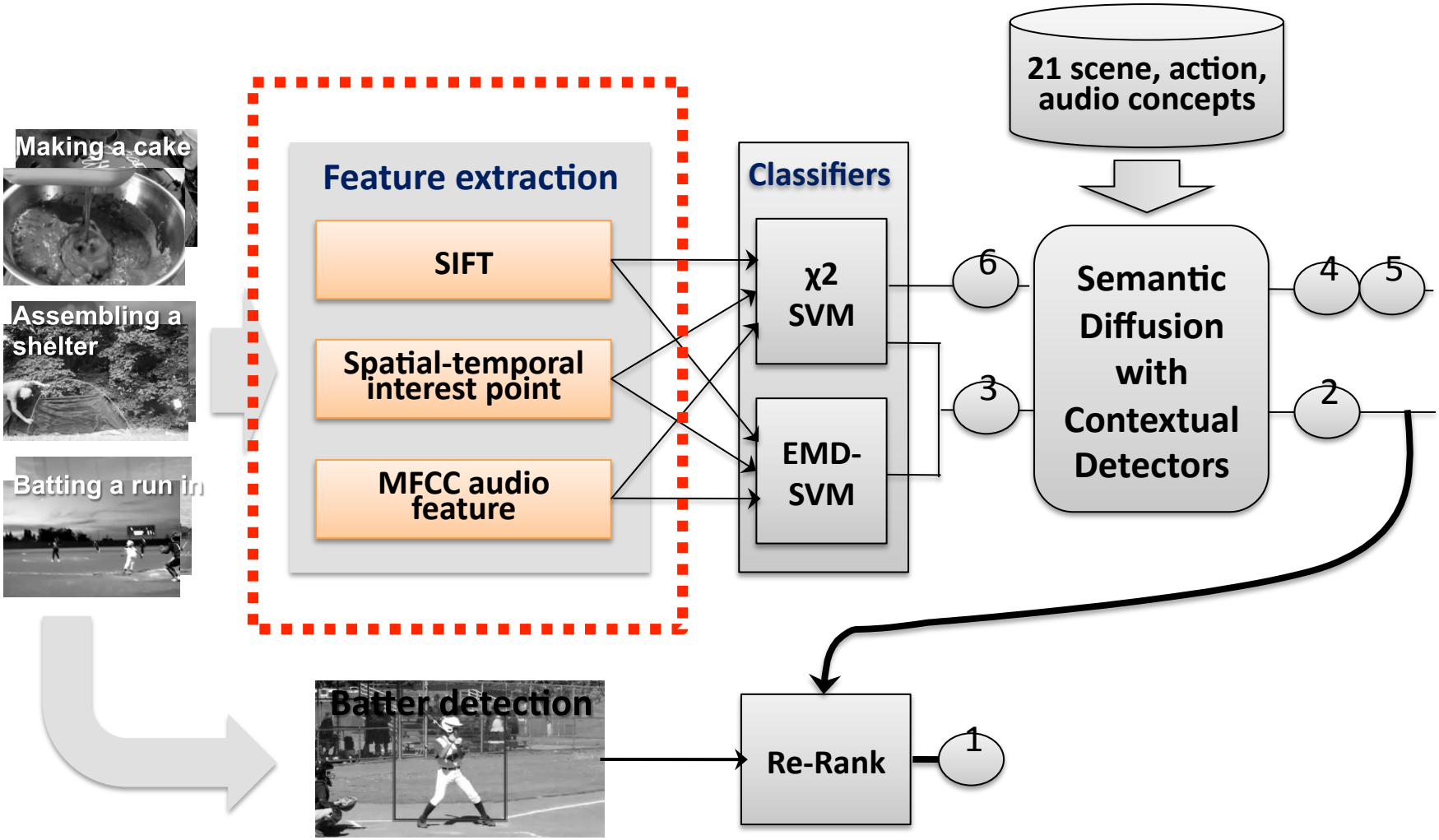


Overview: overall performance



- 45 systems by 8 teams from around the world
- Novel "normalized cost" metric
- Six Columbia systems scored best

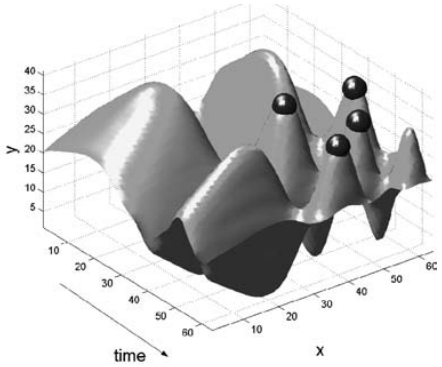
Roadmap > multiple modalities



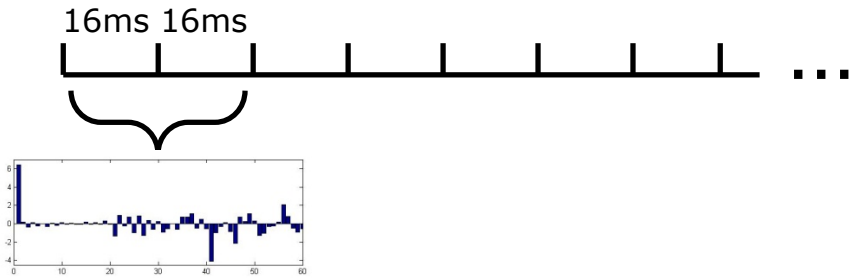
Three Feature Modalities…

- *SIFT (visual)*
  - *D. Lowe, IJCV 04.*

- *STIP (visual)*
  - *I. Laptev, IJCV 05.*

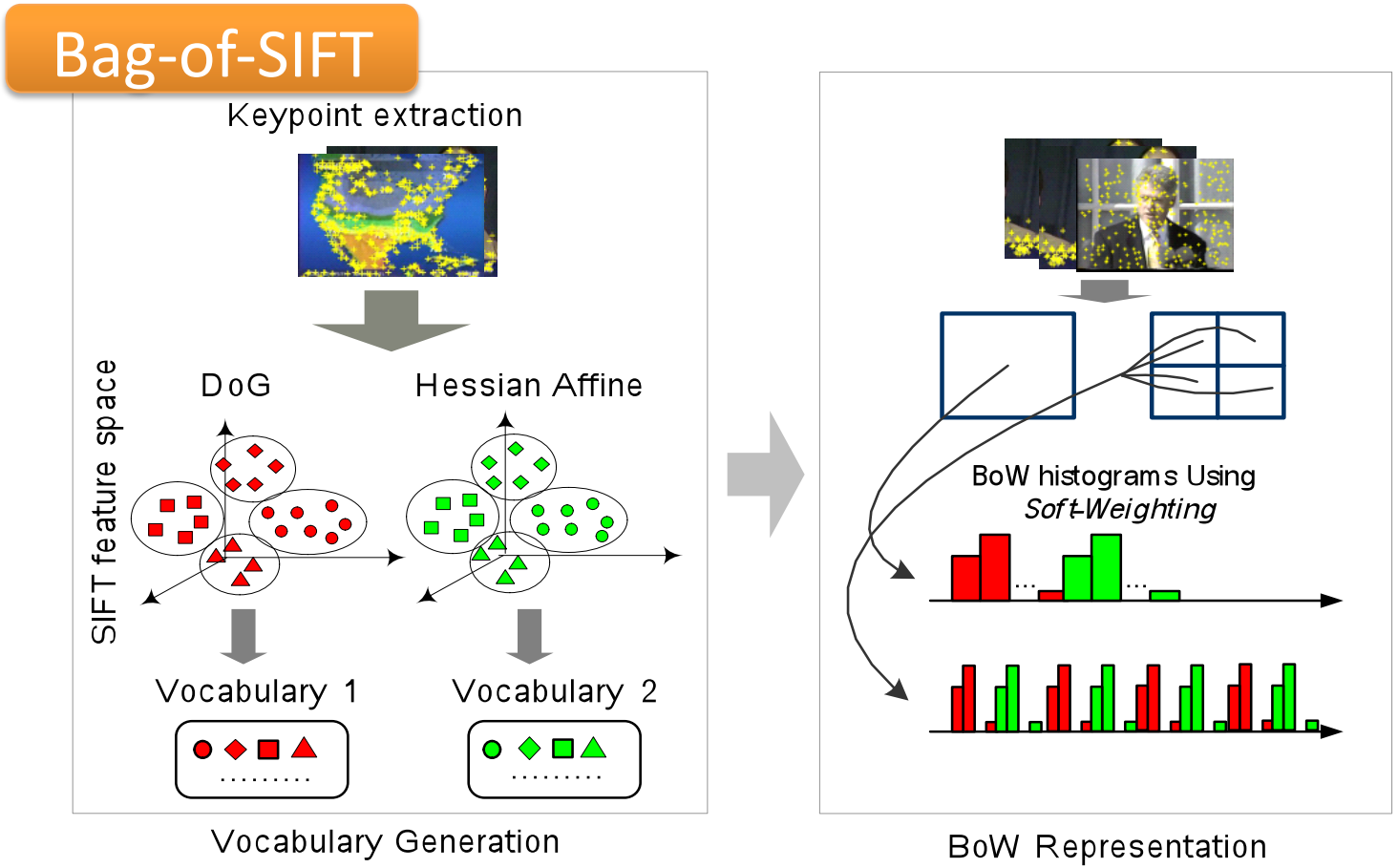


- *MFCC (audio)*



Bag-of-X Representation

- **X = SIFT or STIP or MFCC**
- **Soft weighting** (Jiang, Ngo and Yang, ACM CIVR 2007)



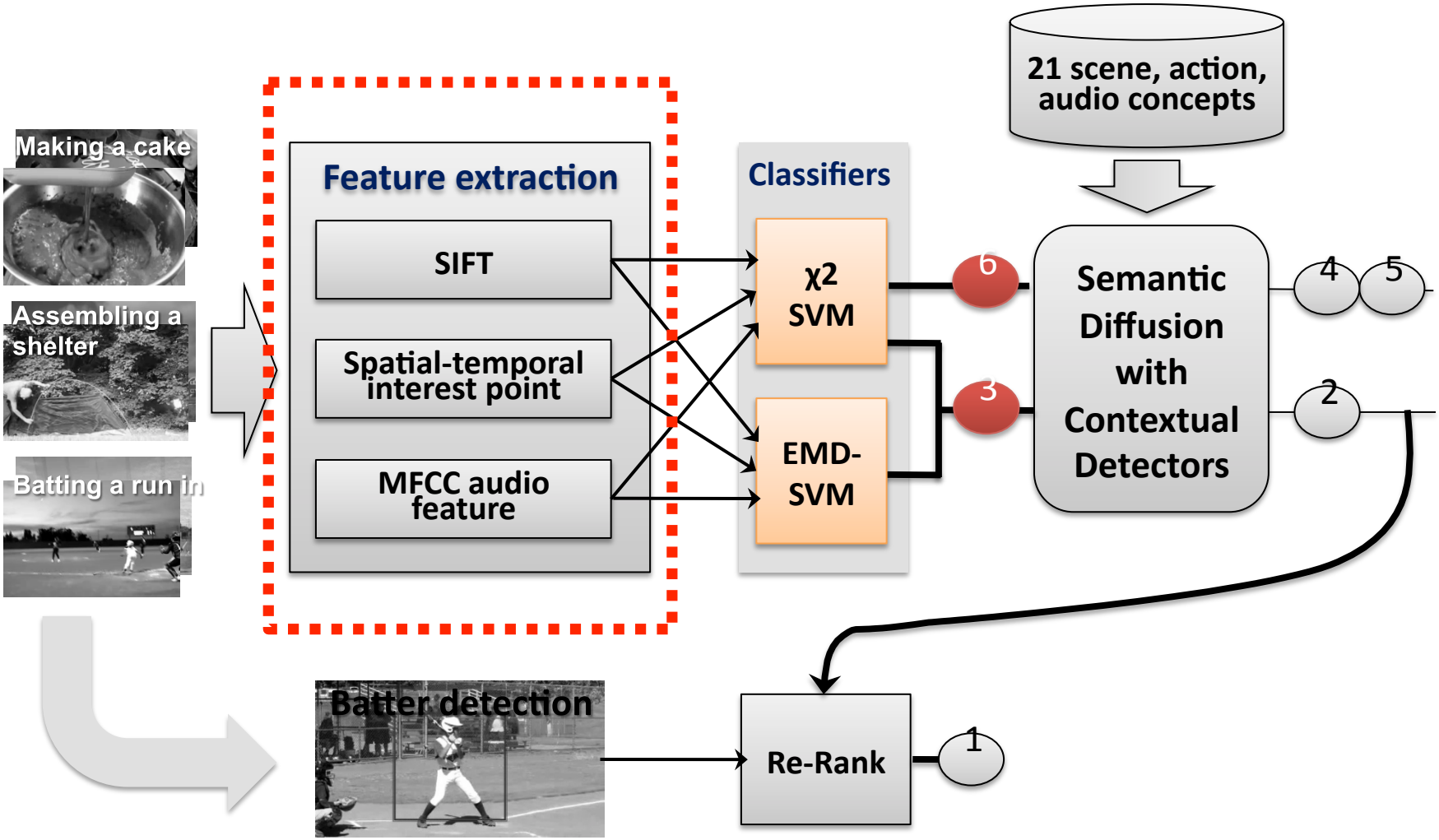
Results on Dry-run Validation Set

- Measured by Average Precision (AP)

	Assembling a shelter	Batting a run in	Making a cake	*Mean AP*
Visual STIP	0.468	0.719	0.476	0.554
Visual SIFT	0.353	0.787	0.396	0.512
Audio MFCC	0.249	0.692	0.270	0.404
STIP+SIFT	0.508	0.796	0.476	0.593
STIP+SIFT+MFCC	**0.533**	**0.873**	**0.493**	**0.633**

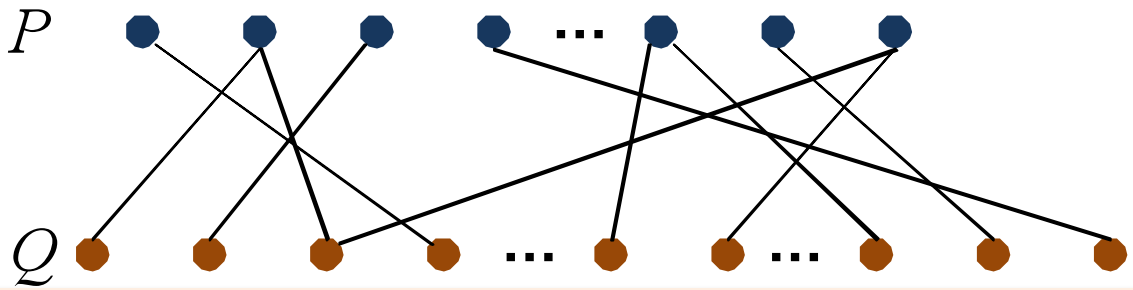
- STIP works best for event detection
- The 3 features are highly complementary!
  - Should be jointly used for multimedia event detection

Roadmap > temporal matching



Temporal Matching With EMD Kernel

- Earth Mover's Distance (EMD)



Given two frame sets $P = \{(p_1, w_{p1}), \ldots, (p_m, w_{pm})\}$ and $Q = \{(q_1, w_{q1}), \ldots, (q_n, w_{qn})\}$, the EMD is computed as

$$\text{EMD}(P, Q) = \Sigma_i \Sigma_j f_{ij} d_{ij} / \Sigma_i \Sigma_j f_{ij}$$

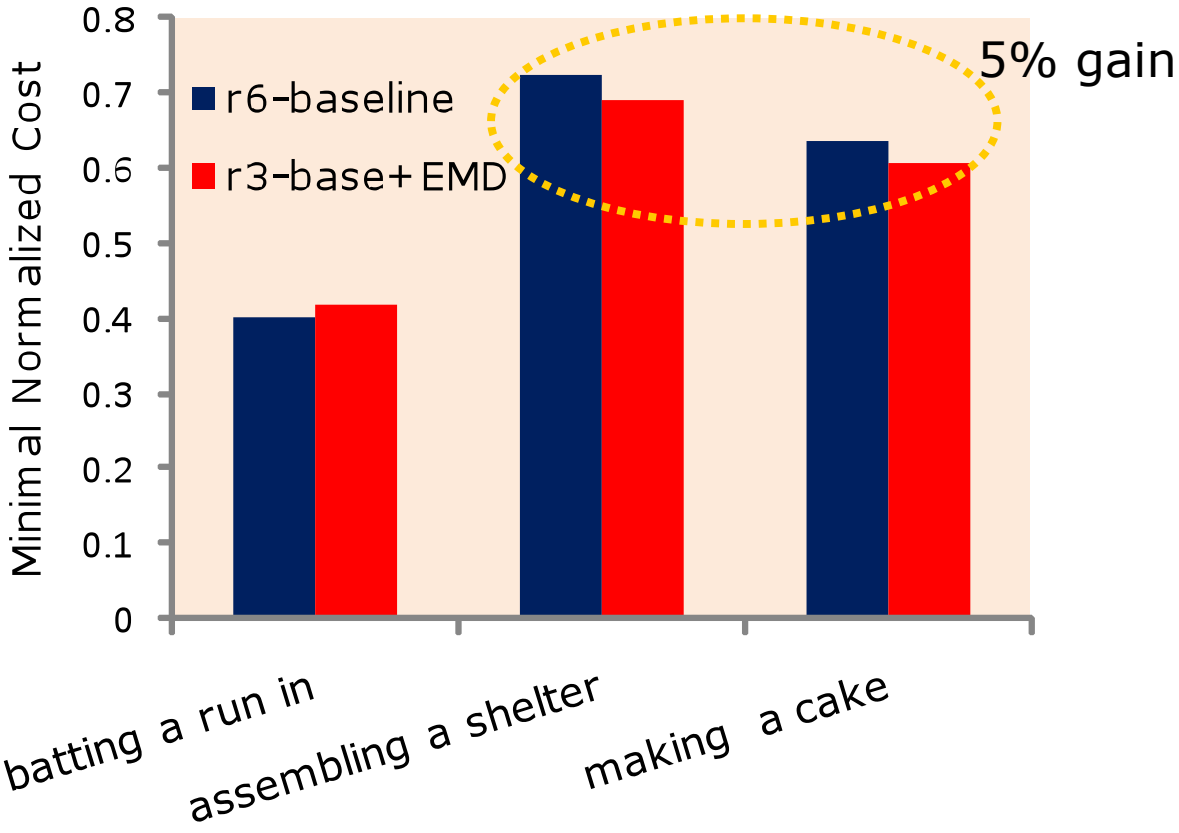
d_{ij} is the χ^2 visual feature distance of frames p_i and q_j. f_{ij} (weight transferred from p_i and q_j) is optimized by minimizing the overall transportation workload $\Sigma_i \Sigma_j f_{ij} d_{ij}$

- EMD Kernel: $K(P,Q) = \exp^{-\rho \text{EMD}(P,Q)}$

Y. Rubner, C. Tomasi, L. J. Guibas, "A metric for distributions with applications to image databases", ICCV, 1998.
D. Xu, S.-F. Chang, "Video event recognition using kernel methods with multi-level temporal alignment", PAMI, 2008.

Temporal Matching Results

- EMD is helpful for two events
  - results measured by minimal normalized cost (lower is better)



Conclusions

- Novel audio features focus on foreground and background
  - Successful combinations
- Large-scale annotation for public data set
  - Columbia Consumer Video
- Multimedia Event Detection is feasible
  - Columbia system came top in TREC evaluation