GENIE

General Engine for Indexing Events

TRECVID MED 2012

26 Nov 2012

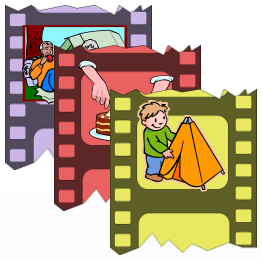
Amitha Perera

Kitware	Honeywell	Stanford	Simon Fraser University	Georgia Tech	SUNY Buffalo
Sangmin Oh	Scott McCloskey	Kevin Tang	Kevin Cannons	Ilseo Kim	Chenliang Xu
Megha Pandey	Ben Miller	Alexandre Alahi	Hossein Hajimirsadeghi	You-Chi Cheng	Gang Chen
Amitha Perera		Fei-Fei Li	Arash Vahdat	Chin-Hui Lee	Pradipto Das
		Daphne Koller	Greg Mori		Jason Corso
					Rohini Srihari

Funded by IARPA ALADDIN contract D11PC20069 BAA-10-01

GENIE MED 2012 System

Multimedia Archive

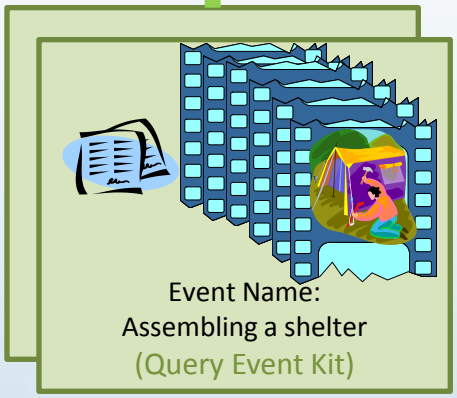


Feature Computation

- HOG3D
- Object Bank
- GIST
- MFCC
- ASM
- ⋮

500 core Linux PC cluster, 4 GB RAM per core

Single quad-core PC, 8GB RAM



Event Name:
Assembling a shelter
(Query Event Kit)

Event Kits

Codebook generation

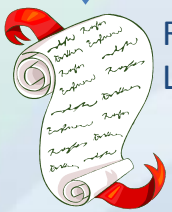
Base Classifiers

- HIK SVM
- NGD SVM
- Latent SVM
- ⋮

Learn classifiers

Score Fusion

- MFoM
- Expert Forest
- ⋮



Result List

Learn fusion model

Testing

Training

Kitware

MED 12 Feature List

Feature	Type	Temporal	Spatial
HOG3D	Video	Every 5th fr	Max 160 pixels
Gist	Video	Every 20th fr	Full
Object Bank	Video	1 fr / 2 secs	Full
MFCC	Audio	10ms	N/A
ASM	Audio	100-300ms	N/A
Color-SIFT	Video	1 fr / 2 secs	Full,
Transformed Color Histogram	Video	1 fr / 2 secs	Full
ISA (Le et al. CVPR 2011)	Video	Full	Max 160
SUN 09	Video	1 fr / 4 secs	Max 400

MED 11

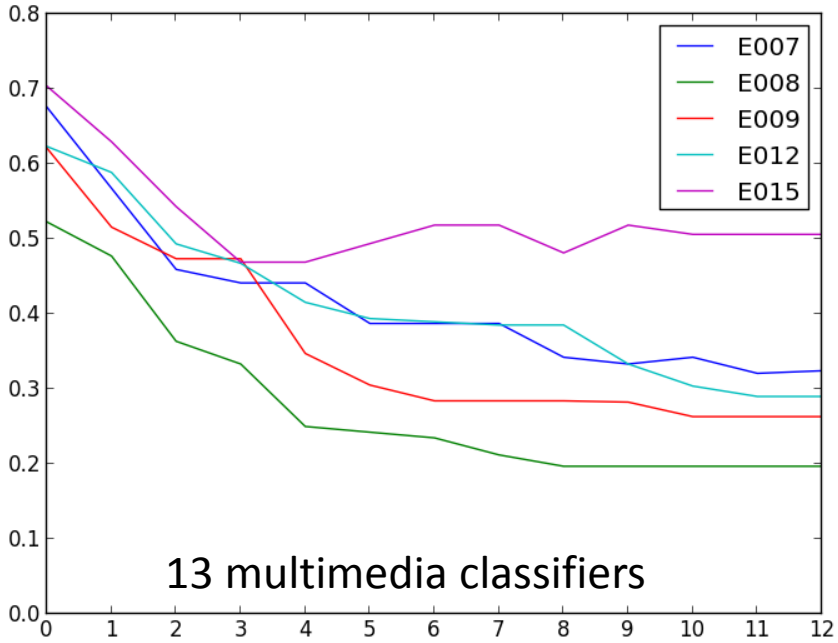
MED 12

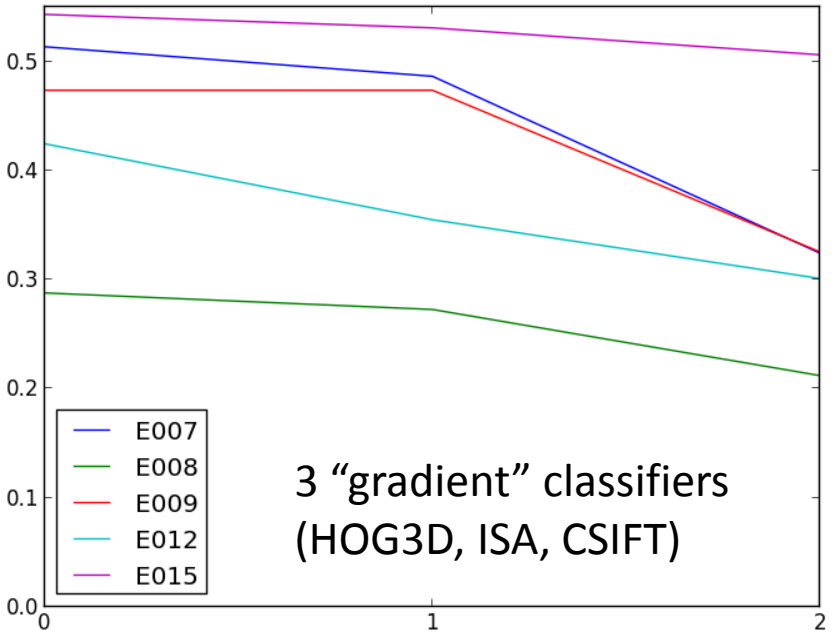
Each feature can be used by more than one event agent

MED As DET Optimization

- DET curves can improve just by fusing more things
  - But does this "solve" MED?

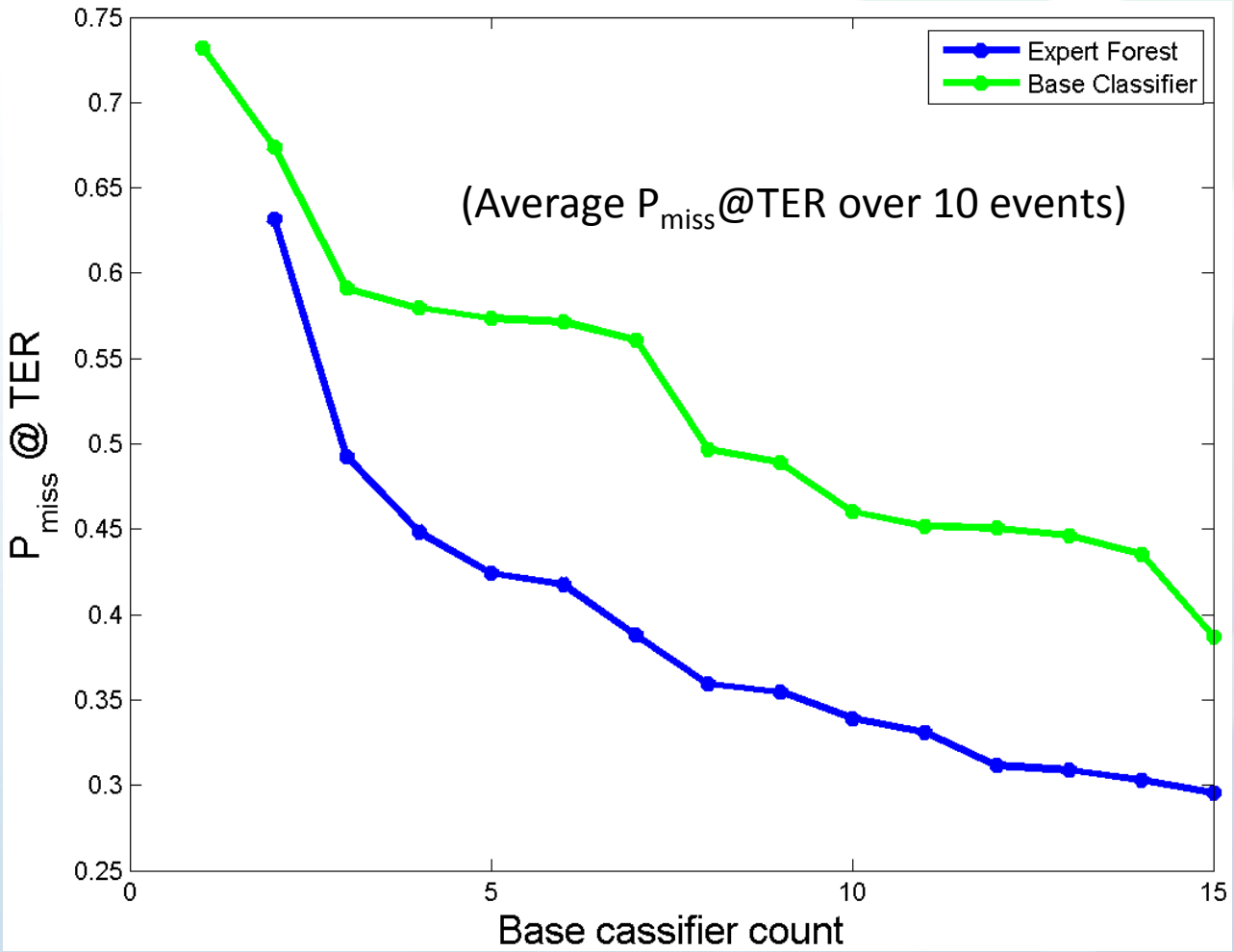
Fusion using Geometric Mean





Using Better Fusion

- With better fusion algorithms, go on for ever?
  - Again, does this "solve" MED?



“Solving” MED

- Scene Types Model
  - Begin to “understand” the constituent elements of the video
- MED <-> text
  - Begin to “understand” semantics (of low-level features, black box classifiers, etc.)

Video Representation

- Bag-of-words model?

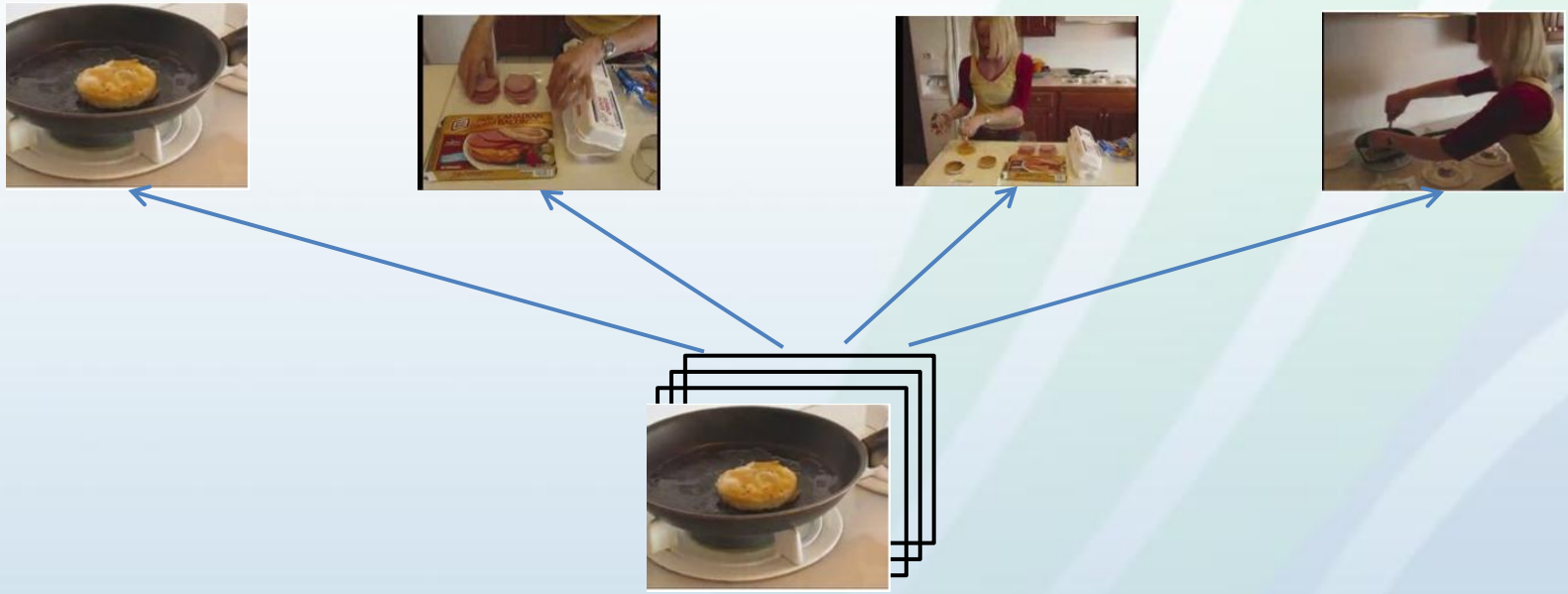
  - Simple model, lose all temporal information

- Temporal model, e.g. HMM?

  - Relatively temporal rigid structure, often model every frame

Key frame Representation

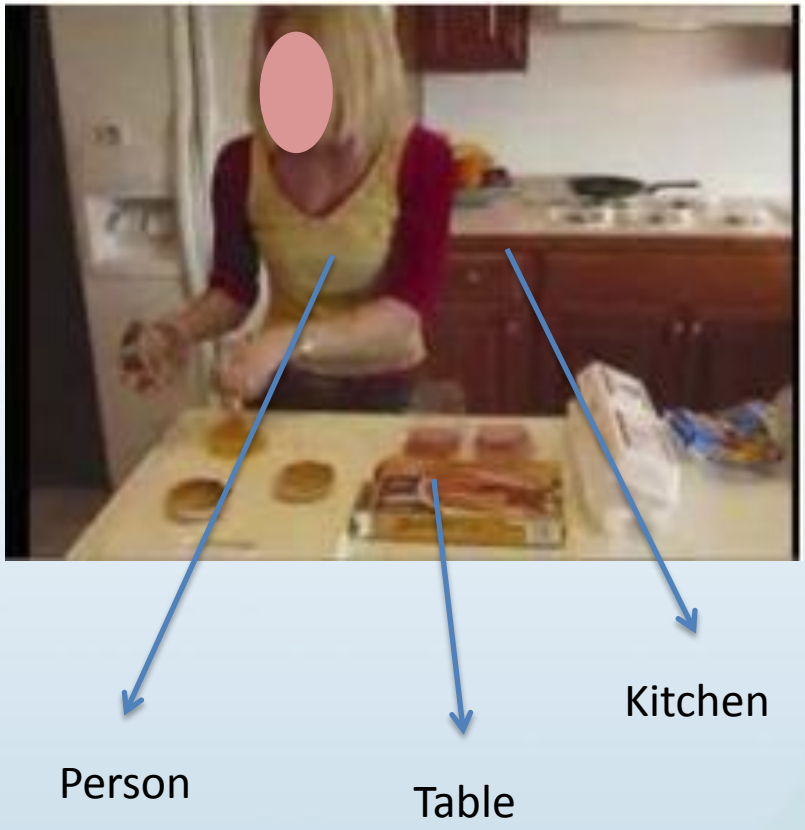
- We use a key frame representation
  - Describe event class by a small set of discriminative sub-events



How to describe a key frame?

Scene Types

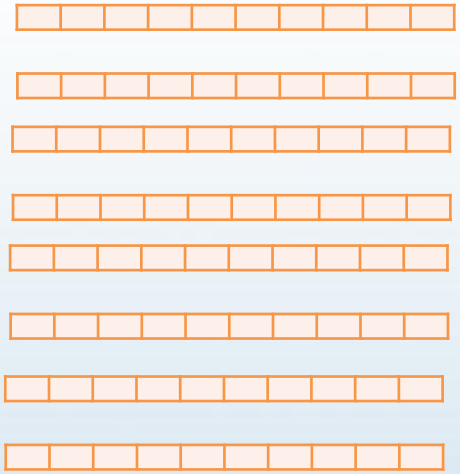
- "Scene types" discrete quantization of individual frames



"This is a scene in a kitchen with a person at a table" (scene type X)

Learning Scene Types

- Scene types are automatically learned by clustering training video frames



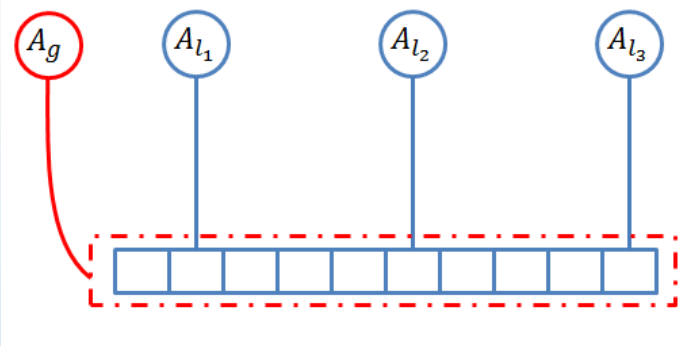
Frames

Object Bank Features

"Scene Type" Clusters

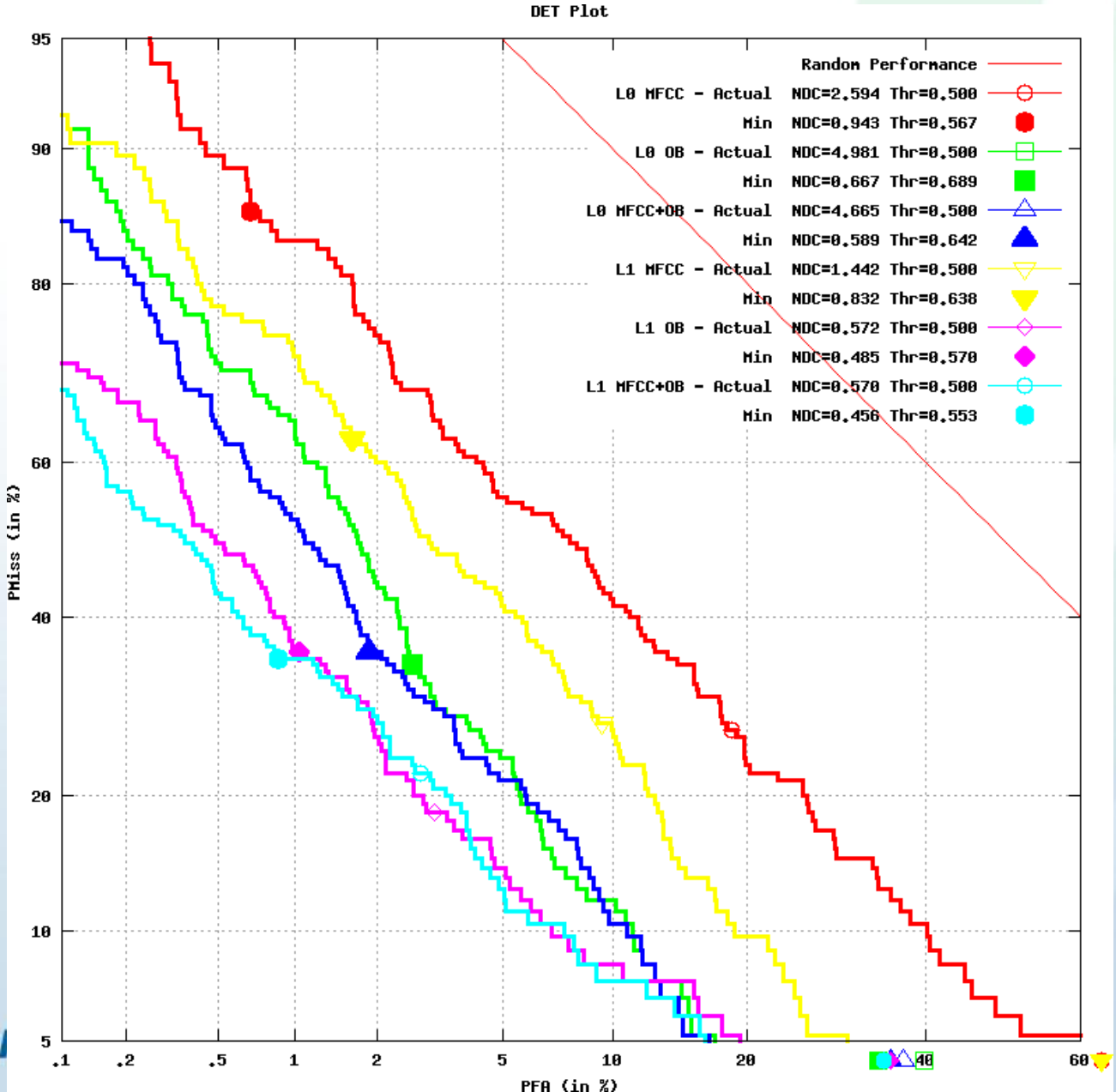
Learning Scene Type Model for an Event

- Scene types contain some useful clusters
  - And lots of slag
- Which are useful for discriminating an event?
- Develop a Latent SVM to automatically learn which scene types are discriminative for the event

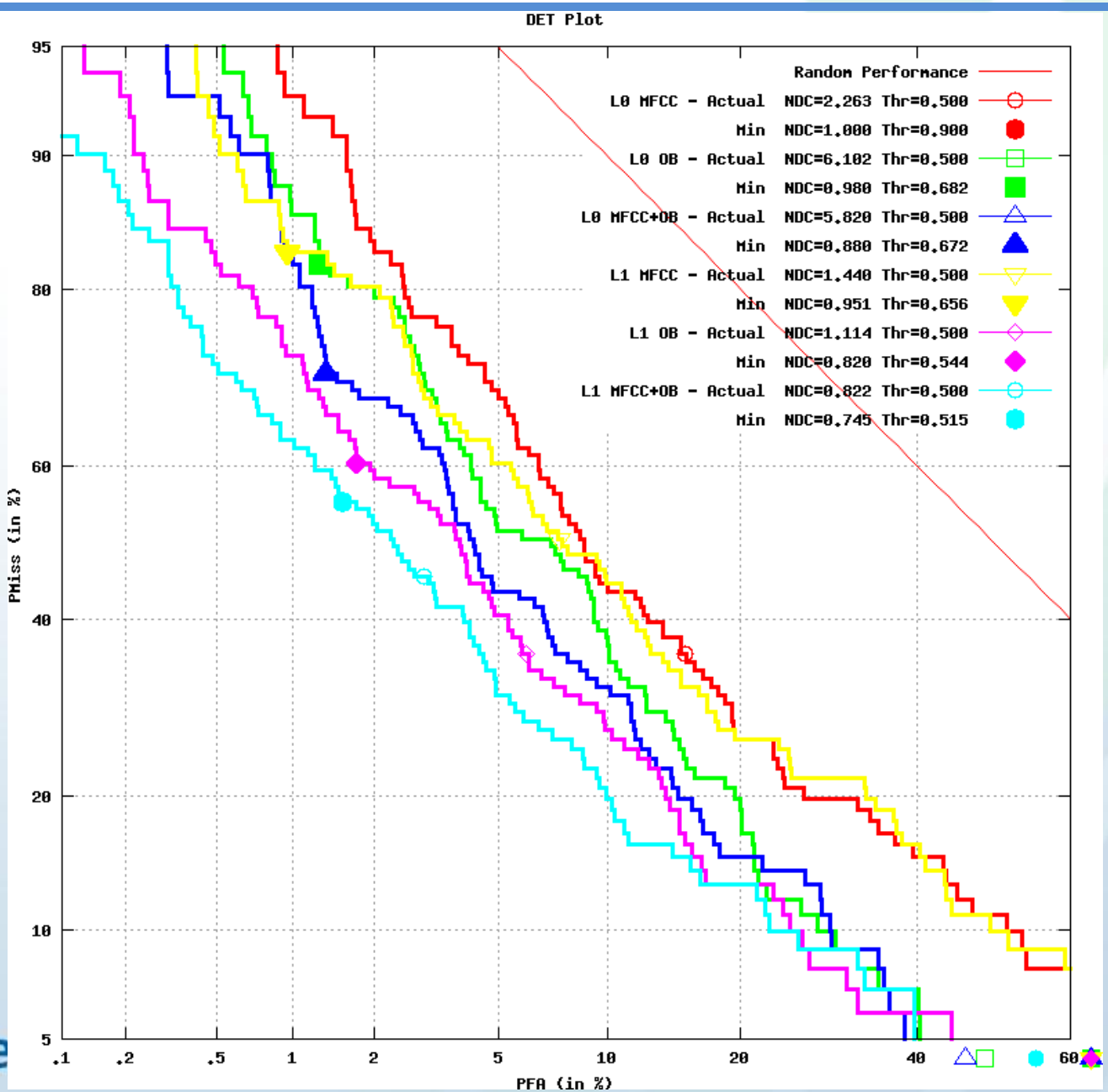


  - Parameters describe which scene types occur in which events
  - Learning only needs single video-level event label
    - All other information is latent, automatically inferred during training/testing

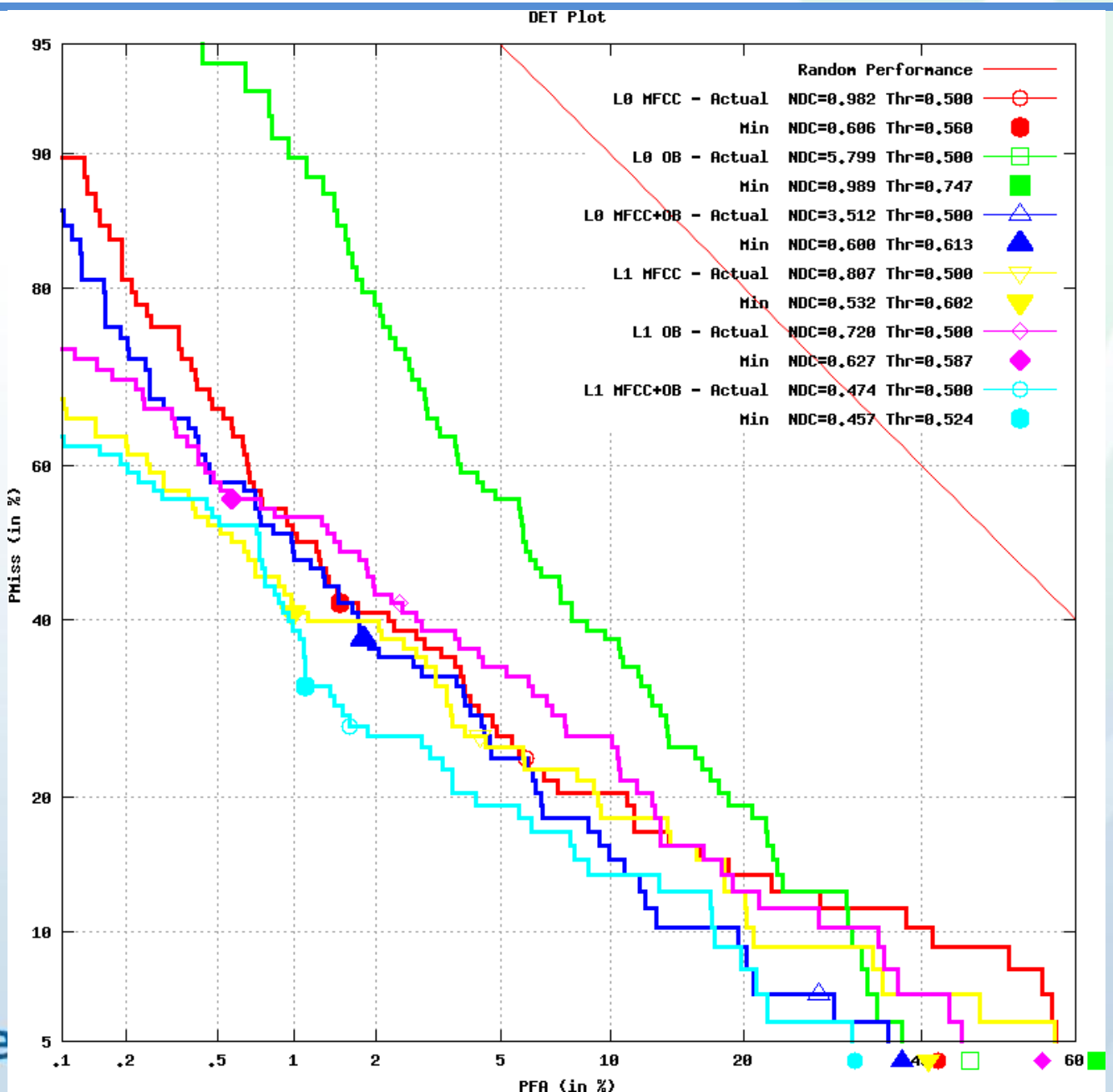
DET Curve (Event 8: Flash Mob)



DET Curves (Event 13: Parkour)



DET Curves (Event 14: Repairing an Appliance)



MED Results

- Probability of missed detection at 5% false positive rate (lower is better)

	Event Class										
System	1	2	3	4	5	6	7	8	9	10	Average
L0:MFCC	52.9	75.0	55.6	67.1	85.0	77.8	64.7	69.3	26.1	74.7	64.8
L0:OB	70.9	48.2	23.7	30.5	66.3	62.2	49.2	51.5	55.7	54.4	51.3
L0:MFCC+OB	50.6	39.3	21.5	**26.8**	65.0	62.2	39.0	43.6	**23.9**	**50.6**	42.3
L1:MFCC	**40.7**	67.9	41.5	64.6	76.3	71.1	59.9	60.4	25.0	67.1	57.5
L1:OB	50.0	**38.4**	**14.1**	32.9	**60.0**	**45.2**	**34.2**	**40.6**	34.1	**49.4**	**39.9**
L1:MFCC+OB (proposed)	**34.3**	**33.9**	**12.6**	**25.6**	**48.8**	**54.1**	**28.3**	**30.7**	**19.3**	**50.6**	**33.8**

Analysis

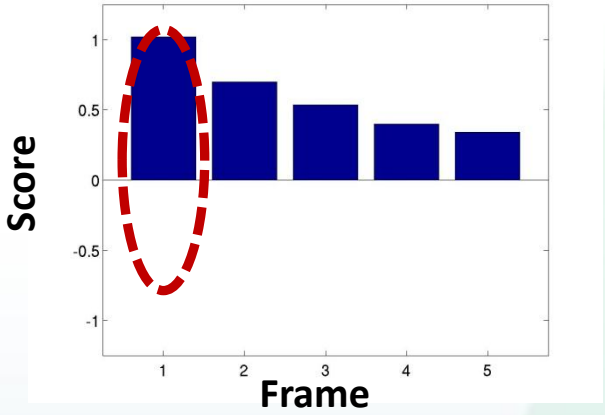
- Following slides show examples from MED11 data

- For each video, set of 5 latent key frames are shown
  - Scores for all key frames
  - Corresponding scene-type cluster for each latent key frame

High-scoring Positives

Making a Sandwich

Latent Frame Scores

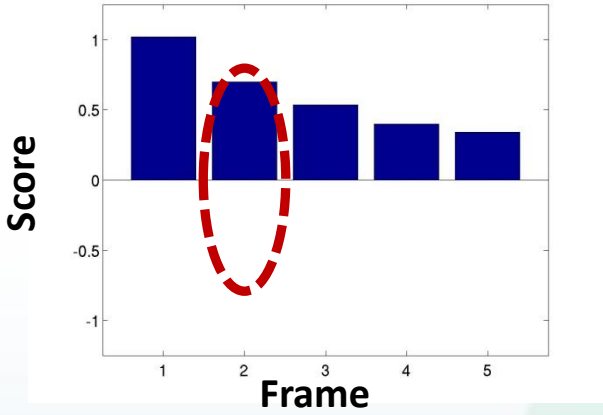


Latent Frames

Corresponding Scene-Type Cluster

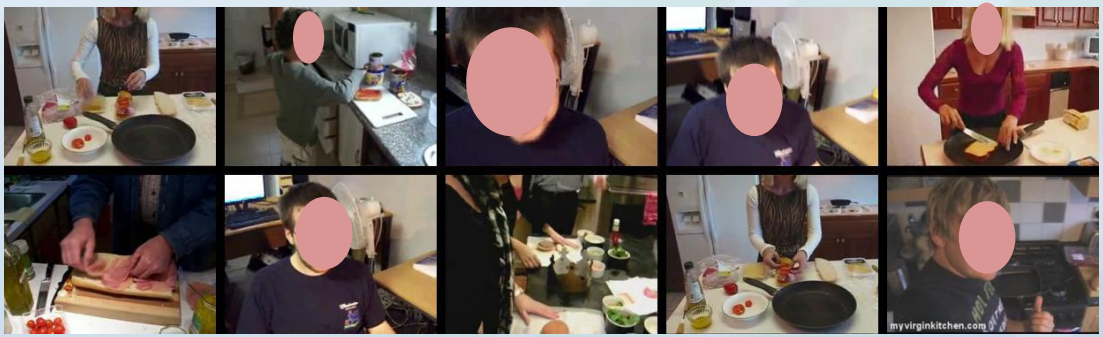
Making a Sandwich

Latent Frame Scores



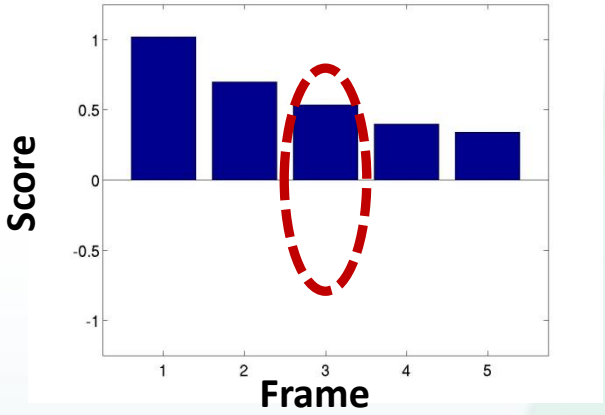
Latent Frames

Corresponding Scene-Type Cluster



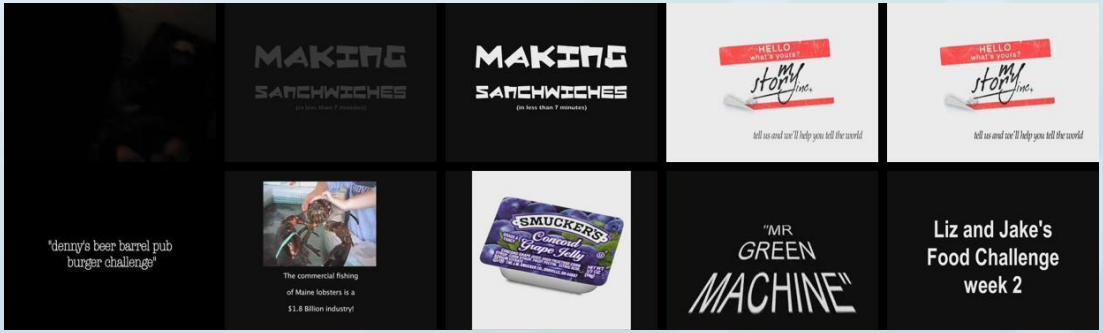
Making a Sandwich

Latent Frame Scores



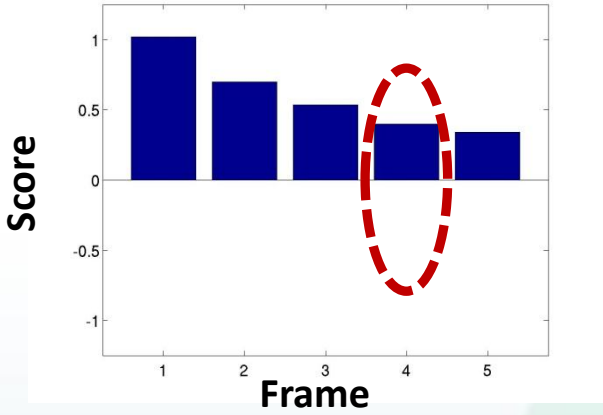
Latent Frames

Corresponding Scene-Type Cluster



Making a Sandwich

Latent Frame Scores

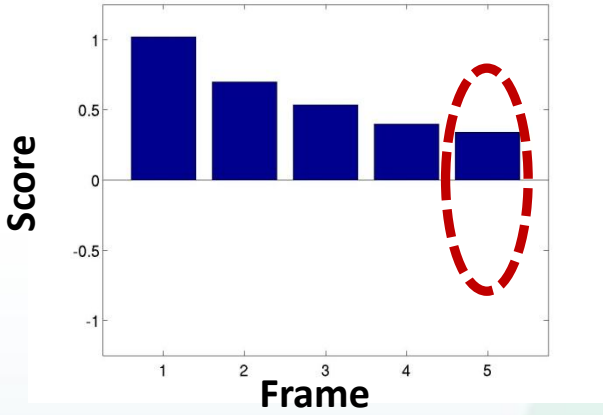


Latent Frames

Corresponding Scene-Type Cluster

Making a Sandwich

Latent Frame Scores



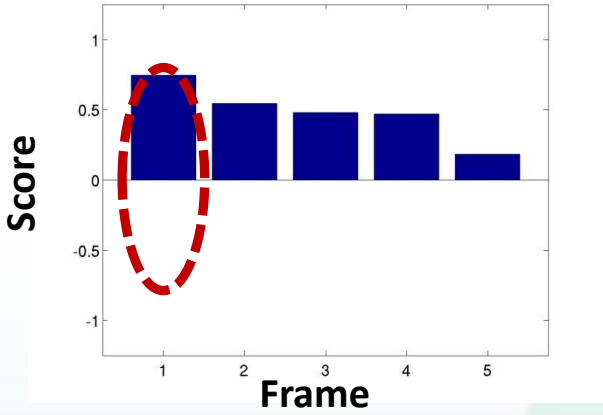
Latent Frames

Corresponding Scene-Type Cluster

Hard Negatives

Making a Sandwich, Hard Negative

Latent Frame Scores

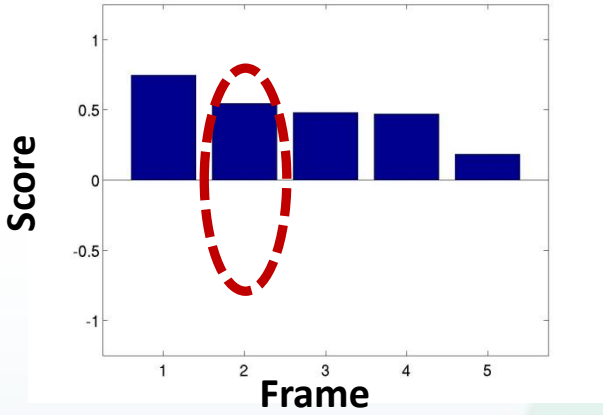


Latent Frames

Corresponding Scene-Type Cluster

Making a Sandwich, Hard Negative

Latent Frame Scores

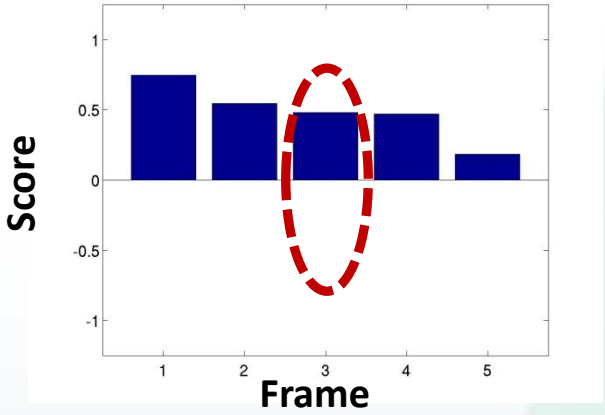


Latent Frames

Corresponding Scene-Type Cluster

Making a Sandwich, Hard Negative

Latent Frame Scores

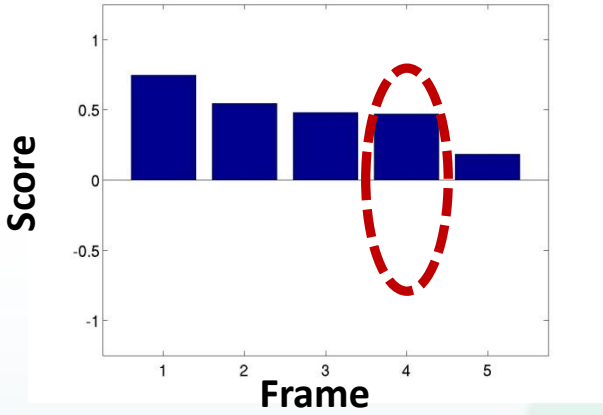


Latent Frames

Corresponding Scene-Type Cluster

Making a Sandwich, Hard Negative

Latent Frame Scores

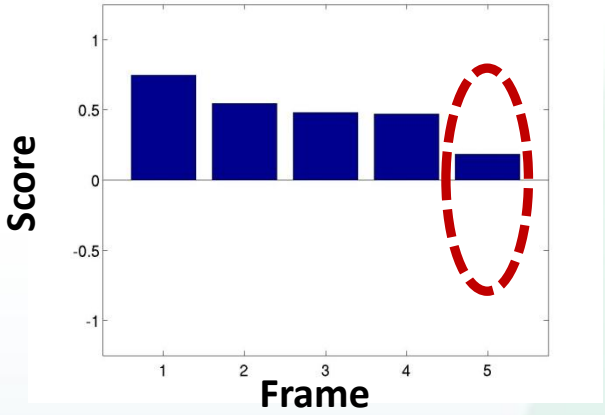


Latent Frames

Corresponding Scene-Type Cluster

Making a Sandwich, Hard Negative

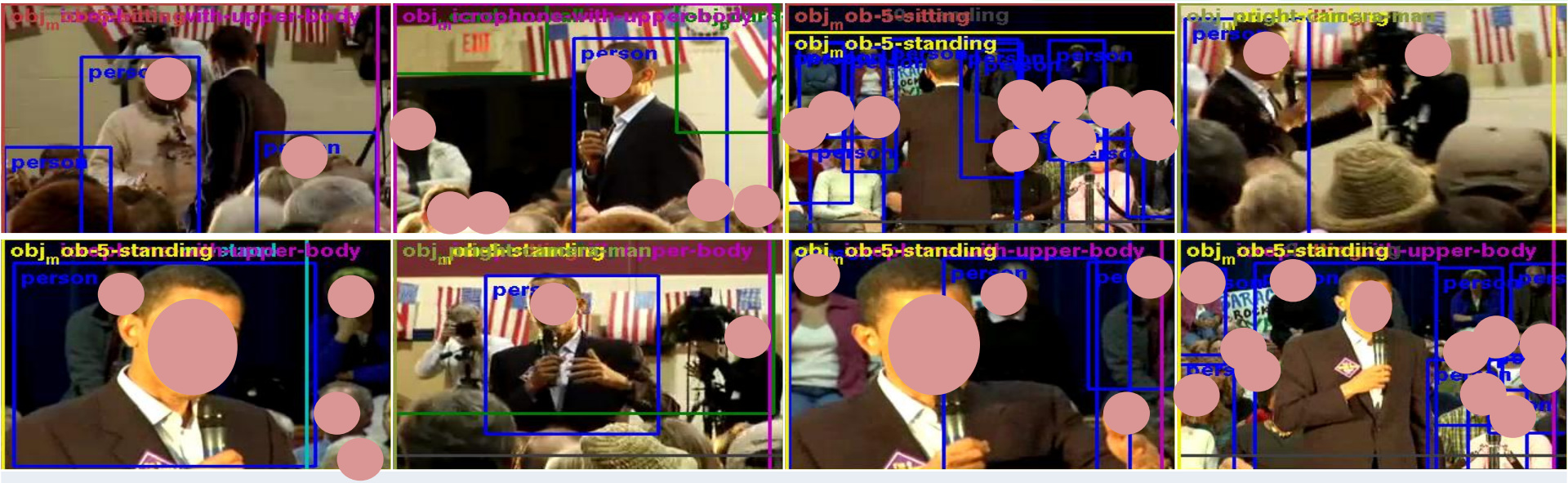
Latent Frame Scores



Latent Frames

Corresponding Scene-Type Cluster

Visualized MER output for HVC585090

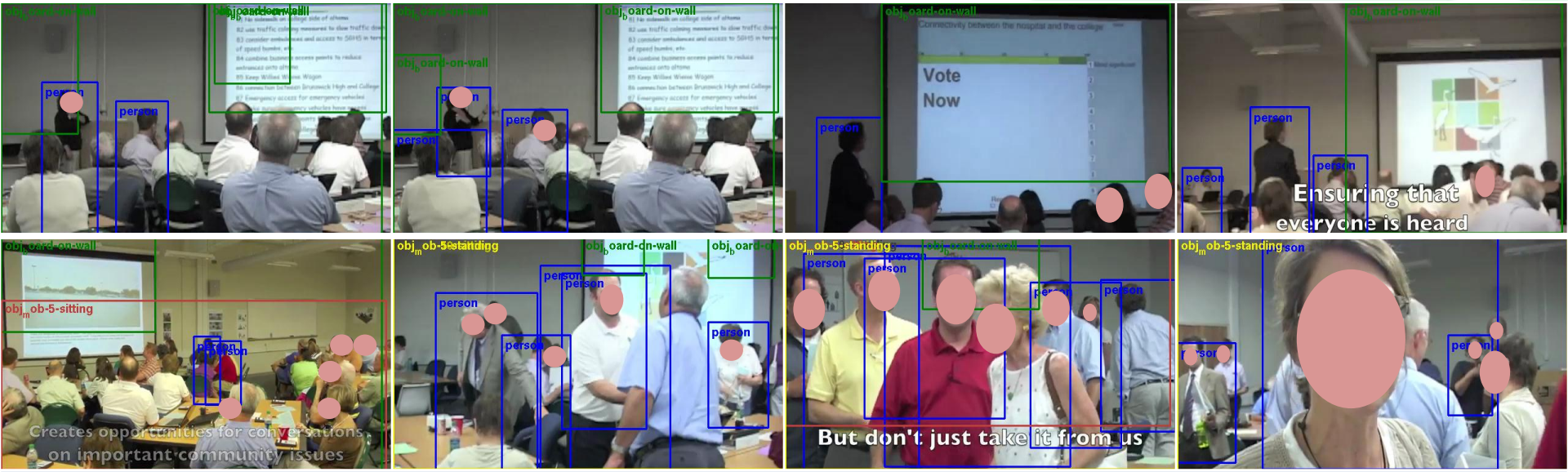


Object Evidence: microphone-with-upper-body, microphone-on-stand, upright-camera-man, mob-5-sitting, mob-5-standing, mob-10-standing, board-on-wall, person

Scene Evidence: crowded indoor

Inferred Evidence Descriptions: *Labels from topic and Part-of-speech models*
meeting/VERB town/NOUN hall/OBJ microphone/OBJ man/SUBJ-HUMAN people/OBJ speaks/VERB woman/SUBJ-HUMAN chairs/NOUN talking/VERB standing/VERB cameras/OBJ politician/SUBJ-HUMAN podium/OBJ speaking/VERB
(**Human Summary** - the president answers questions at a town hall meeting in New Hampshire)

Visualized MER output for HVC104842



Object Evidence: mob-5-sitting, board-on-wall, mob-5-standing, mob-10-sitting, mob-10-standing, person

Scene Evidence: crowded indoor

Inferred Evidence Descriptions: *Labels from topic and Part-of-speech models*
meeting/VERB hall/OBJ town/NOUN woman/SUBJ-HUMAN people/OBJ speaks/VERB question/VERB microphone/OBJ audience/SUBJ representative/SUBJ-HUMAN man/SUBJ-HUMAN talking/VERB asks/VERB podium/OBJ chairs/NOUN
(**Human Summary** - amateur ad for an institute that instructs and hosts town hall meetings)

Beyond DET Curves

- Demonstration of exploration tool

IARPA

Thanks!

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069 and by the Defence Advanced Research Projects Agency (DARPA) under contract number HR0011-08-C-0135. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, DARPA, or the U.S. Government.

Kitware