GENIE

General Engine for Indexing Events

TRECVID MED 2011

5 Dec 2011

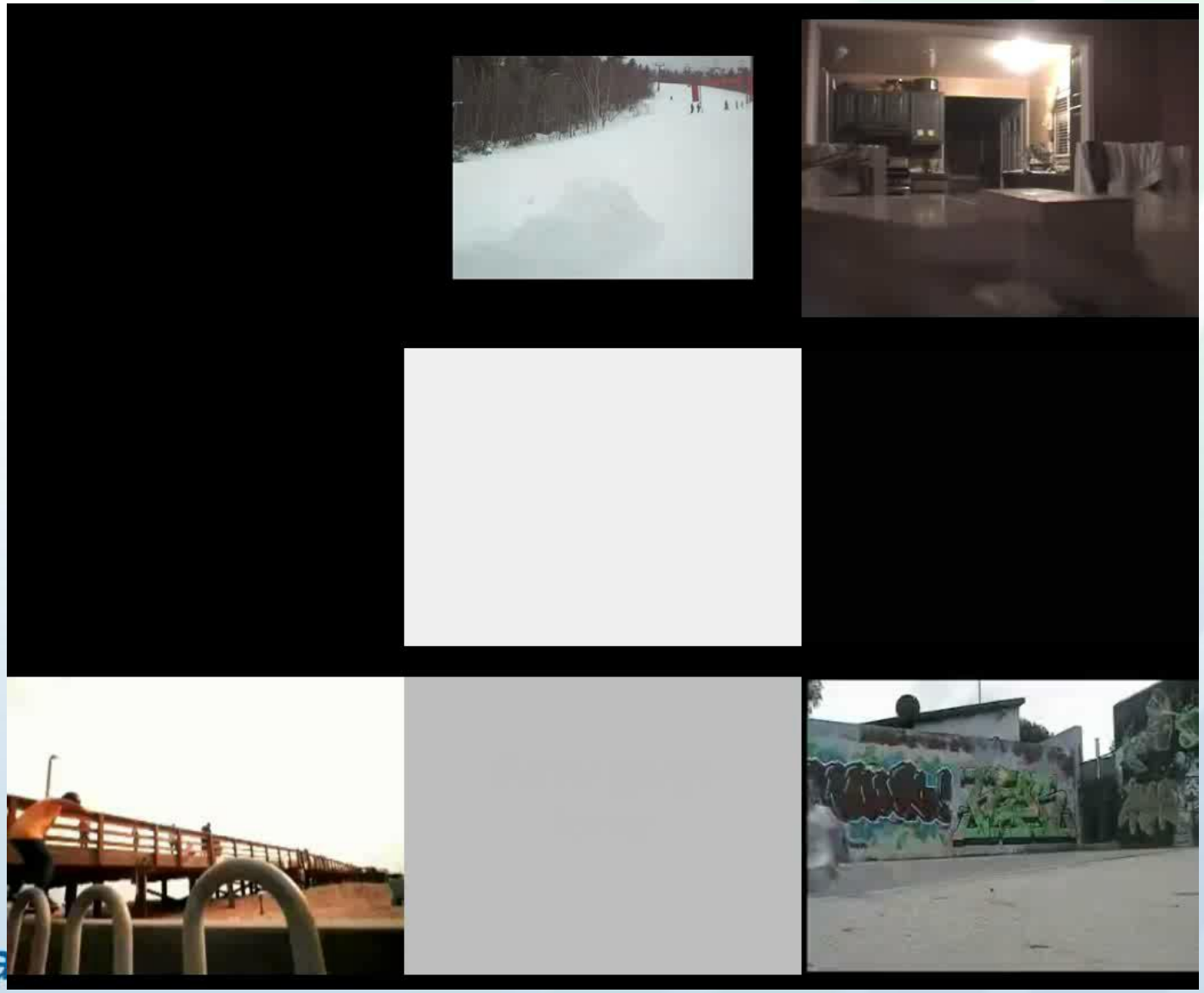
Amitha Perera

Kitware	Honeywell	Stanford	Simon Fraser University	Georgia Tech	SUNY Buffalo
Sangmin Oh	Scott McCloskey	Kevin Tang	Zhi Feng Huang	Ilseo Kim	Kang Li
Matt Leotta	Ben Miller	Daphne Koller	Arash Vahdat	Byungki Byun	Gang Chen
Ben Boeckel	Jingchen Liu	Fei-fei Li	Weilong Yang	Chin-Hui Lee	Jason Corso
Amitha Perera			Greg Mori		Yun Fu
					Rohini Srihari

Funded by IARPA ALADDIN contract D11PC20069 BAA-10-01

Intraevent Diversity

Attempting a Board Trick (first 90 seconds, at 3x speed)



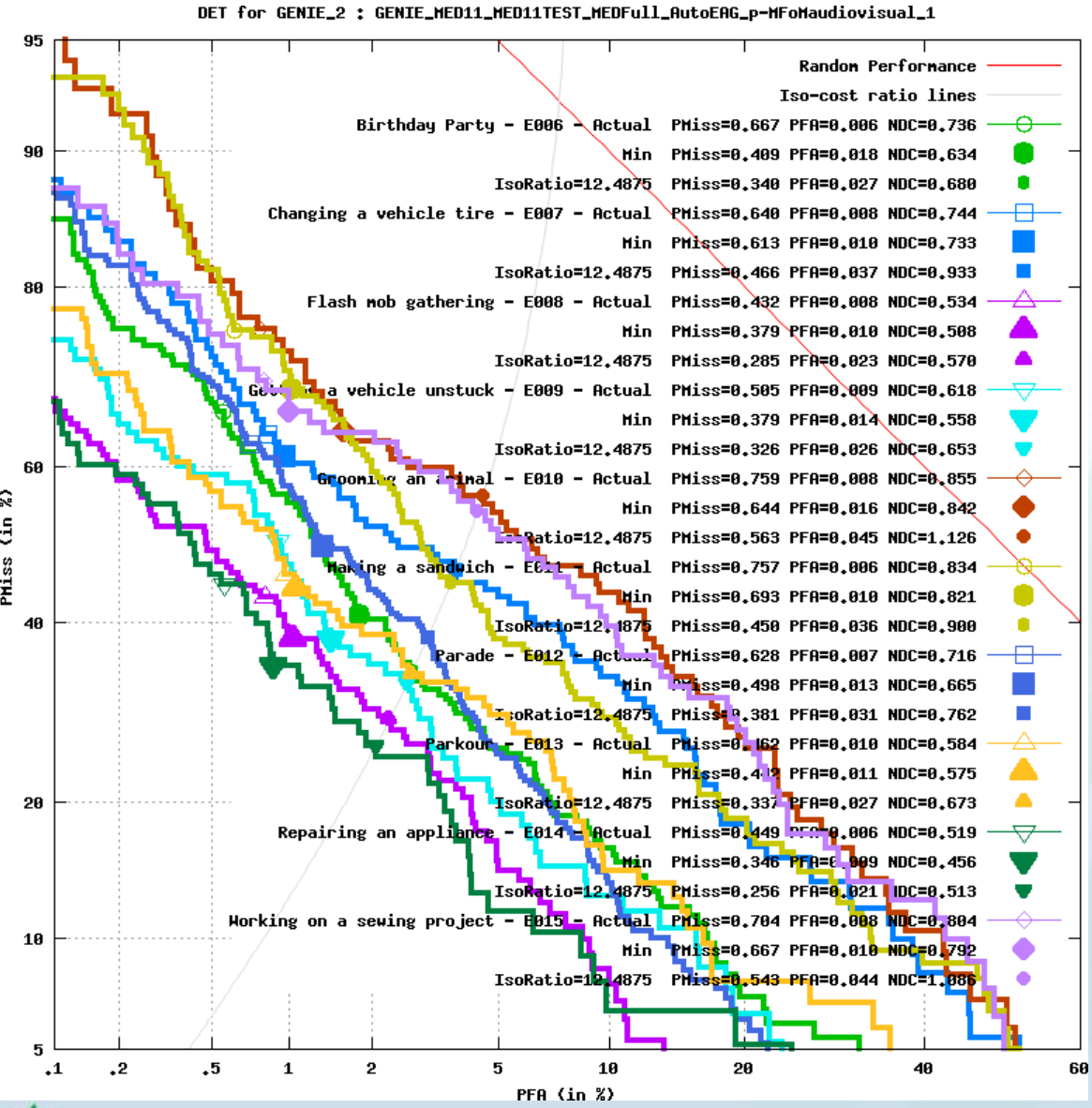
Intraevent Diversity

Wedding Ceremony (first 90 seconds, at 3x speed)

Goals

- Obtain baseline result
  - Measure performance of current vision/machine learning techniques
  - Approach MED problem from data side (instead of high-level modeling)
- Ad-hoc
  - No specific tuning of detectors to events
  - No extra annotation

MED 2011 Result



E006: Birthday Party

E007: Changing a vehicle tire

E008: Flash mob gathering

E009: Getting a vehicle unstuck

E010: Grooming an animal

E011: Making a sandwich

E012: Parade

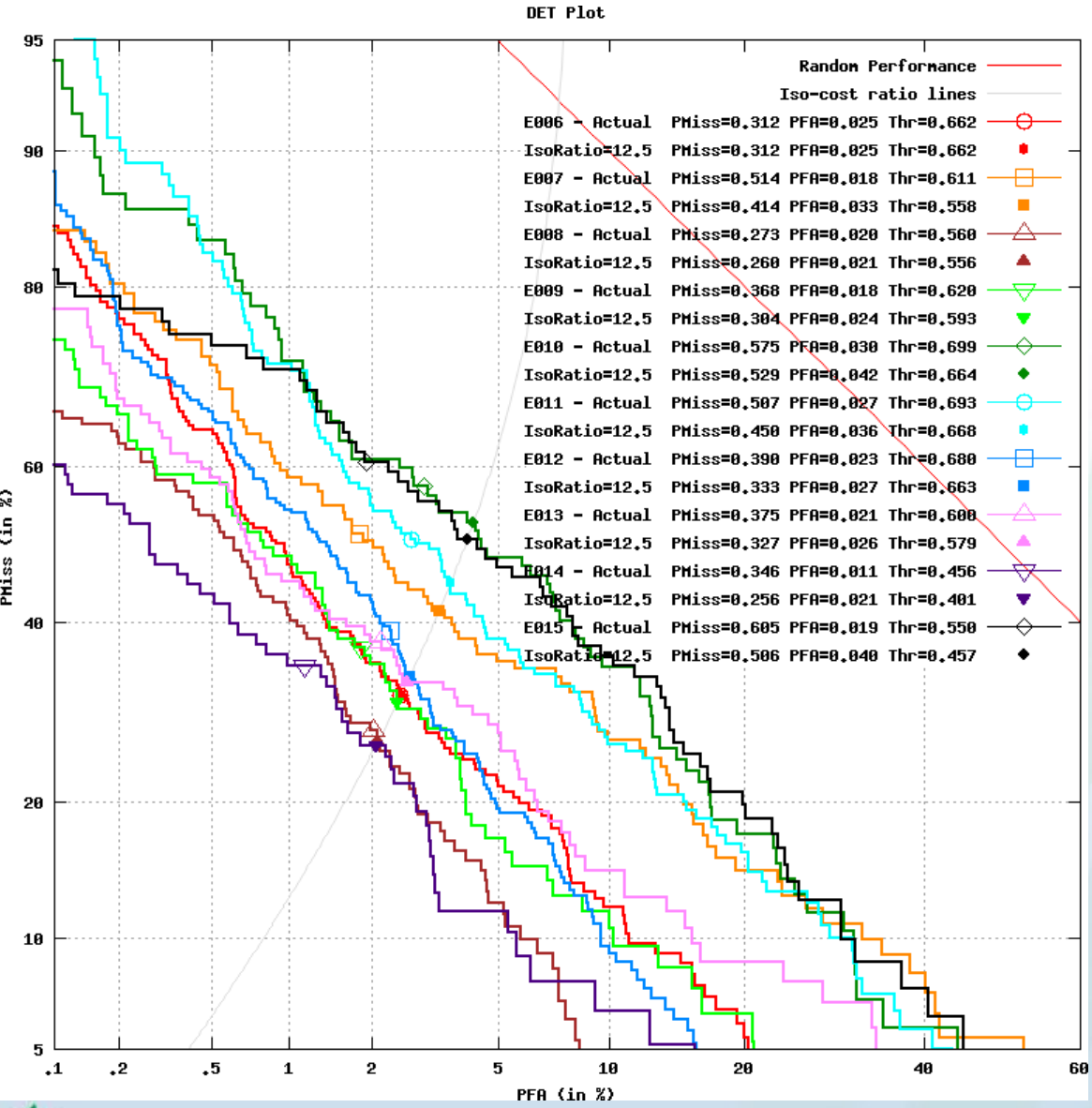
E013: Parkour

E014: Repairing an appliance

E015: Working on a sewing project

Nov 10 result, by NIST

MED 2011 Corrected Result



E006: Birthday Party

E007: Changing a Vehicle Tire

E008: Flash mob gathering

E009: Getting a vehicle unstuck

E010: Grooming an animal

E011: Making a sandwich

E012: Parade

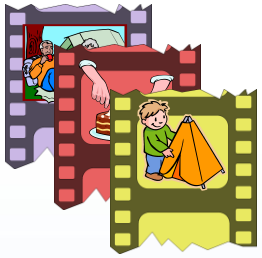
E013: Parkour

E014: Repairing an appliance

E015: Working on a sewing project

Data Flow

Multimedia Archive



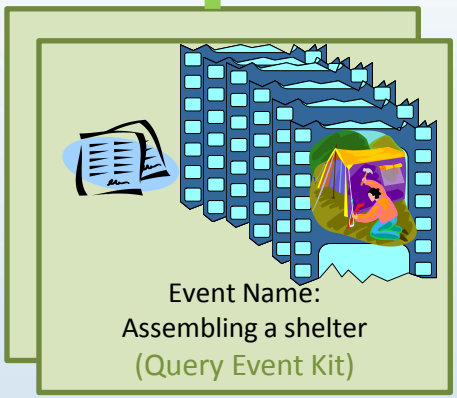
Feature Computation

500 core Linux PC cluster, 4 GB RAM per core

© ChrisDag, flickr, CC-by-2.0

- HOG3D
- Object Bank
- GIST
- MFCC
- ASM

Single quad-core PC, 8GB RAM



Event Name:
Assembling a shelter
(Query Event Kit)

Event Kits

Codebook generation

Base Classifiers

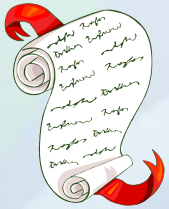
- SVM
- SVM
- SVM
- SVM
- SVM

Learn non-linear SVM

Learn fusion model

Score Fusion

Model



Result List

Testing

Training

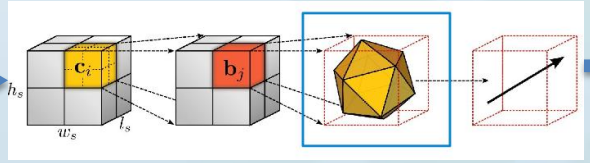
Kitware

Feature 1: HOG3D

- Alexander Kläser, Marcin Marszałek, and Cordelia Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients". BMVC 2008.
- Resize videos to max. 160 pixels wide/tall, maintain aspect ratio
- Sample HOG3D descriptors every 9x9 pixels spatially, 5 frames temporally
- Dense sampling, no interest point operator
- Vector quantize: 1024 codewords obtained from k-means
- Histogram Intersection distance

Codebook

Sampled Points

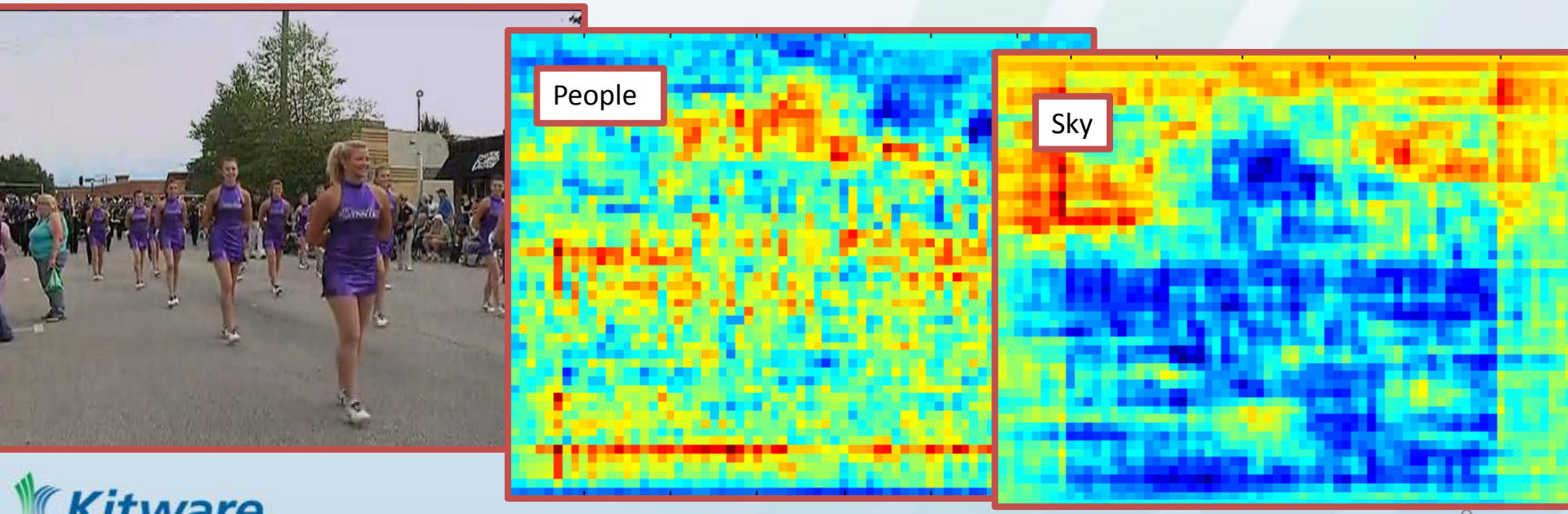


HOG3D Descriptors

Codeword Histogram

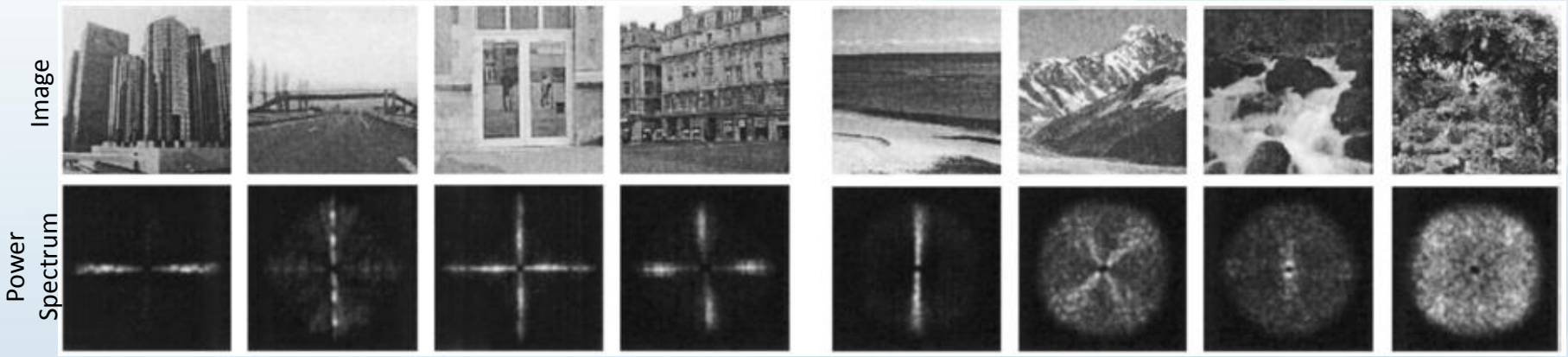
Feature 2: Object Bank

- Li-Jia Li, Hao Su, Eric P. Xing and Li Fei-Fei, "Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification". NIPS, 2010.
- 177 object detectors run at different scales over each frame
  - Computed at key frames
  - 44604-d feature vector, reduced to 177-d by choosing max response per object type
  - Max pooling over all frames
- L2 distance



Feature 3: GIST

- Oliva and Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope". IJCV 2001.
- Gist describes scene type, which correlates with activity.
  - Indoor (sandwich, appliance) / Outdoor (parade, board trick)
  - Man-made (parkor, parade) / Natural (landing fish)
- Full frame descriptor
- L2 distance



Feature 4: MFCC

- P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental". Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed., 1976.
- 32-dimensional feature is extracted at every 10ms with a 25ms window.
- Vector quantized to 1024 codewords
- BoW (Bag-of-Words)
  - Constructed using entropy normalized word count w_{ij}

$$w_{ij} = (1-\varepsilon_i)\frac{c_{ij}}{n_j}, \text{ where } \varepsilon_i = -\frac{1}{\ln(N)}\sum_{j=1}^{N}\left(\frac{c_{ij}}{t_i}\right)\ln\left(\frac{c_{ij}}{t_i}\right), t_i = \sum_{j=1}^{M} c_{ij}, n_j = \sum_{i=1}^{N} c_{ij}$$

- Histogram Intersection distance

Feature 5: ASM

- Acoustic Segment Models computed over phone-sized units
- General audio can be represented
- Feature vector is unigram and bigram staistics,

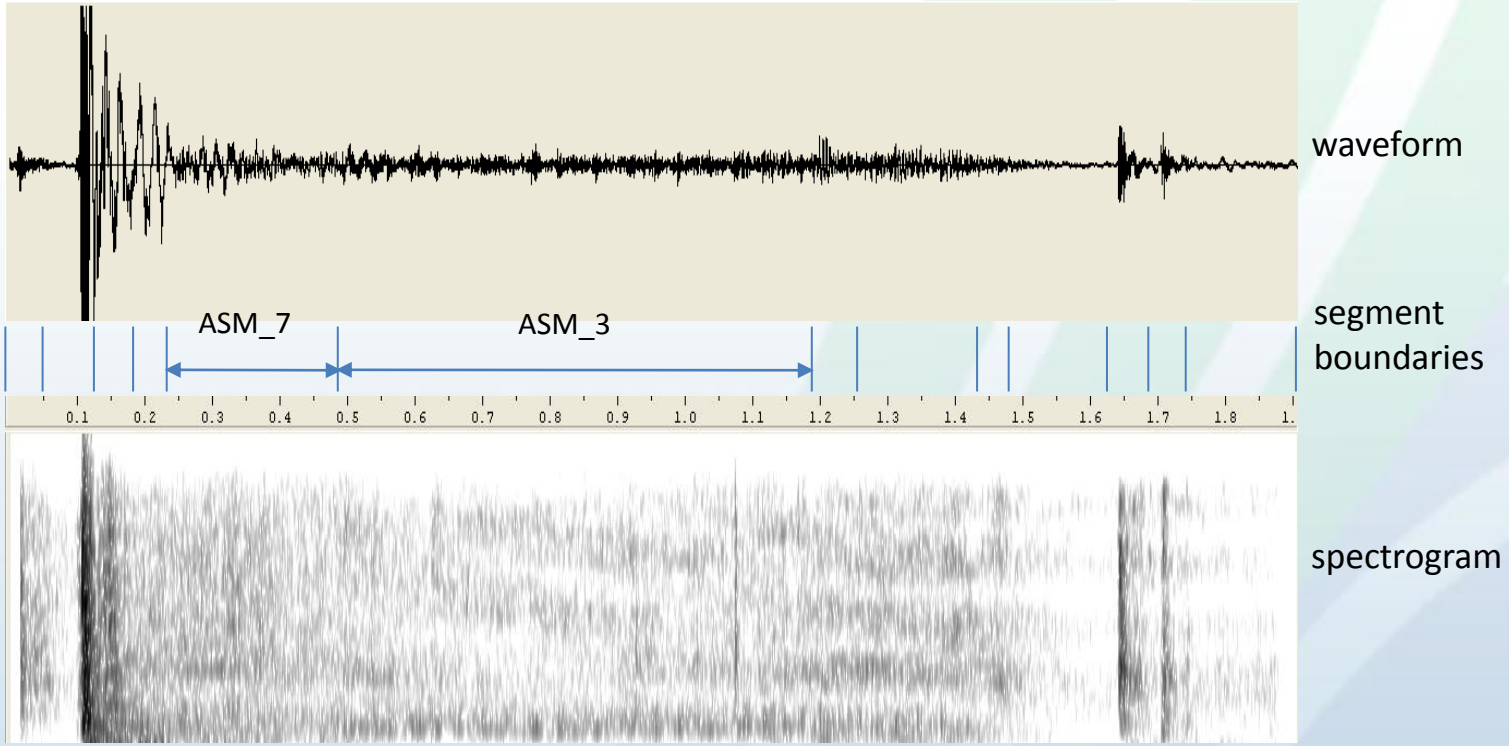
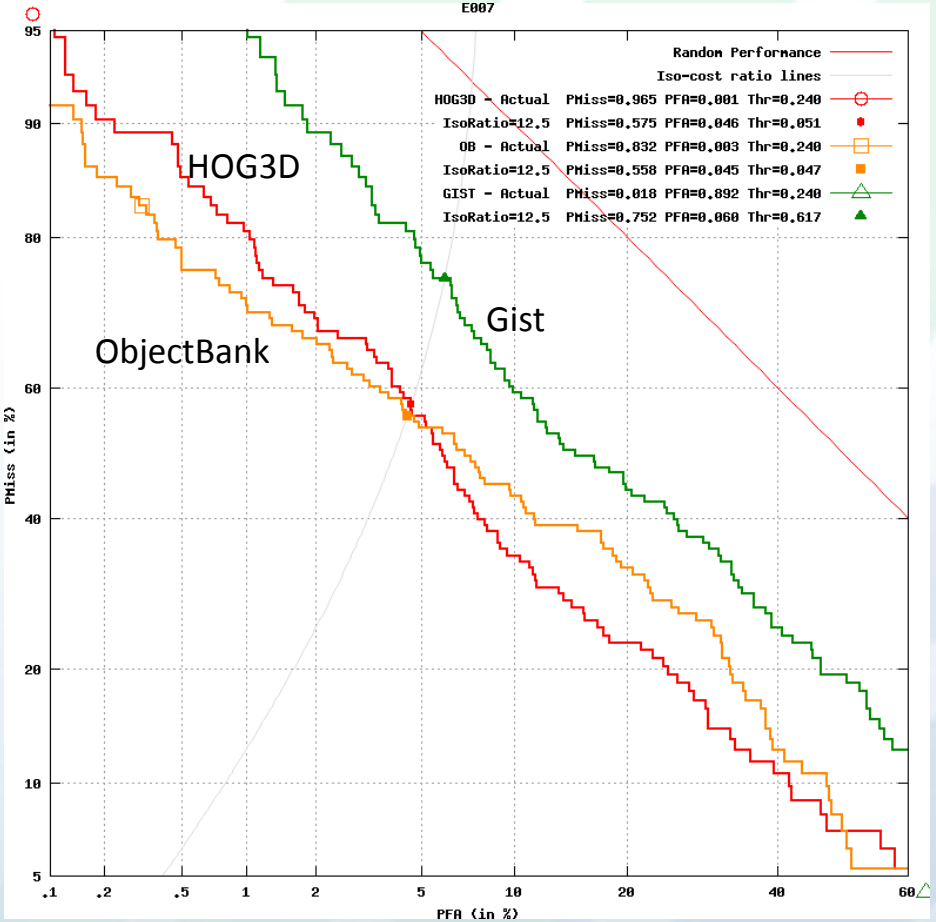


Fig. An example of an audio event, skateboarding, segmented with ASMs

Feature Fusion

- Combined base classifiers using score fusion
- Simple linear fusion is not effective
  - Different base classifiers are better as different parts of the score space / ROC space
- Use different weights in different parts of the space



Non-linear Feature Fusion

- Explored two non-linear fusion algorithms:
  - Mixture of Local Experts (MLE)
  - Maximal Figure of Merit (MFoM) optimization with a non-linear SVM
- MLE
  - Partition score space into segments and learn a different expert for each segment
- MFoM
  - Iteratively learns classifier parameters to optimize an objective function
  - Not just maximize margin in score space
  - Currently using a single non-linear SVM

MLE Score Fusion

Threshold estimated by looking at performance on pseudo-test data extracted from event kit.

E006: Birthday Party

E007: Changing a Vehicle Tire

E008: Flash mob gathering

E009: Getting a vehicle unstuck

E010: Grooming an animal

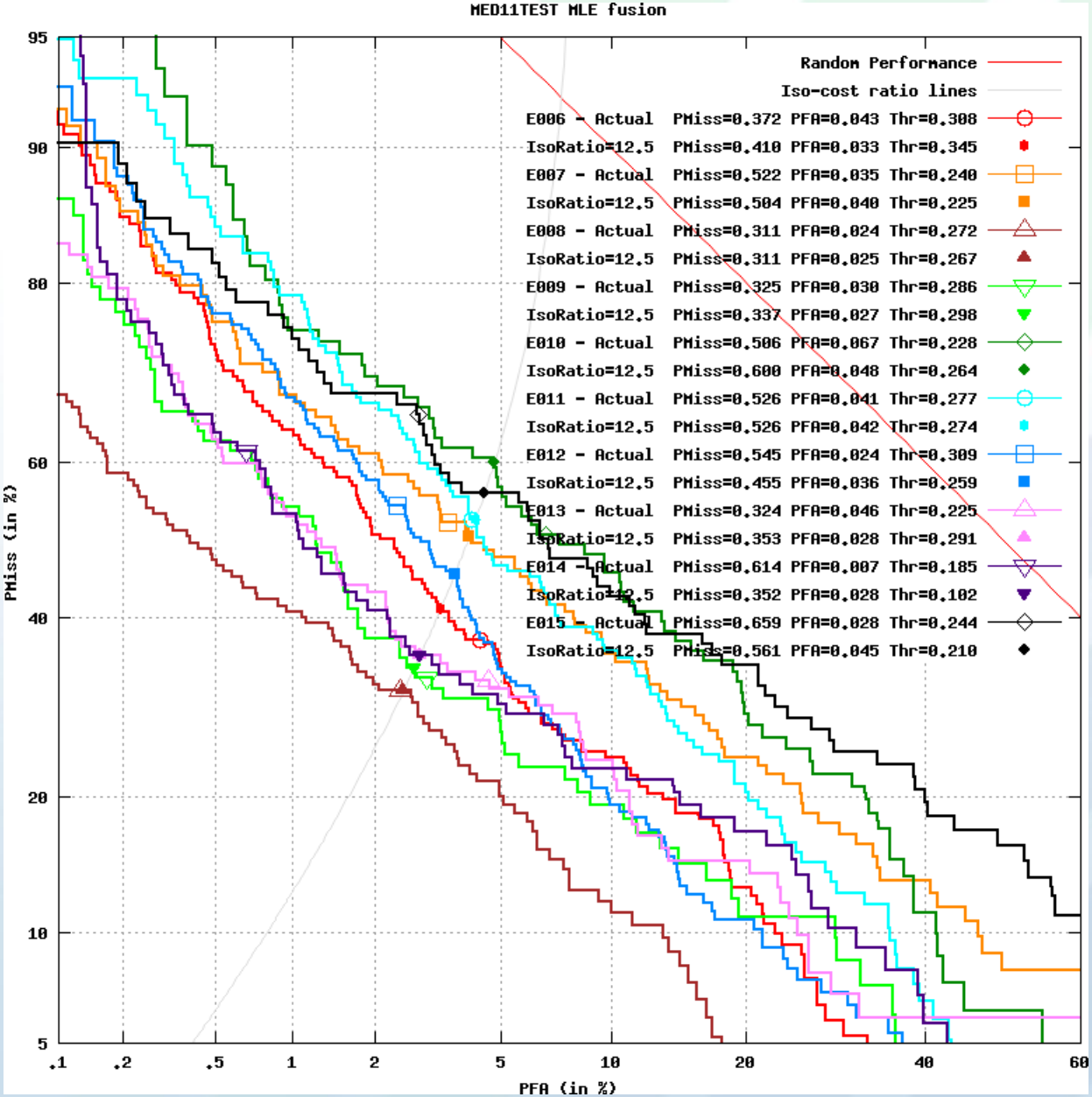
E011: Making a sandwich

E012: Parade

E013: Parkour

E014: Repairing an appliance

E015: Working on a sewing project



MFoM Optimization

Threshold estimation is part of the optimization process, based on the margins of the SVM

E006: Birthday Party

E007: Changing a Vehicle Tire

E008: Flash mob gathering

E009: Getting a vehicle unstuck

E010: Grooming an animal

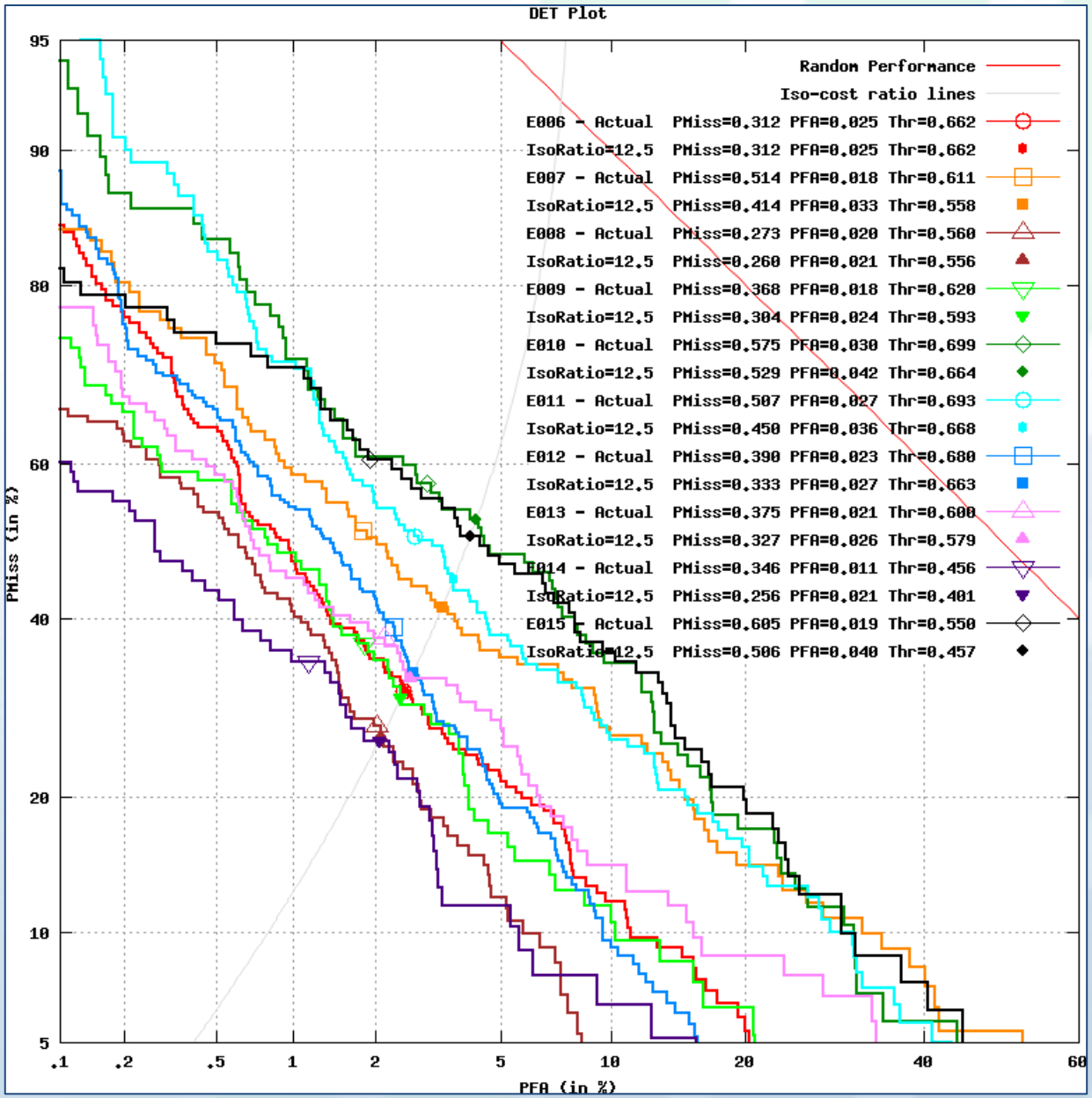
E011: Making a sandwich

E012: Parade

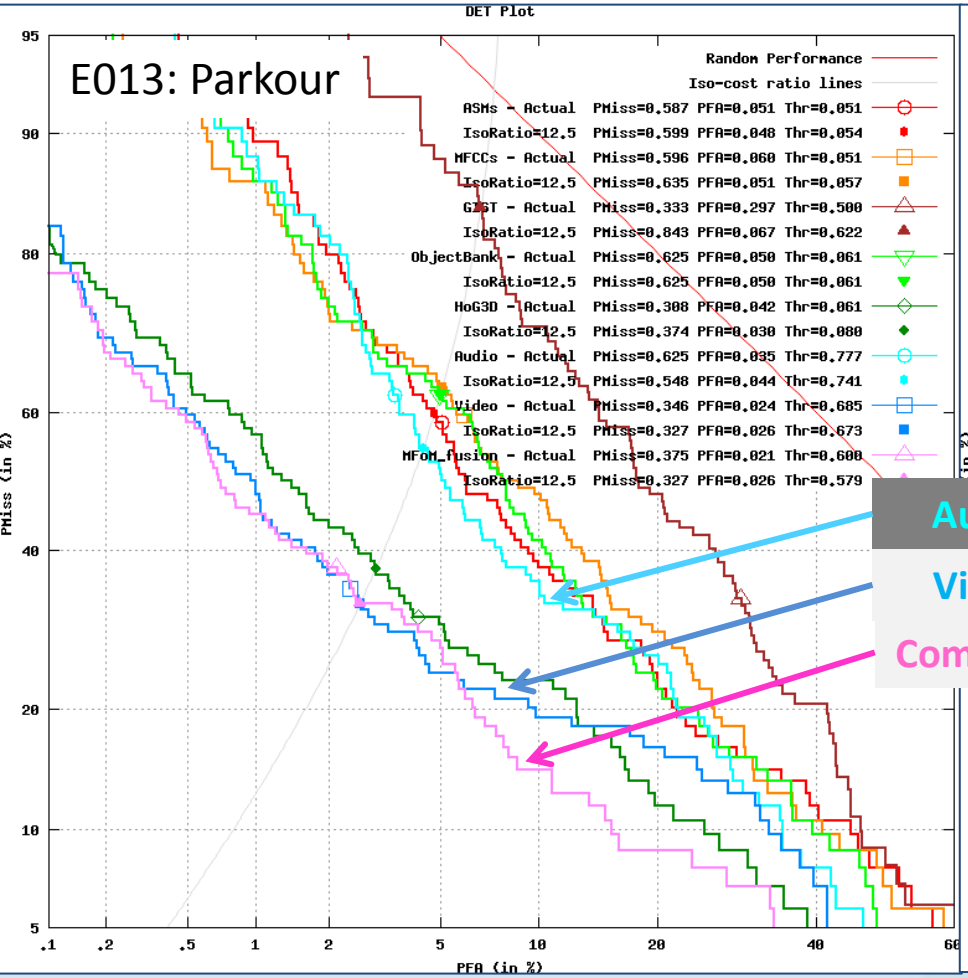
E013: Parkour

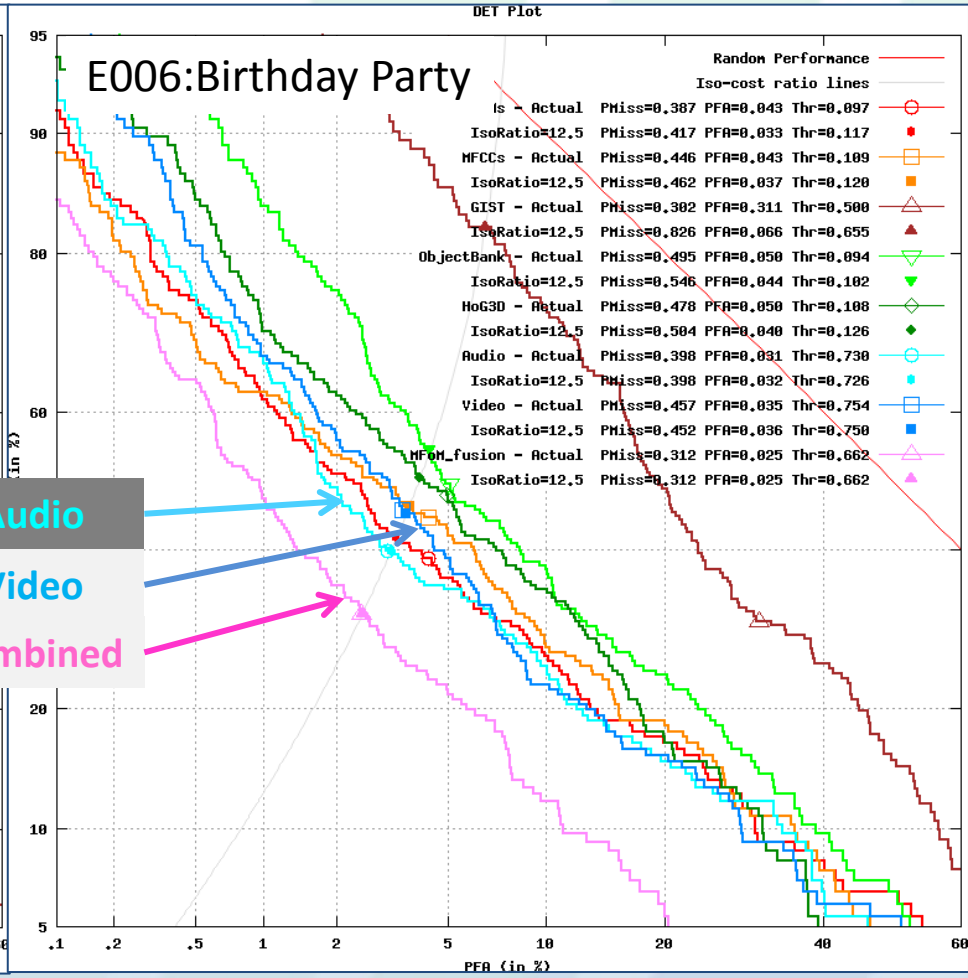
E014: Repairing an appliance

E015: Working on a sewing project



Contributions of Base Classifiers





Audio

Video

Combined

ASMs

MFCCs

GIST

ObjectBank

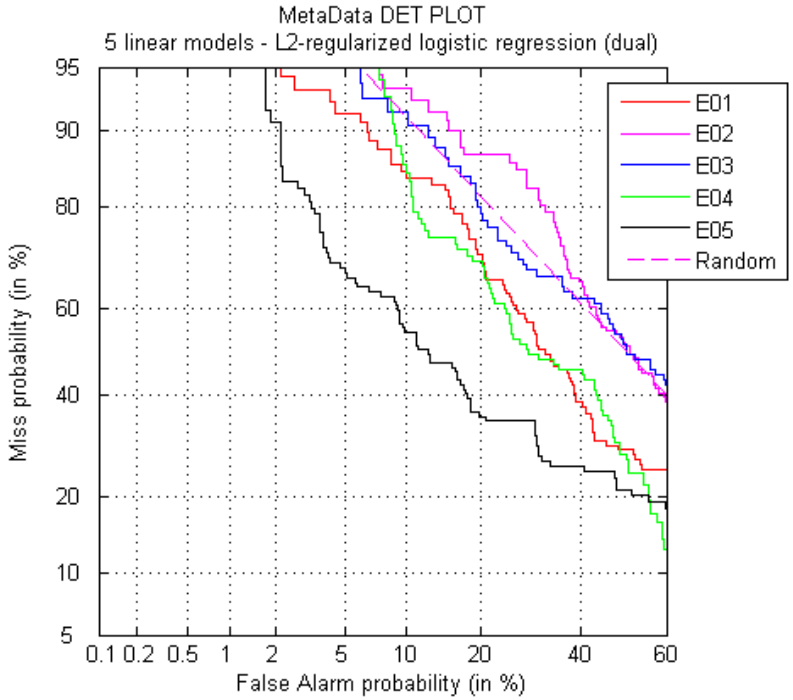
HoG3D

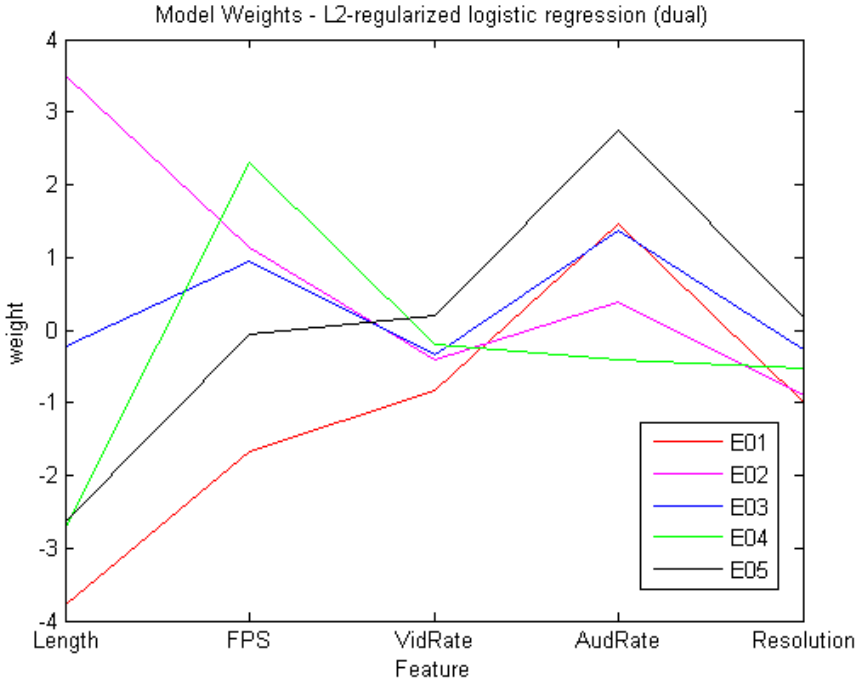
MFoM Fusion (MFCCs + ASMs)

MFoM Fusion (HoG3D + OB + GIST)

MFoM Fusion (HoG3D + OB + GIST + MFCCs + ASMs)

Metadata-based Classification





Events:

1. Board trick – short, low frame rate
2. Feeding animal – long, high frame rate
3. Landing fish – high frame rate, high audio bitrate
4. Wedding ceremony – short, high frame rate
5. Woodworking – short, high audio bitrate

Summary

- Established a strong baseline result on MED11 showing the performance of state-of-the-art, data-driven techniques
- Video-word (HOG3D), noisy semantic (Object Bank), and gross (Gist) video features are complementary
- Low-level (MFCC) and intemediate-level (ASM) audio features are complementary
- Audio and video features are complementary
- Need non-linear fusion (for score fusion)
- Metadata carries a signal (but should we use it?)