Joint Audio-Visual Signatures for Web Video Analysis

Dan Ellis & Shih-Fu Chang
Wei Jiang, Courtenay Cotton, Yu-Gang Jiang, Xiaohong Zeng

Electrical Engineering, Columbia University, NY USA
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Web Video Analysis
2. Labeled Data Gathering
3. Scene / Object Context
4. Future Work

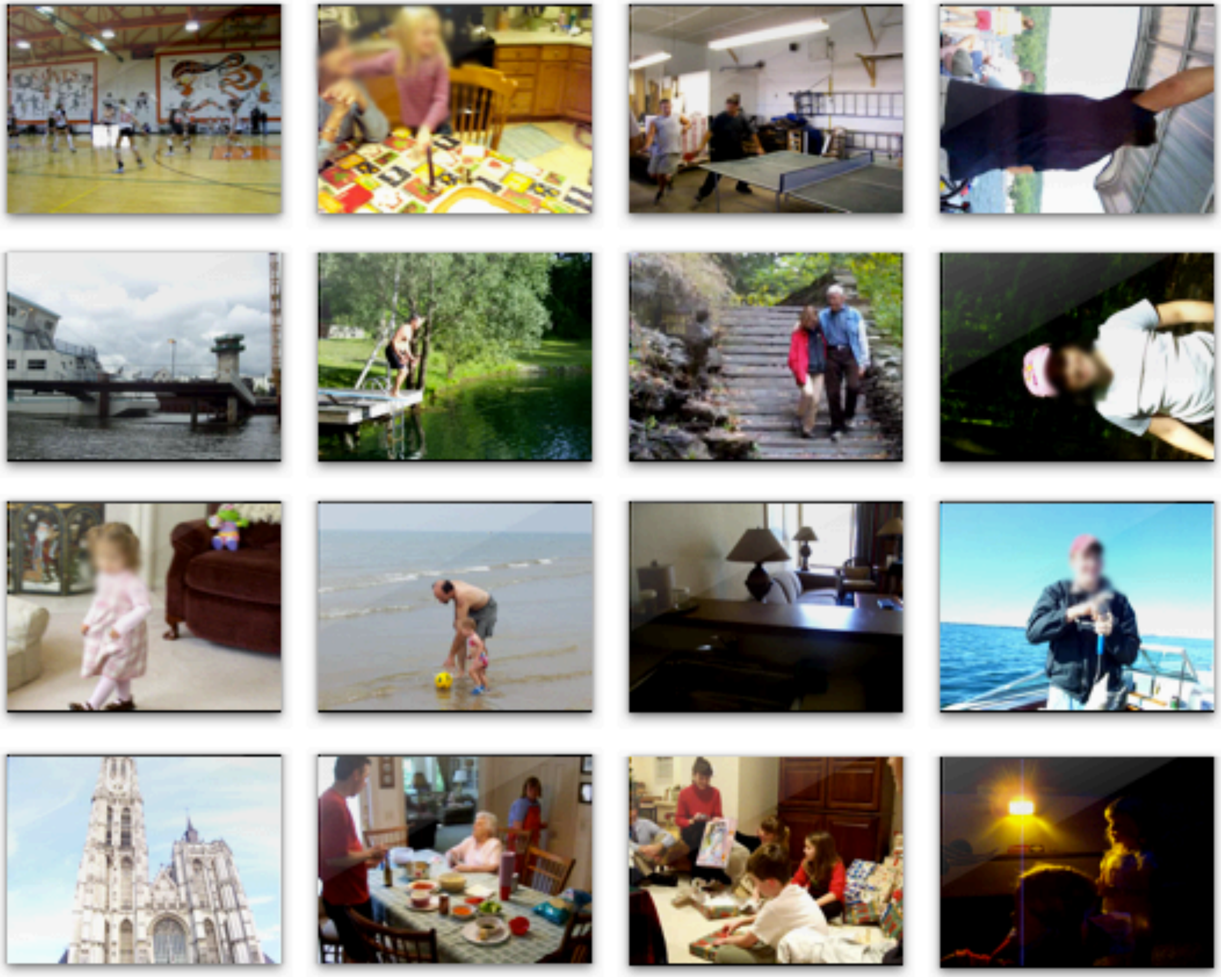
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

DVMM lab
digital video | multimedia laboratory

Columbia University
in the city of New York

1. Web Video Analysis

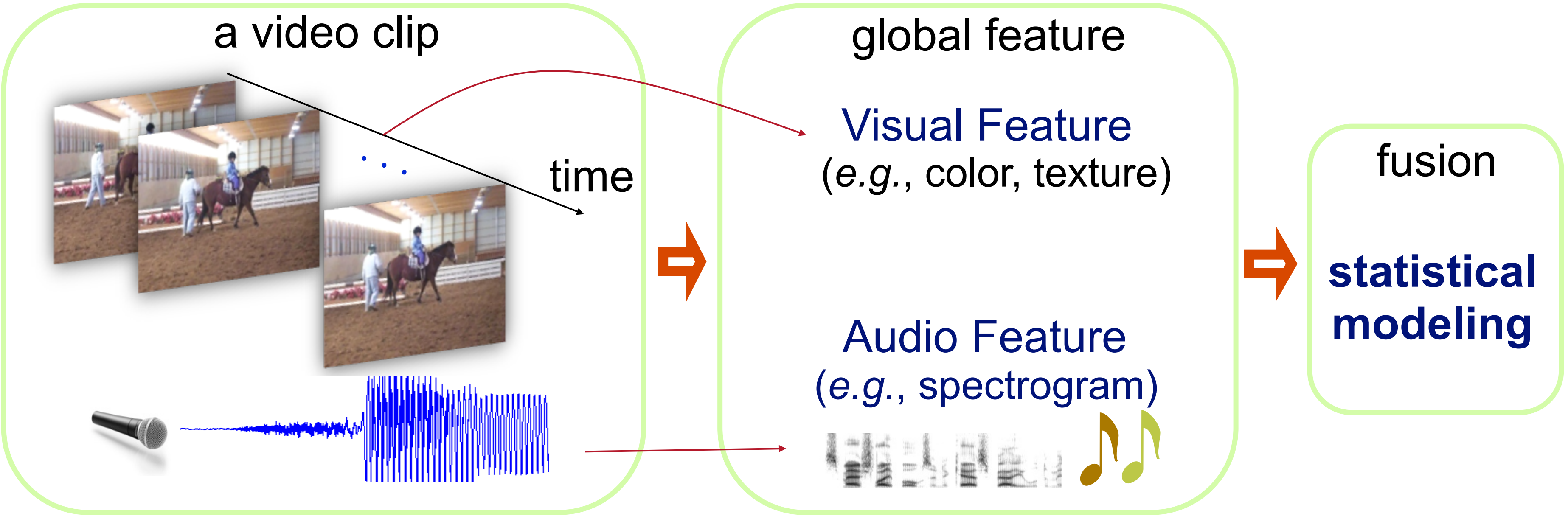
- Web video
  - Huge volume
  - Poor labels
  - Low quality



- The goal:
  - Automatic, efficient, human-like labeling
    - into categories
    - by events/objects

Conventional Approach: Global Features

- Train classifiers for predefined categories based on statistics of whole clip

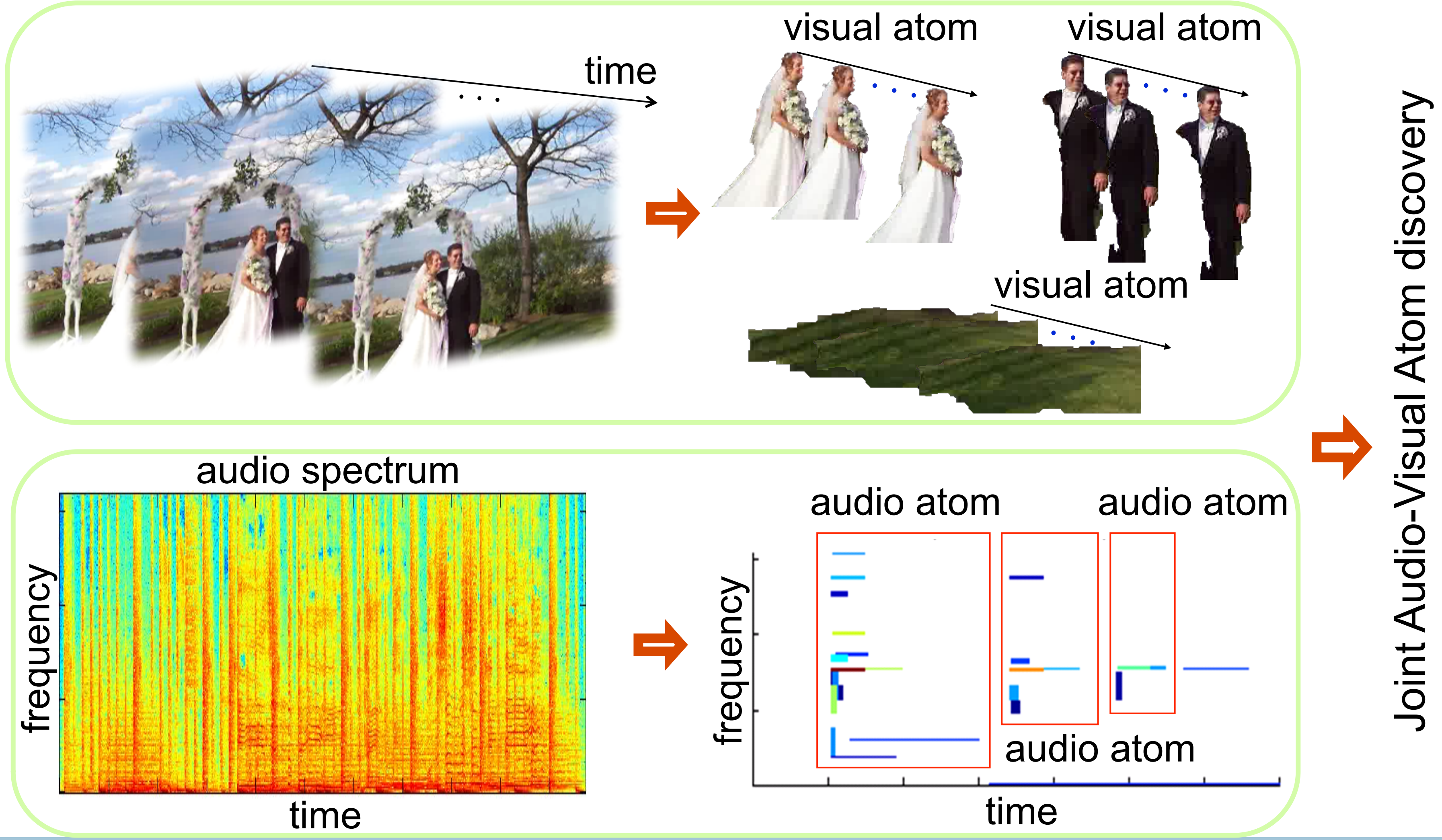


  - no object-level description

[Chang et al. MIR 2007]
[Cristani et.al., TMM 2007]

Novel Approach: Audio-Visual Atoms

- Decompose aud/vid into object-like atoms
  - statistical models of their combinations



Joint Audio-Visual Atom discovery

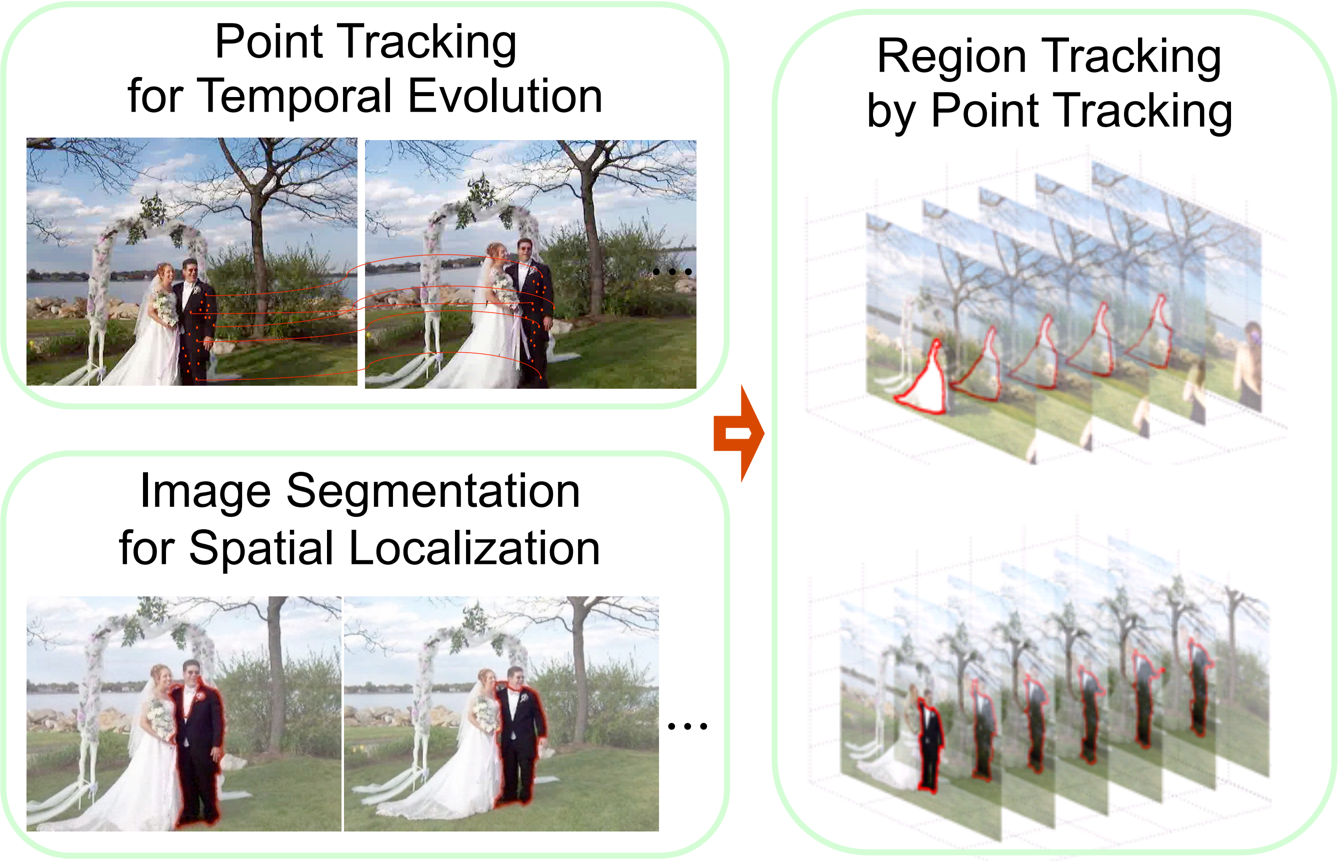
Challenges in Unconstrained Video

- Poor quality
  - focus
  - lighting
  - camera motion
  - occlusions
  - ambient noise
  - handling noise

- Poor A-V correlation
  - sounds from unobserved objects
  - sound-producing motions are slight

Visual Atom Formation

- Point tracking
  - sift points link successive frames
- + Region Segmentation
  - color / texture define regions
- Link shorter tracks across time



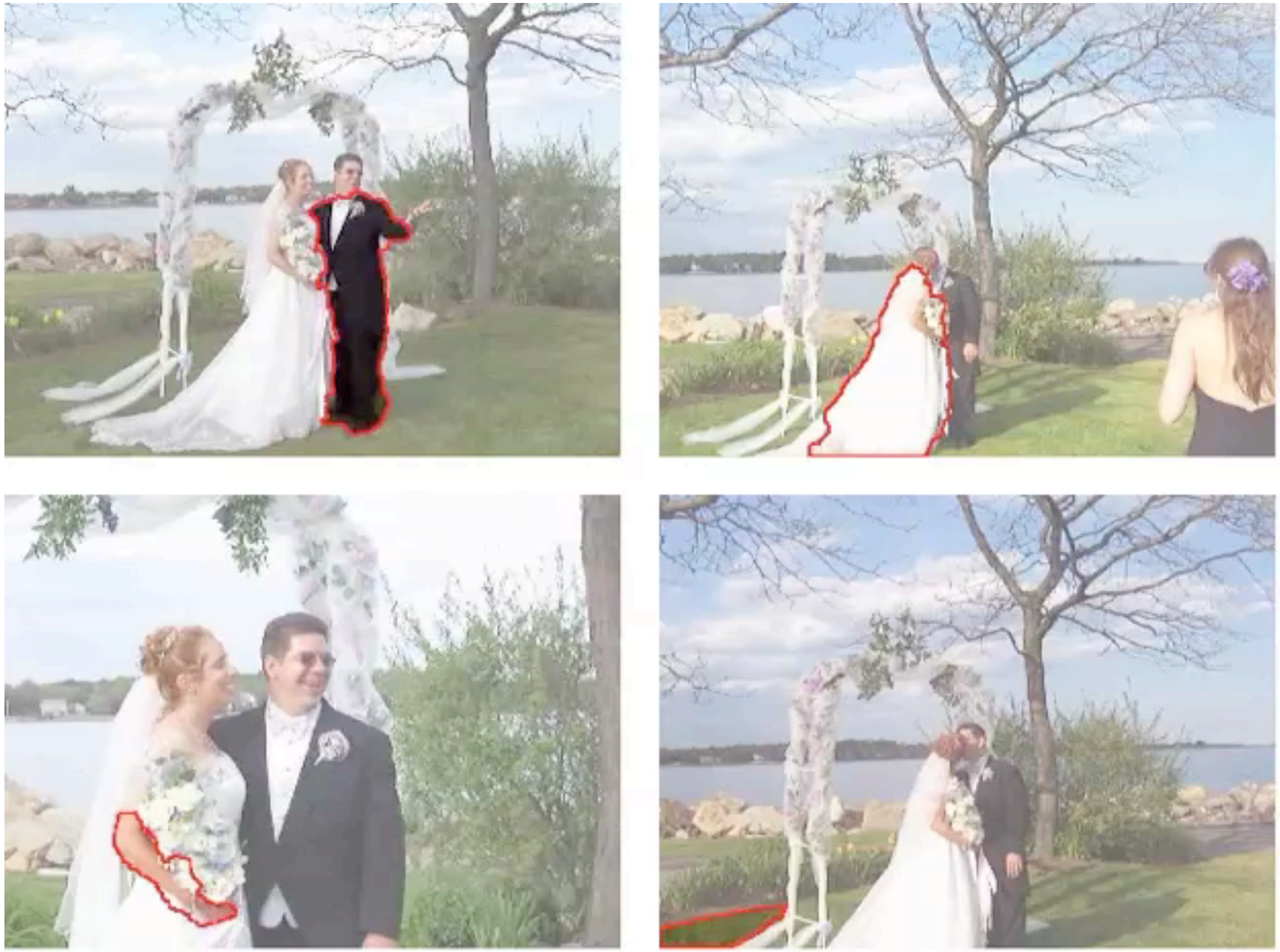
Point Tracking for Temporal Evolution

Image Segmentation for Spatial Localization

Region Tracking by Point Tracking

[Jiang et.al., MM 2009]

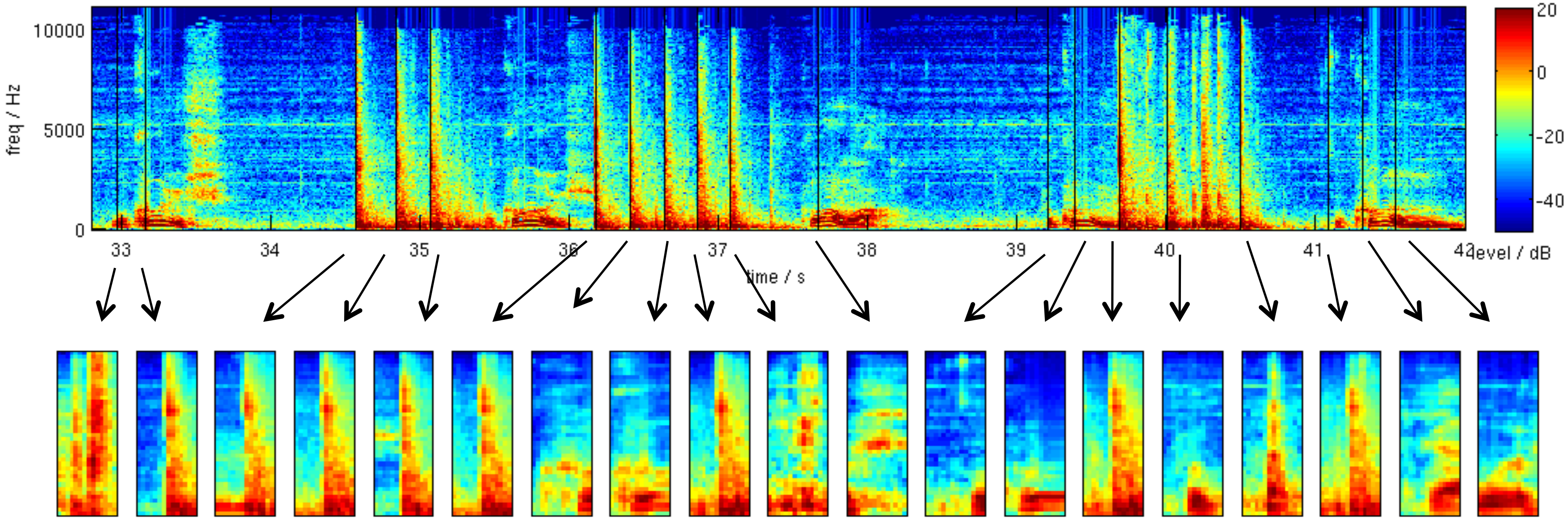
Visual Atom Examples



- A few examples out of 100+ for “wedding”
  - build codebook based on appearance, shape

Audio Atom Formation

- Extract & describe Transients
  - for maximum robustness against noise

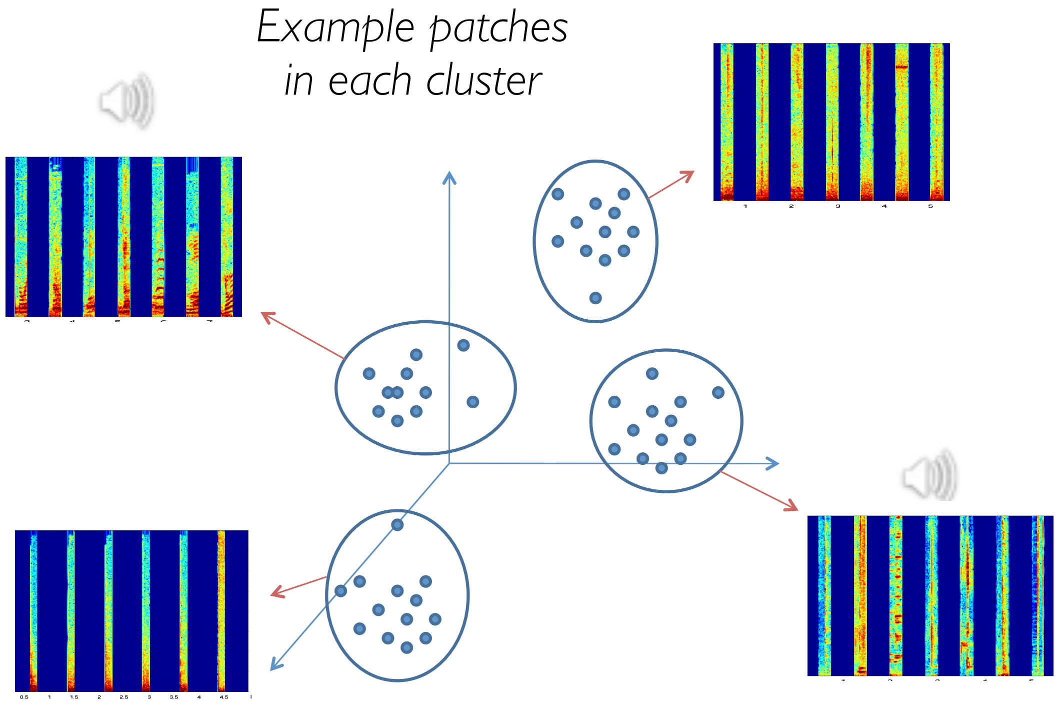


- Multiscale analysis to find energy "bursts"
  - extract 250 ms mel-spectrum window
  - describe with PCA

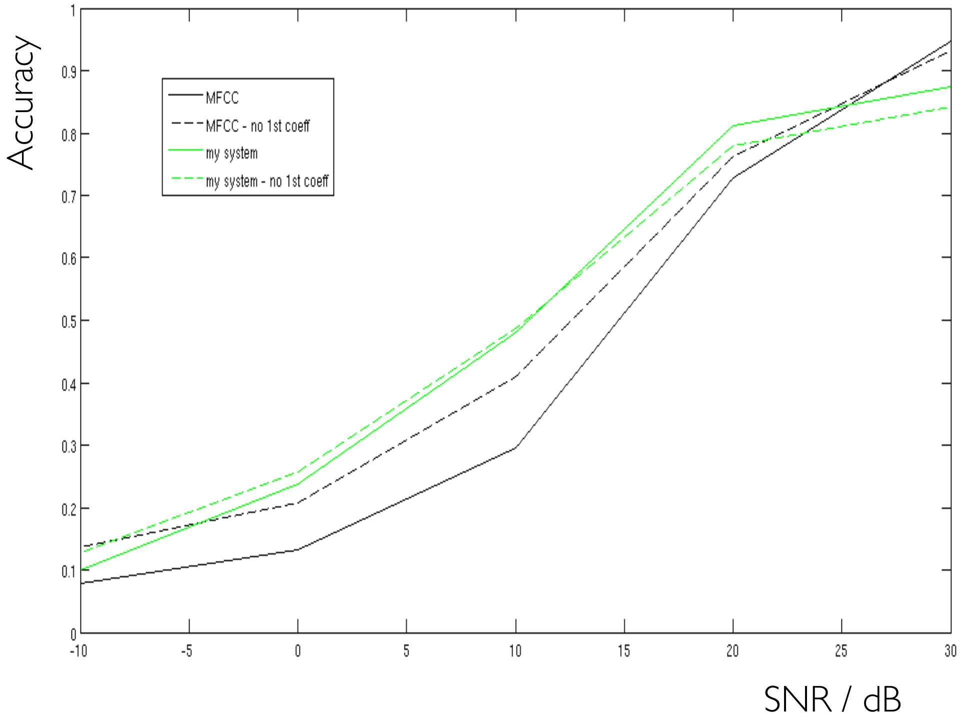
Audio Atom Results

- K-means clustering to form codewords

- Better noise resistance than MFCCs

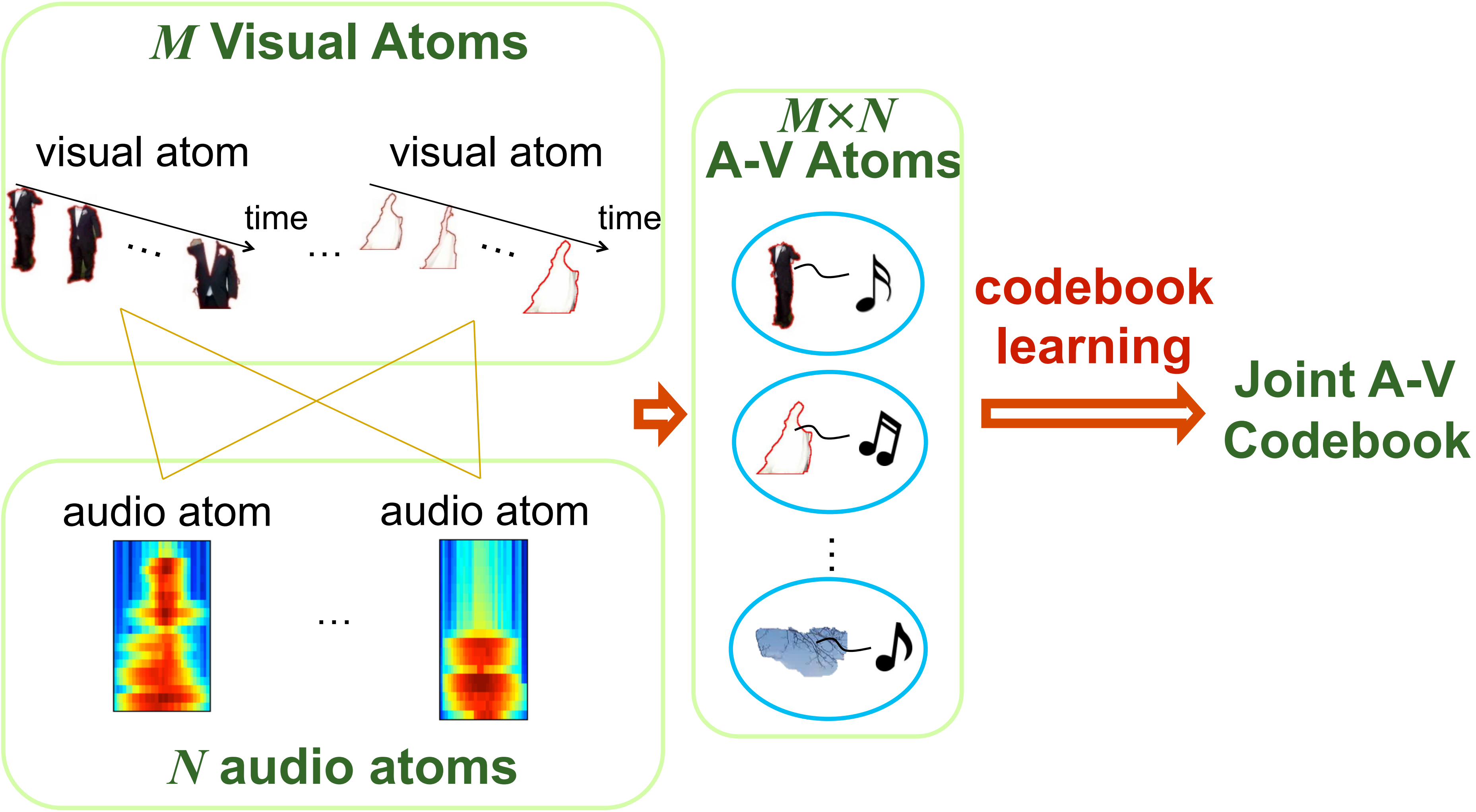


Example patches in each cluster



Joint Audio-Visual Atoms

- Consider all possible A-V combinations



M Visual Atoms

visual atom

time

visual atom

time

audio atom

audio atom

N audio atoms

M×*N*
A-V Atoms

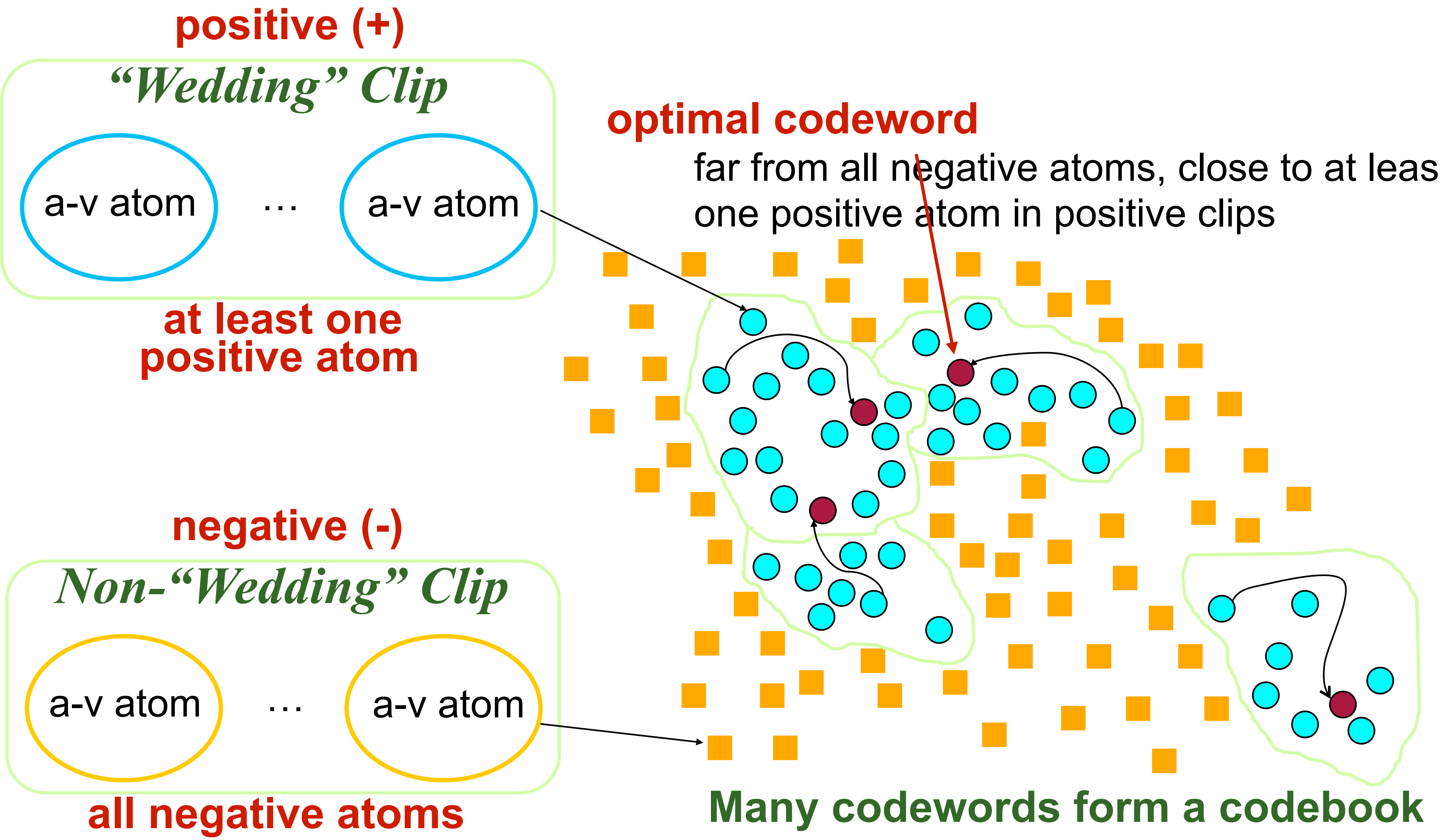
codebook learning

Joint A-V Codebook

Concept Codebook Learning

- by Multiple-Instance Learning (MIL)
  - only have clip-level labels

[Maron et al., NIPS, 1998]



positive (+)

“Wedding” Clip

a-v atom … a-v atom

at least one positive atom

negative (-)

Non-“Wedding” Clip

a-v atom … a-v atom

all negative atoms

optimal codeword

far from all negative atoms, close to at least one positive atom in positive clips

Many codewords form a codebook

Evaluation on Consumer Video

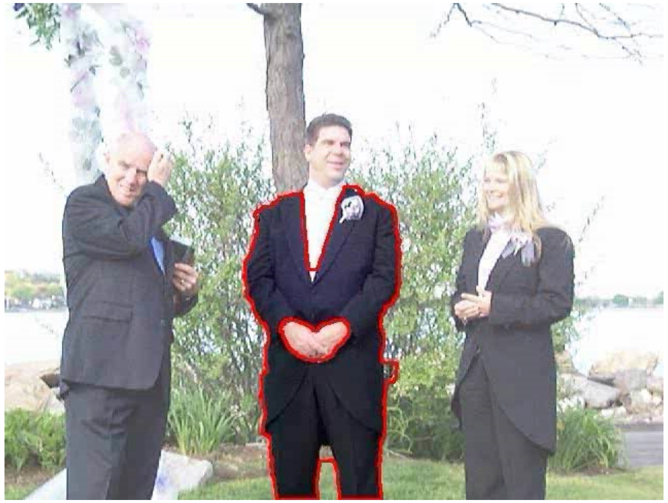
- Kodak consumer video benchmark set
  - 1358 videos (813 training)
  - 25 labels

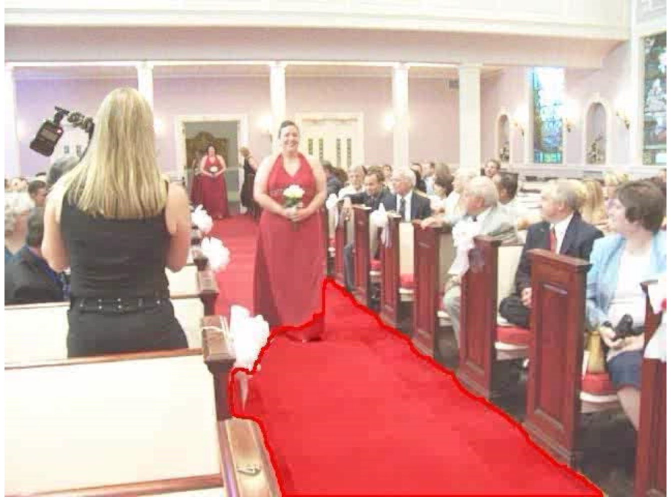
[Loui et al. MIR 2007]

Example A-V Codewords

- “Wedding” class

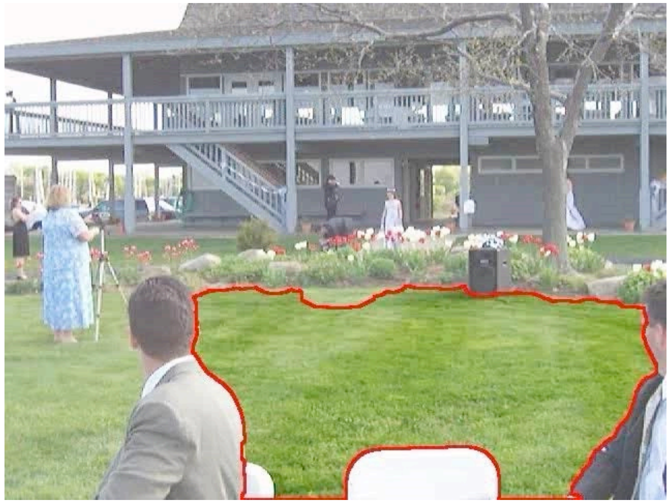
black suit
+
romantic
music

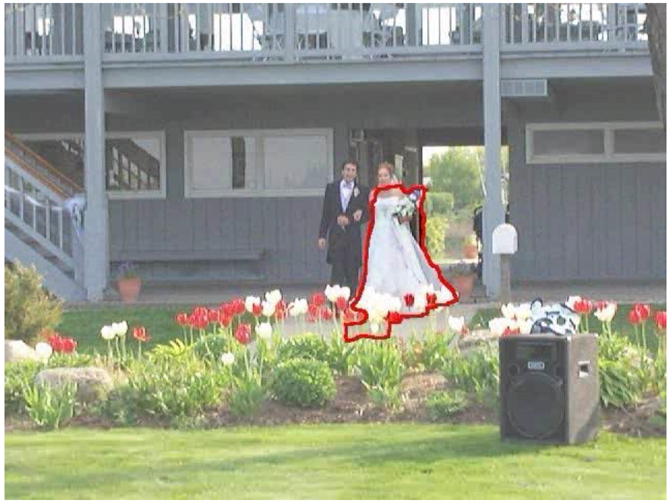




red carpet
+
romantic
music

green grass
+
romantic
music



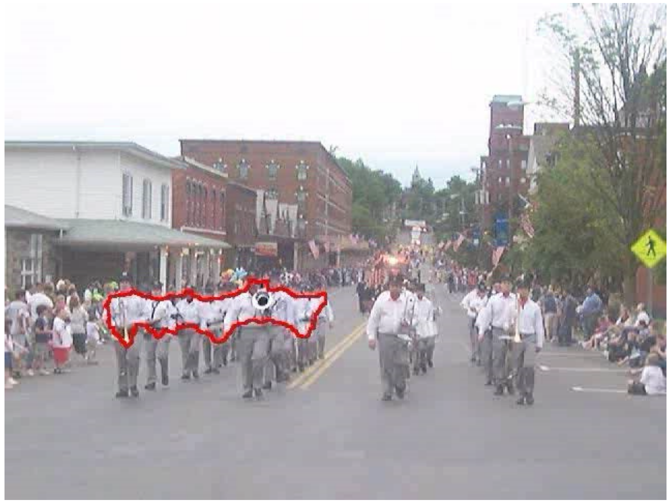


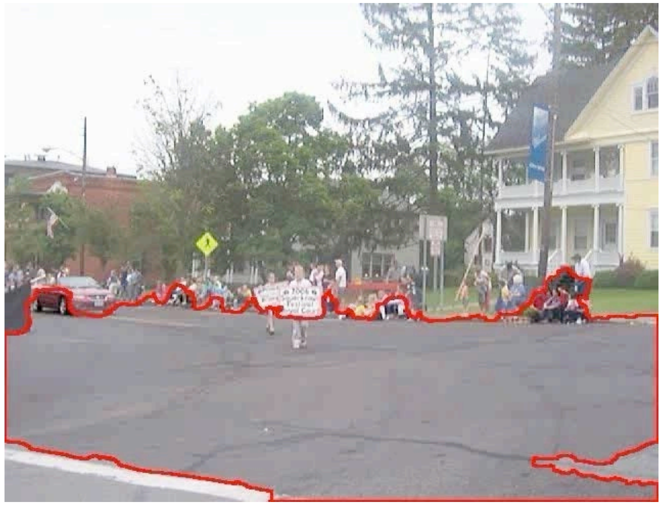
white gown
+
romantic
music

Example A-V Codewords

- “Parade” class

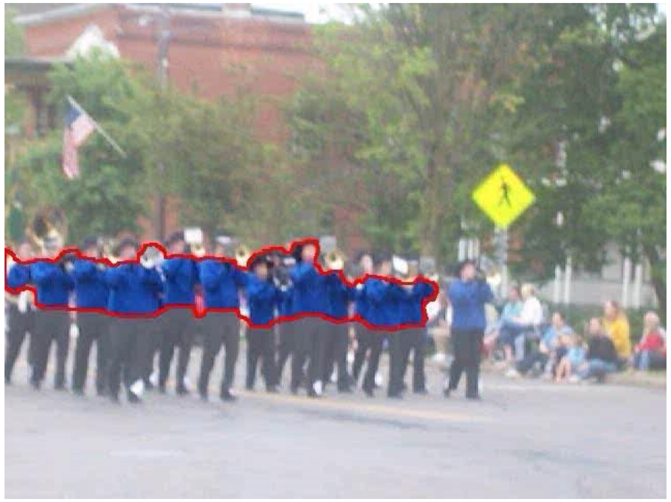
marching
people
+
parade
sound

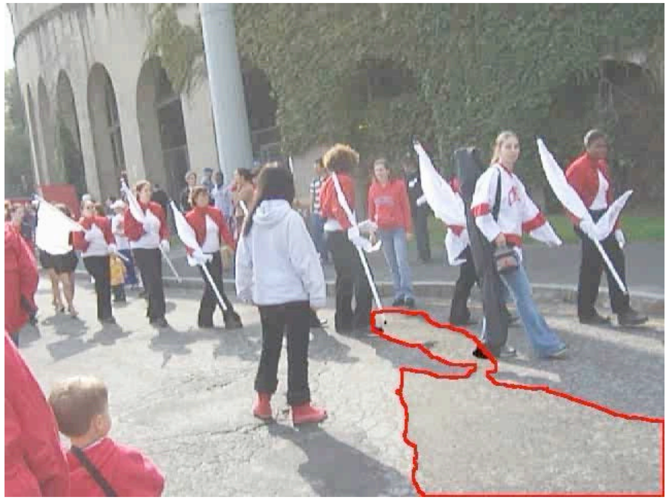




road
+
parade
sound

marching
people
+
parade
sound

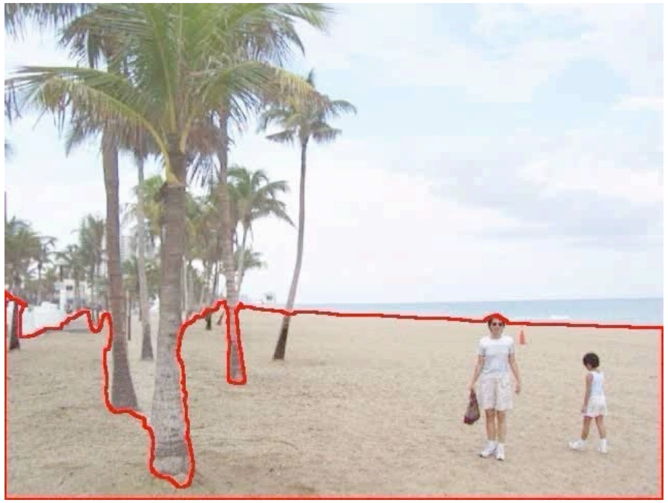




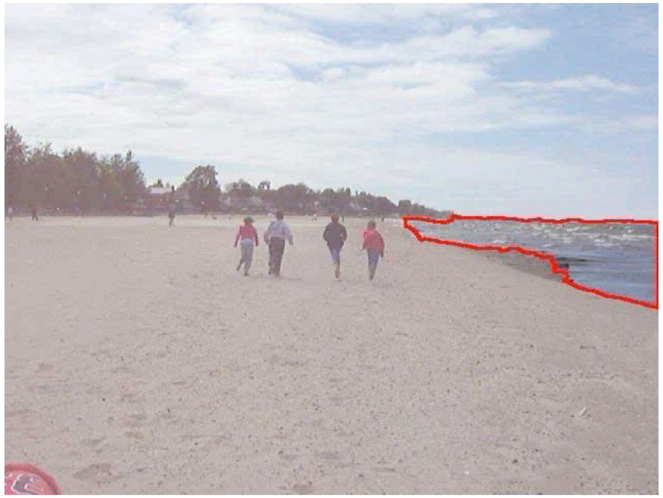
road
+
parade
sound

Example A-V Codewords

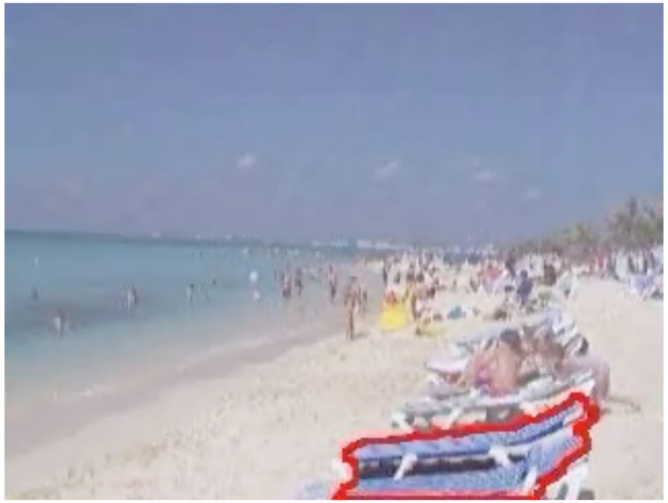
- "Beach" class



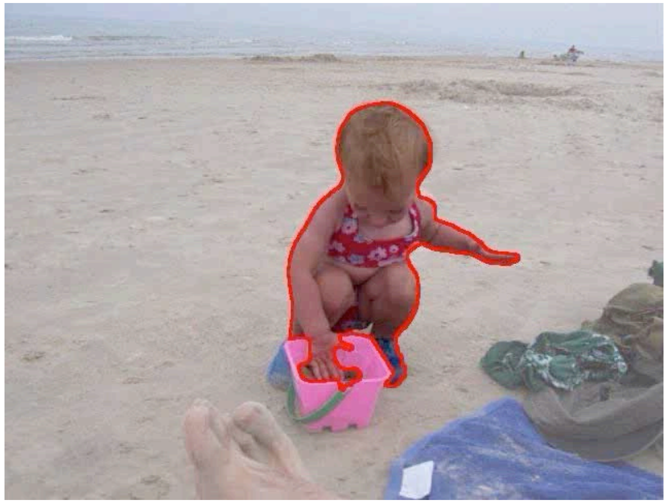
sand
+
beach
sound



water
+
beach
sound



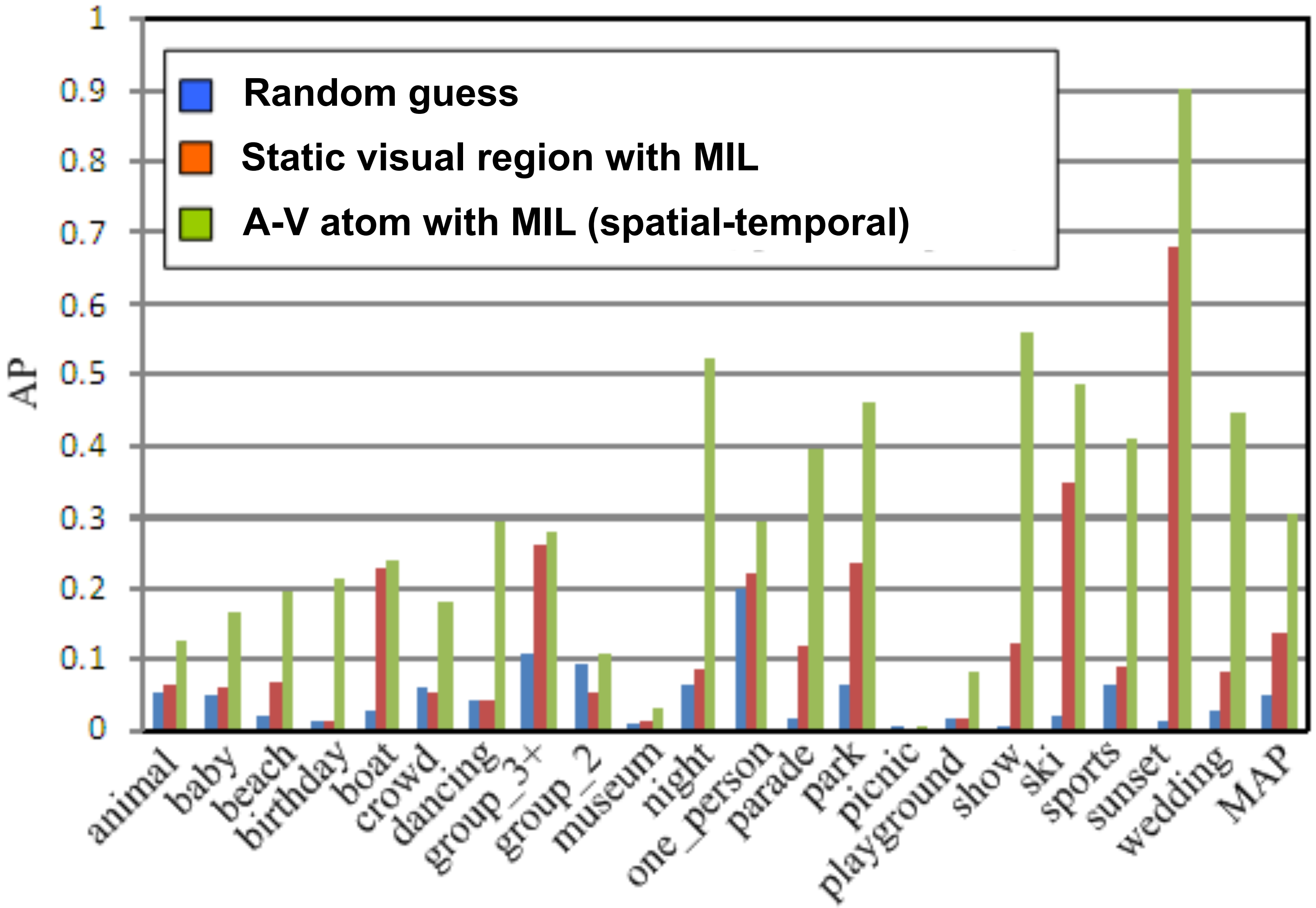
beach
chair
+
beach
sound



people with
swim suit
+
beach
sound

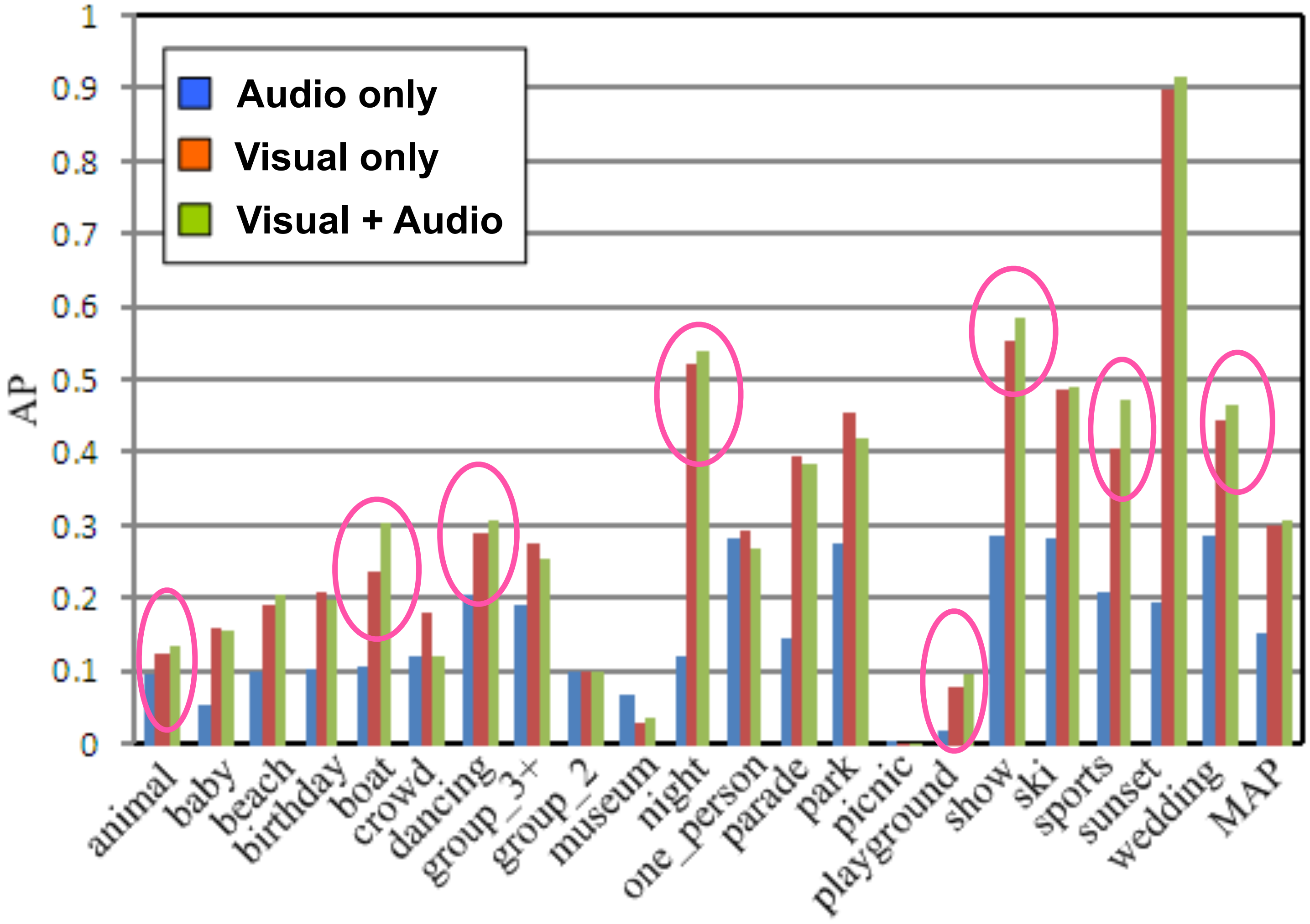
Consumer Video Evaluation

- A-V atoms vs. Static regions
  - Average Precision on test set



Consumer Video Evaluation

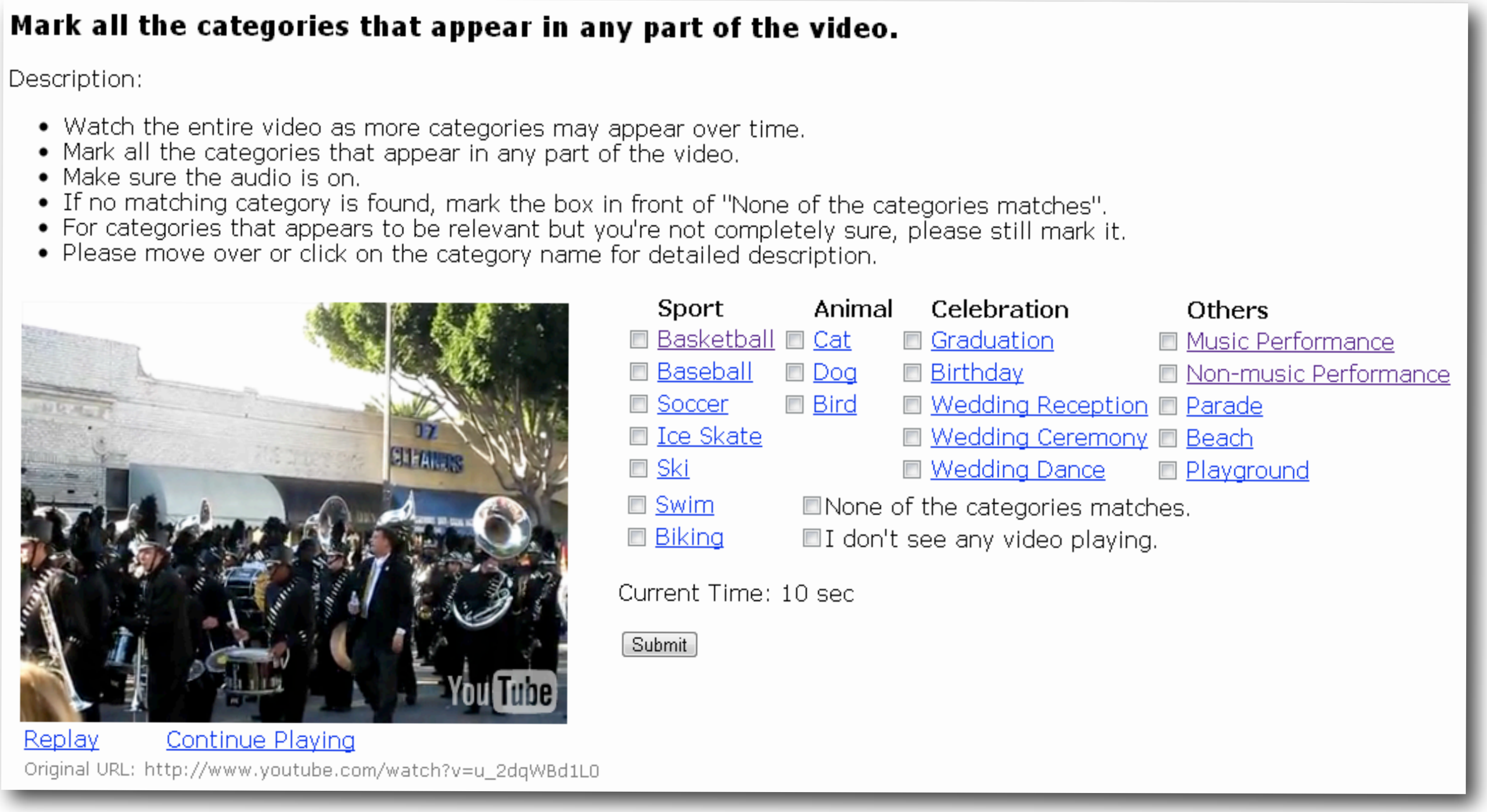
- Audio / Video / both
  - audio useful for many classes



?

2. Labeled Data Gathering

- Learning codebooks requires labeled data
  - novel concept → need new labels
  - more labels → better performance
- Amazon “Mechanical Turk”



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

Description:

- 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 the 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 move over or click on the category name for detailed description.

Sport	Animal	Celebration	Others
□ Basketball	□ Cat	□ Graduation	□ Music Performance
□ Baseball	□ Dog	□ Birthday	□ Non-music Performance
□ Soccer	□ Bird	□ Wedding Reception	□ Parade
□ Ice Skate		□ Wedding Ceremony	□ Beach
□ Ski		□ Wedding Dance	□ Playground

□ Swim □ None of the categories matches.

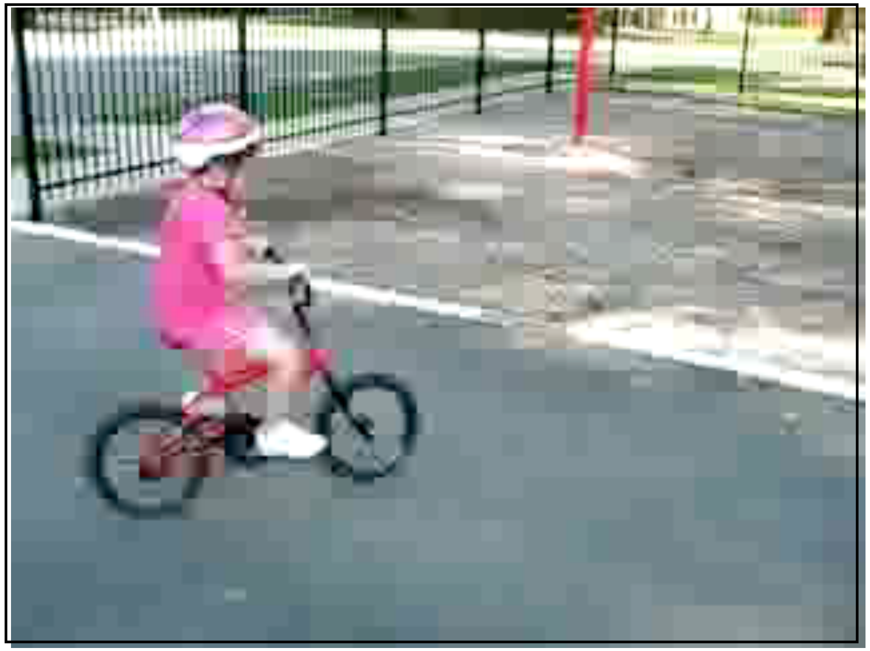
□ Biking □ I don't see any video playing.

Current Time: 10 sec

Original URL: http://www.youtube.com/watch?v=u_2dqWBd1L0

MTurk Results

- Data: YouTube raw camera uploads
  - based on keyword search for 20 categories
- MTurk Human Intelligence Tasks (HITs)
  - paid $0.02 per 10s clip (~$7/hr)
  - 4 labelers/clip, finished 9,641 videos in 2 weeks

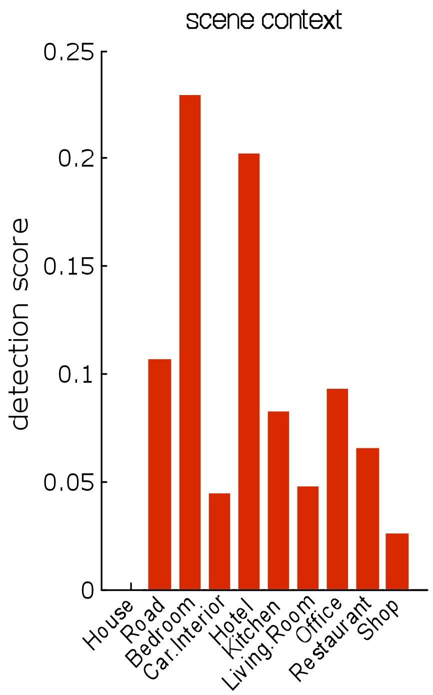


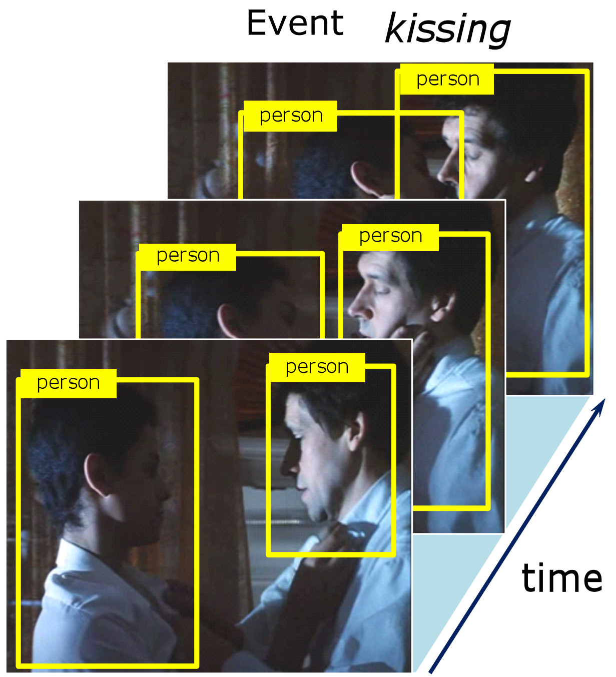
Playground, Biking

Non-music Performance, Ice Skating

3. Scene / Object Context

- How to identify events in video?
  - not objects, not locations
  - e.g. "people kissing"
- Traditional approach:
  - get low level features for large training set
  - statistical classifier
- Our approach
  - use specialized mid-level detectors (faces, cars)
  - learn context, relationships
  - e.g. "kissing" = 2 faces moving together

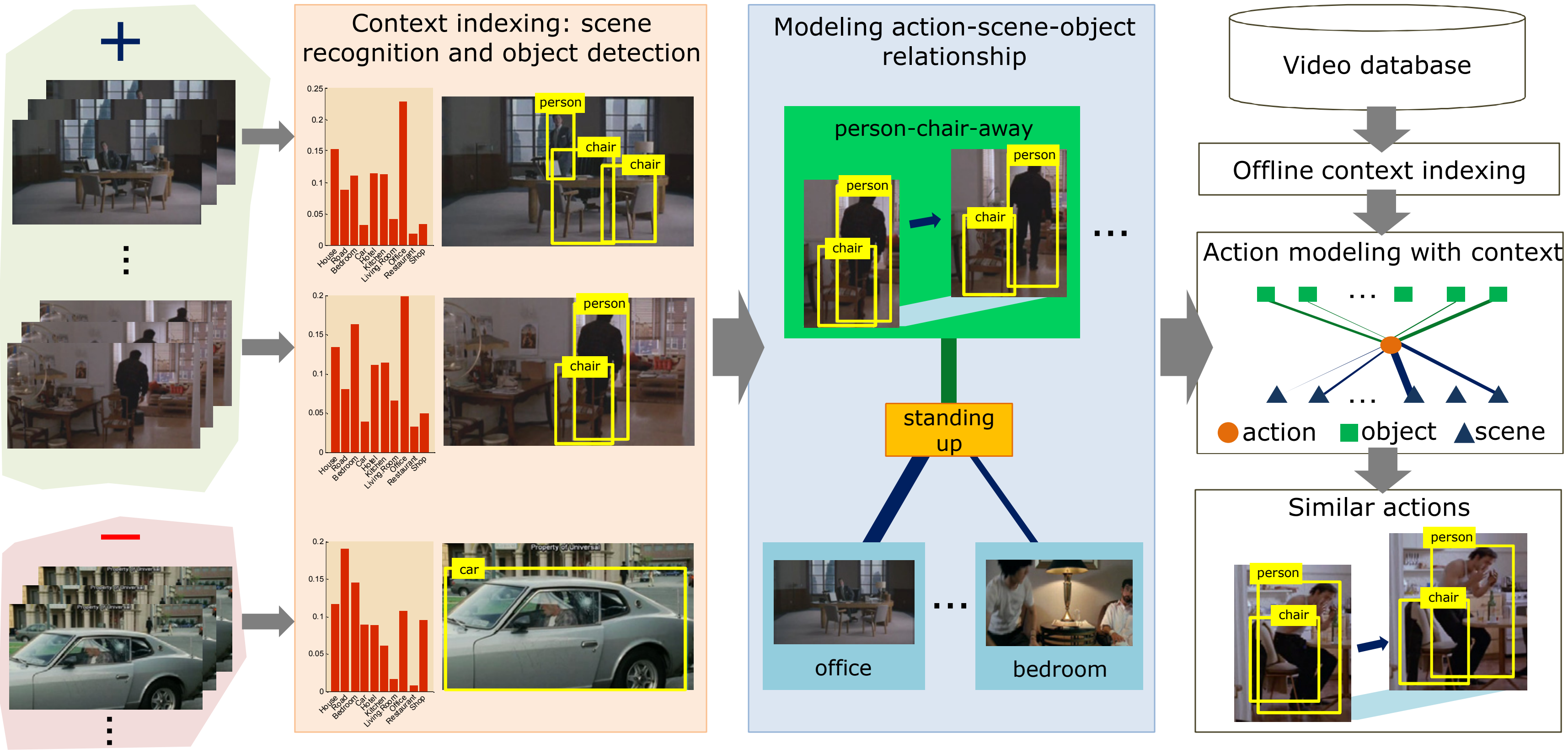




[Jiang, Li, Chang, TSCVT 2011]

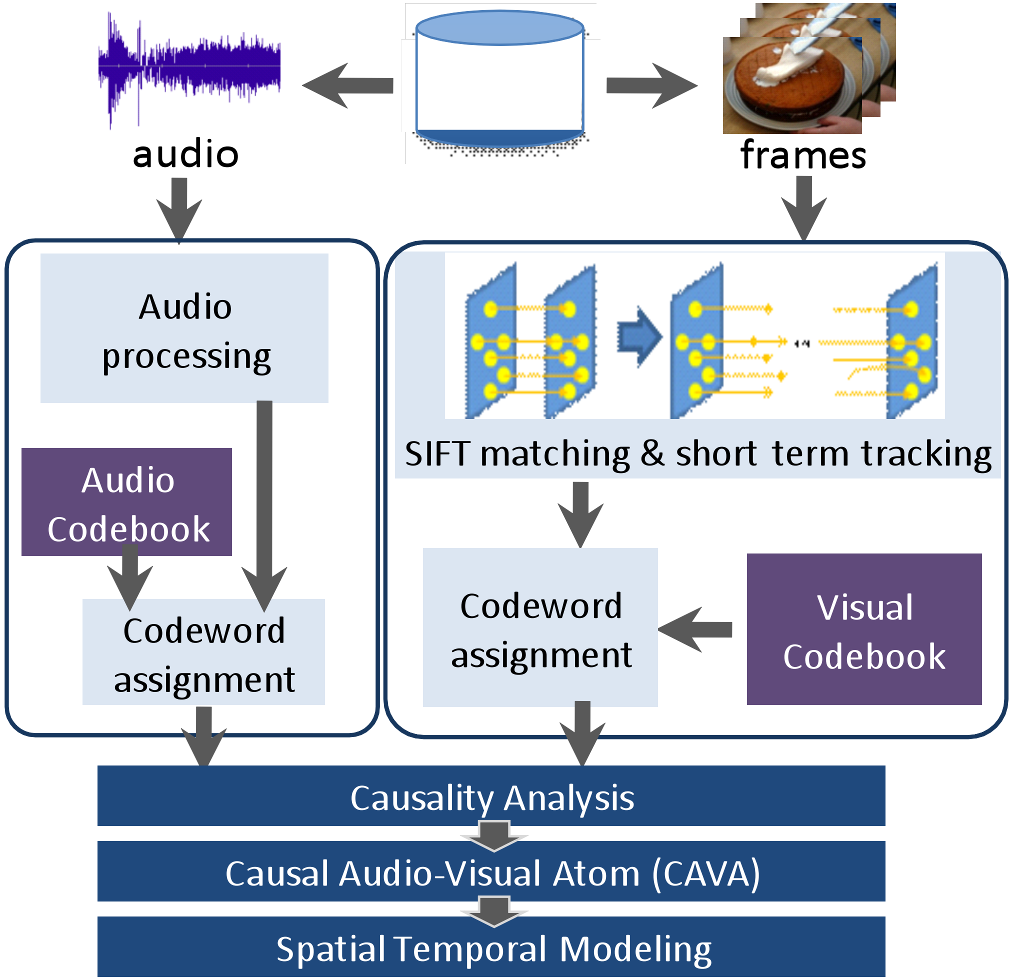
Action-Scene-Object

- Identify relevant objects, scenes from a few training examples (~ 10)
- Learn relationships for action
  - accuracy much better than raw classifier



4. Future Work

- Existing joint Audio-Visual atoms are based on simple co-occurrence
  - no temporal structure
- Synchrony?
  - too hard to detect in web video
- Causality?
  - e.g. simple ordering
  - “Causal Audio-Video Atoms’ CAVAs



Summary

- Web video analysis
  - desperate need for automatic analysis
  - must be in terms of objects, scenes, actions
- Joint Audio-Visual Atoms
  - object-related codebooks for audio, video
  - MIL of all possible combinations to find cues
- Labeled Data
  - Mechanical Turk quickly labels web video examples
- Context-Based Action Detection
  - uses mature existing object and scene detectors
- Better "Causal Audio-Visual Atoms"