

# Recognizing Complex Events in Internet Videos with Audio-Visual Features

**Yu-Gang Jiang**

[yjiang@ee.columbia.edu](mailto:yjiang@ee.columbia.edu)

In collaboration with **Xiaohong Zeng<sup>1</sup>**, **Guangnan Ye<sup>1</sup>**, **Subh Bhattacharya<sup>2</sup>**,  
**Dan Ellis<sup>1</sup>**, **Mubarak Shah<sup>2</sup>**, **Shih-Fu Chang<sup>1</sup>**, **Alexander C. Loui<sup>3</sup>**

Columbia University<sup>1</sup>

University of Central Florida<sup>2</sup>

Kodak Research Labs<sup>3</sup>

We take photos/videos  
everyday/everywhere...



Barack Obama Rally, Texas, 2008. <http://www.paulridenour.com/Obama14.JPG>

# Outline

- A System for Recognizing Events in Internet Videos
  - Best performance in TRECVID 2010 Multimedia Event Detection Task
  - Features, Kernels, Context, etc.
- Internet Consumer Video Analysis
  - A Benchmark Database
  - An Evaluation of Human & Machine Performance

# Outline

- A System for Recognizing Events in Internet Videos
  - Best performance in TRECVID 2010 Multimedia Event Detection Task
  - Features, Kernels, Context, etc.
- Internet Consumer Video Analysis
  - A Benchmark Database
  - An Evaluation of Human & Machine Performance



# The TRECVID Multimedia Event Detection Task

- Target: Find videos containing an event of interest
- Data: unconstrained Internet videos
  - 1700+ training videos (~50 positive each event); 1700+ test videos

## Making a cake



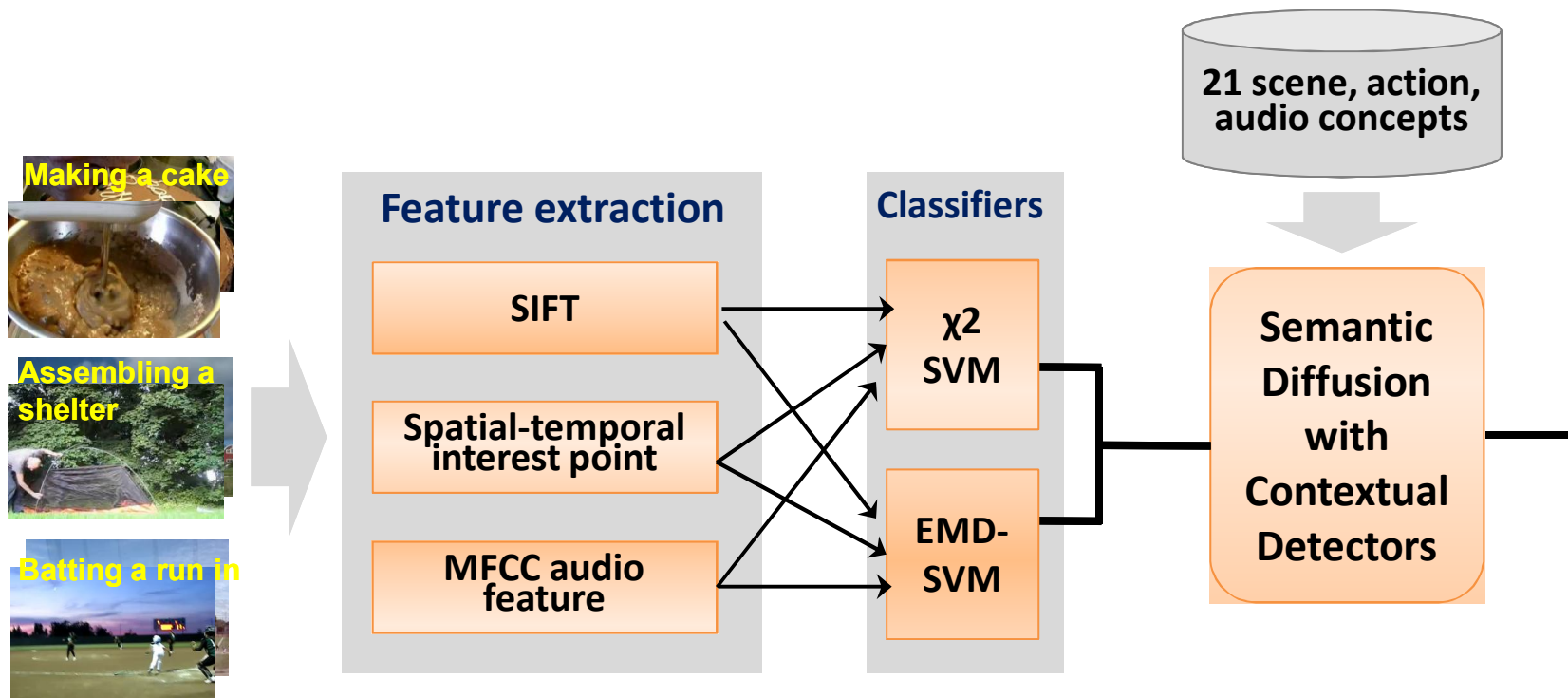
## Assembling a shelter



## Batting a run in

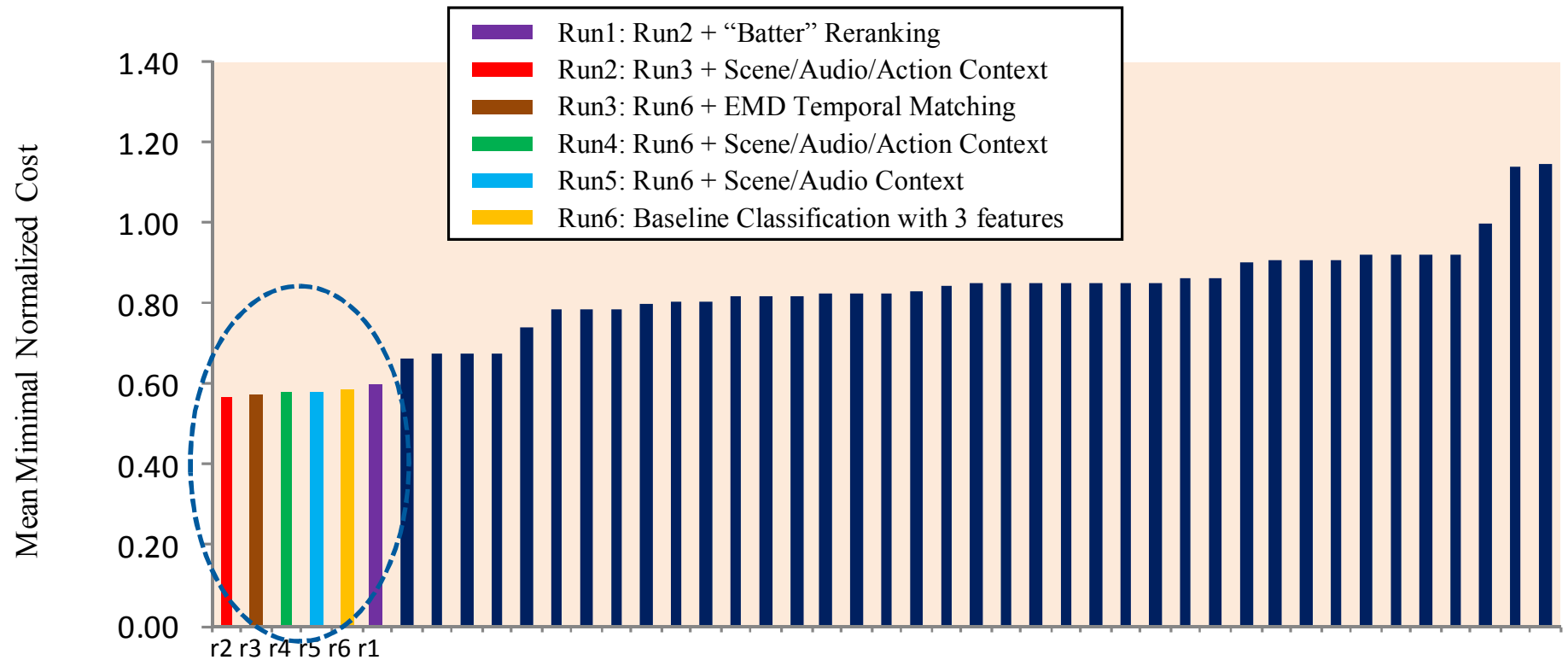


# The system: 3 major components



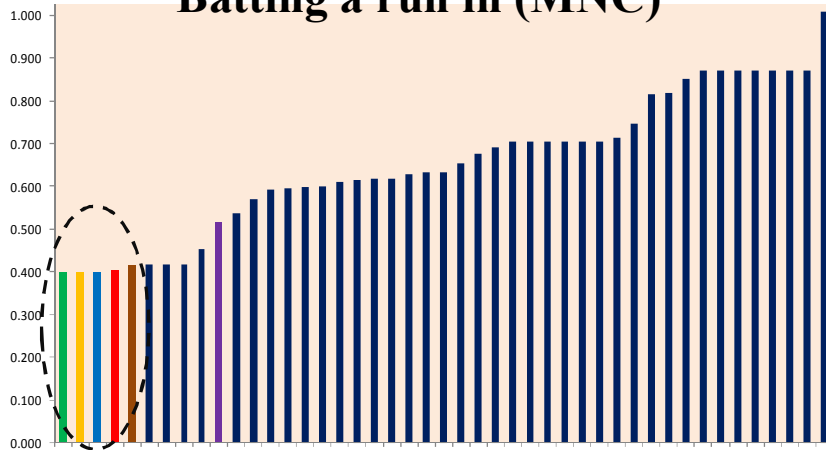
Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, S. Bhattacharya, Dan Ellis, Mubarak Shah, Shih-Fu Chang, **Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching**, in TRECVID 2010.

# Best performance in TRECVID2010 *Multimedia event detection (MED) task*

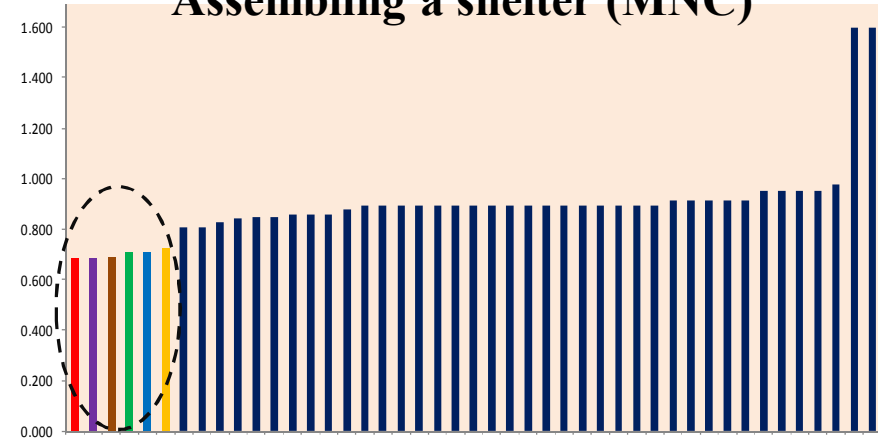


# Per-event performance

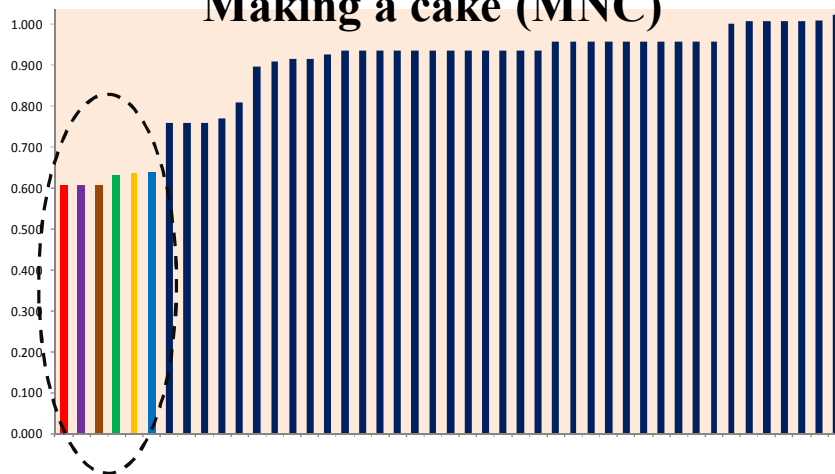
## Batting a run in (MNC)



## Assembling a shelter (MNC)



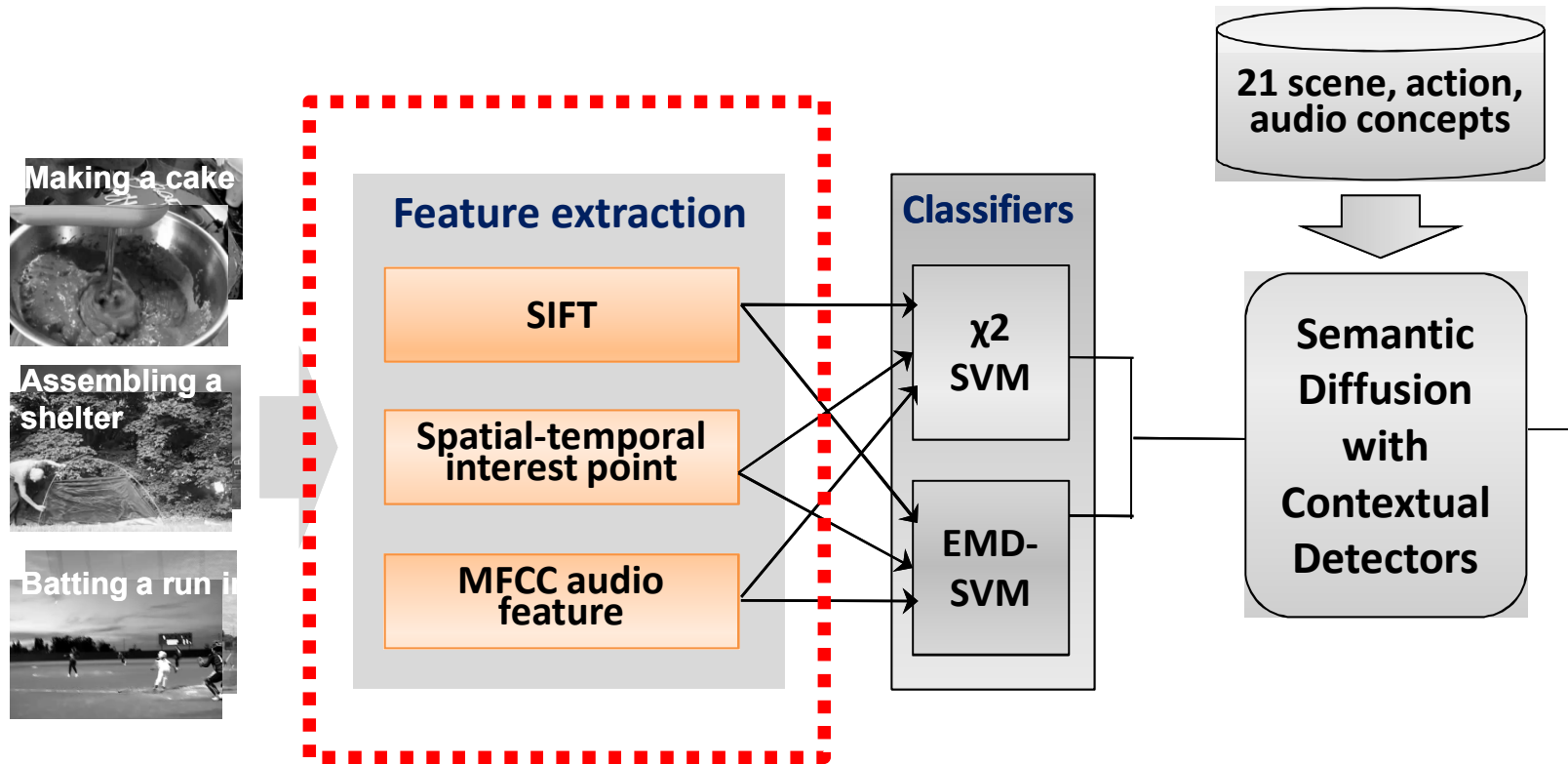
## Making a cake (MNC)



- Run1: Run2 + "Batter" Reranking
- Run2: Run3 + Scene/Audio/Action Context
- Run3: Run6 + EMD Temporal Matching
- Run4: Run6 + Scene/Audio/Action Context
- Run5: Run6 + Scene/Audio Context
- Run6: Baseline Classification with 3 features

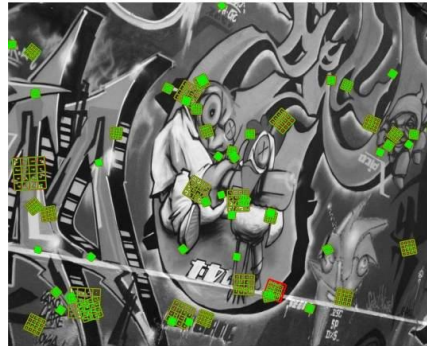


# Roadmap > audio-visual features

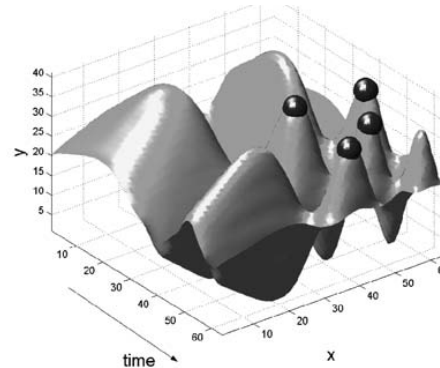


# Three audio-visual features...

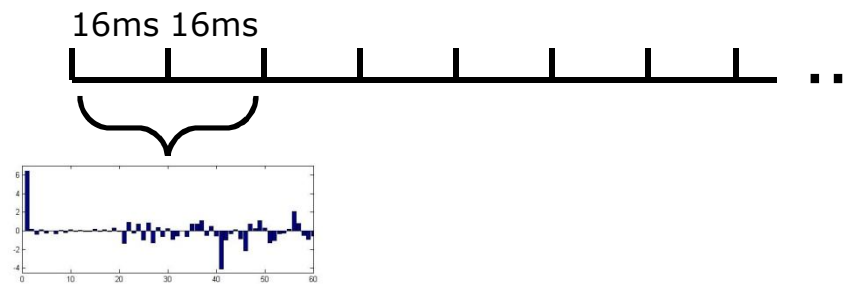
- SIFT (visual)  
– D. Lowe, IJCV 04.



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



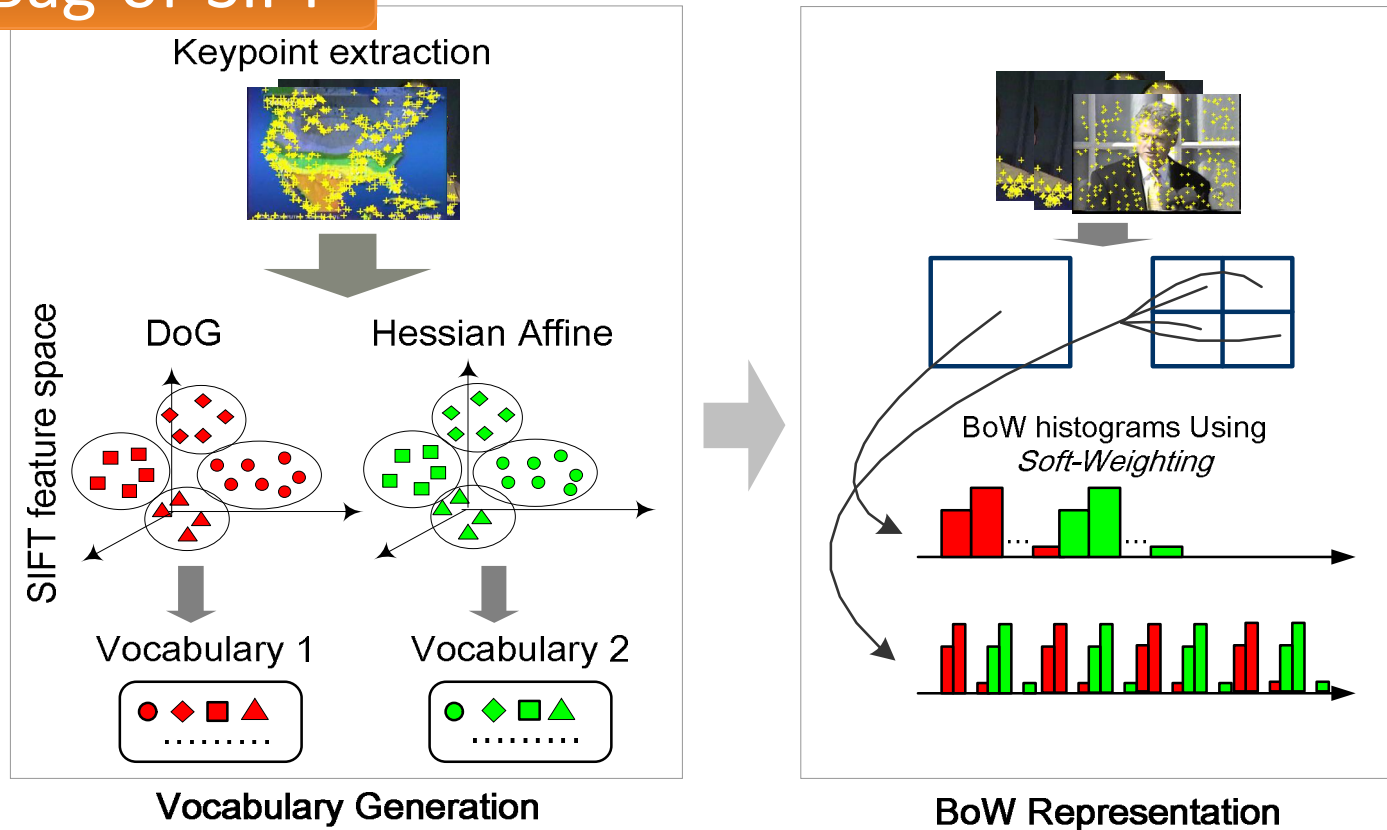
- MFCC (audio)



# Bag-of-~~X~~ representation

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

## Bag-of-SIFT



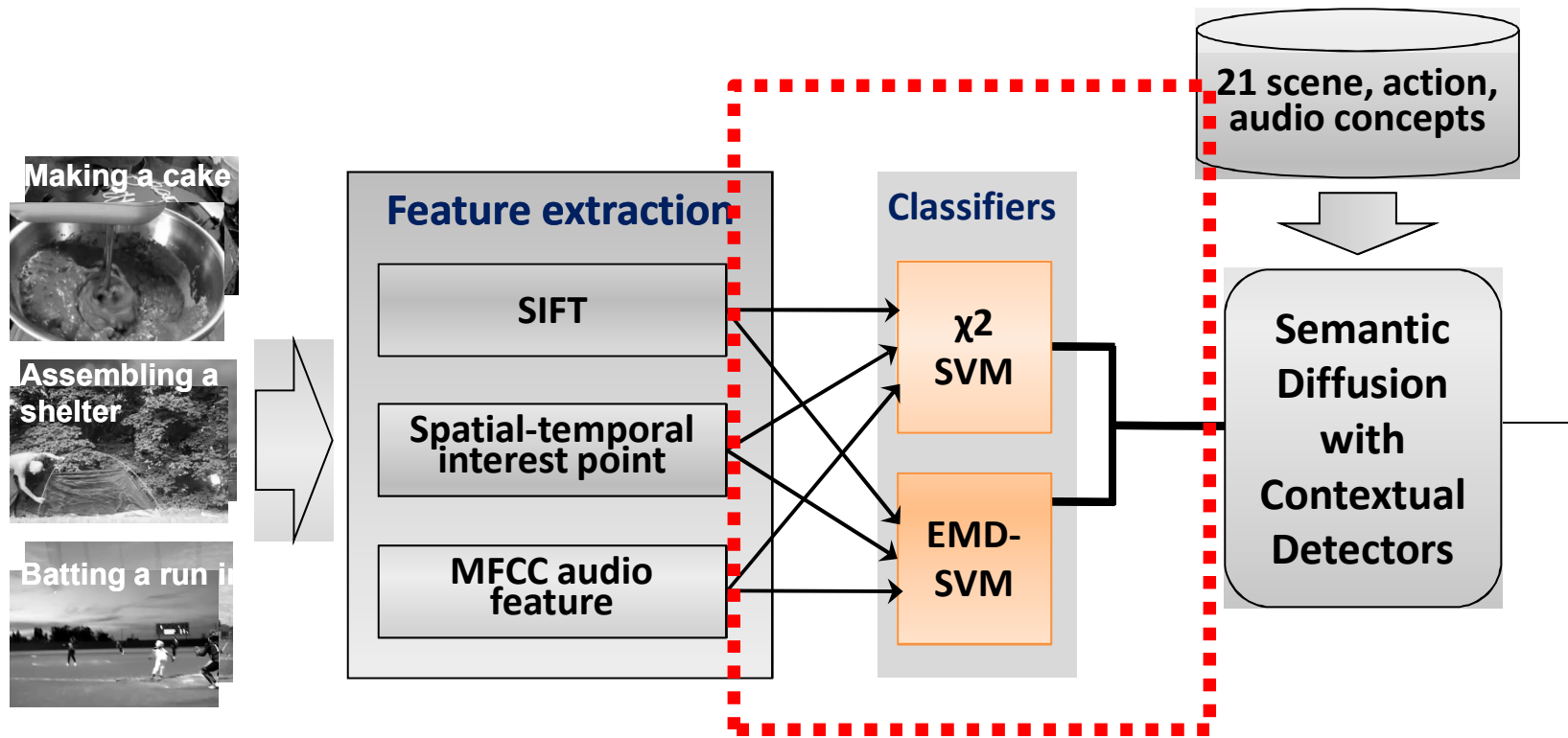
# Results of audio-visual features

- 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>

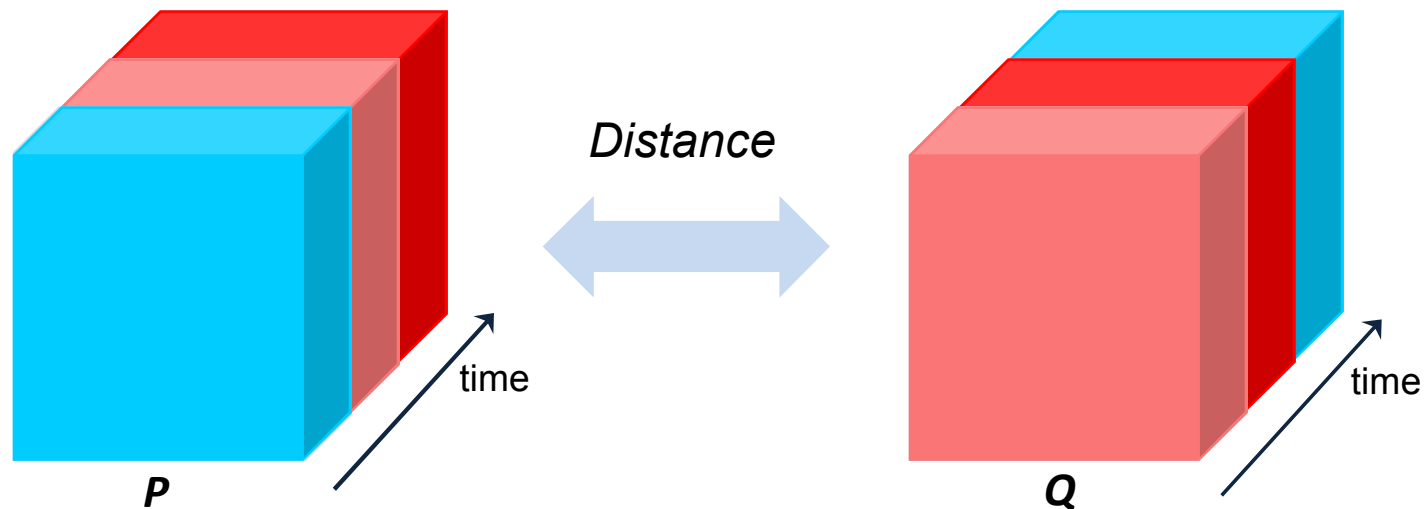
- STIP works the best for event detection
- The 3 features are **highly complementary!**

# Roadmap > temporal matching



# Temporal matching with EMD kernel

- Earth Mover's Distance (EMD)



Given two clip 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

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

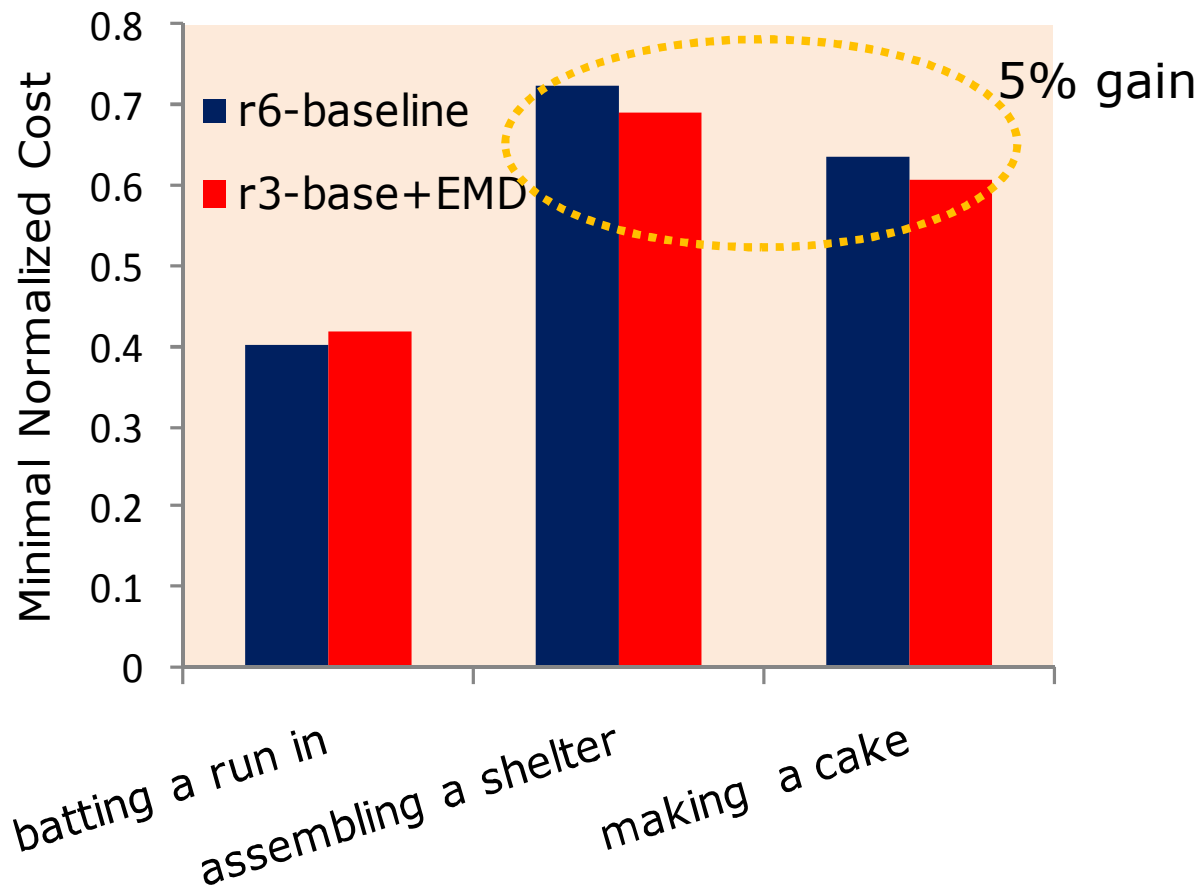
$d_{ij}$  is the  $\chi^2$  visual feature distance of video clips  $p_i$  and  $q_j$ .  $f_{ij}$  (weight transferred from  $p_i$  and  $q_j$ ) is optimized by minimizing the overall transportation workload  $\sum_i \sum_j f_{ij} d_{ij}$

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

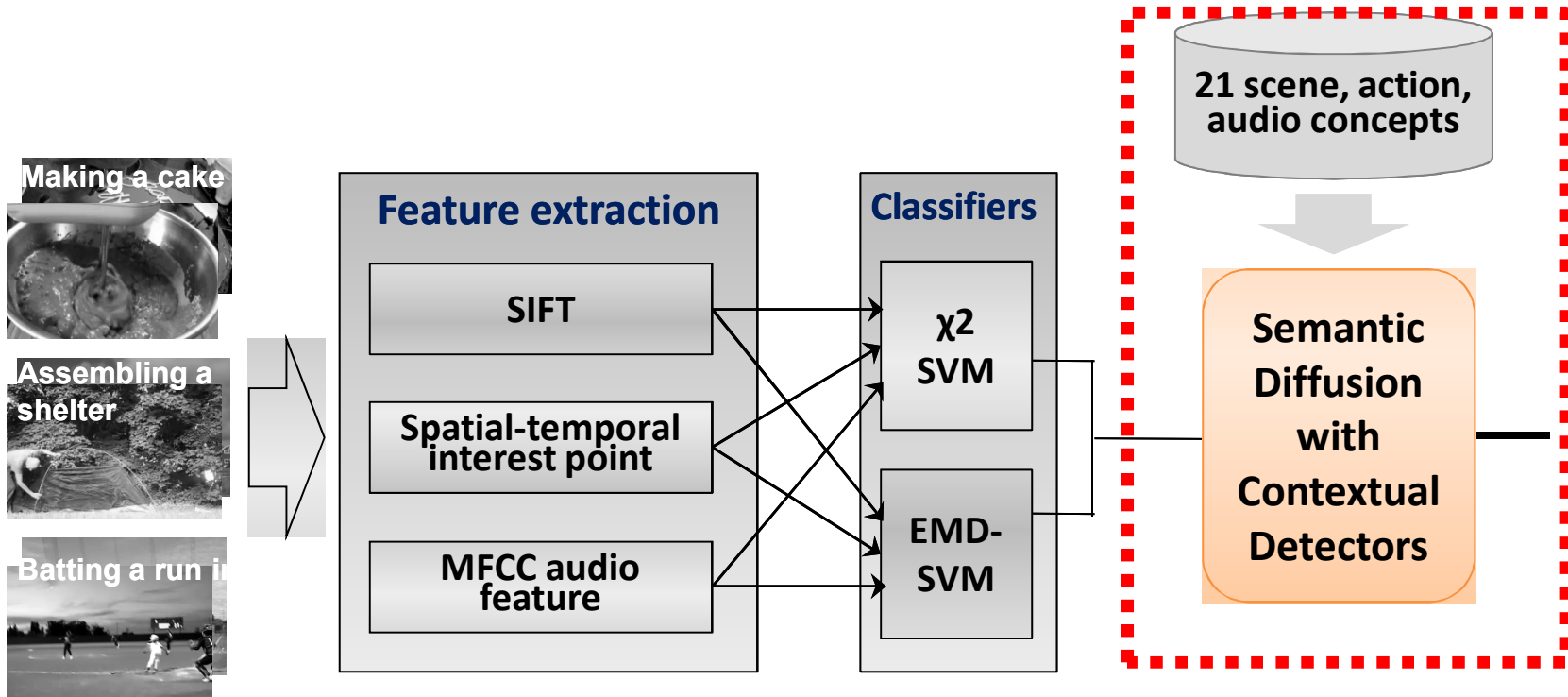


# Temporal matching results

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

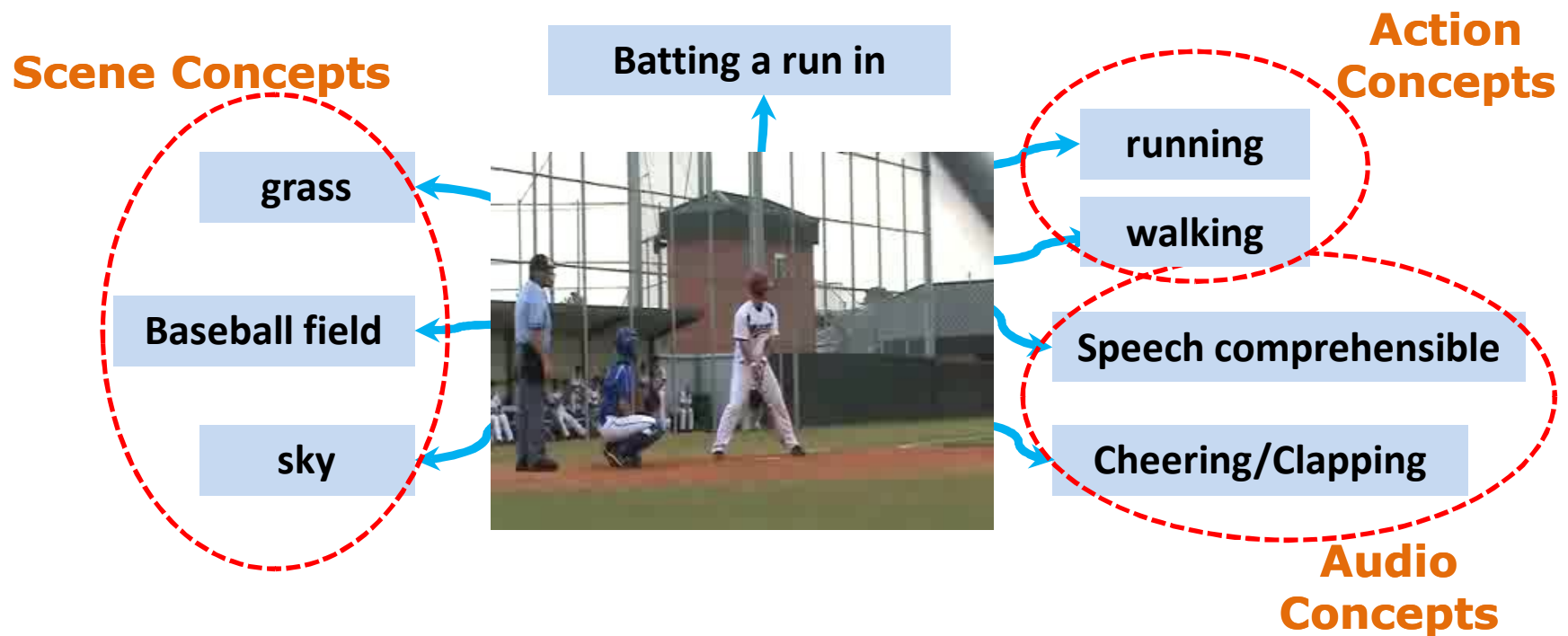


# Roadmap > contextual diffusion



# Event context

- Events generally occur under particular scene settings with certain audio sounds!
  - Understanding contexts may be helpful for event detection



# Contextual concepts

- 21 concepts are defined and annotated over TRECVID MED development set.

Human Action Concepts	Scene Concepts	Audio Concepts
<ul style="list-style-type: none"><li>▪ Person walking</li><li>▪ Person running</li><li>▪ Person squatting</li><li>▪ Person standing up</li><li>▪ Person making/assembling stuffs with hands (hands visible)</li><li>▪ Person batting baseball</li></ul>	<ul style="list-style-type: none"><li>▪ Indoor kitchen</li><li>▪ Outdoor with grass/trees visible</li><li>▪ Baseball field</li><li>▪ Crowd (a group of 3+ people)</li><li>▪ Cakes (close-up view)</li></ul>	<ul style="list-style-type: none"><li>▪ Outdoor rural</li><li>▪ Outdoor urban</li><li>▪ Indoor quiet</li><li>▪ Indoor noisy</li><li>▪ Original audio</li><li>▪ Dubbed audio</li><li>▪ Speech comprehensible</li><li>▪ Music</li><li>▪ Cheering</li><li>▪ Clapping</li></ul>

- SVM classifier for concept detection
  - STIP for action concepts, SIFT for scene concepts, and MFCC for audio concepts

# Concept detection: example results

Baseball field



Cakes  
(close-up view)



Crowd  
(3+ people)



Grass/trees



Indoor kitchen



# Contextual diffusion model

- Semantic diffusion

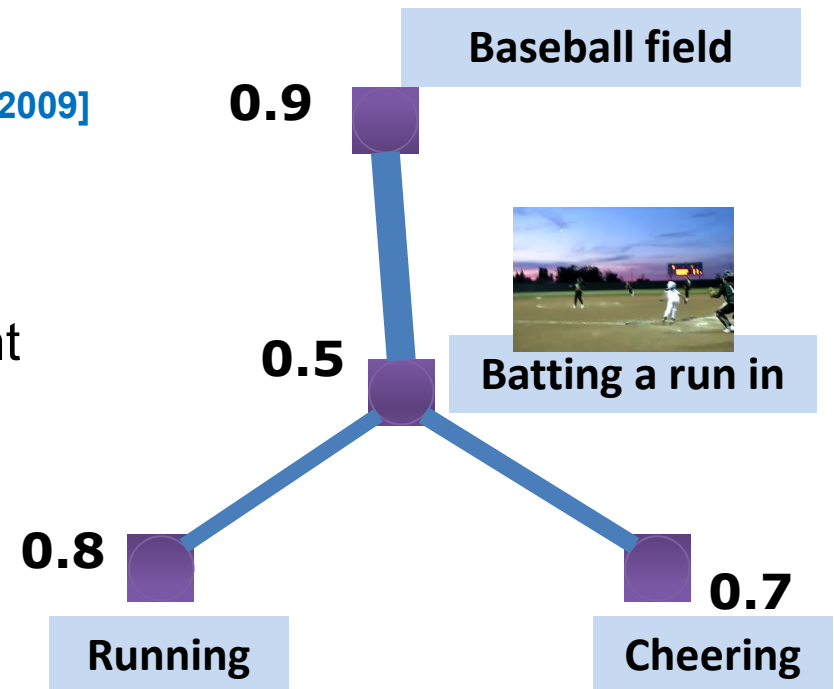
[Y.-G. Jiang, J. Wang, S.F. Chang & C.W. Ngo, ICCV 2009]

- Semantic graph

- Nodes are concepts/events
    - Edges represent concept/event correlation

- Graph diffusion

- Smooth detection scores w.r.t. the correlation



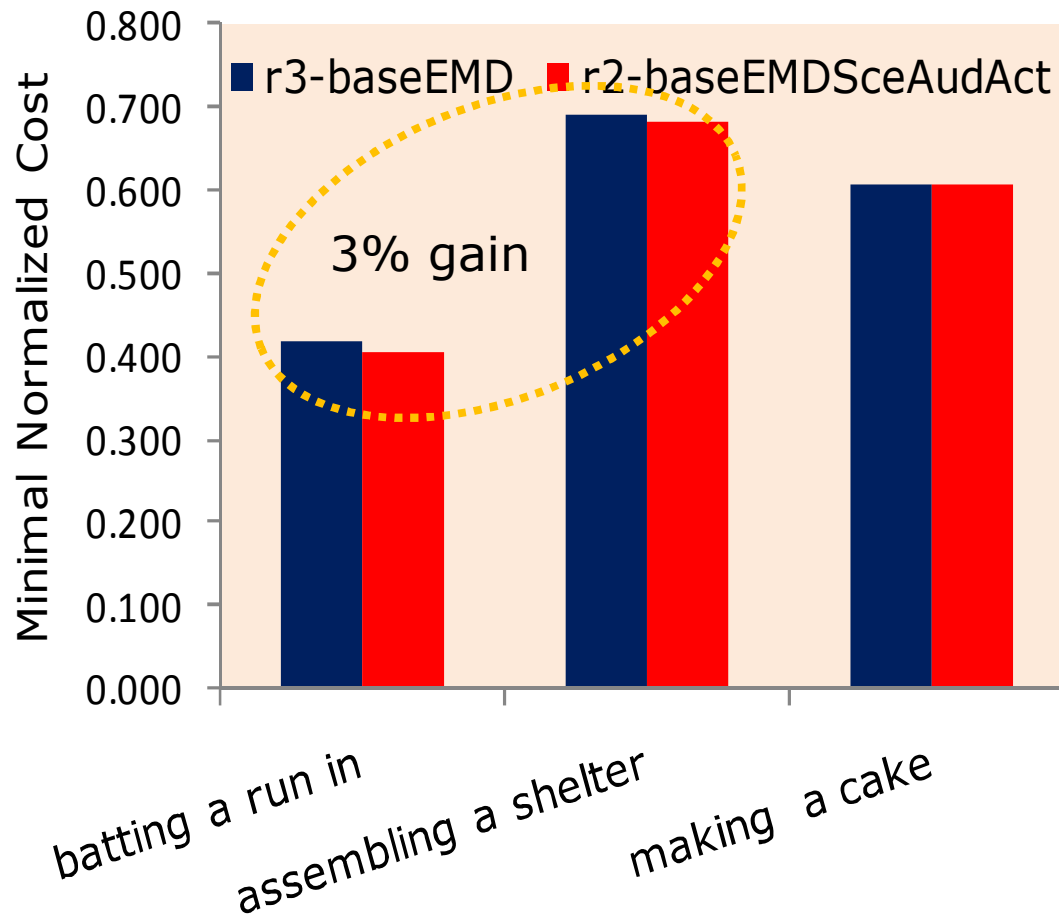
**Project page and source code:**

<http://www.ee.columbia.edu/ln/dvmm/researchProjects/MultimediaIndexing/DASD/dasd.htm>



# Contextual diffusion results

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



# Outline

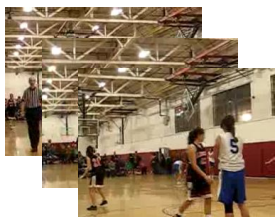
- A System for Recognizing Events in Internet Videos
  - Best performance in TRECVID 2010 Multimedia Event Detection Task
  - Features, Kernels, Context, etc.
- Internet Consumer Video Analysis
  - A Benchmark Database
  - An Evaluation of Human & Machine Performance

# What are Consumer Videos?

- Original unedited videos captured by ordinary consumers
  - Interesting and very diverse contents
  - Very weakly indexed
    - On average, 3 tags per consumer video on YouTube **vs.** 9 tags each YouTube video has
  - Original audio tracks are preserved; good for audio-visual joint analysis



# Columbia Consumer Video (CCV) Database



Basketball



Skiing



Dog



Wedding Reception



Non-music Performance



Baseball



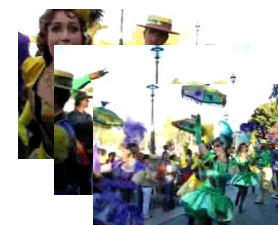
Swimming



Bird



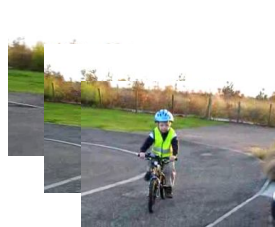
Wedding Ceremony



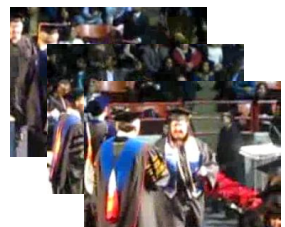
Parade



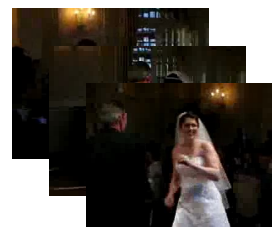
Soccer



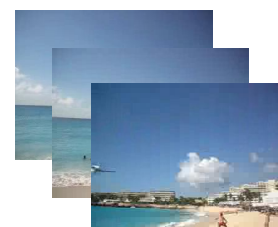
Biking



Graduation



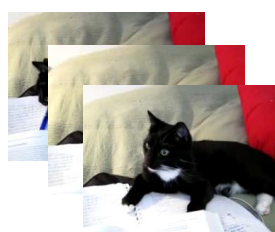
Wedding Dance



Beach



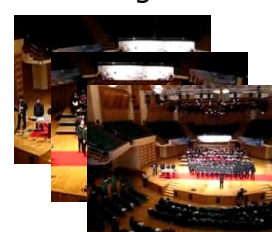
Ice Skating



Cat



Birthday Celebration



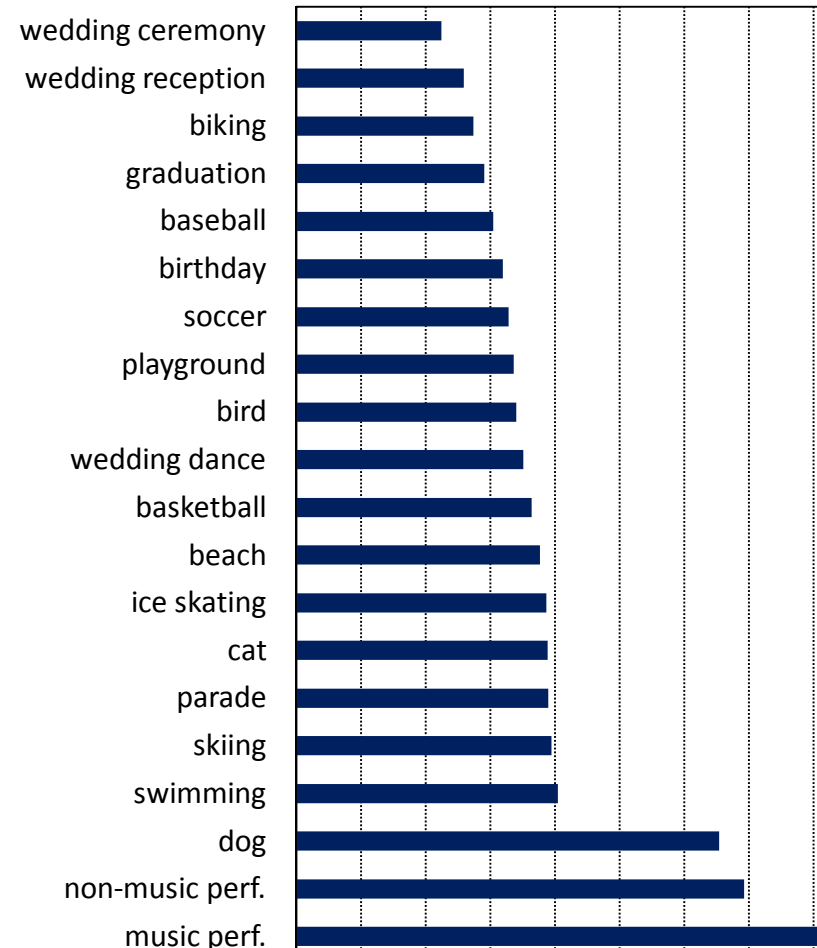
Music Performance



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".

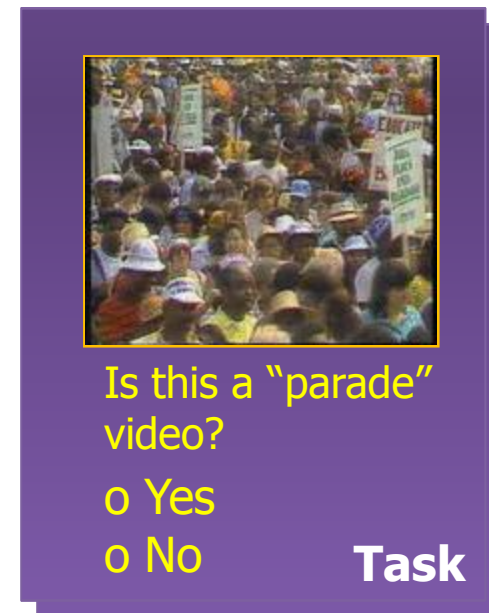
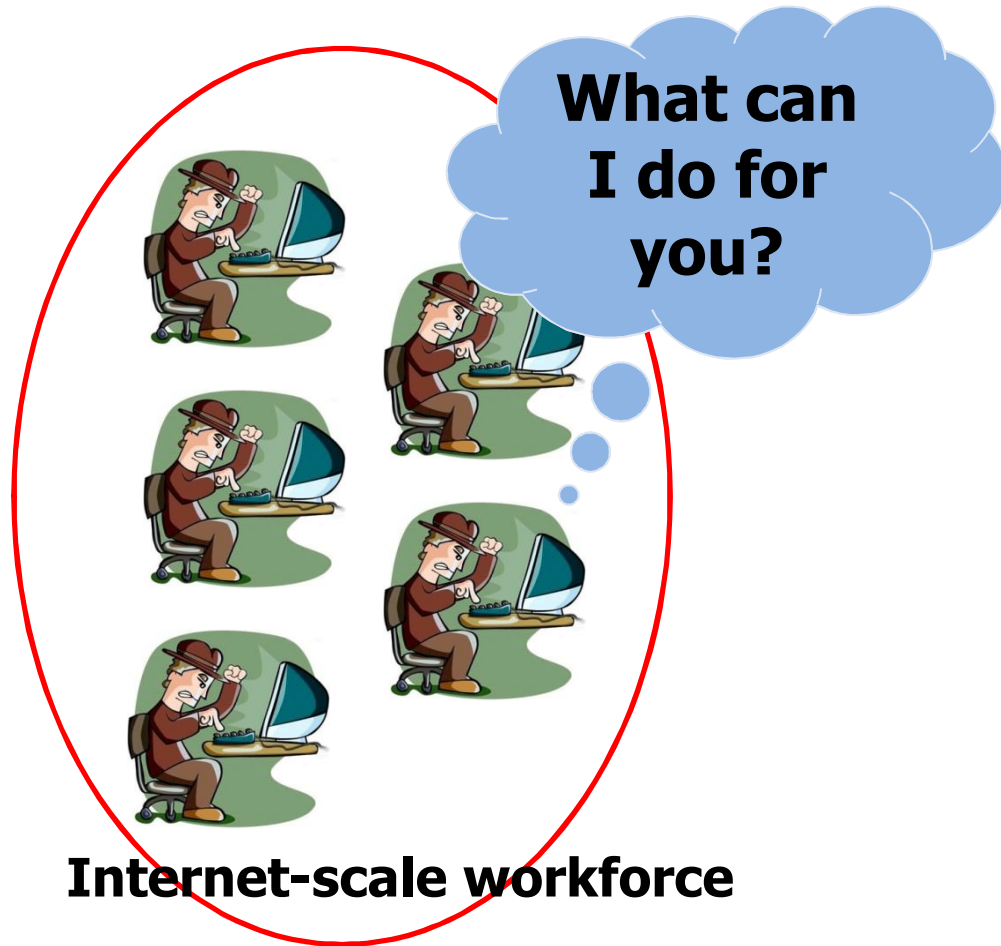
# Existing Database?

	<b><u>CCV Database</u></b>
<ul style="list-style-type: none"><li>• Human Action Recognition<ul style="list-style-type: none"><li>– KTH &amp; Weizmann<ul style="list-style-type: none"><li>• (constrained environment) <u>2004-05</u></li></ul></li><li>– Hollywood Database<ul style="list-style-type: none"><li>• (12 categories, movies) <u>2008</u></li></ul></li><li>– UCF Database<ul style="list-style-type: none"><li>• (50 categories, YouTube Videos) <u>2010</u></li></ul></li></ul></li></ul>	<b>Unconstrained YouTube videos</b>
<ul style="list-style-type: none"><li>• Kodak Consumer Video<ul style="list-style-type: none"><li>• (25 classes, 1300+ videos) <u>2007</u></li></ul></li></ul>	<b>More videos &amp; better defined categories</b>
<ul style="list-style-type: none"><li>• LabelMe Video<ul style="list-style-type: none"><li>• (many classes, 1300+ videos) <u>2009</u></li></ul></li></ul>	<b>More videos &amp; larger content variations</b>
<ul style="list-style-type: none"><li>• TRECVID MED 2010<ul style="list-style-type: none"><li>• (3 classes, 3400+ videos) <u>2010</u></li></ul></li></ul>	<b>More videos &amp; categories</b>



# Crowdsourcing: Amazon Mechanical Turk

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



**\$?.??**  
**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
<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.		

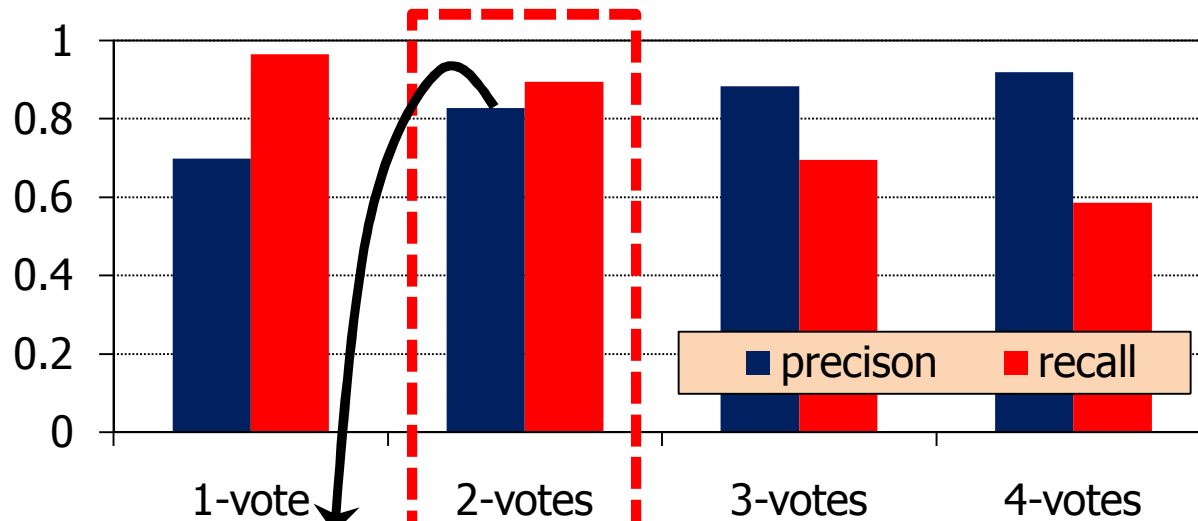
Current Time: 10 sec

**\$ 0.02**

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

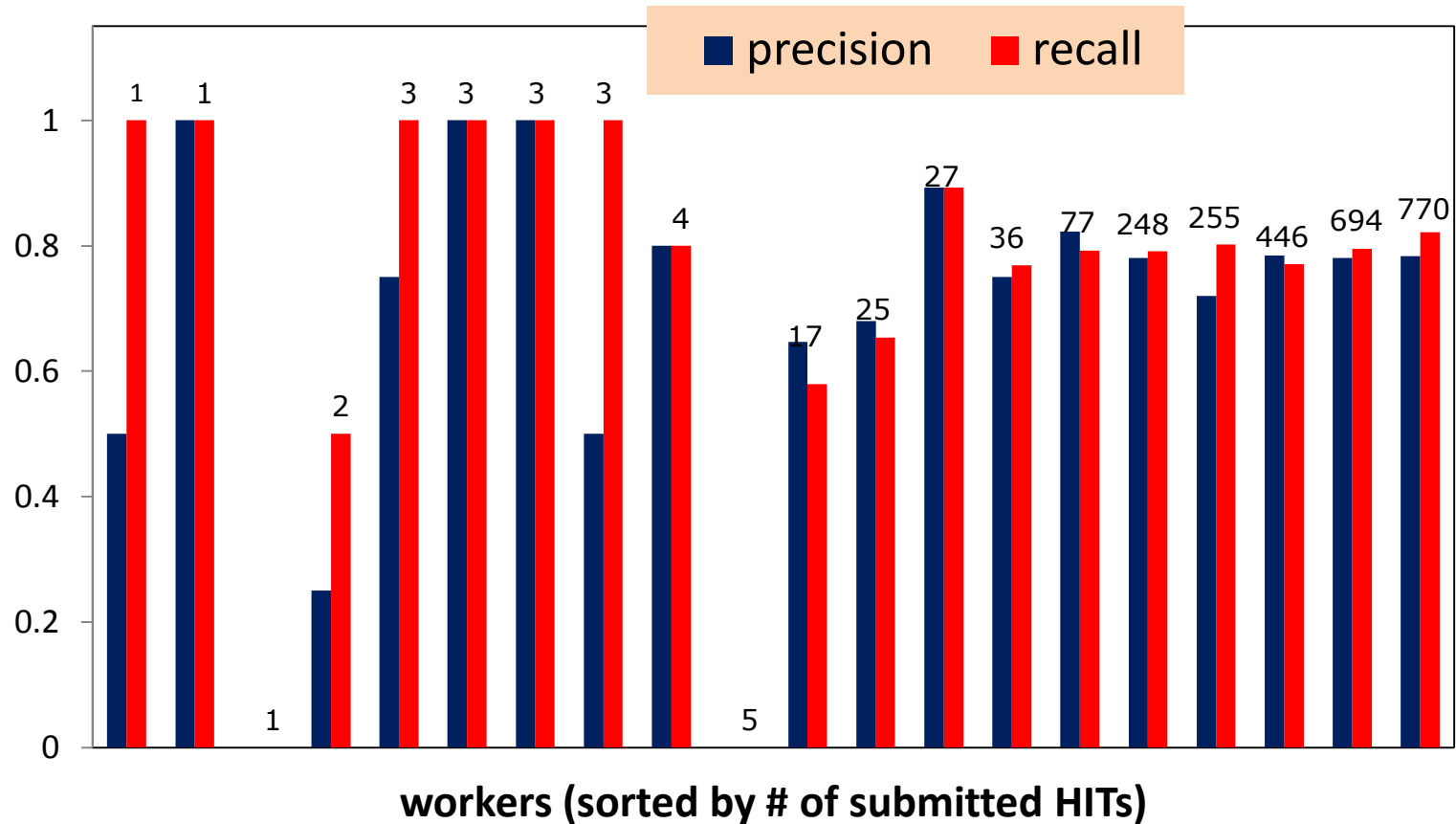
# Human Recognition Performance

- 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

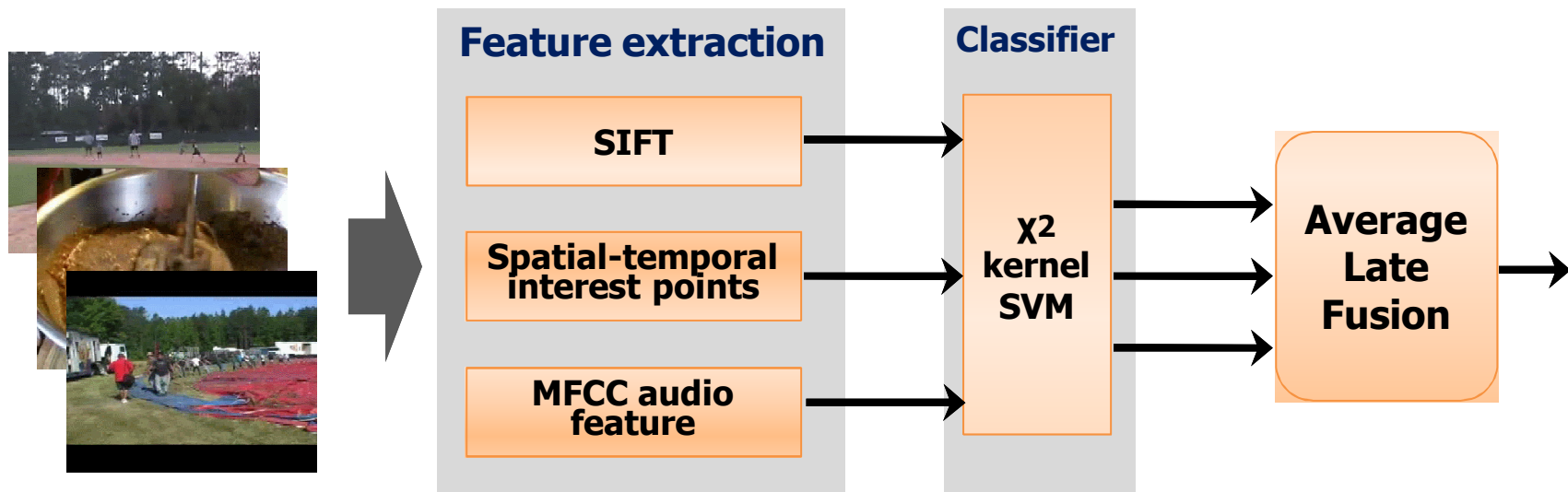


Plus additional manual filtering of 6 positive sample sets: 94% final precision

# Human Recognition Performance (cont.)



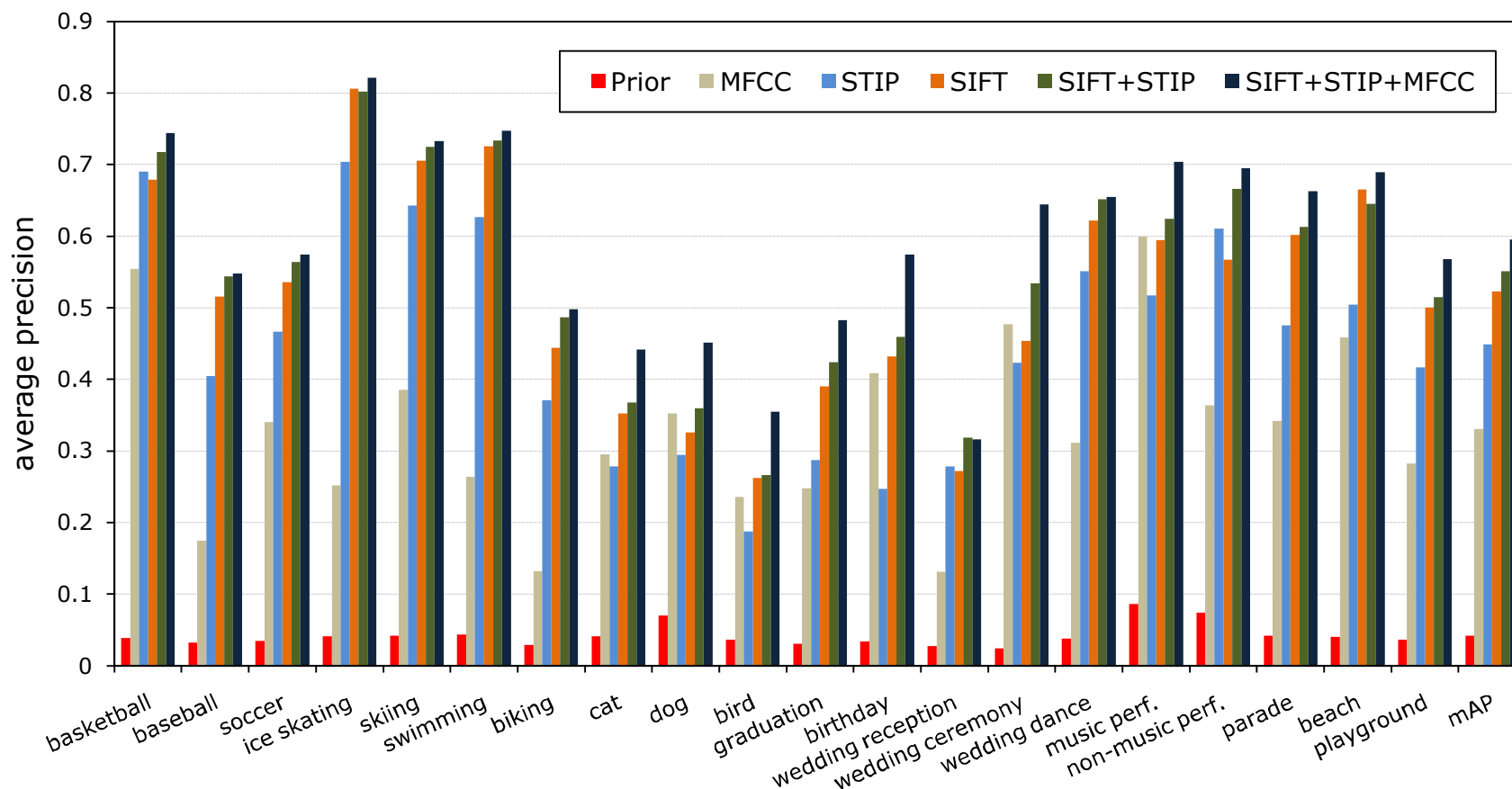
# Machine Recognition System



Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Subh Bhattacharya, Dan Ellis, Mubarak Shah, Shih-Fu Chang, **Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching**, NIST TRECVID Workshop, 2010.

# Machine Recognition Accuracy

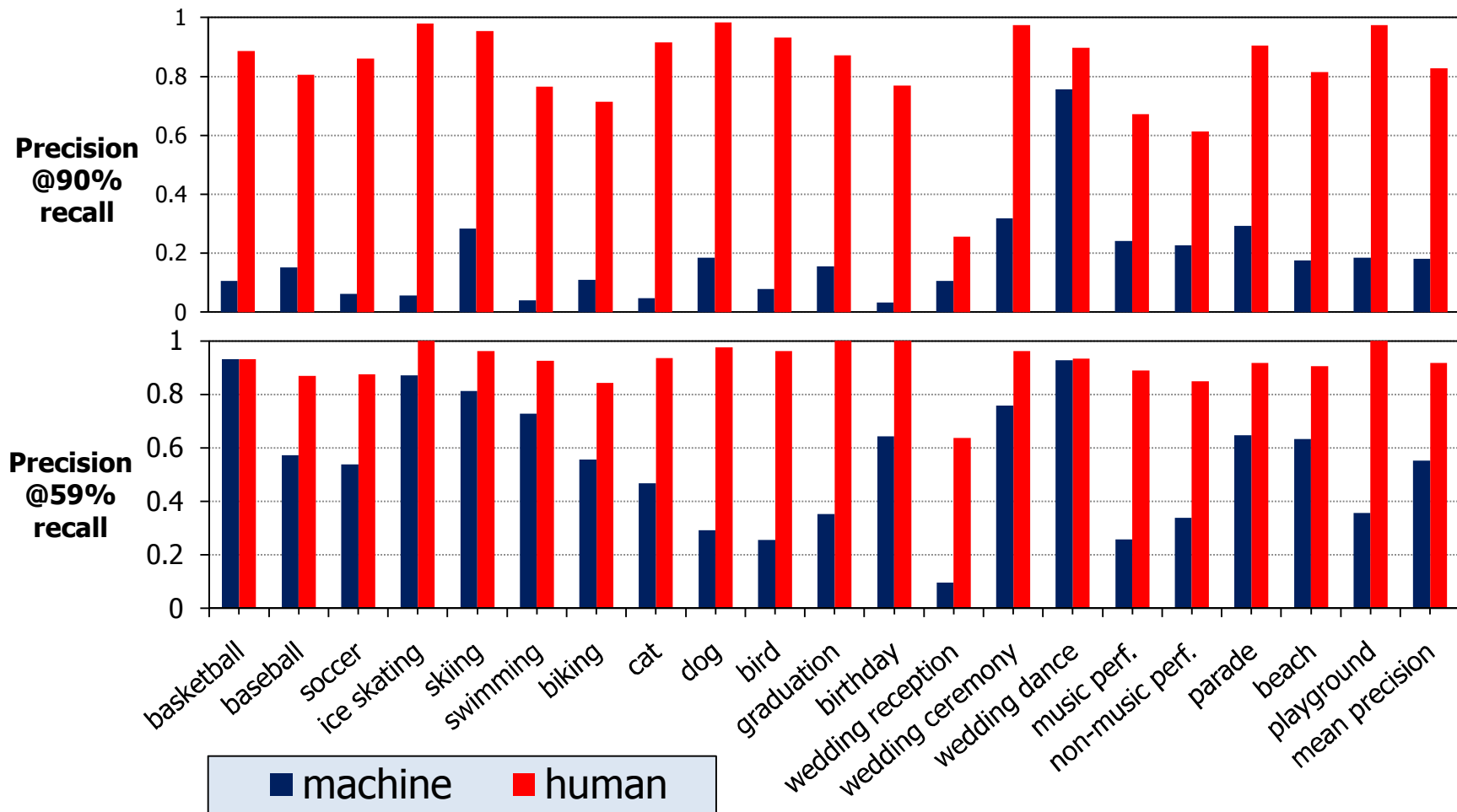
- Measured by average precision
  - SIFT works the best for event detection
  - The 3 features are **highly complementary!**





# Human vs. Machine

- Human has much **better recall**, and is much **better for non-rigid objects**
- Machine is **close to human on top-list precision**



# Human vs. Machine: Result Examples

	true positives			false positives	
	found by human&machine	found by human only	found by machine only	found by human only	found by machine only
wedding dance	 	 	 		 
soccer	 	 	n/a	 	 
cat	 	 	n/a		 

# Summary

- The combination of the three audio-visual features is key for good video event recognition performance
- Temporal matching is useful for some complex events
- Current automatic event recognition methods are not that bad
- A new dataset (CCV) for consumer video analysis

# Dataset download

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

<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 audio/visual feature representations to the community for research usage.

**CCV Snapshot**

- # videos: 9,317 (210 hrs in total)
- # categories: 20 (semantic categories)

**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 (CMR), Trento, Italy, April 2011.

**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.

**THANK YOU!**  
THANK YOU!

email: [yjiang@ee.columbia.edu](mailto:yjiang@ee.columbia.edu)