Recognizing Complex Events in Internet Videos with Audio-Visual Features

Yu-Gang Jiang

yjiang@ee.columbia.edu

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

Columbia University[1] University of Central Florida[2] Kodak Research Labs[3]

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

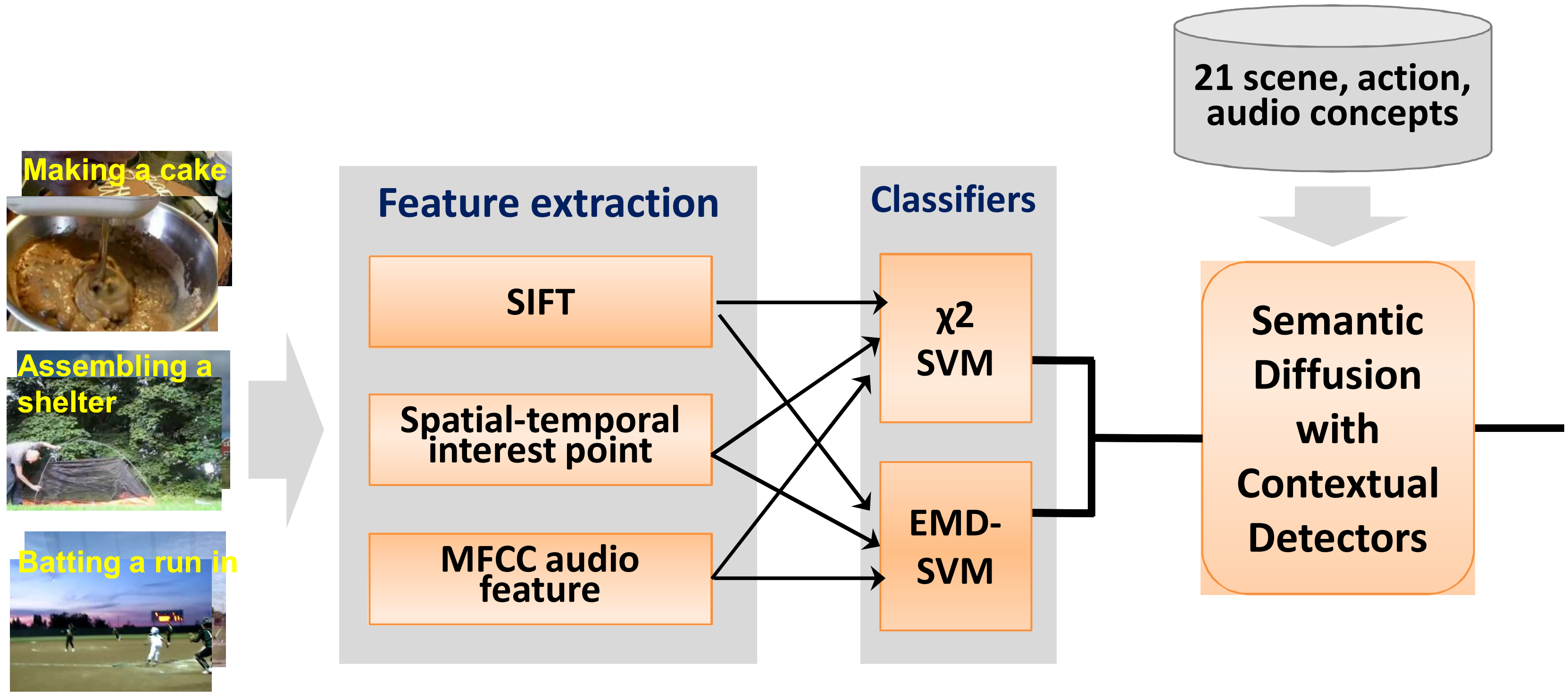
- Target: Find videos containing an event of interest
- Data: unconstrained Internet videos
  - 1700+ training videos (~50 positive each event); 1700+ test videos

Making a cake

Assembling a shelter

Batting a run in

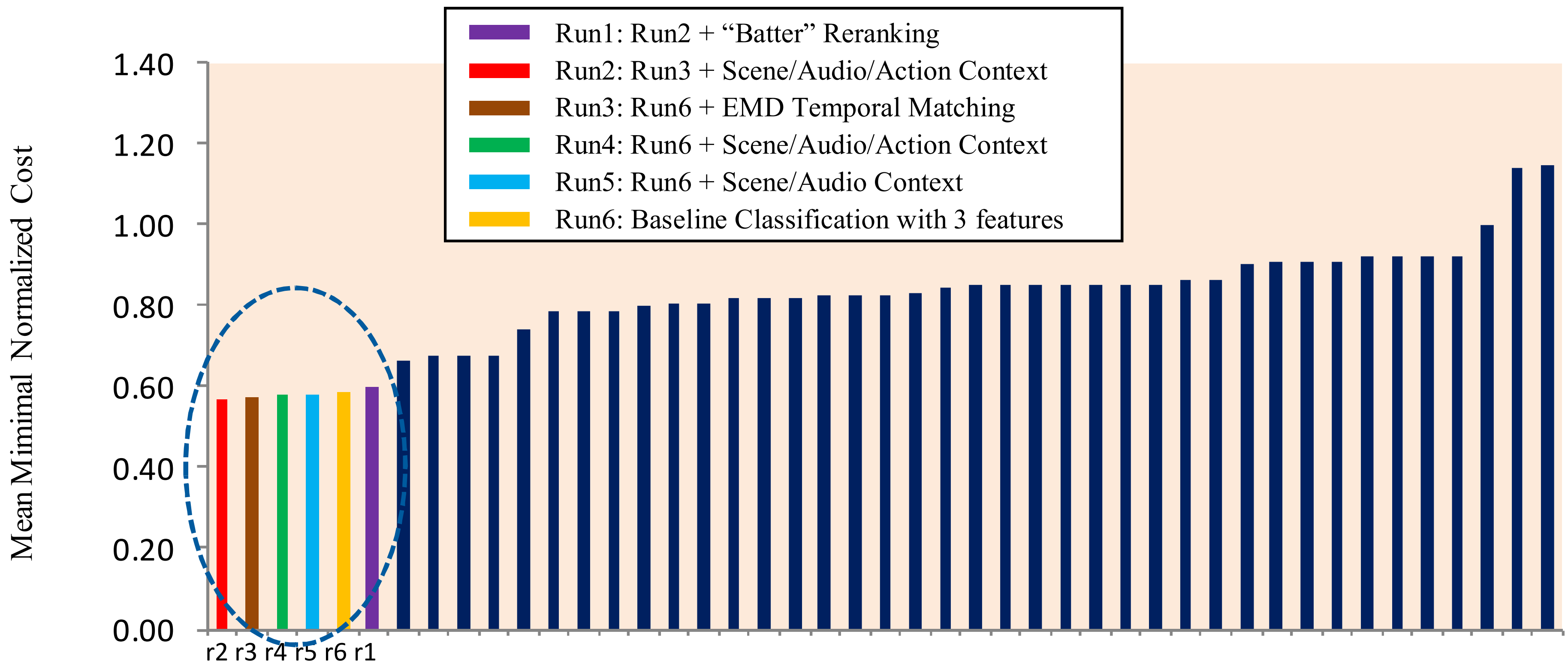
The system: 3 major components



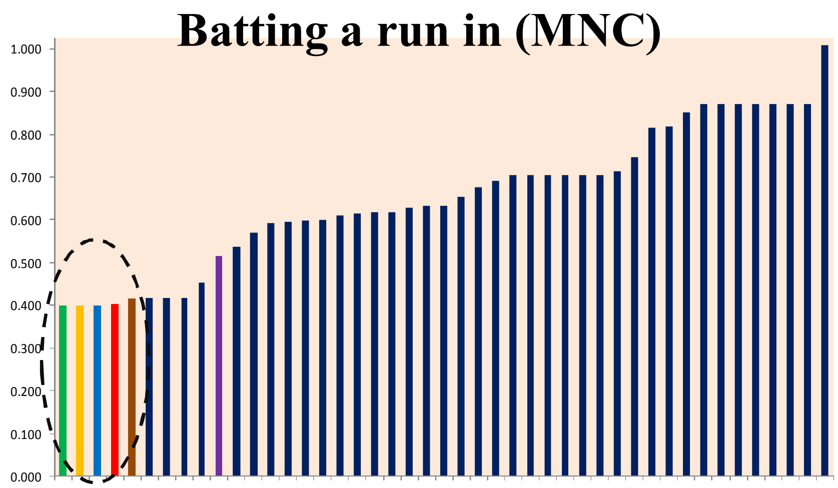
Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, S. Bhattacharya, Dan Ellis, Mubarak Shah, Shih-Fu Chang, **Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching**, in TRECVID 2010.

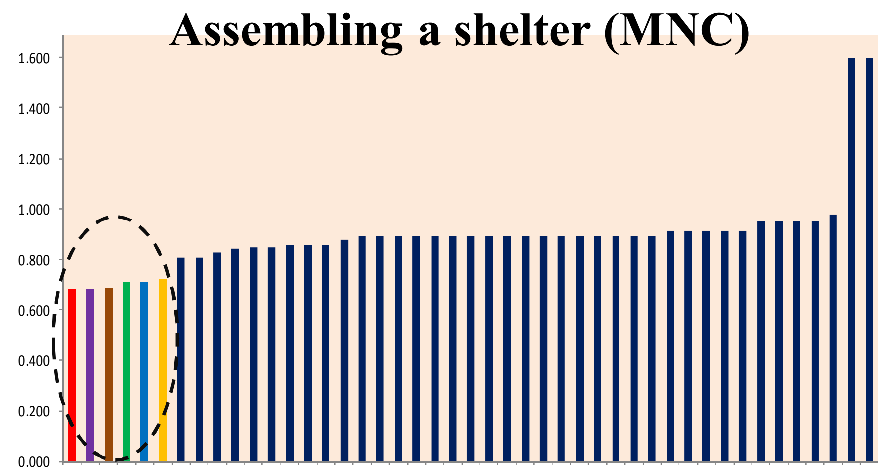
Best performance in TRECVID2010

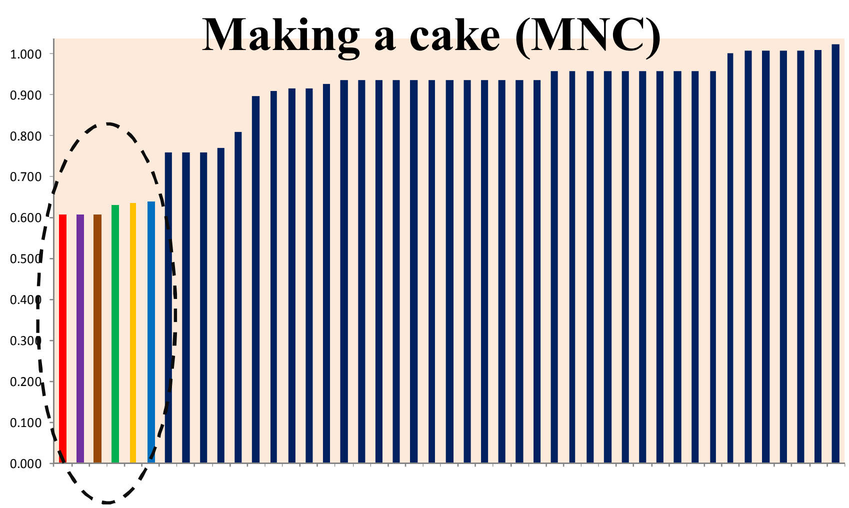
Multimedia event detection (MED) task



Per-event performance

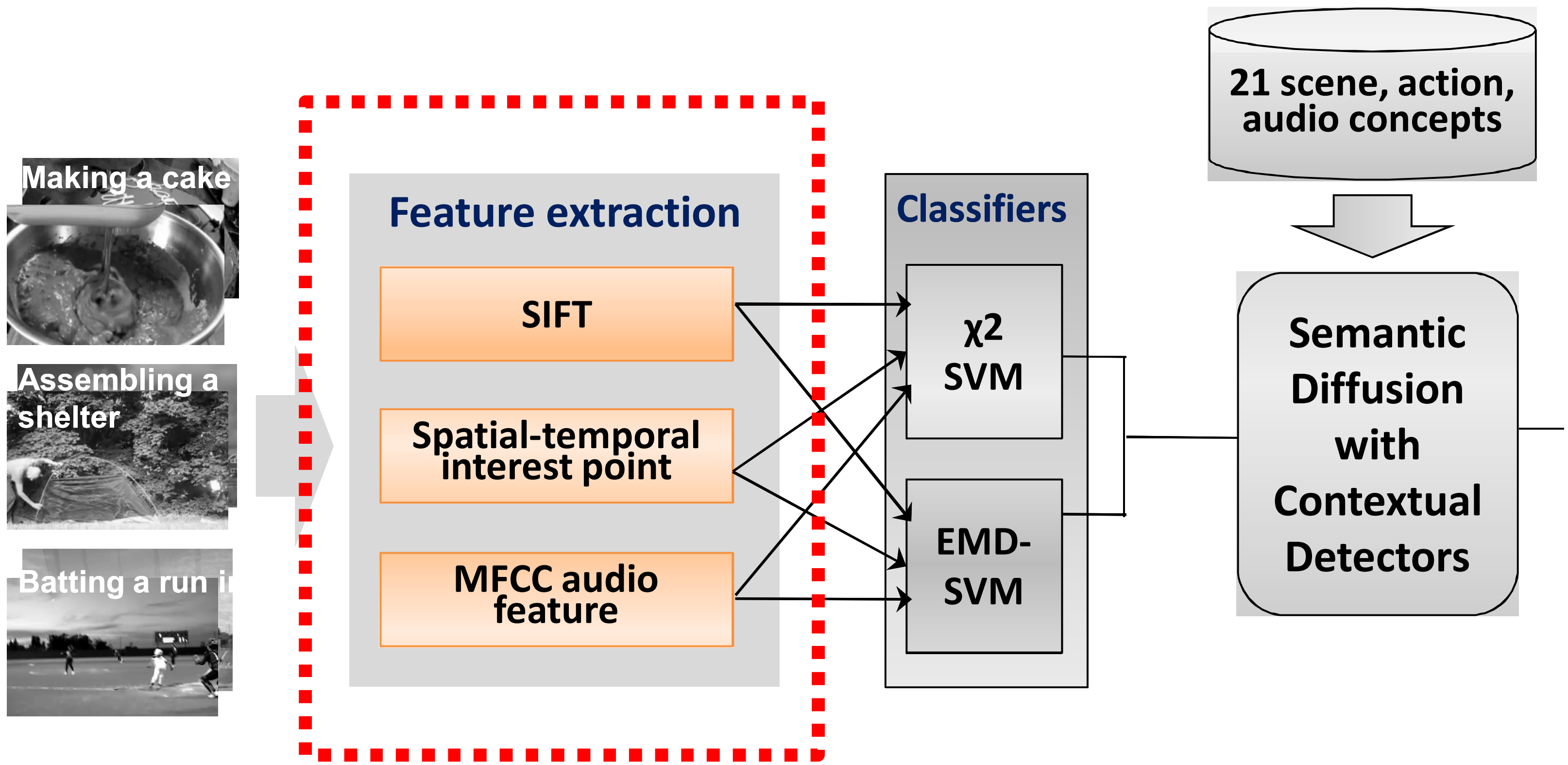






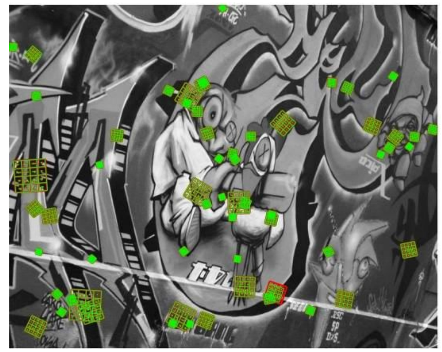
- Run1: Run2 + "Batter" Reranking
- Run2: Run3 + Scene/Audio/Action Context
- Run3: Run6 + EMD Temporal Matching
- Run4: Run6 + Scene/Audio/Action Context
- Run5: Run6 + Scene/Audio Context
- Run6: Baseline Classification with 3 features

Roadmap > audio-visual features

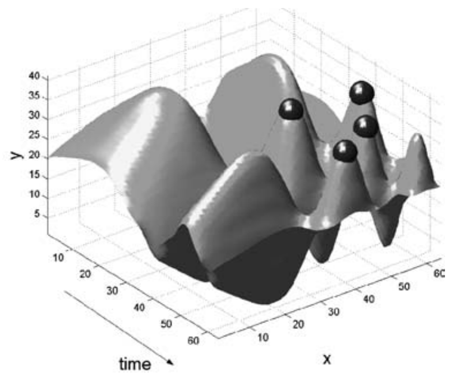


Three audio-visual features…

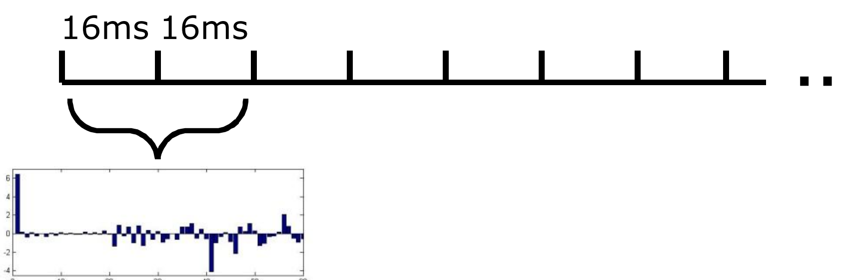
- *SIFT (visual)*
  - – *D. Lowe, IJCV 04.*



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



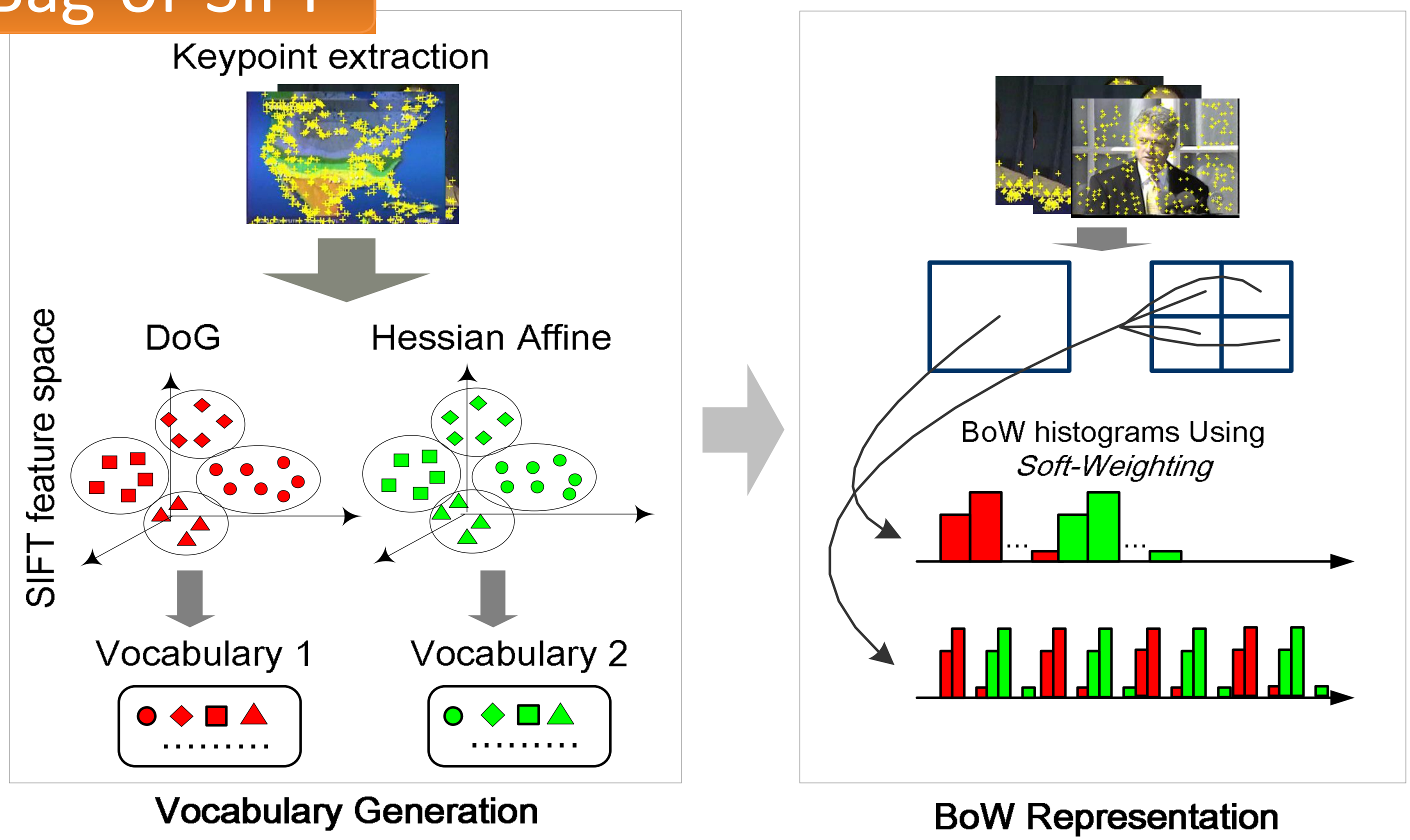
- *MFCC (audio)*



Bag-of-X representation

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

Bag-of-SIFT



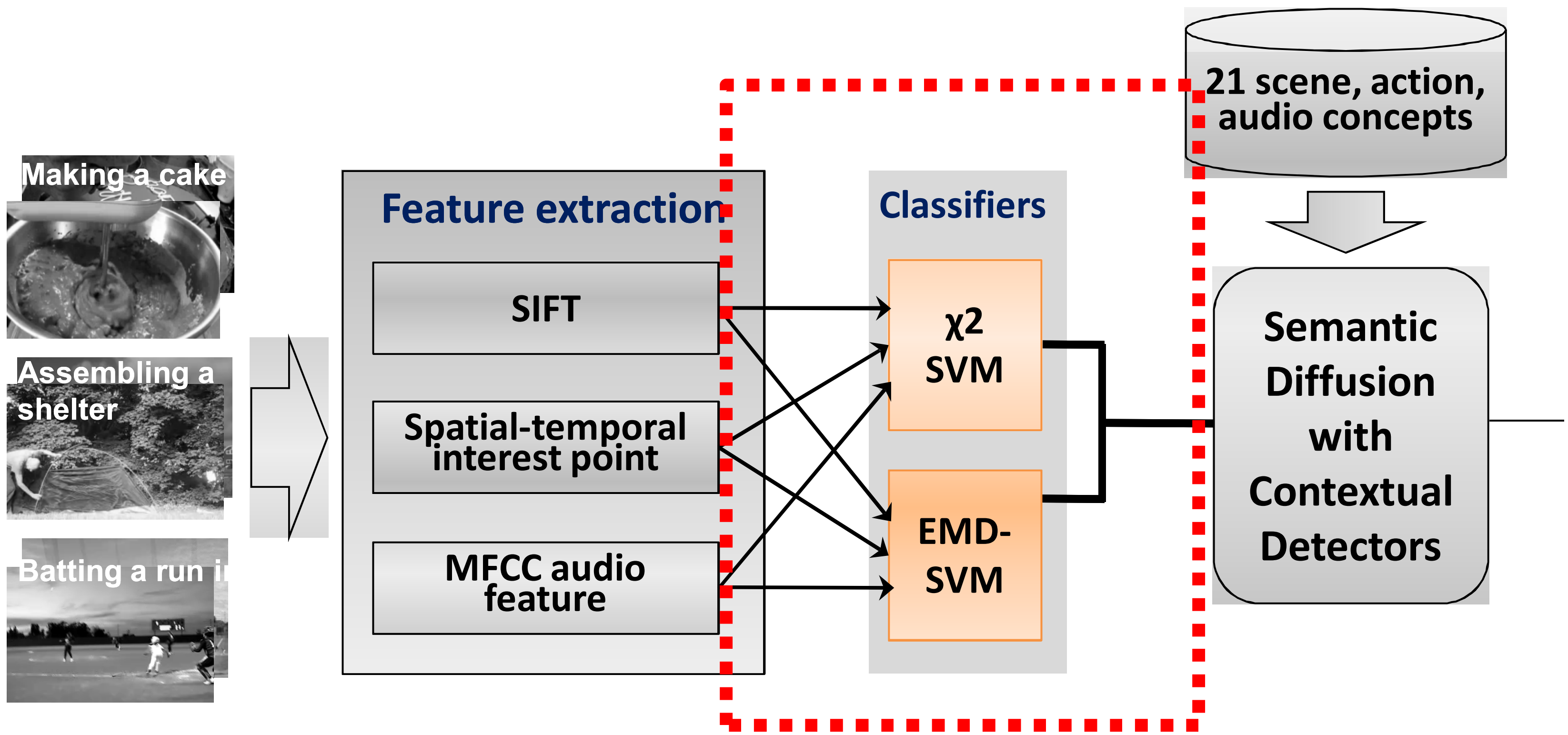
Results of audio-visual features

- Measured by Average Precision (AP)

	Assembling a shelter	Batting a run in	Making a cake	*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**

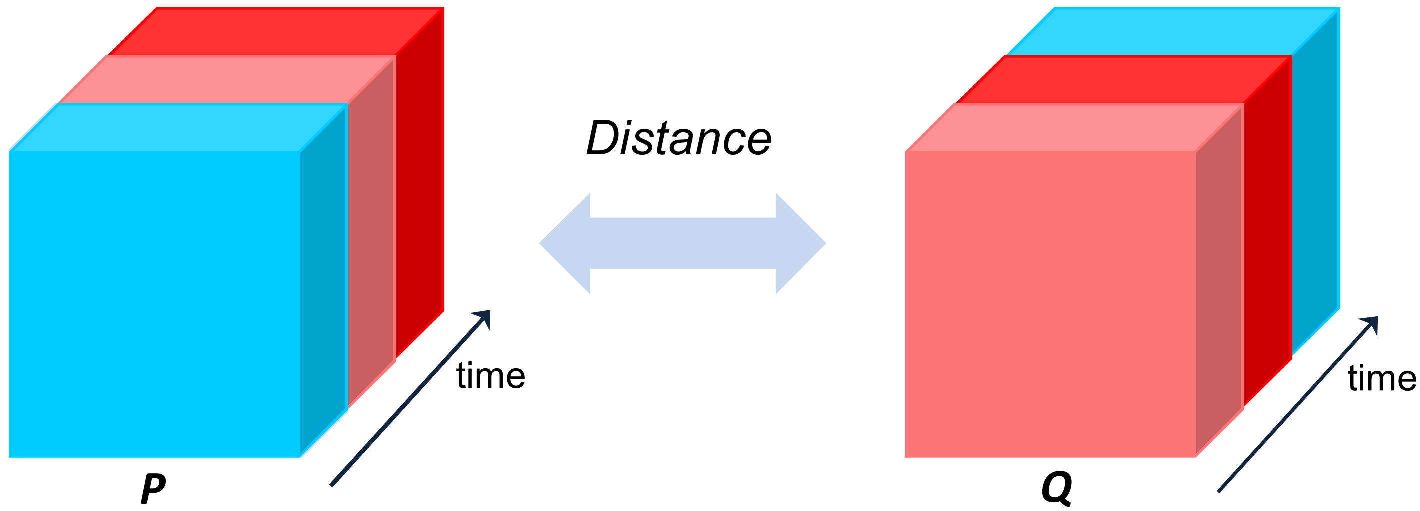
- STIP works the best for event detection
- The 3 features are highly complementary!

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

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

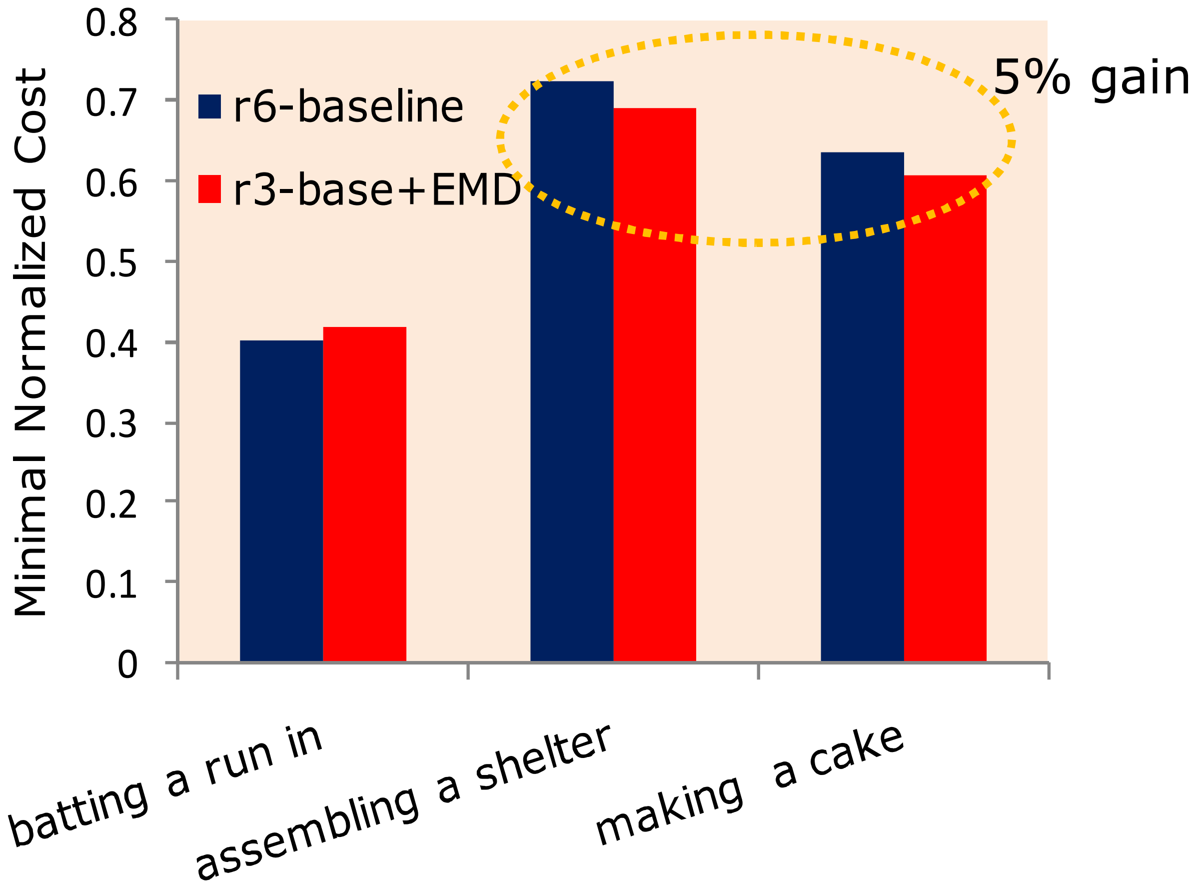
d_{ij} is the χ^2 visual feature distance of video clips p_i and q_j. f_{ij} (weight transferred from p_i and q_j) is optimized by minimizing the overall transportation workload $\Sigma_i \Sigma_j f_{ij} d_{ij}$

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

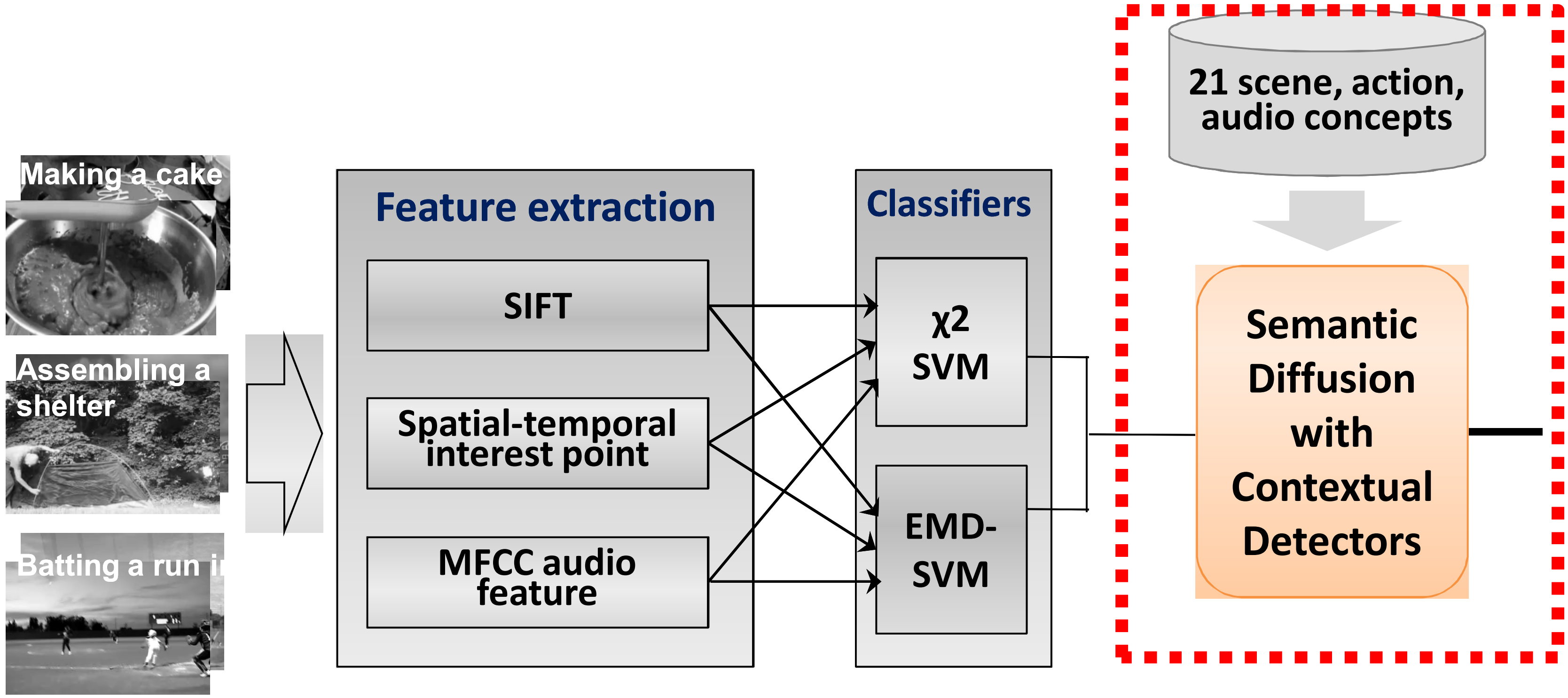
Y. Rubner, C. Tomasi, L. J. Guibas, "A metric for distributions with applications to image databases", ICCV, 1998.
D. Xu, S.-F. Chang, "Video event recognition using kernel methods with multi-level temporal alignment", PAMI, 2008.

Temporal matching results

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

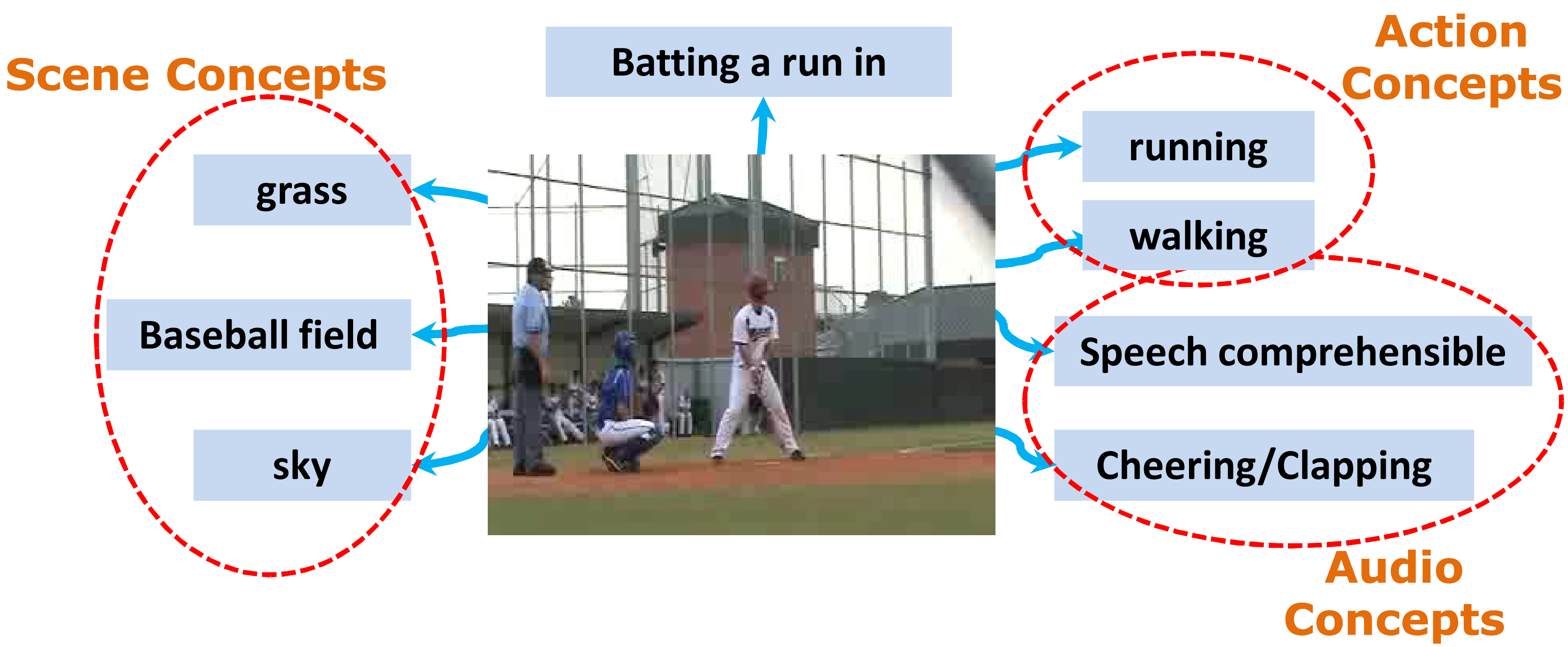


Roadmap > contextual diffusion



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
▪ Person walking ▪ Person running ▪ Person squatting ▪ Person standing up ▪ Person making/assembling stuffs with hands (hands visible) ▪ Person batting baseball	▪ Indoor kitchen ▪ Outdoor with grass/trees visible ▪ Baseball field ▪ Crowd (a group of 3+ people) ▪ Cakes (close-up view)	▪ Outdoor rural ▪ Outdoor urban ▪ Indoor quiet ▪ Indoor noisy ▪ Original audio ▪ Dubbed audio ▪ Speech comprehensible ▪ Music ▪ Cheering ▪ Clapping

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

Concept detection: example results

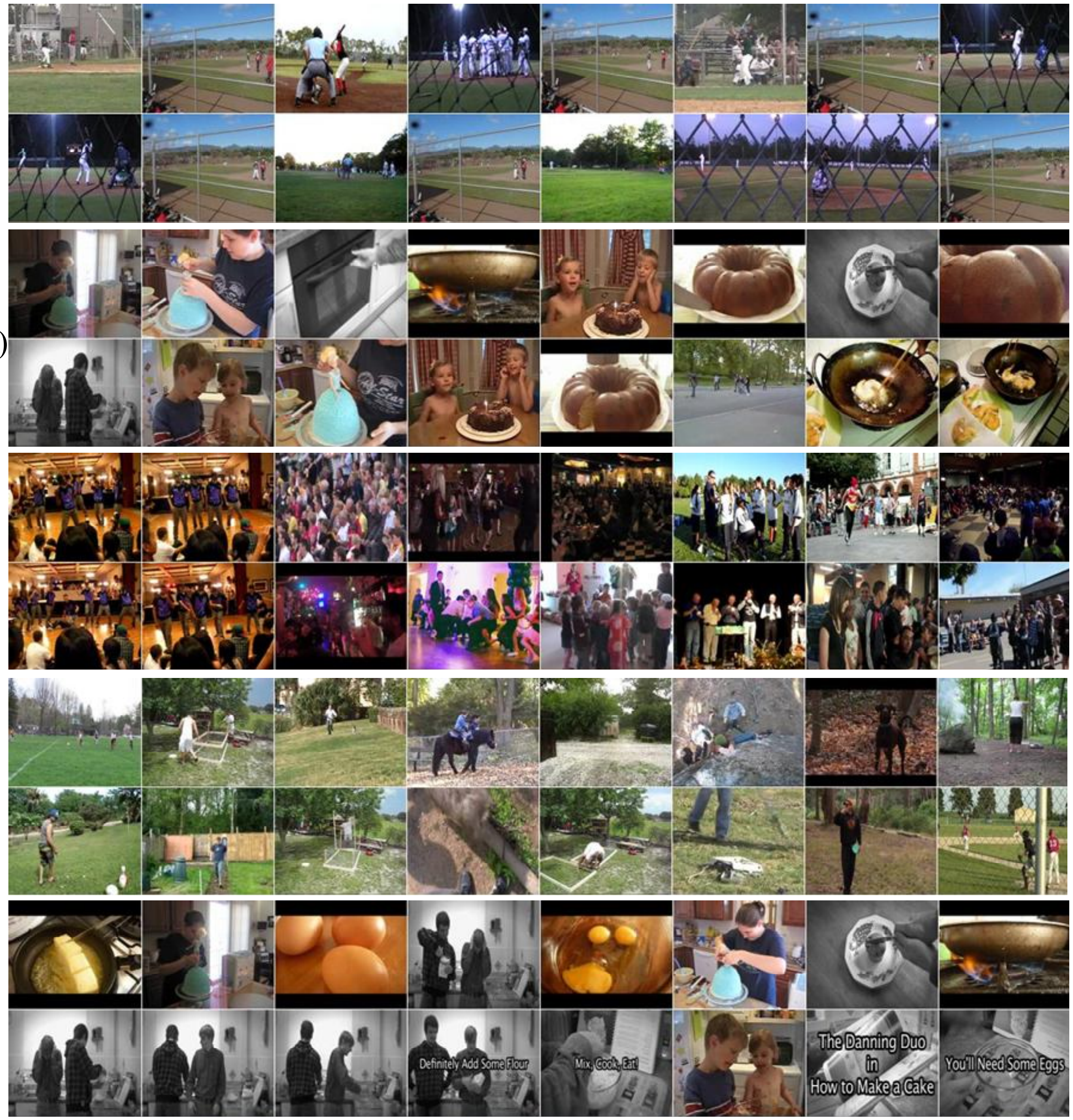
Baseball field

Cakes
(close-up view)

Crowd
(3+ people)

Grass/trees

Indoor kitchen

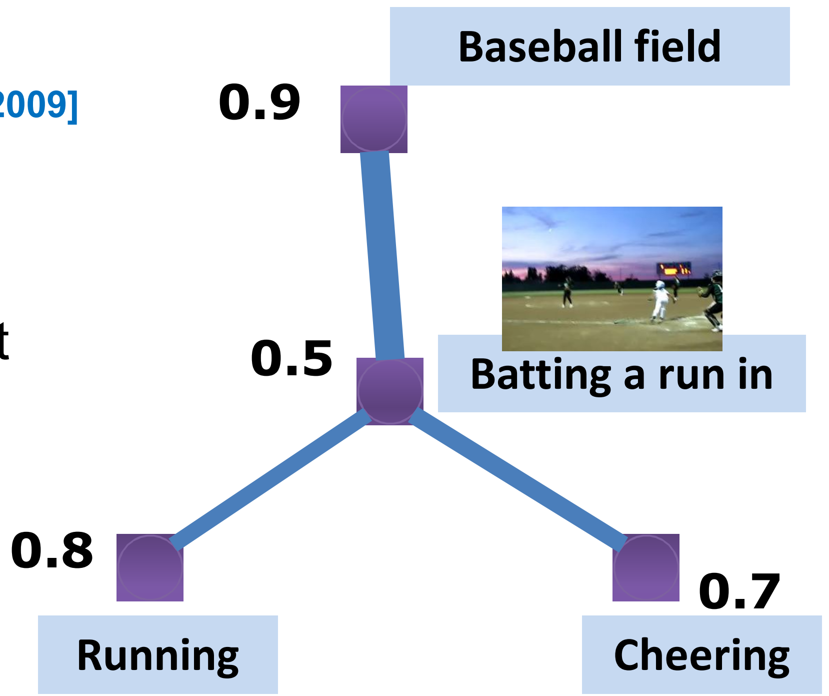


Contextual diffusion model

- Semantic diffusion

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

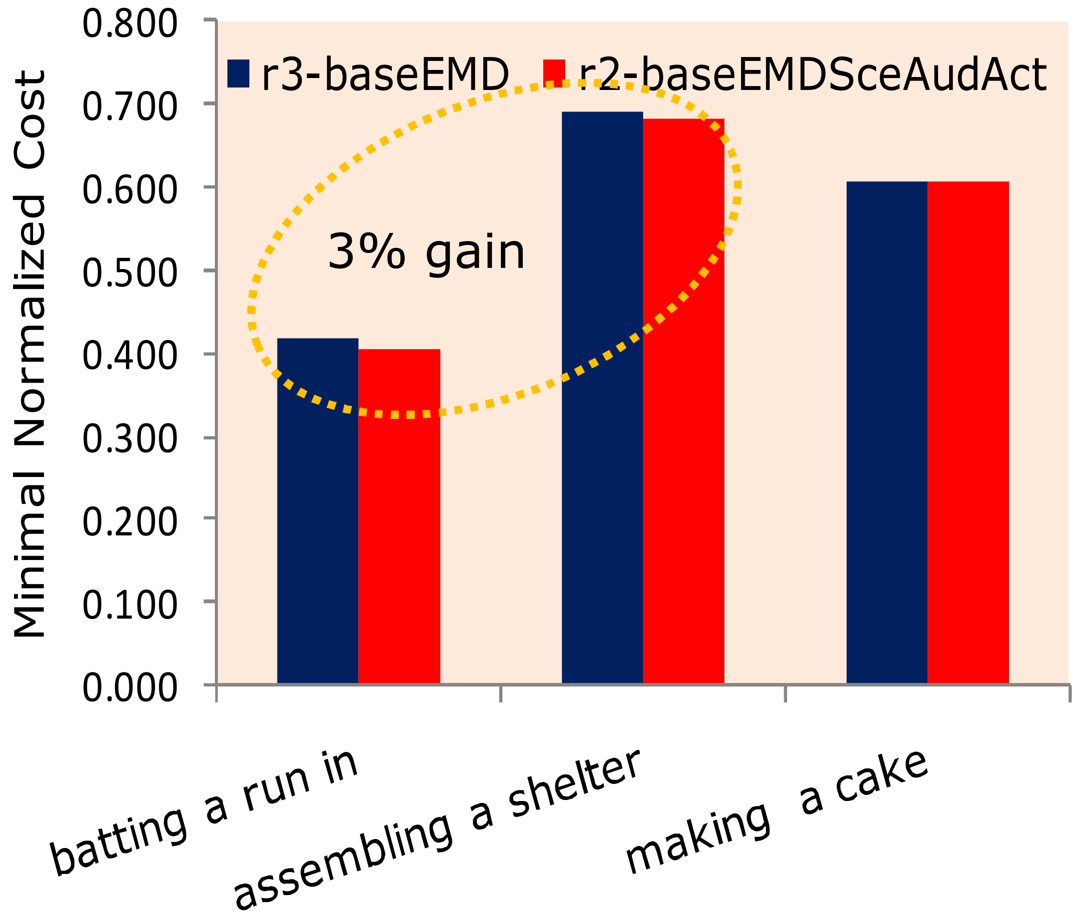
  - Semantic graph
    - Nodes are concepts/events
    - Edges represent concept/event correlation
  - Graph diffusion
    - Smooth detection scores
      w.r.t. the correlation



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

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**, in ACM ICMR 2011.

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

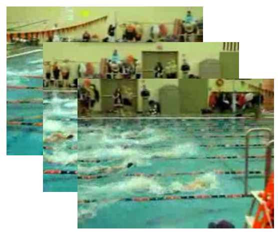

Skiing


Dog


Wedding Reception

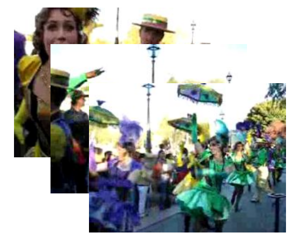

Non-music Performance


Baseball

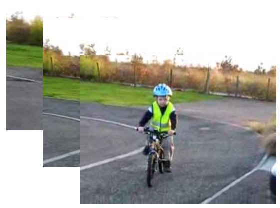

Swimming


Bird

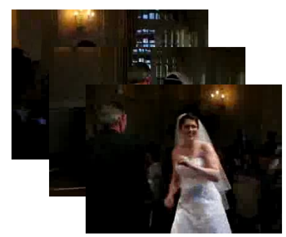

Wedding Ceremony


Parade


Soccer

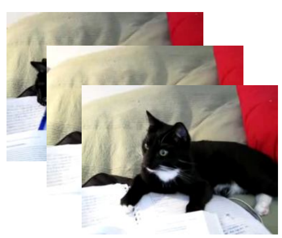

Biking


Graduation

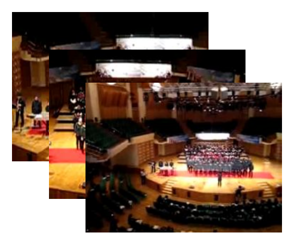

Wedding Dance


Beach


Ice Skating


Cat

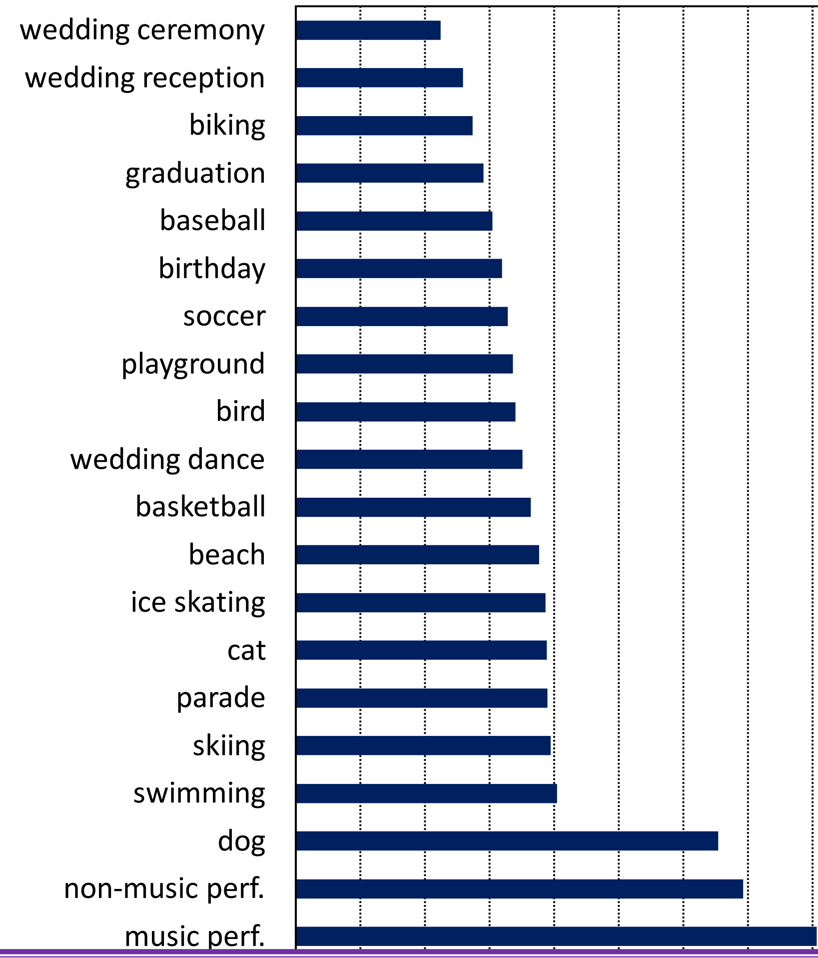

Birthday Celebration


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



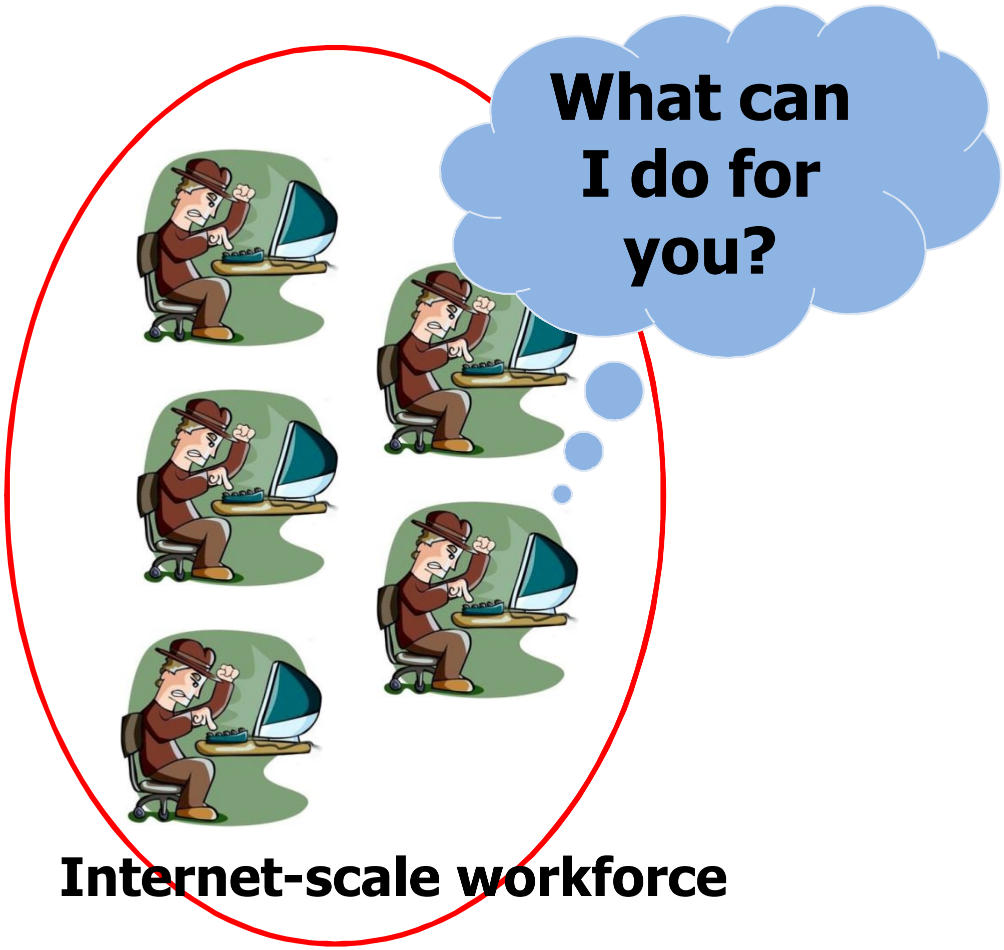
The trick of digging out consumer videos from YouTube:
Use default filename prefix of many digital cameras: "**MVI** and parade".

Existing Database?

	CCV Database
• Human Action Recognition – KTH & Weizmann • (constrained environment) 2004-05	**Unconstrained YouTube videos**
– Hollywood Database • (12 categories, movies) 2008 – UCF Database • (50 categories, YouTube Videos) 2010	**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



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

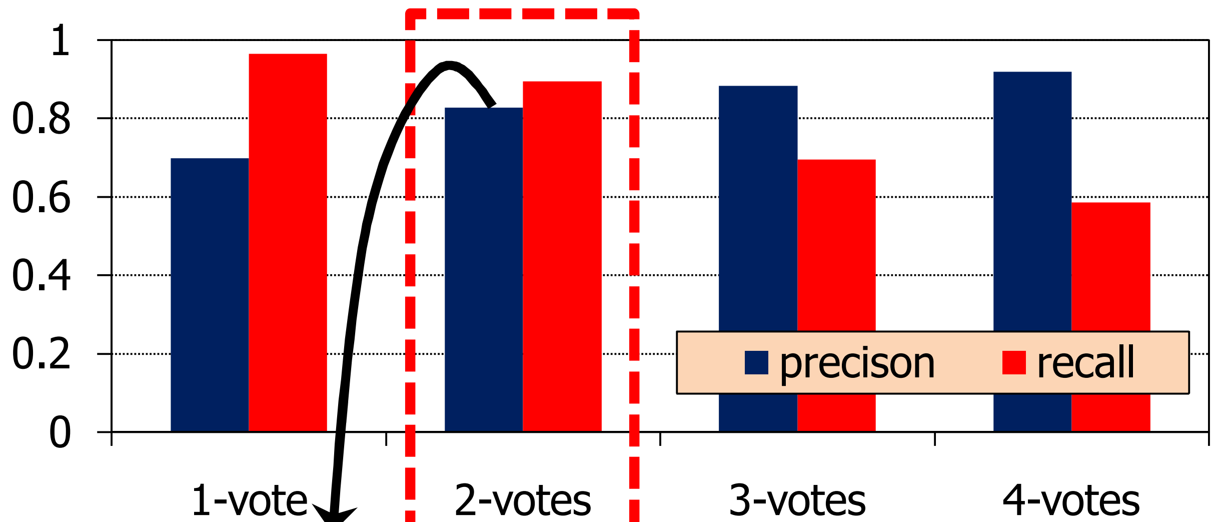
Submit

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

$ 0.02

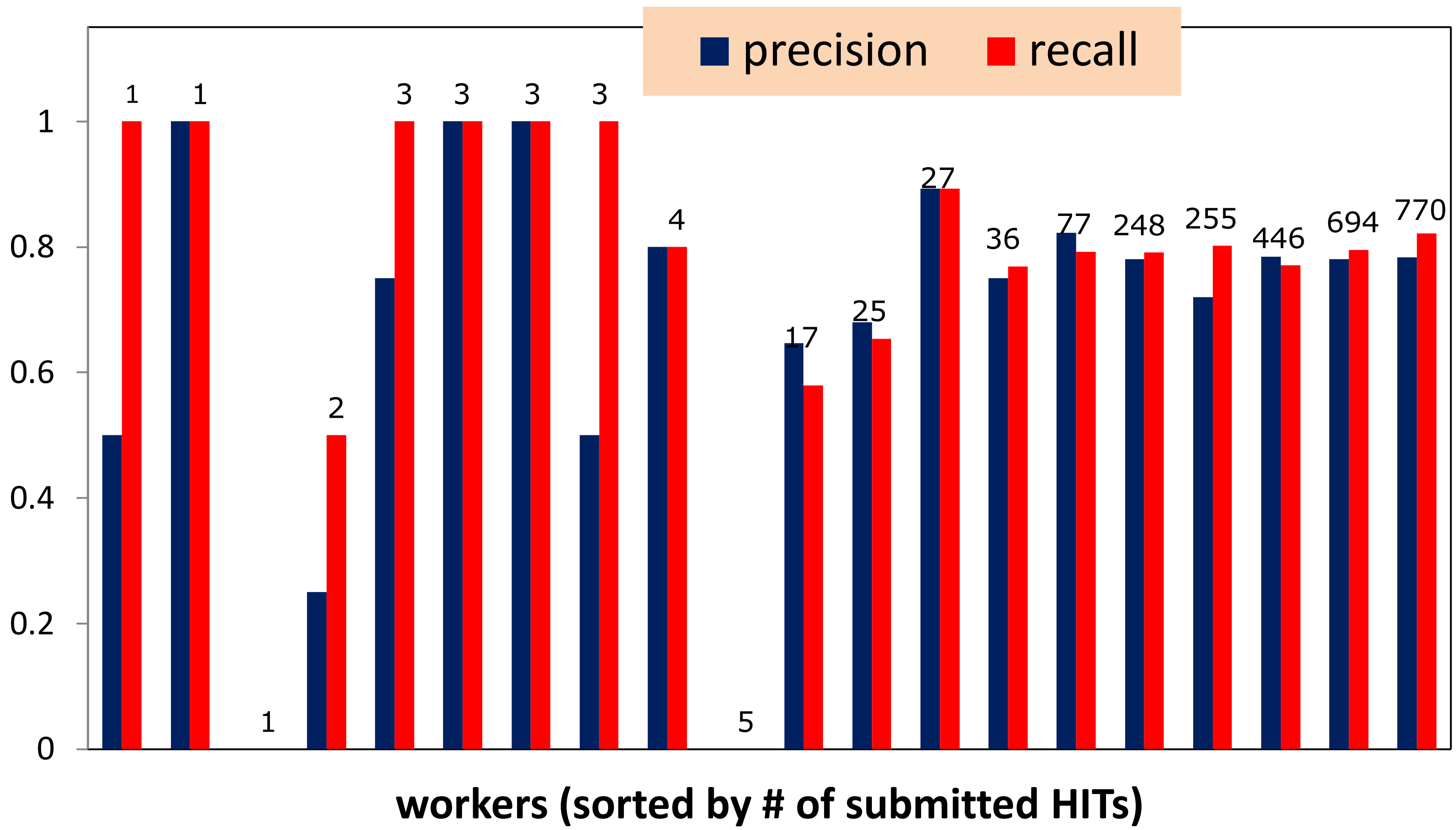
Human Recognition Performance

- 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

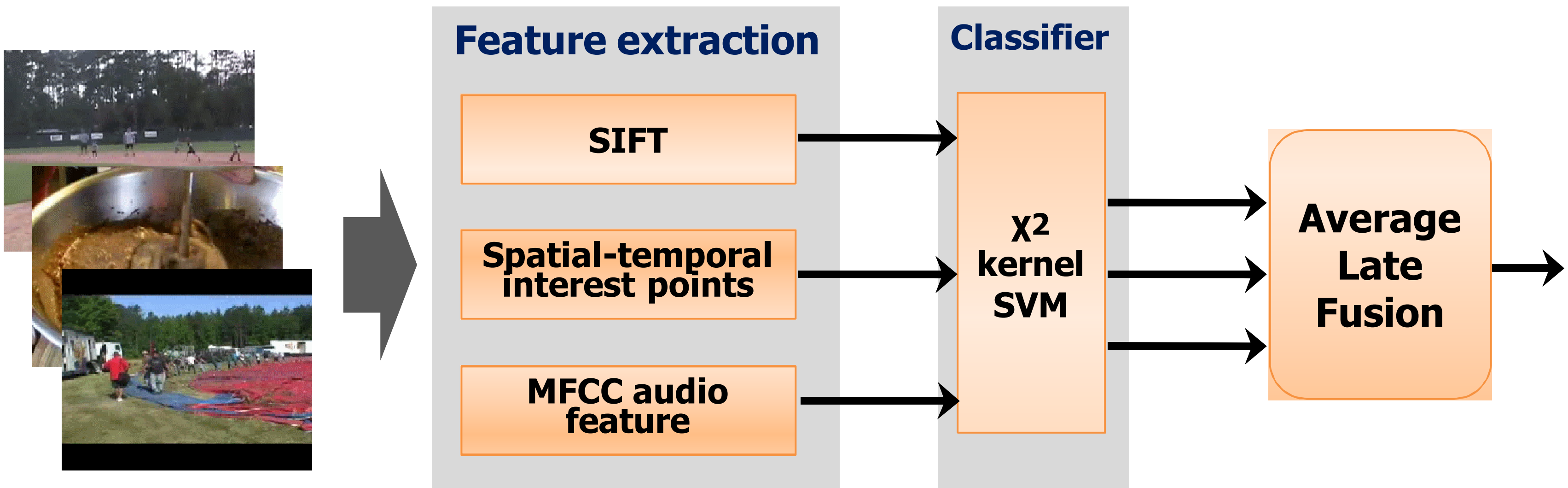


Plus additional manual filtering of 6 positive sample sets: 94% final precision

Human Recognition Performance (cont.)



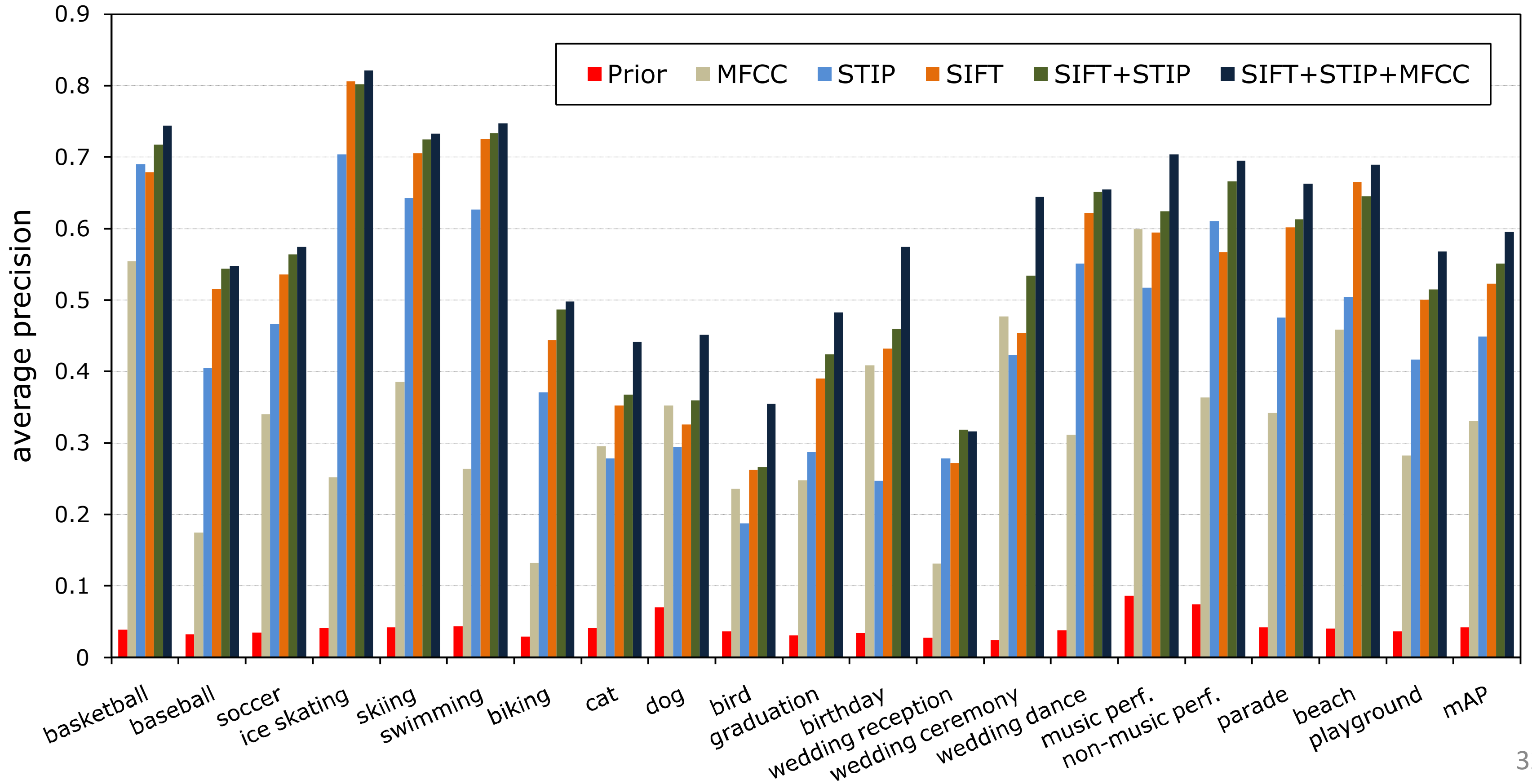
Machine Recognition System



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.

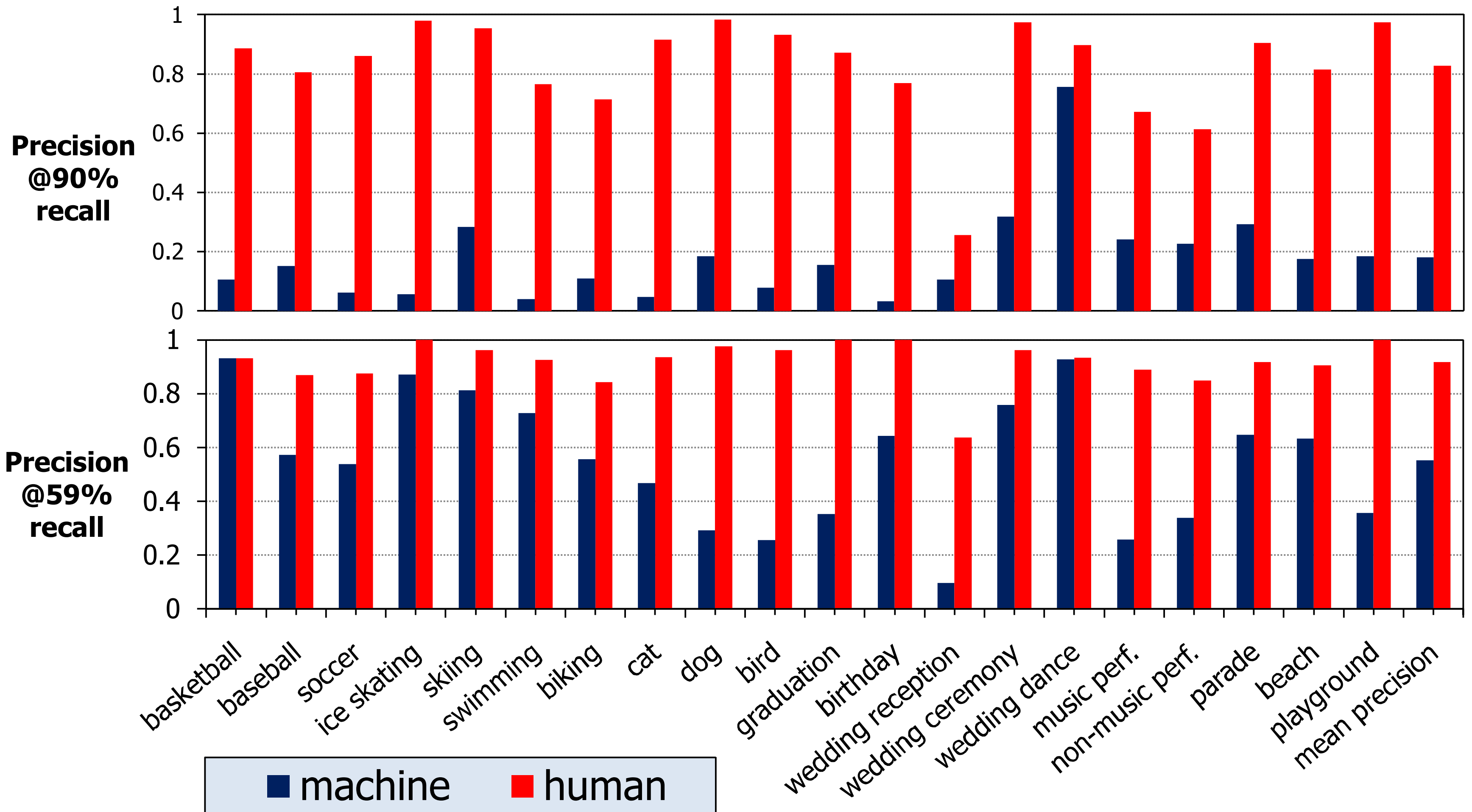
Machine Recognition Accuracy

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

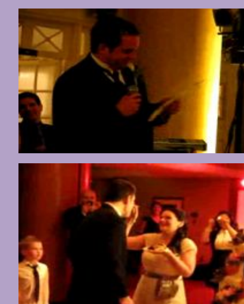
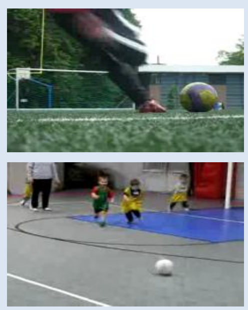
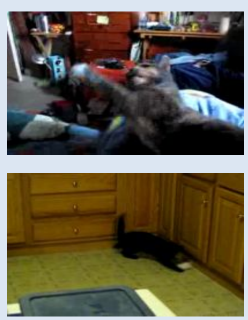
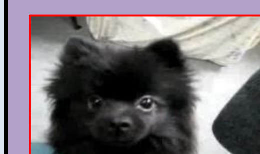
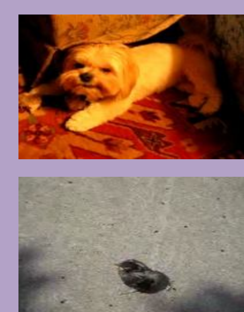


Human vs. Machine

- 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 … →

THANK YOU!

email: yjiang@ee.columbia.edu