



# 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

# 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



...

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

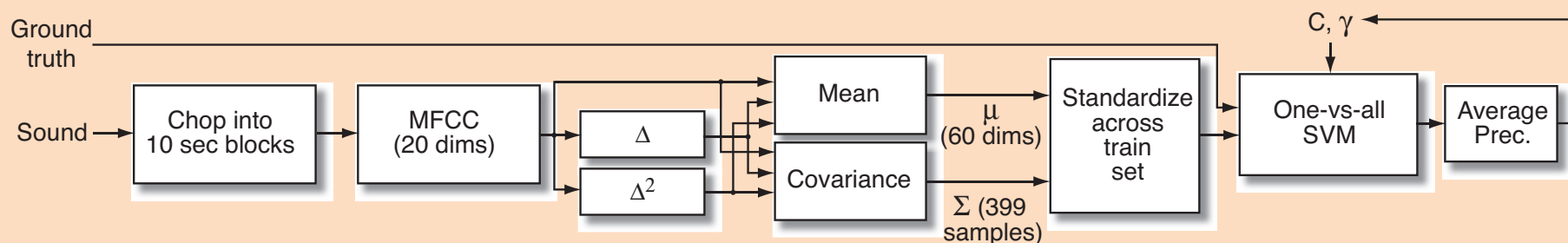


# Highlights 2010-2011

- **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:

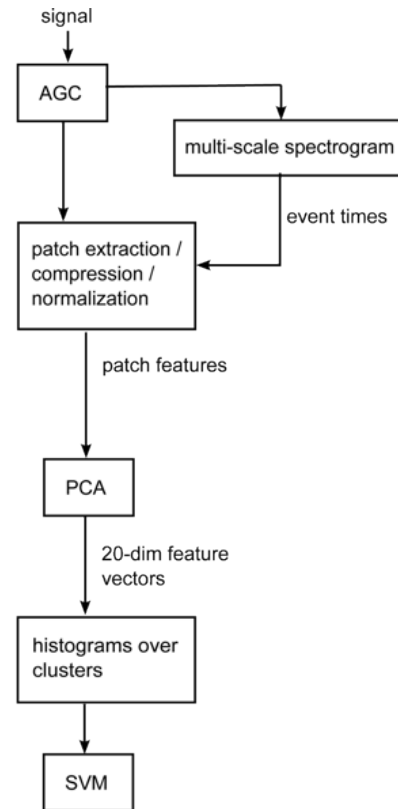
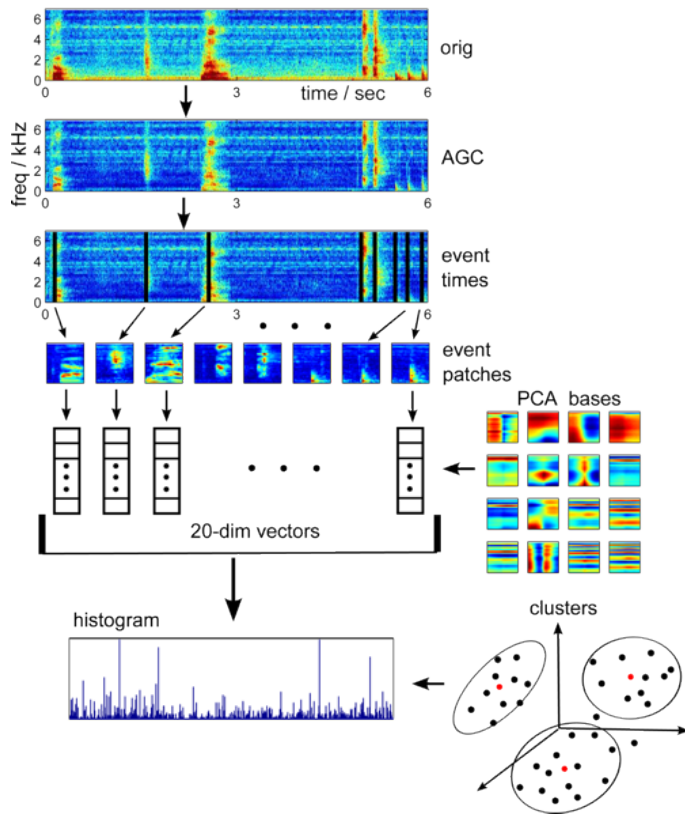


– everything mixed in together

- Can we differentiate foreground and background?



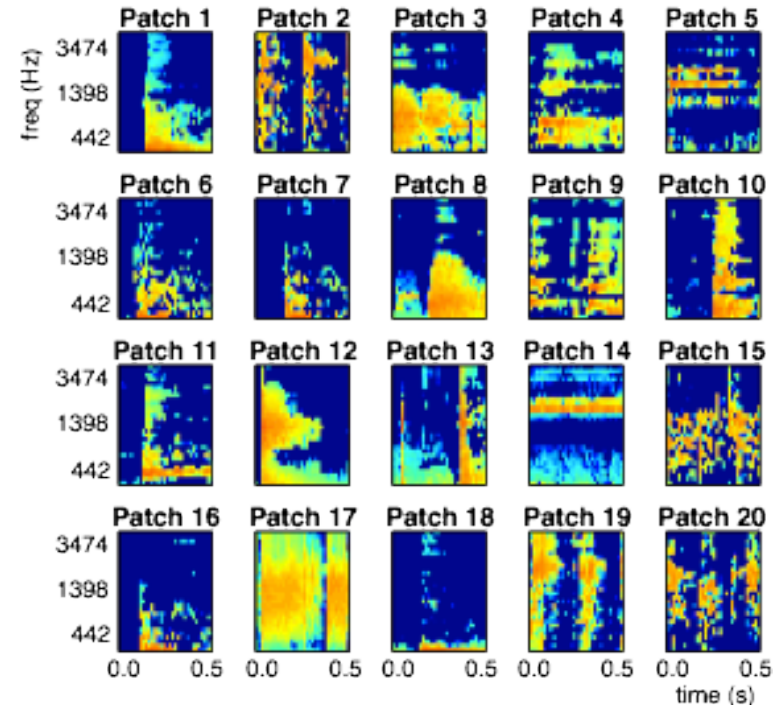
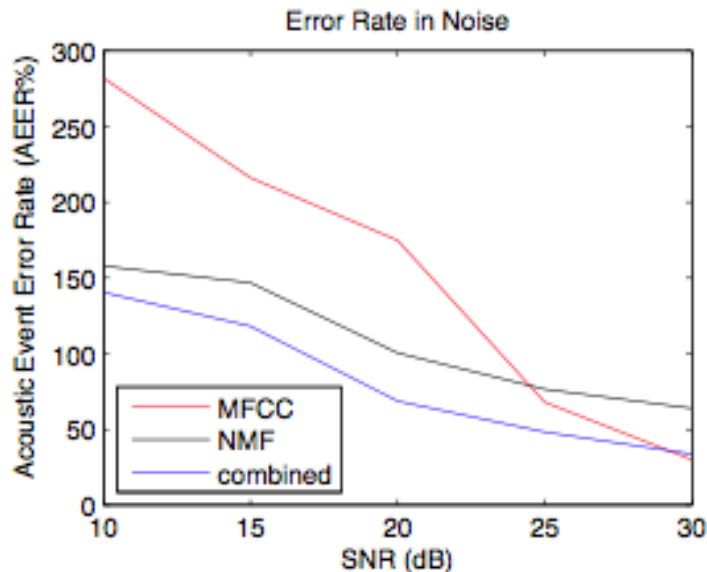
# Foreground: Transient Features



- 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

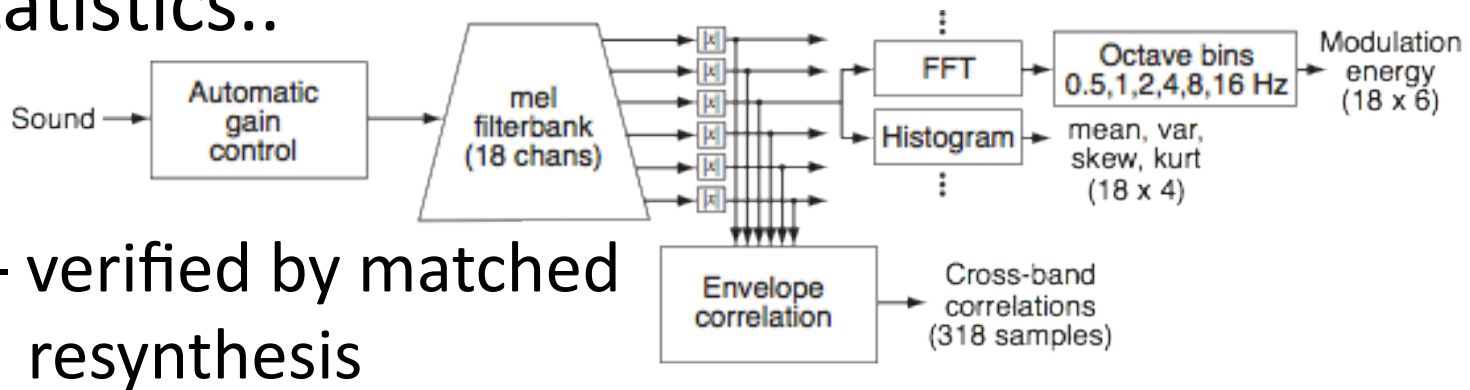
- Learn 20 patches by Nonnegative Matrix Factorization
- Compare to MFCC-HMM



- NMF more noise-robust
  - combines well

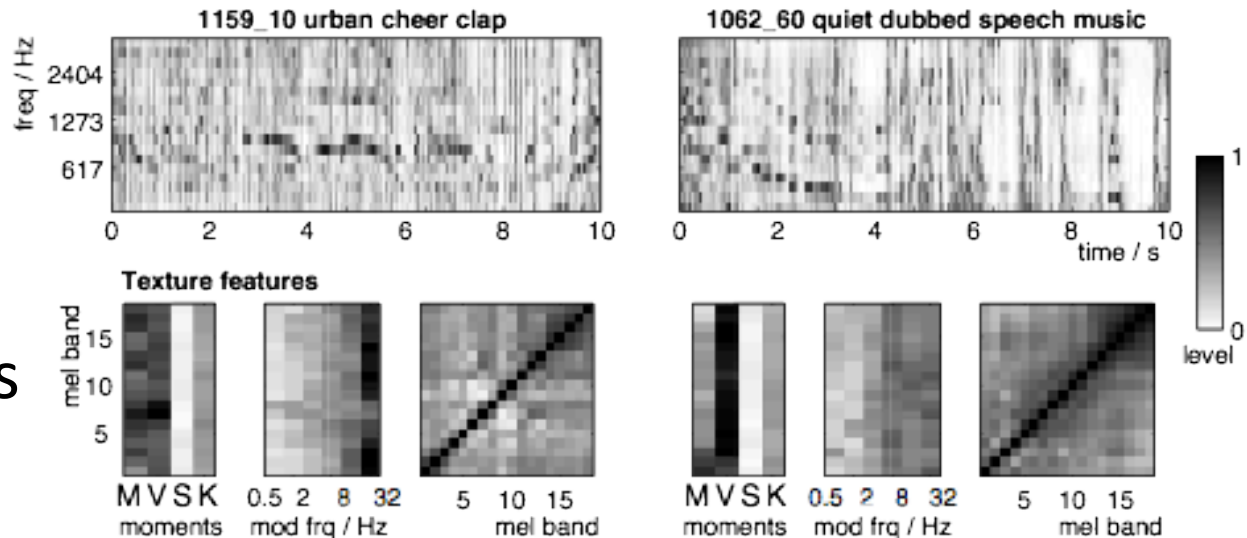
# Background: Texture features

- Characterize sounds by perceptually-sufficient statistics..



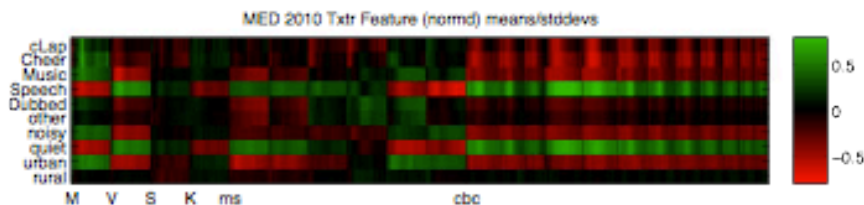
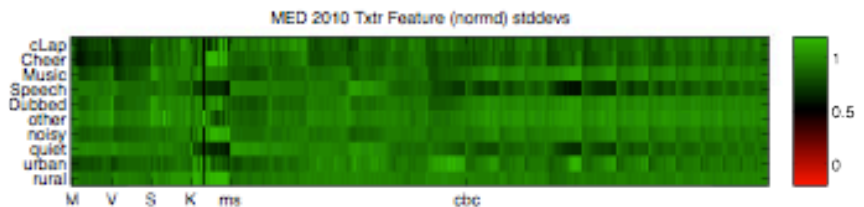
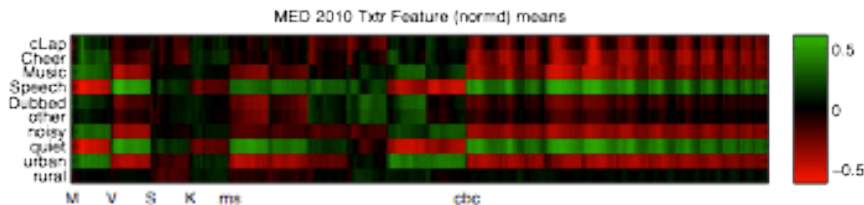
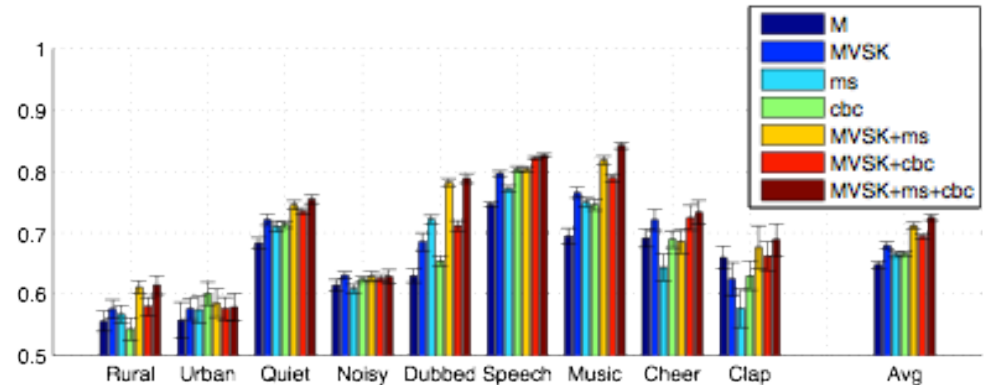
- verified by matched resynthesis

- Subband distributions & env x-corrs
- Mahalanobis distance ...



# Texture Feature Results

- Test on MED 2010 development data
  - 10 labels

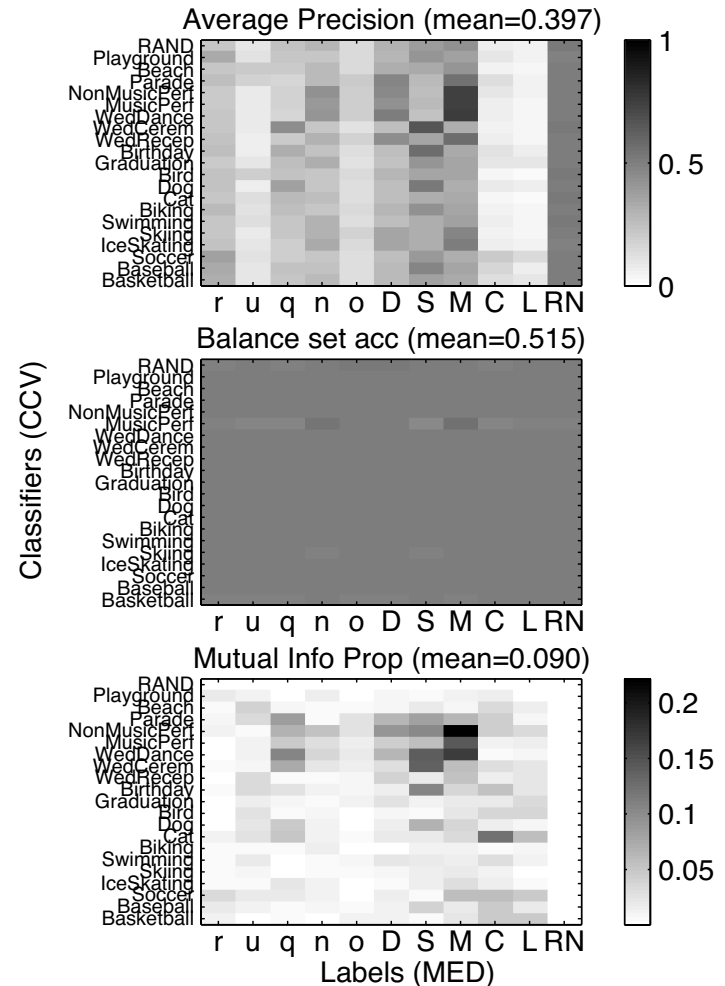
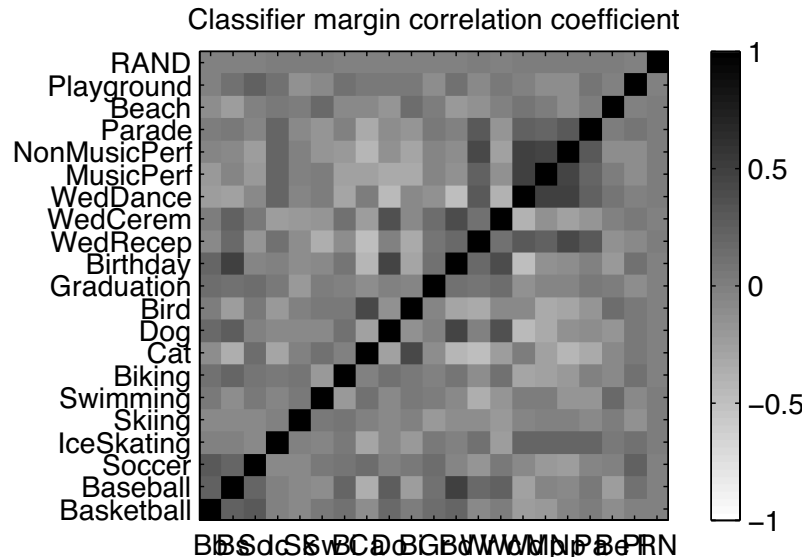


- Contrasts in feature sets
  - correlation of labels
- Perform ~ same as MFCCs
  - combine well

# Audio Classifier Evaluation

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

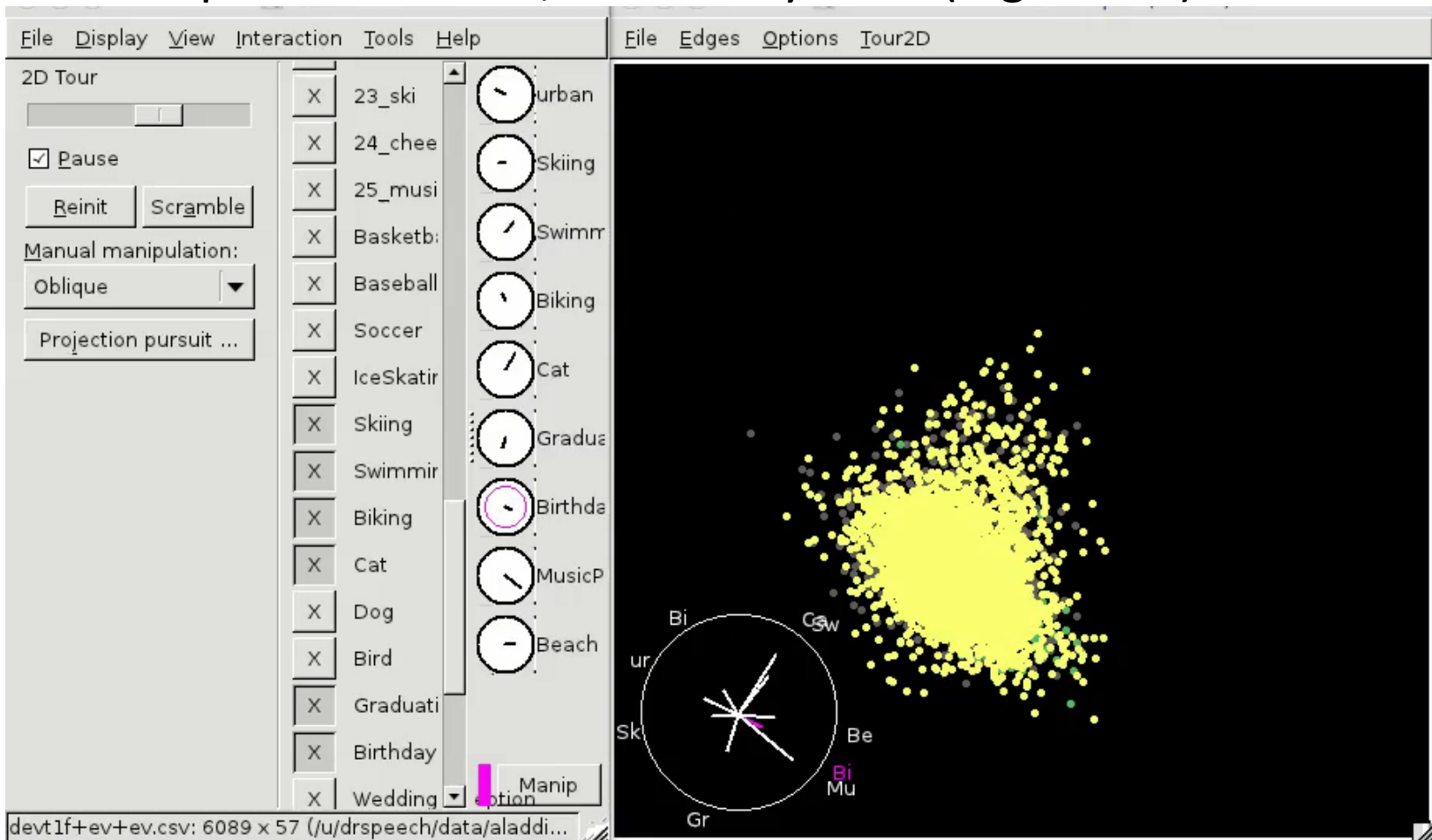
$$MIP = \frac{I(\text{classifier}; \text{label})}{H(\text{label})}$$



- 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)





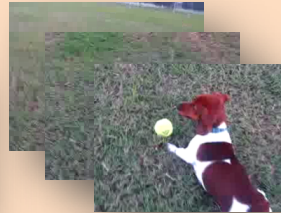
# Columbia Consumer Video (CCV) Database



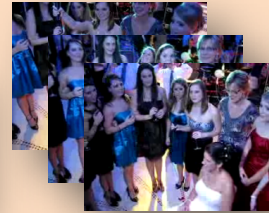
Basketball



Skiing



Dog



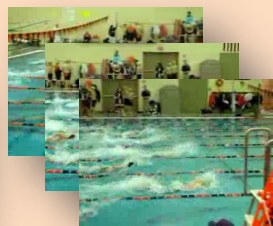
Wedding Reception



Non-music Performance



Baseball



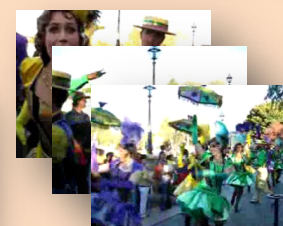
Swimming



Bird



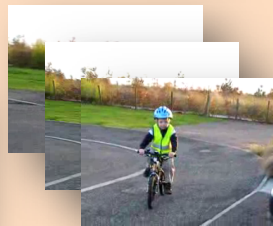
Wedding Ceremony



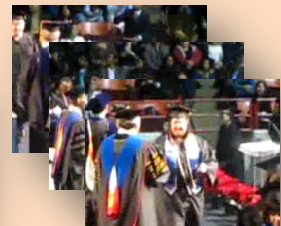
Parade



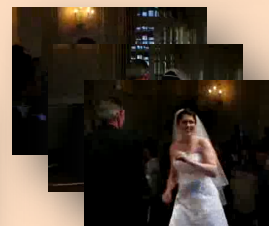
Soccer



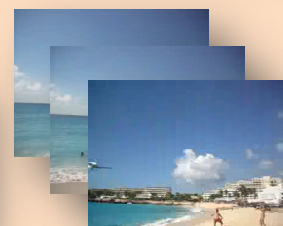
Biking



Graduation



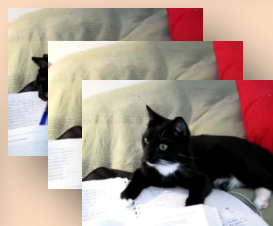
Wedding Dance



Beach



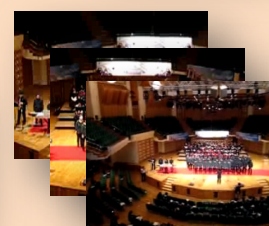
Ice Skating



Cat



Birthday Celebration



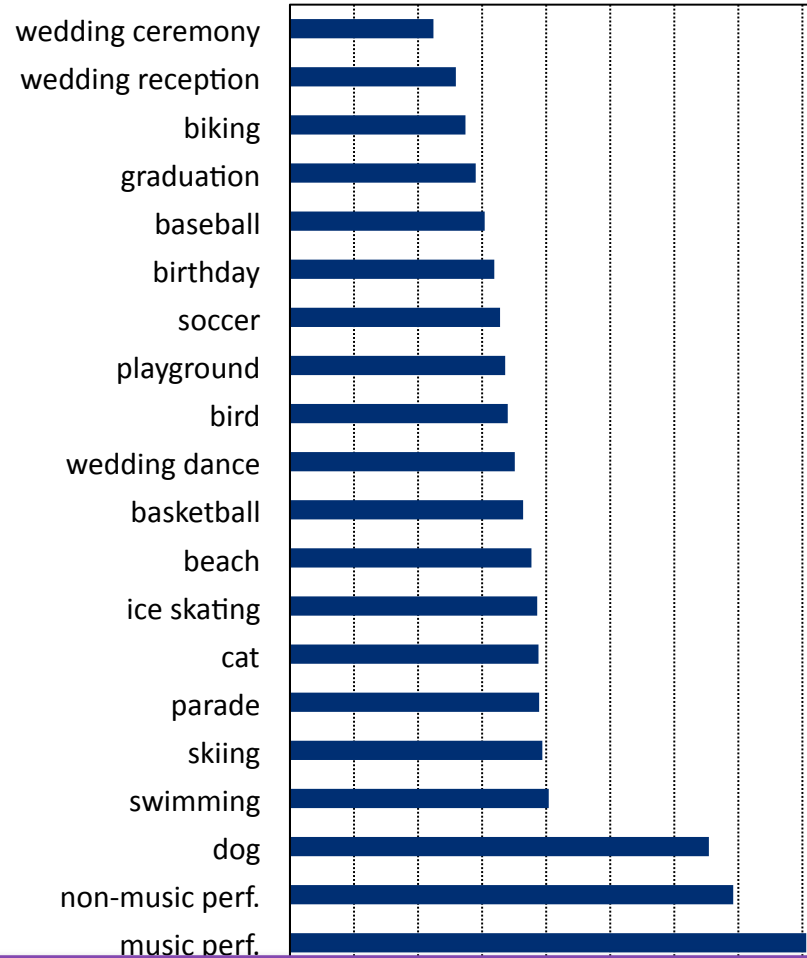
Music Performance



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?

## CCV Database

- Human Action Recognition

- KTH & Weizmann

- (constrained environment) 2004-05

- Hollywood Database

- (12 categories, movies) 2008

- UCF Database

- (50 categories, YouTube Videos) 2010

**Unconstrained YouTube videos**

**Higher-level complex events**

- Kodak Consumer Video

- (25 classes, 1300+ videos) 2007

**More videos & better defined categories**

- LabelMe Video

- (many classes, 1300+ videos) 2009

**More videos & larger content variations**

- TRECVID MED 2010

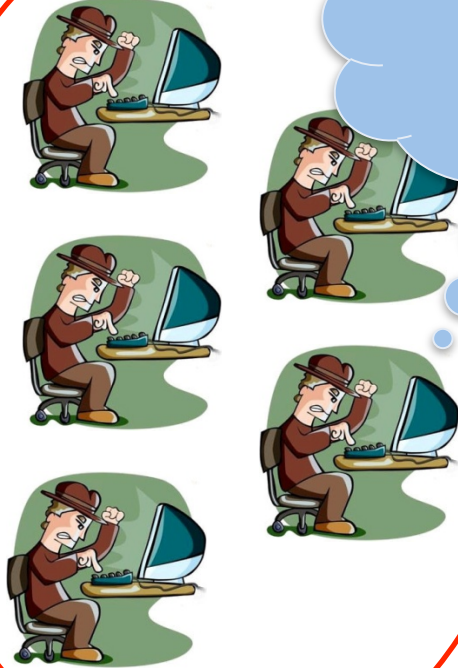
- (3 classes, 3400+ videos) 2010

**More videos & categories**

# Crowdsourcing: Amazon Mechanical Turk

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

What can  
I do for  
you?



**Internet-scale workforce**



Is this a "parade"  
video?

- Yes
- No

**Task**

**\$?.??**

**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.



Sports	Animal	Celebration	Others
<input type="checkbox"/> <a href="#">Basketball</a>	<input type="checkbox"/> <a href="#">Cat</a>	<input type="checkbox"/> <a href="#">Graduation</a>	<input type="checkbox"/> <a href="#">Music Performance</a>
<input type="checkbox"/> <a href="#">Baseball</a>	<input type="checkbox"/> <a href="#">Dog</a>	<input checked="" type="checkbox"/> <a href="#">Birthday</a>	<input type="checkbox"/> <a href="#">Non-music Performance</a>
<input type="checkbox"/> <a href="#">Soccer</a>	<input type="checkbox"/> <a href="#">Bird</a>	<input type="checkbox"/> <a href="#">Wedding Reception</a>	<input type="checkbox"/> <a href="#">Parade</a>
<input type="checkbox"/> <a href="#">Ice Skating</a>		<input type="checkbox"/> <a href="#">Wedding Ceremony</a>	<input type="checkbox"/> <a href="#">Beach</a>
<input type="checkbox"/> <a href="#">Skiing</a>		<input type="checkbox"/> <a href="#">Wedding Dance</a>	<input type="checkbox"/> <a href="#">Playground</a>
<input type="checkbox"/> <a href="#">Swimming</a>	<input type="checkbox"/> None of the categories matches.		
<input type="checkbox"/> <a href="#">Biking</a>	<input type="checkbox"/> I don't see any video playing.		

Current Time: 10 sec

[Replay](#) [Continue Playing](#)

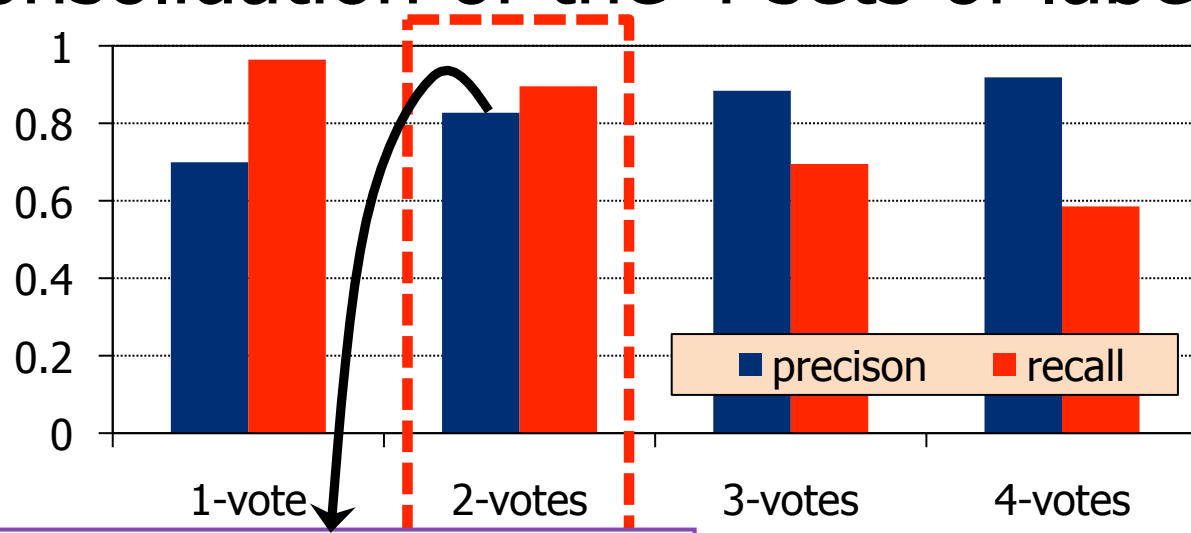
Original URL: <http://www.youtube.com/watch?v=-0n50a7seNI>

**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/](http://www.ee.columbia.edu/dvmm/CCV/)

Fill out this ...



**Columbia Consumer Video (CCV) Database**  
--- A Benchmark for Consumer Video Analysis

**Summary**

Recognizing visual content in unconstrained videos has become a very important problem for many applications. Existing corpora for video analysis lack scale and/or content diversity, and thus limited the needed progress in this critical area. To stimulate innovative research on this challenging issue, we constructed a new database called CCV, containing 9,317 YouTube videos over 20 semantic categories. The database was collected with extra care to ensure relevance to consumer interest and originality of video content without post-editing. Such videos typically have very little textual annotation and thus can benefit from the development of automatic content analysis techniques.

We used Amazon MTurk platform to perform manual annotation, and implemented automatic classifiers using state-of-the-art multi-modal approach that achieved top performance in 2010 TRECVID multimedia event detection task. These automatic classifiers produce a decent baseline performance. We release unique YouTube IDs of CCV videos, ground-truth annotations, a standard training and testing partition, and three audiovisual feature representations to the community for research usage.

**CCV Snapshot**

- # Videos: 9,317 (210 hrs in total)
- [www.ee.columbia.edu/dvmm/CCV/](http://www.ee.columbia.edu/dvmm/CCV/)

**CCV Citation**

Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, Alexander C. Loui, *Consumer Video Understanding: A Benchmark Database and An Evaluation of Human and Machine Performance*, ACM International Conference on Multimedia Retrieval (ICMR), Trento, Italy, April 2011.

**# positive videos per category**

Category	# positive videos
wedding ceremony	~750
wedding reception	~700
biking	~650
graduation	~600
baseball	~550
beach	~500
ice skating	~450
cat	~400
parade	~350
skiing	~300
swimming	~250
dog	~200
non-music perf.	~150
music perf.	~100

**Download**

To download the CCV database, please fill out the following form. We will send you download instructions via email immediately. **People who request and use this database should agree that 1) the use of the data is restricted to research purpose only, and 2) the authors of the above ICMR paper and their affiliated organizations make no warranties regarding this database, such as (not limited to) non-infringement.**

Name:  Affiliation:  Email Address:

**Baseline Evaluation**

We implemented a baseline system using three popular audio/visual features, namely SIFT, STIP, and MFCC. For all the three features, videos are represented by bag-of-word framework. Classification results are given in the following figure, where the performance is measured by average precision. The combination of multiple features is done by averaging separate SVM prediction scores. For more details of our baseline classifier design, please refer to the *CCV paper*. All the three features are included in the released package.

**More results: Per-category precision-recall curves and example frames.**

# TRECVID MED 2010

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

**Making a  
cake**



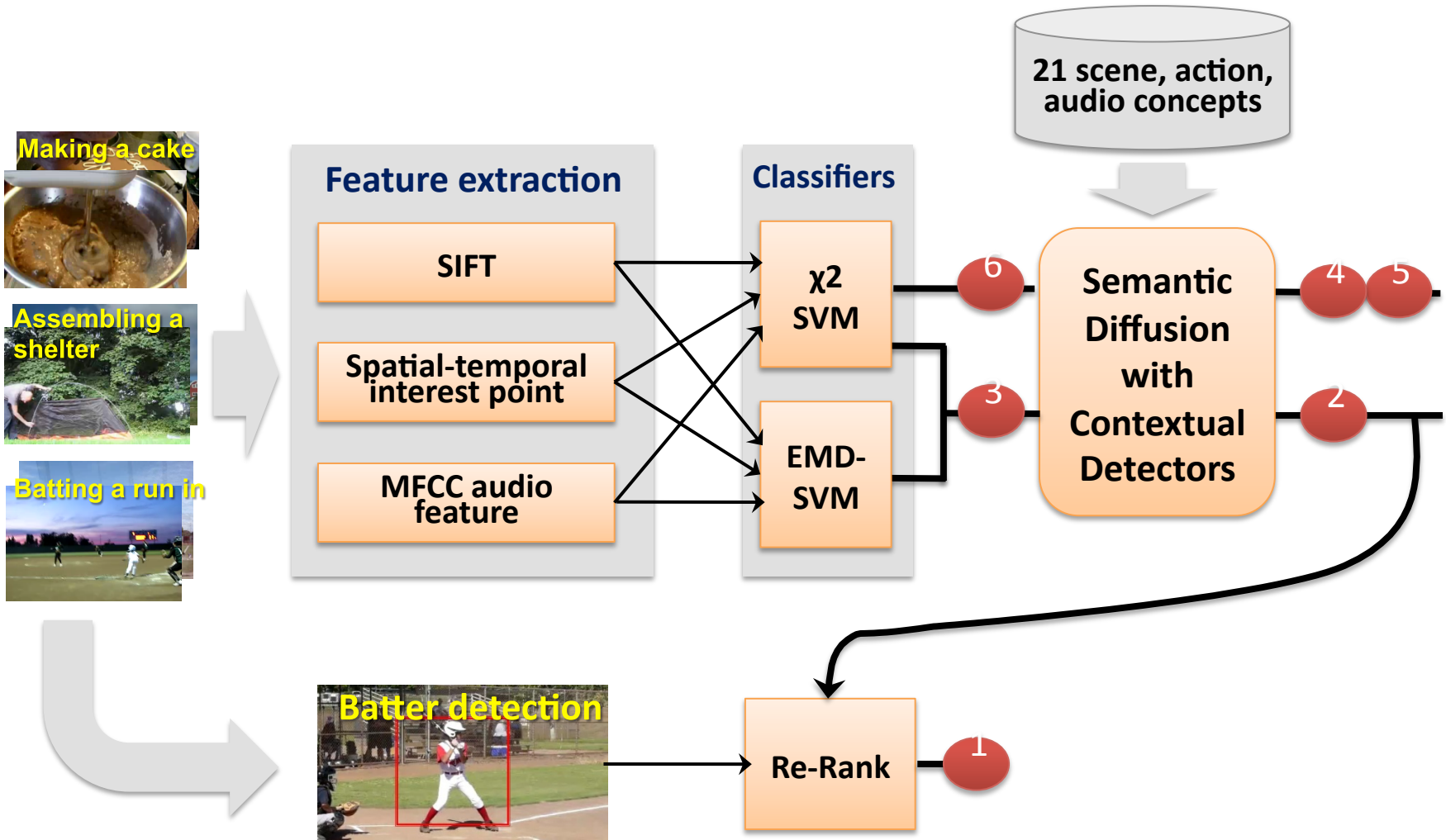
**Assembling  
a shelter**



**Batting a  
run in**

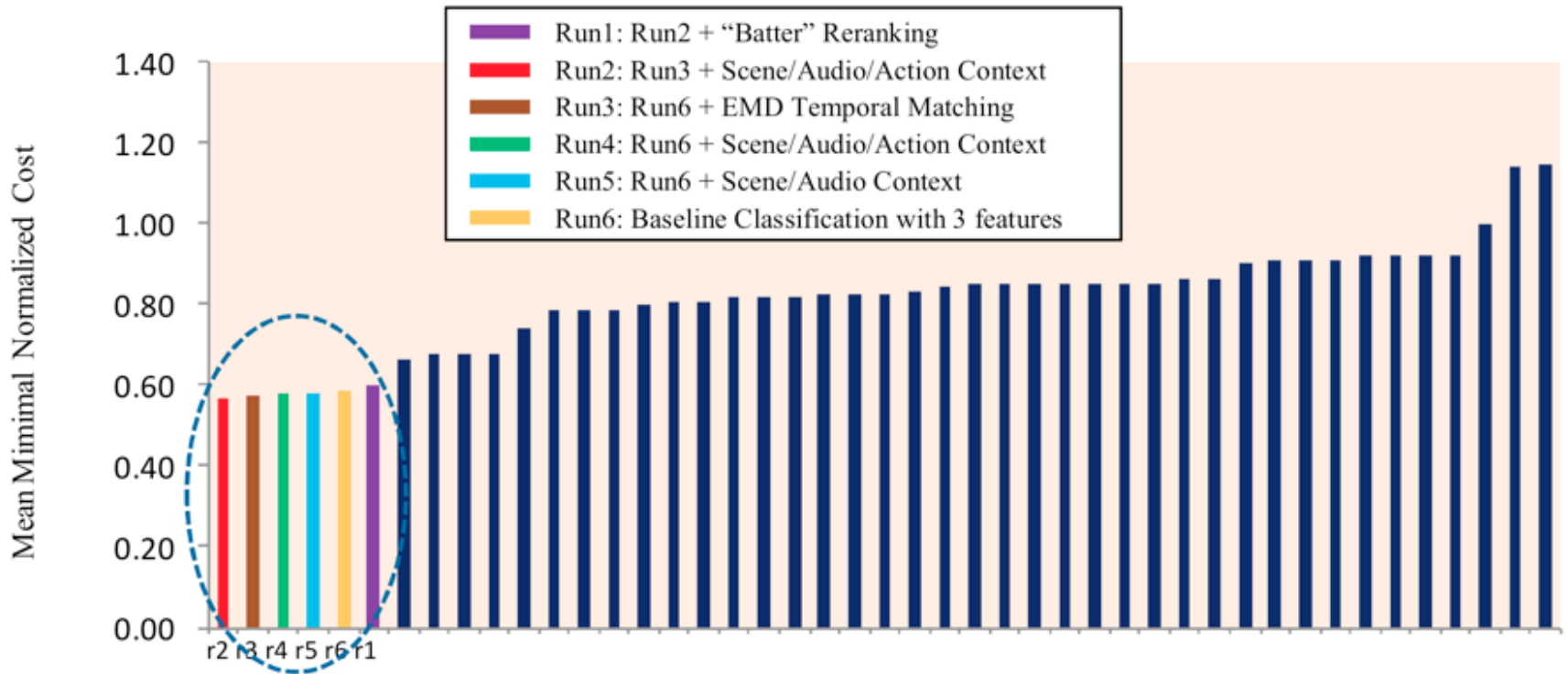


# Overview: 4 major components & 6 runs





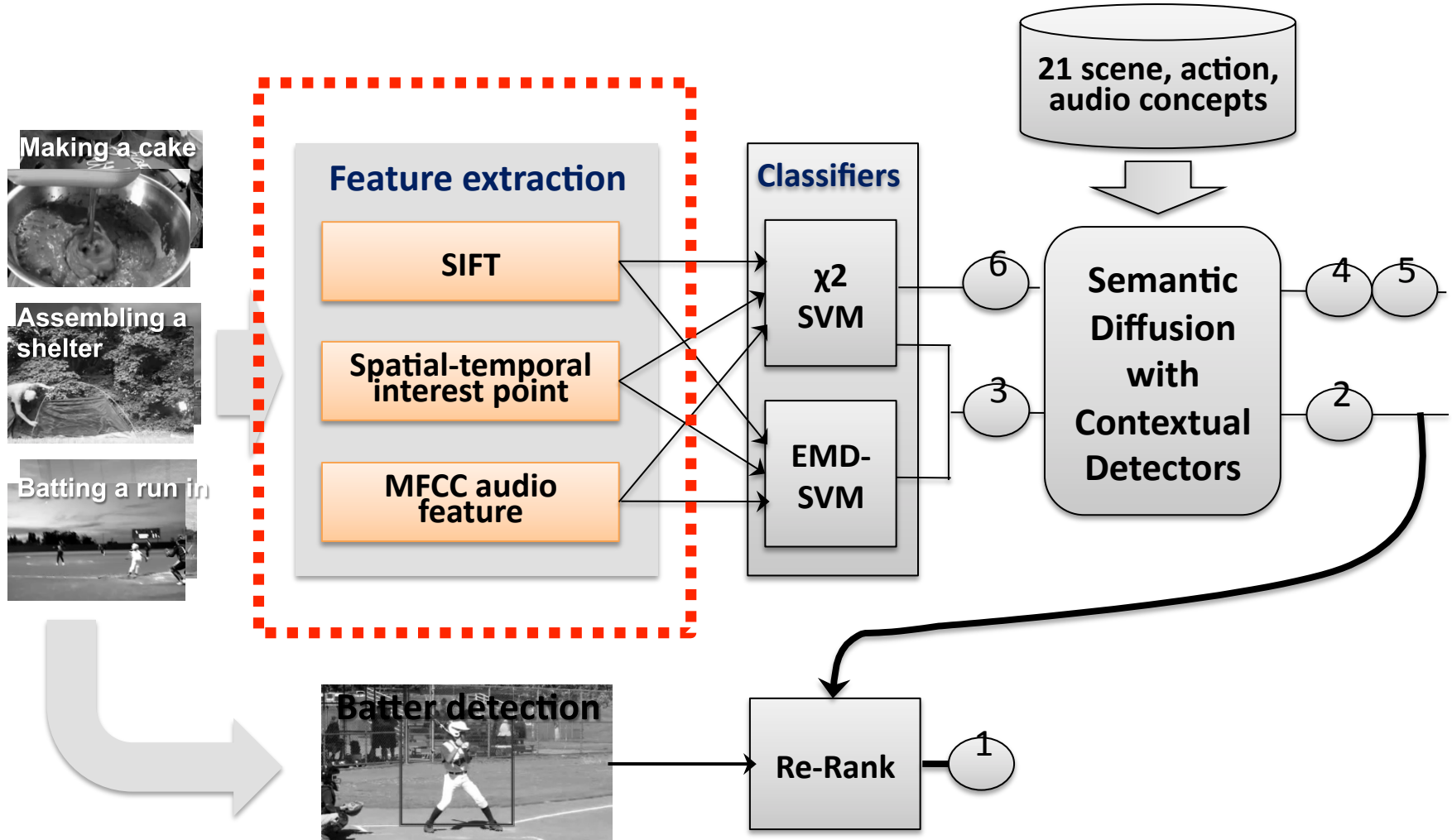
# Overview: overall performance



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

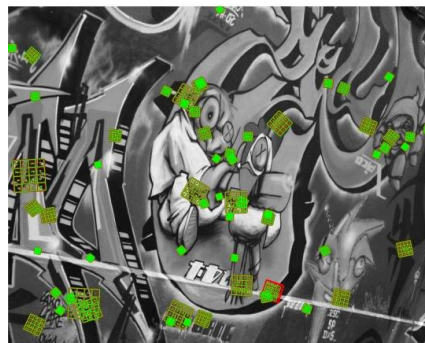


# Roadmap > multiple modalities

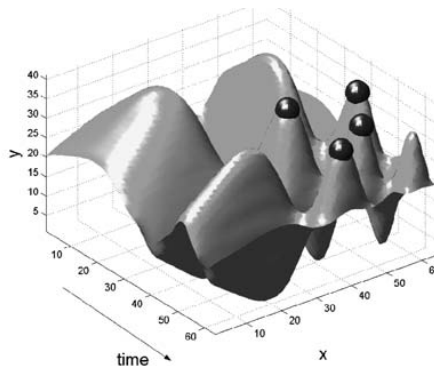


# Three Feature Modalities...

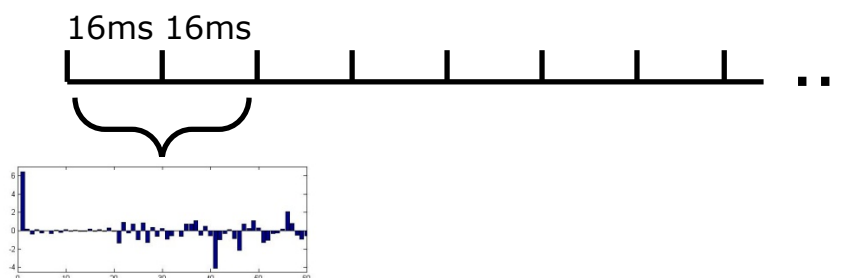
- SIFT (visual)  
– D. Lowe, IJCV 04.



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



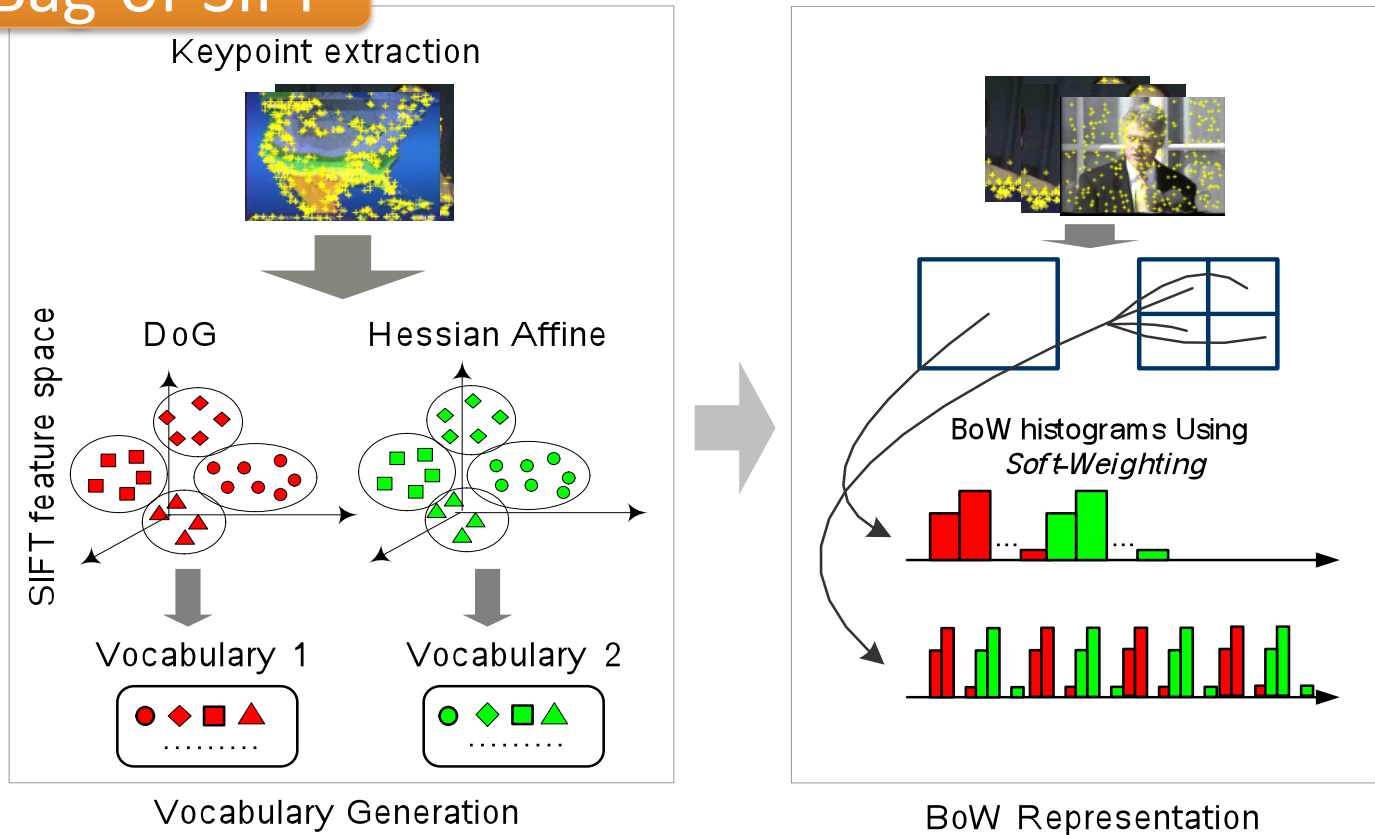
- MFCC (audio)



# Bag-of-~~X~~ Representation

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

## Bag-of-SIFT



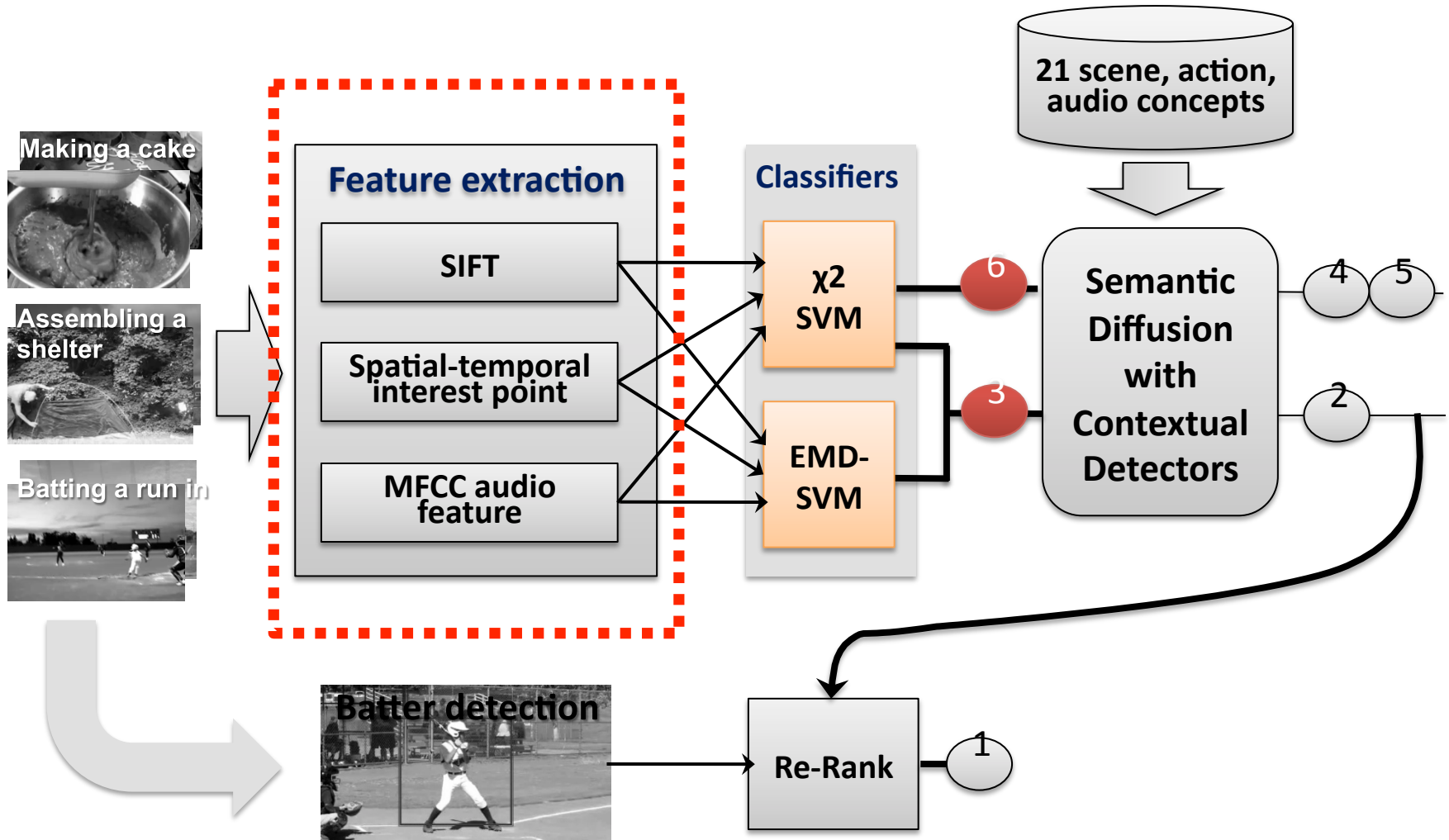
# Results on Dry-run Validation Set

- Measured by Average Precision (AP)

	Assembling a shelter	Batting a run in	Making a cake	<i>Mean AP</i>
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	<b><u>0.533</u></b>	<b><u>0.873</u></b>	<b><u>0.493</u></b>	<b><u>0.633</u></b>

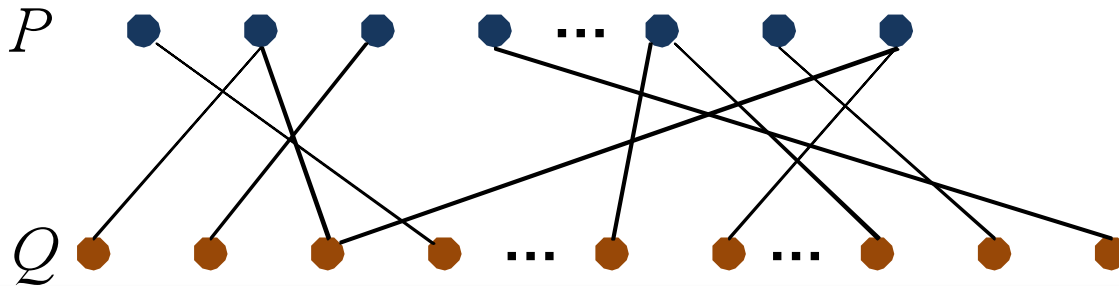
- STIP works best for event detection
- The 3 features are **highly complementary!**
  - Should be jointly used for multimedia event detection<sup>24</sup>

# Roadmap > temporal matching



# Temporal Matching With EMD Kernel

- Earth Mover's Distance (EMD)



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

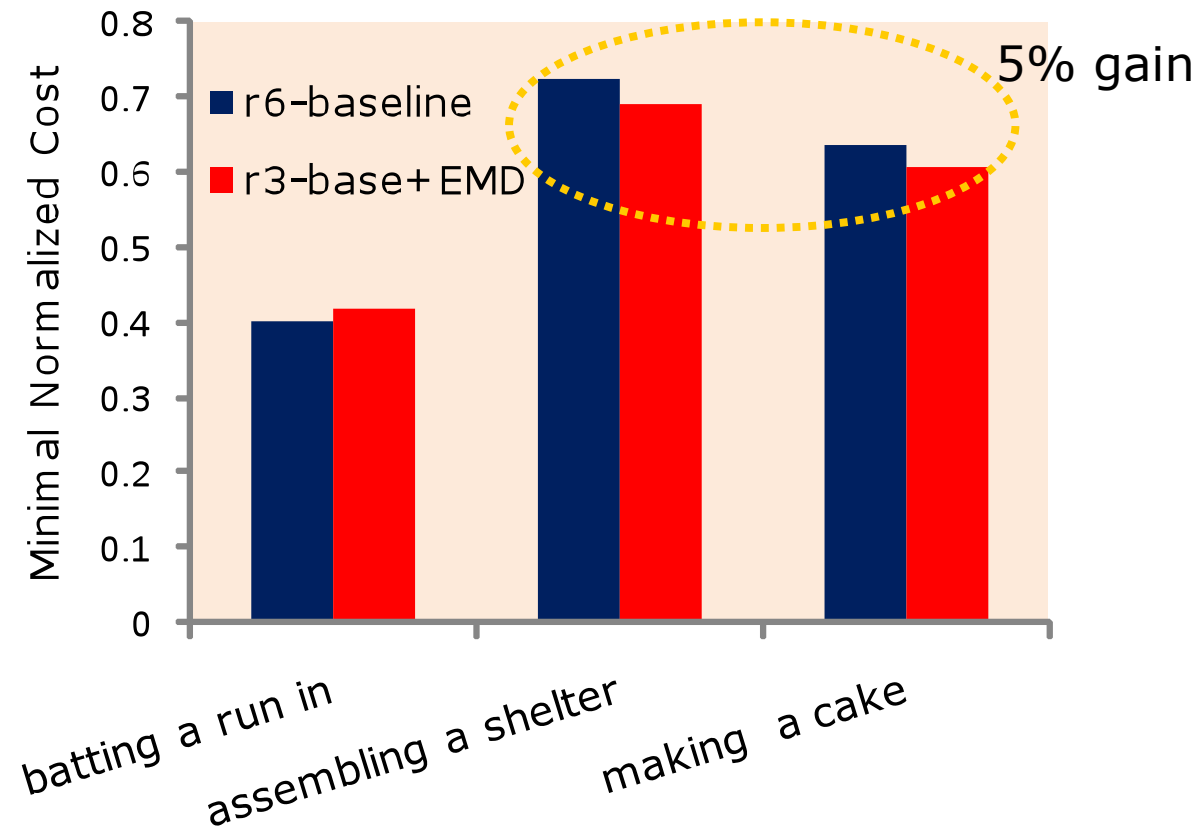
$$\text{EMD}(P, Q) = \sum_i \sum_j f_{ij} d_{ij} / \sum_i \sum_j f_{ij}$$

$d_{ij}$  is the  $\chi^2$  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  $\sum_i \sum_j f_{ij} d_{ij}$

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

# 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