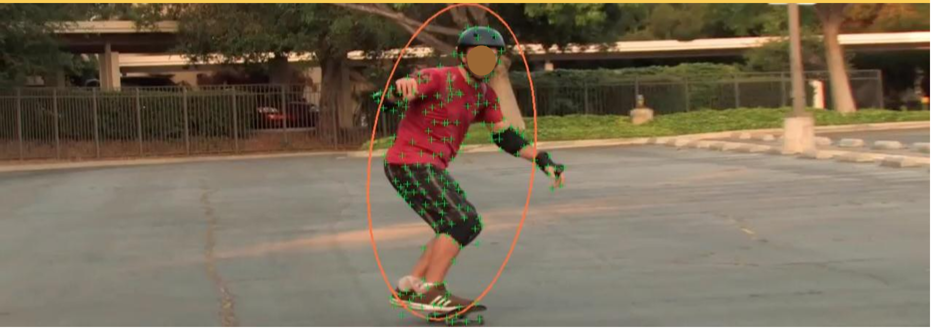
SRI International



SESAME

(Video SEarch with Speed and Accuracy for Multimedia Events)
Multimedia Event Detection (MED) System

November 26, 2012

SRI International

UNIVERSITY OF AMSTERDAM

UNIVERSITY OF SOUTHERN CALIFORNIA 1880

Overview

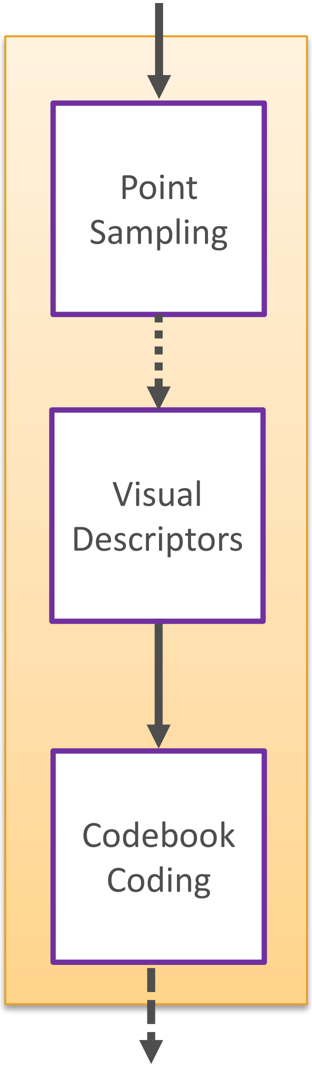
- Event classifiers
- Fusion and threshold selection
- Waypoint experiments on development set
- MED12 evaluation results

14 Low-level and High-level Event Classifiers

- **Low-level features**
  - visual features (2)
  - motion features (5)
  - audio features (1)

- **Concept-level features:**
  - visual concepts (2)
  - ASR (2)
  - video OCR (2)

Visual Features: Bag-of-Words and Difference Coding



Bag of words for

event agents *and*

visual scenes, objects, persons, actions

State-of-the-art

ColorSIFT [Van de Sande et al. TPAMI 2010]

Soft-Assignment [Van Gemert et al. TPAMI 2010]

Real-time Bag-of-Words [CIVR09 best paper]

TSIFT [under review]

Two event classifiers based on visual features

1 frame sampled in every 2 seconds of video

	Sampling	Descriptors	Codebook	Aggregation	Kernel
1.Color average coding	Dense Harris	PCA reduced: SIFT, CSIFT, TSIFT	4096, hard 1x1,1x3	Average	Fast HIK
2. Color difference soft coding	Dense	PCA reduced: SIFT, CSIFT, TSIFT	1024, soft 1x1,1x3	Average	Linear

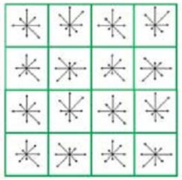
Waypoint experiments showed:

- average coding outperformed difference coding
- difference coding complemented average coding in late fusion experiments

Low-level motion features

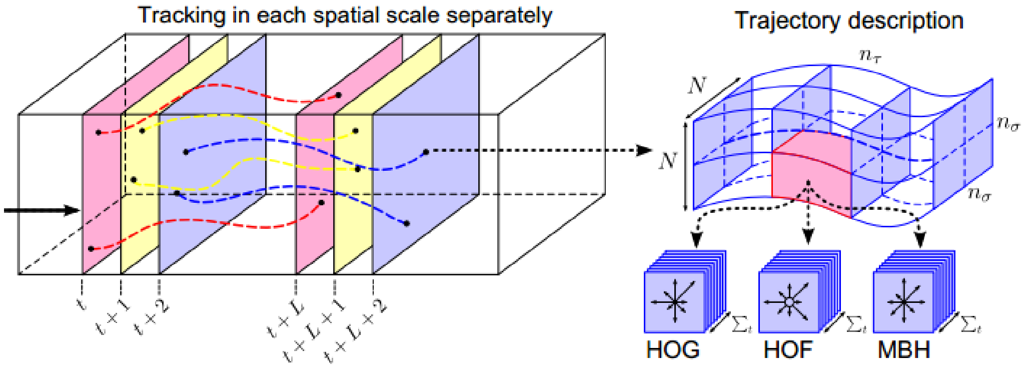
- **STIP:**
  - Corner like detectors in 3D
  - 72-dim HOG + 90-dim HOF

- **MoSIFT**
  - SIFT like detectors in 2D, filtered by motion
  - Extracted, quantized and pooled by CMU



Alexander Kläser et al,
BMVC 2008

- **Dense Trajectories (DT):**
  - Generate tracklets for densely sampled points
  - Describe each tracklet by shape, HoG, HoF and MBH of the volume around it



Wang et al.,
CVPR 2011

Event classifiers using low-level motion features

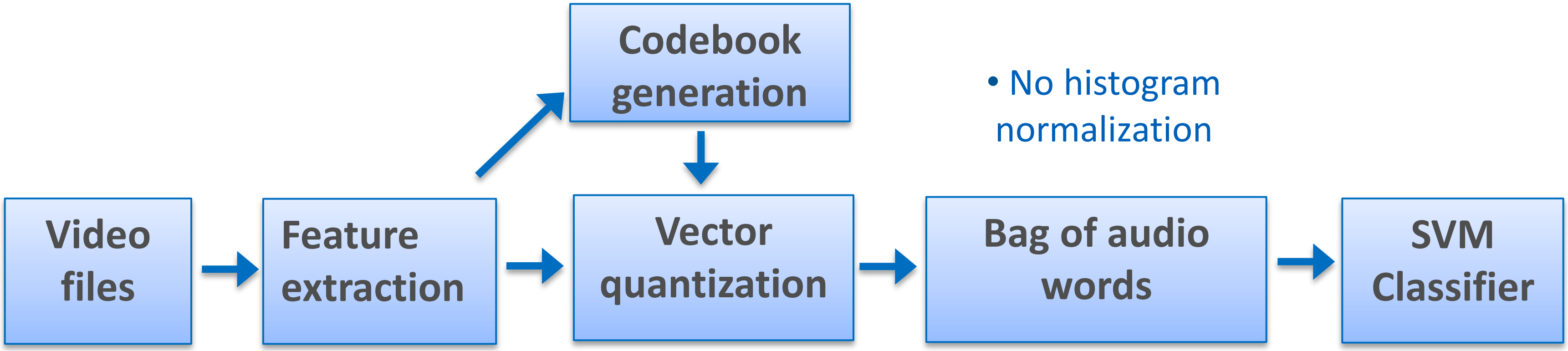
- **5 event classifiers:**

Event Classifier	Feature	Descriptor	Aggregation	Kernel
1	STIP	1st-order Fisher	Average	Gaussian
2	STIP	2nd-order Fisher	Average	Gaussian
3	DT	1st-order Fisher	Average	Gaussian
4	DT	2nd-order Fisher	Average	Gaussian
5	MoSIFT	MoSift	Average	χ2

- **Waypoint experiment showed:**
  - Dense Trajectory gives the best performance
  - 2nd order Fisher vector is better than 1st order
  - All 3 motion features are complementary in late fusion experiments

Event Classifier using Low-level Audio Features

- Codebook size = 1000



Video files

Feature extraction

Codebook generation

Vector quantization

Bag of audio words

SVM Classifier

- No histogram normalization

- 16 kHz sampling rate
- MFCCs every 10 ms
- 12 coeff.+ log-energy + Δ+ Δ-Δ of each = 39 dim total
- No MFCC normalization

- Soft quantization

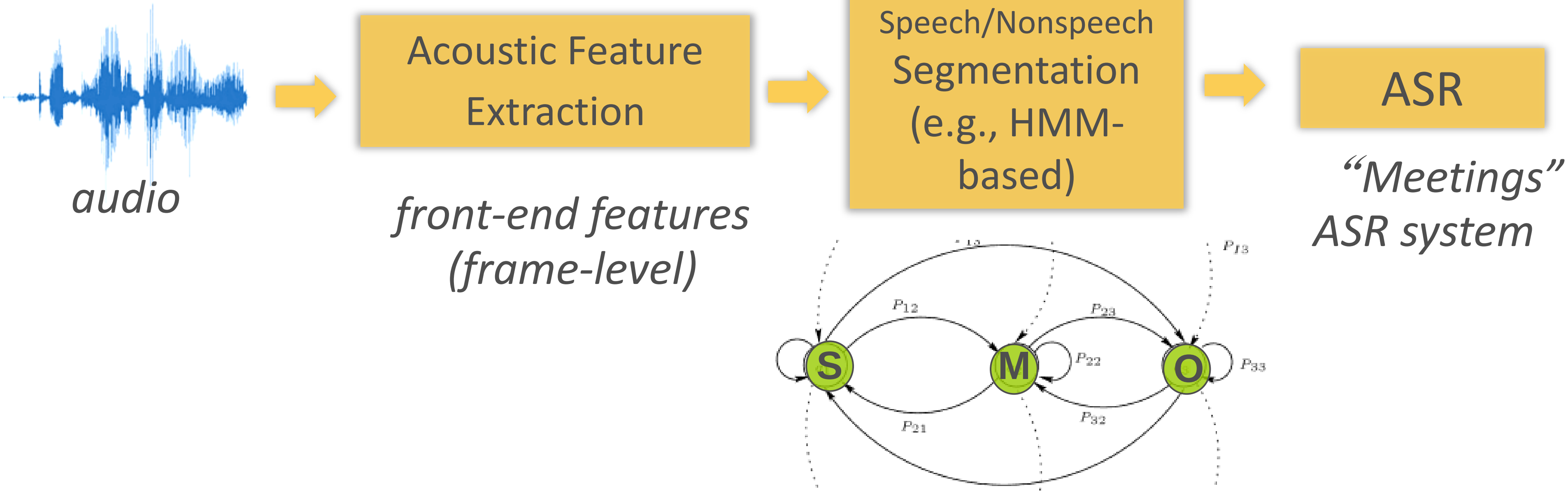
  → add distance from nearest codeword instead of +1 to the histogram for each quantized vector

- Histogram intersection kernel

Event Classifiers using Visual Concept Detectors

- **1346 concept detectors**
  - 346 concepts from the TRECVID 2012 SIN task
  - 1,000 concepts from ImageNet
  - All trained using color difference coding with linear SVM
- **Two event classifiers**
  - One used random forests
  - One used a non-linear SVM

Automatic Speech Recognition (ASR)



audio

Acoustic Feature Extraction

front-end features (frame-level)

Supervised Speech/Nonspeech Segmentation (e.g., HMM-based)

ASR

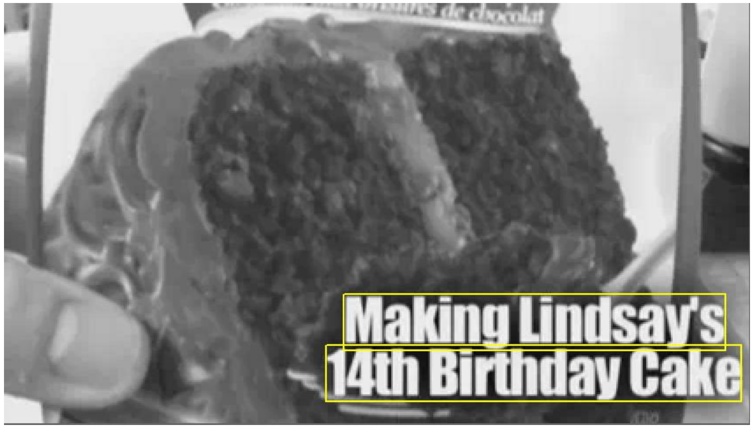
"Meetings" ASR system

- **Un-adapted ASR system trained on far-field microphone meetings data**
- **3-state ergodic HMM for audio segmentation (speech, music, other)**
- **ASR configured to recognize spoken English**

Video OCR

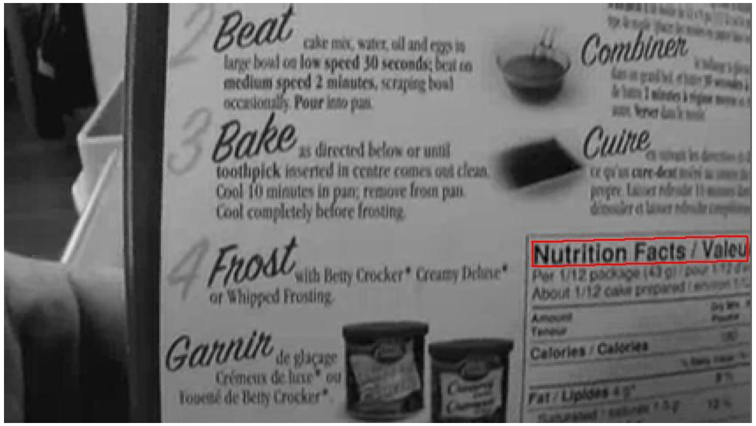
- **SRI’s video optical character recognition (video OCR) for detection, tracking, and recognition of text**
  - **recognizes both overlay text and in-scene text**
  - **configured to recognize English language text**

Text captions



“Making Lindsay s 14th Birthday Cake”

In-scene text


Beat
cake mix, water, oil and eggs in large bowl on low speed 30 seconds; beat on medium speed 2 minutes, scraping bowl occasionally. Pour into pan.
Bake
as directed below or until toothpick inserted in centre comes out clean. Cool 10 minutes in pan; remove from pan. Cool completely before frosting.
Frost
with Betty Crocker® Creamy Deluxe® or Whipped Frosting.
Garnir
Combiner
Cuire
Nutrition Facts / Valeu
Calories / Calories
Fat / Lipides

“Nutrition Facts – Valeu”

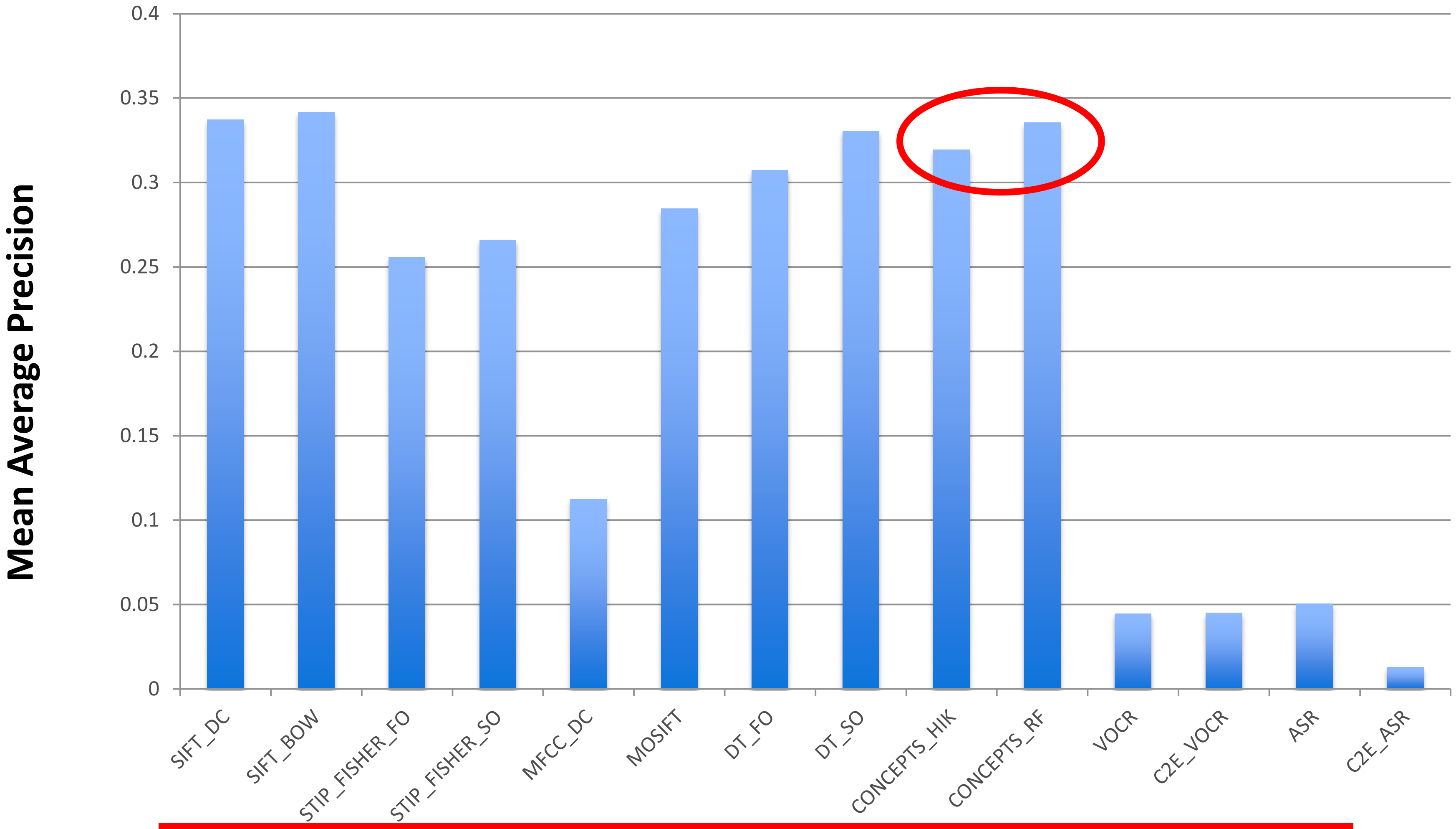
4 Event Classifiers for ASR and OCR Text

- **Each classifier measures the overlap of text in the test video with text in the event model using logistic regression**

- **Two event classifiers (one for ASR and one for OCR) based on text found in training set clips**
  - Unigram bag-of-words event models

- **Two event classifiers (one for ASR and one for OCR) based on text found in the event explications**
  - Identified the top-most relevant terms from the event explication using inverse document frequency (IDF) on a large English language text corpus
  - Augmented the terms with associated concepts found in WordNet

Performance of Individual Event Classifiers: EKFull

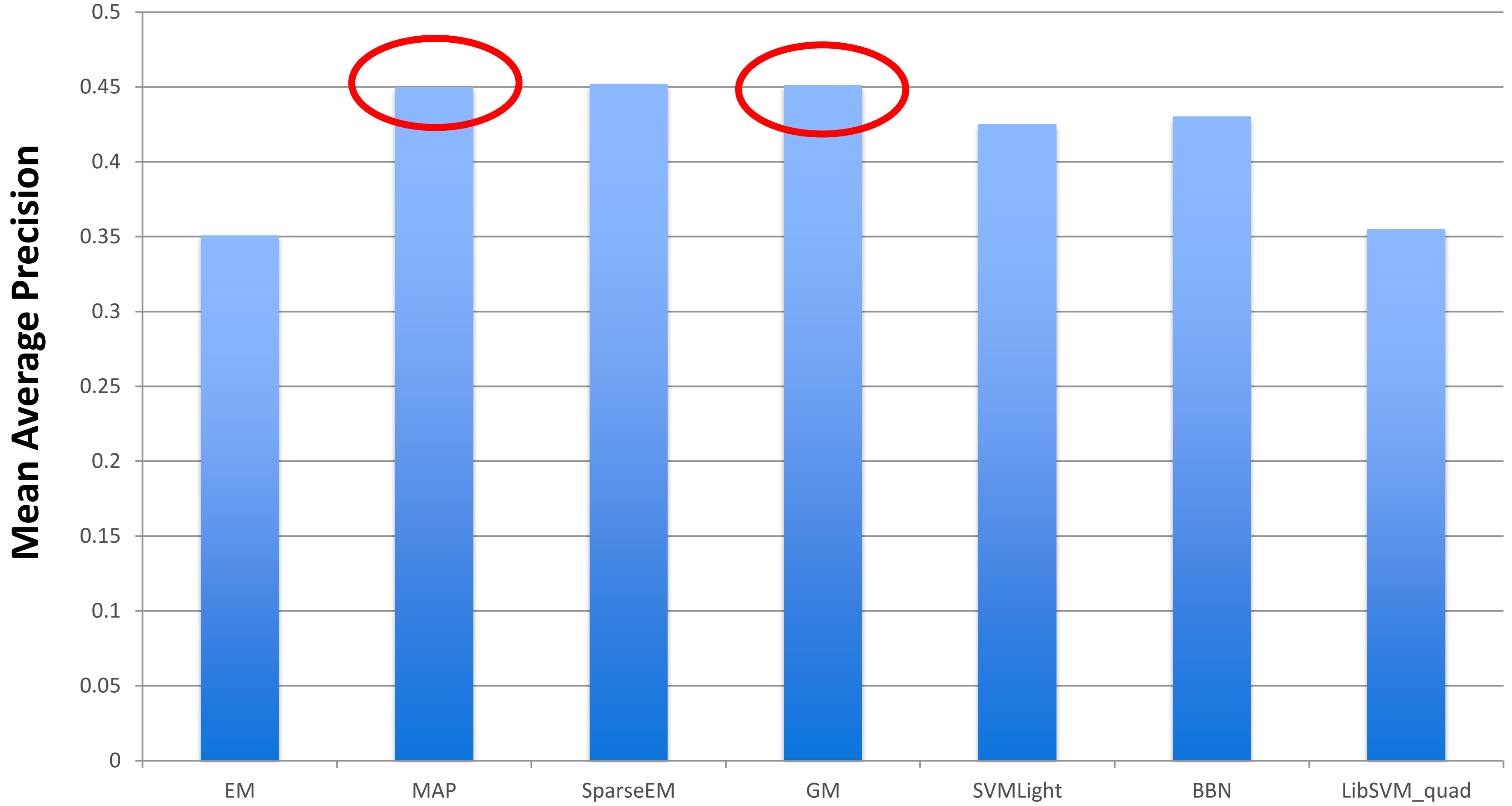


High-level features comparable to low-level features

Late Fusion Models

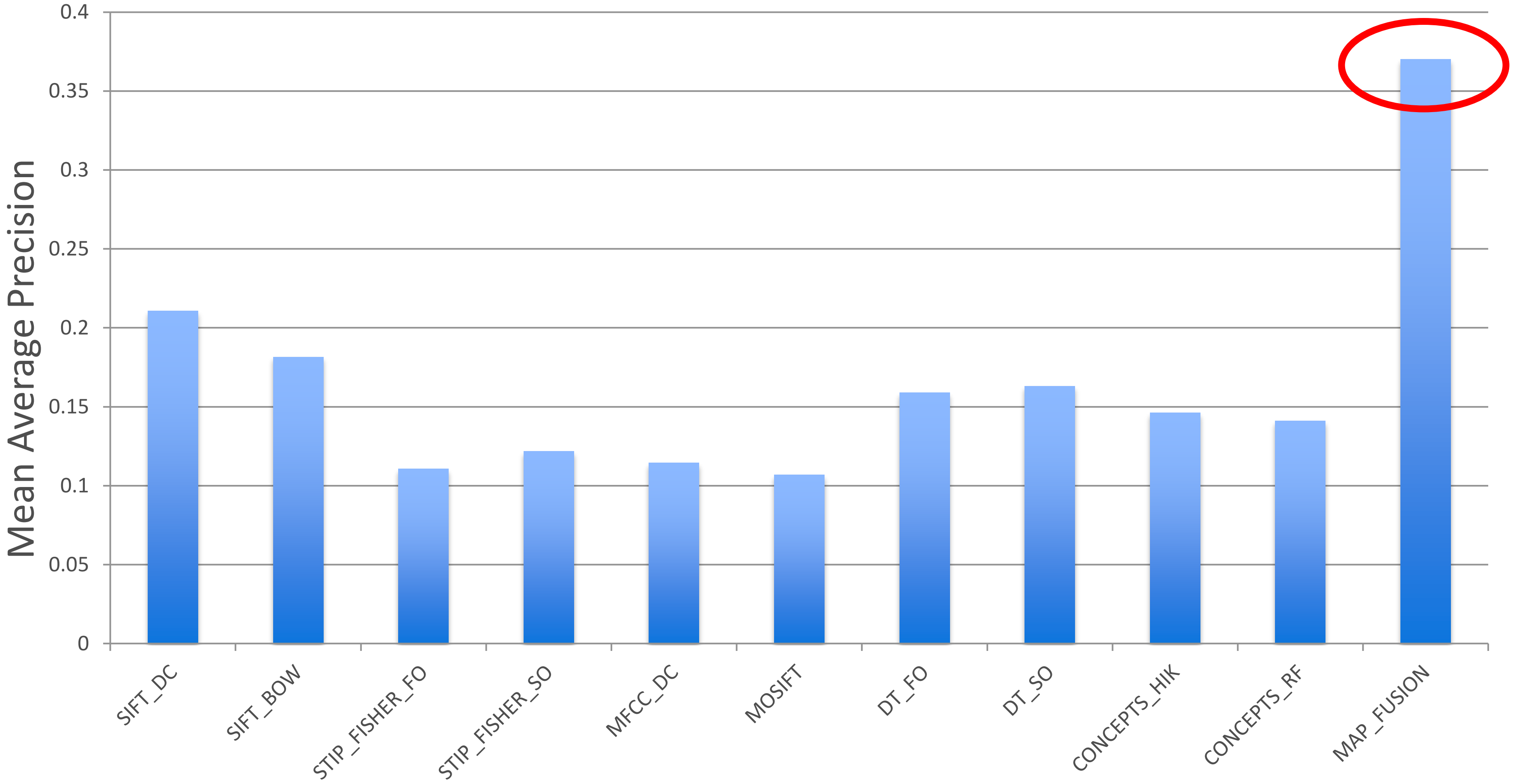
- **No weights**
  - Arithmetic mean (AM)
  - Geometric mean (GM)
- **Fixed weights**
  - Mean average precision-weighted fusion (MAP)
  - Conditional mixture model (EM)
- **Dynamic weights**
  - Sparse conditional mixture model (SparseEM)
  - Weighted mean root
  - SVMLight
  - LibSVM
  - BBN weighting (BBN)

Performance of Late Fusion Models: EKFull



Simple fusion models are good enough

Performance of Individual Event Classifiers : EK10Ex



Fusion produces big gain in performance

Threshold Selection Methods

- **Score@TER**
  - determined by the threshold that achieves the Target Error Ratio
- **Median score@TER**
  - for the ad hoc Ek10Ex condition only
  - median of the score@TER thresholds learned on the pre-specified events for the EK10Ex condition
- **Box-average – the average of two thresholds:**
  - the threshold that achieves P(Miss) = 50%
  - the threshold that achieves P(FA) = 4%

SESAME MED Evaluation Runs on Progress Set

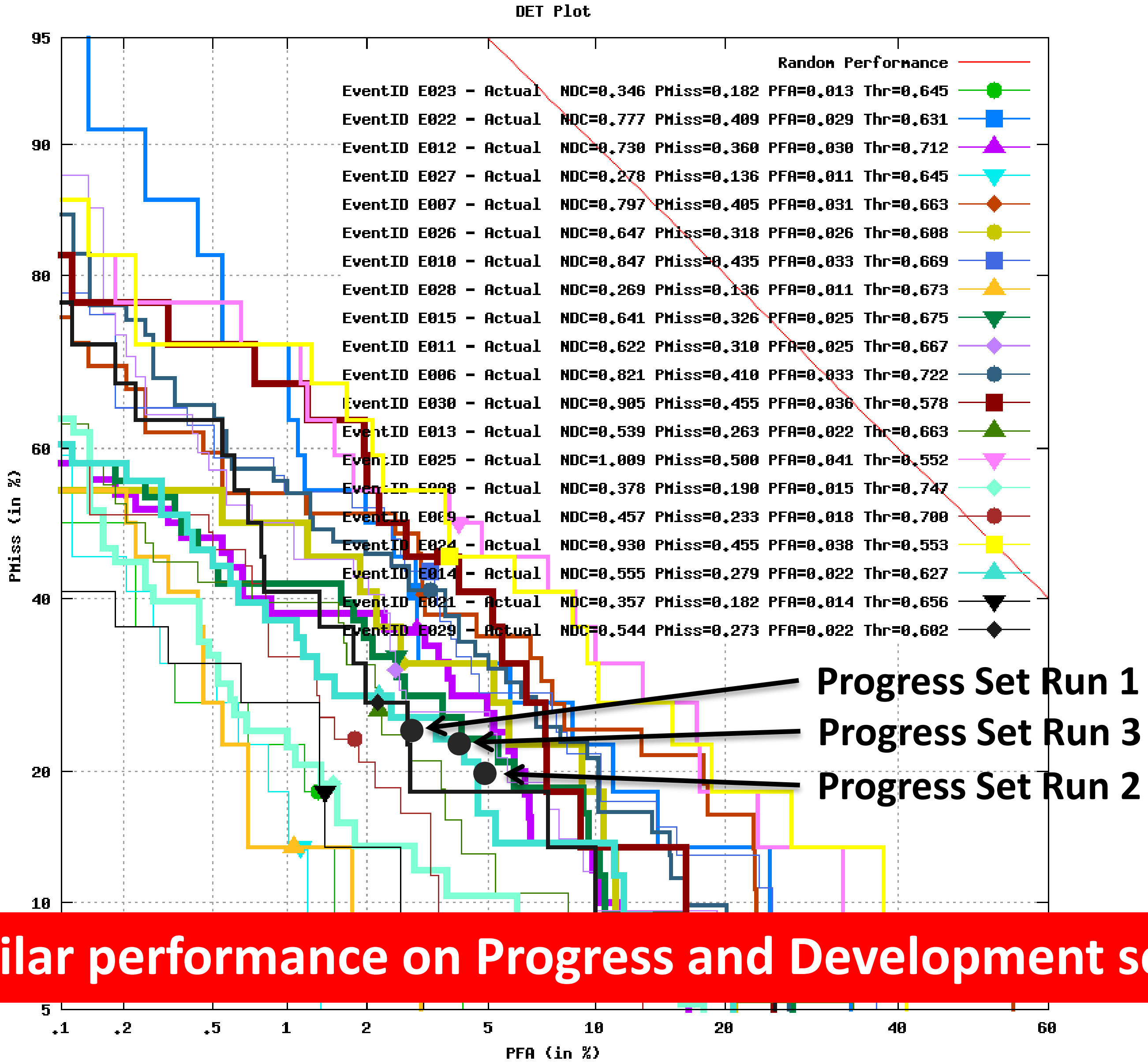
Runs 1, 2, and 3: Pre-specified events; EKFull; mix of extracted metadata, fusion methods, and thresholding methods

Run 4: Pre-specified events; EK10EX

Run 5: Ad hoc events; EKFull

Run 6: Ad hoc events; EK10Ex

Pre-specified Events, EKFull



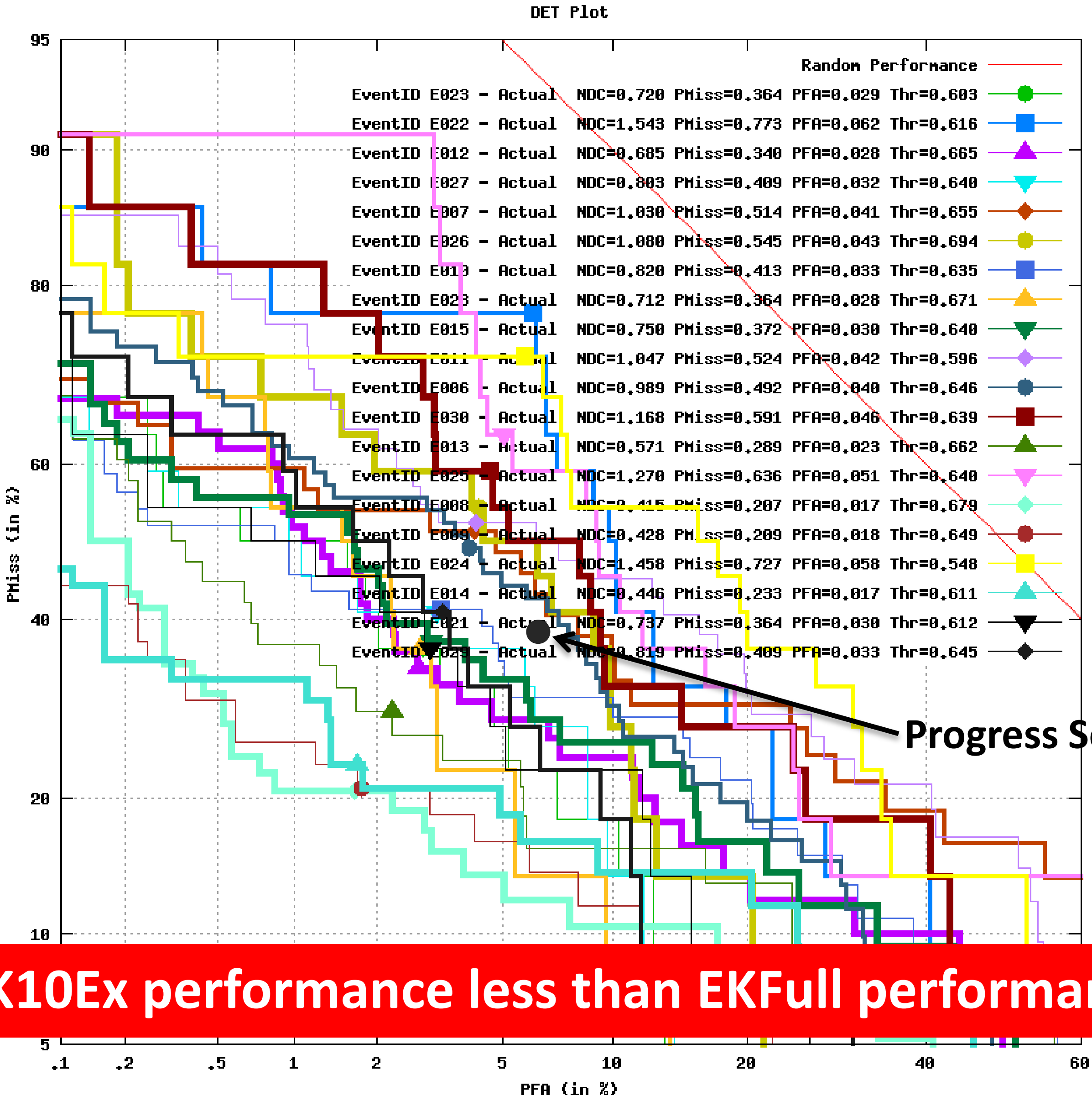
Progress Set Run 1

Progress Set Run 3

Progress Set Run 2

Similar performance on Progress and Development sets

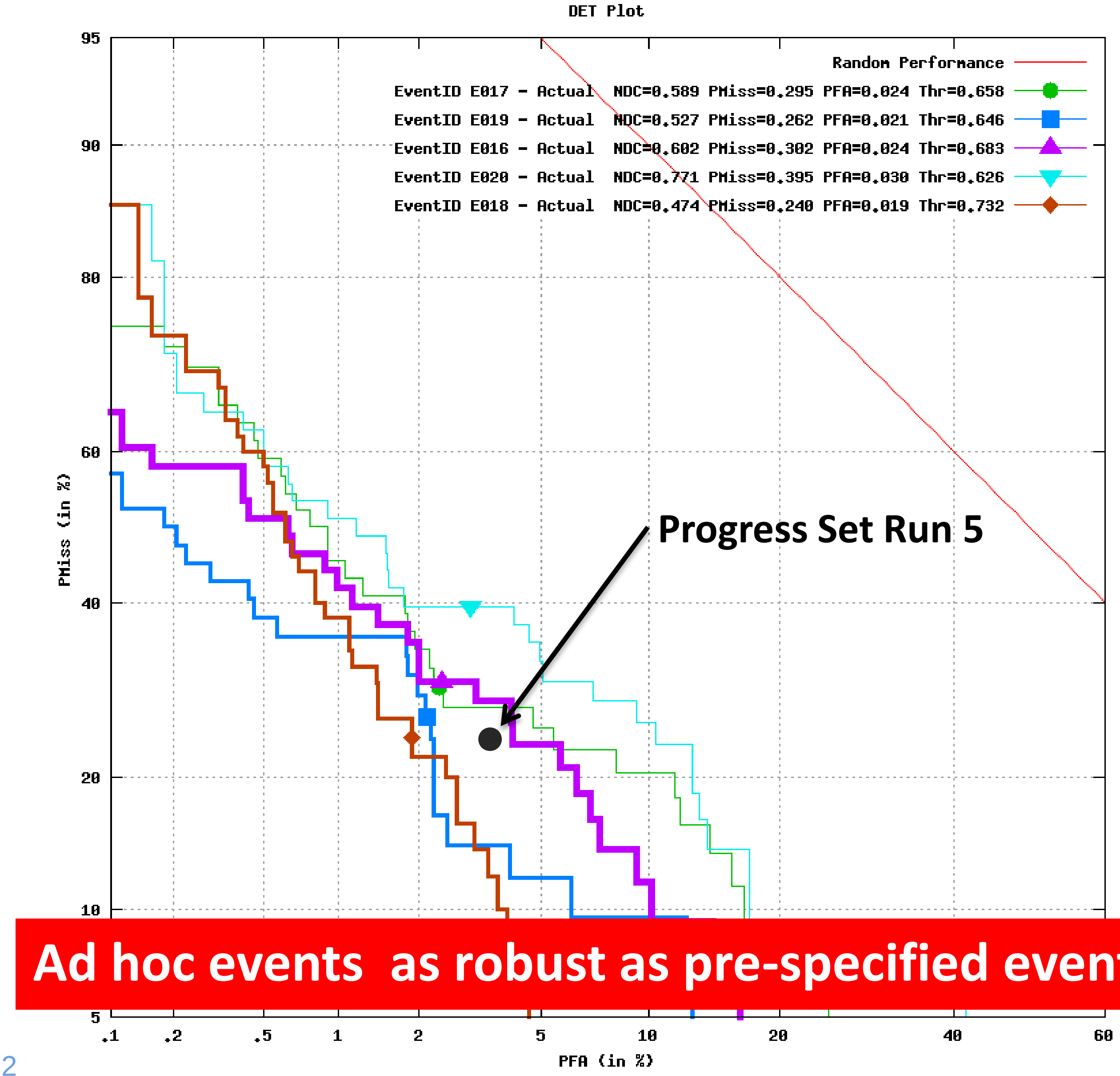
Pre-specified Events, EK10Ex



Progress Set Run 4

EK10Ex performance less than EKFull performance

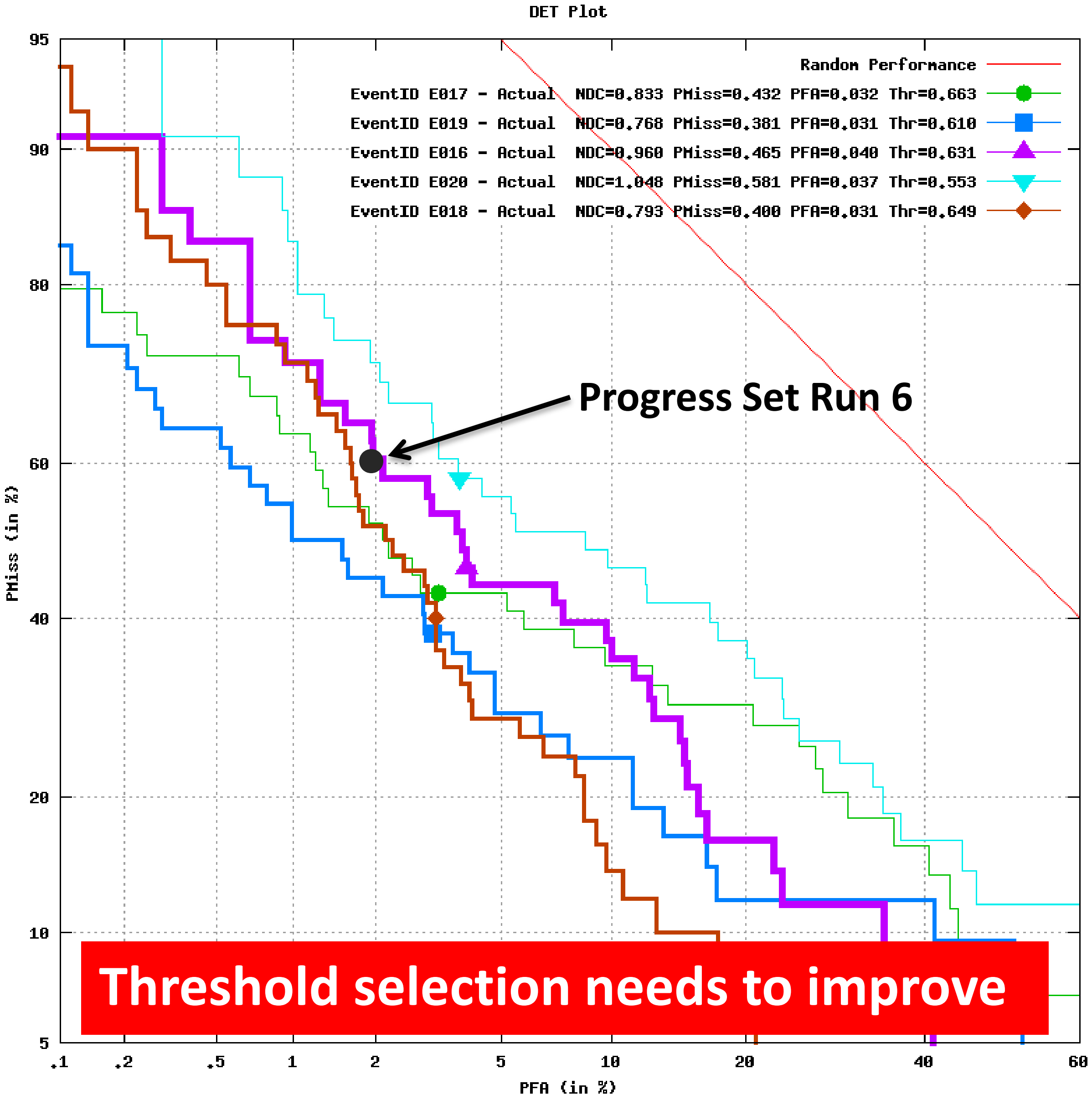
Ad Hoc Events, EKFull



Progress Set Run 5

Ad hoc events as robust as pre-specified events

Ad Hoc Events, EK10Ex



Progress Set Run 6

Threshold selection needs to improve

Conclusions

- High-level features comparable to low-level
- Simple average fusion good enough
- Similar results on Progress Set and our internal development set
- Ad hoc events as robust as pre-specified events
- Threshold selection needs to improve

SESAME Team for MED12

- **SRI International**
  - Gregory K. Myers, Murat Akbacak, Robert C. Bolles, J. Brian Burns, Mark Eliot, Aaron Heller, James A. Herson, Ramesh Nallapati, Stephanie Pancoast, Julien van Hout, Eric Yeh
- **University of Amsterdam**
  - Cees G.M. Snoek, Amirhossein Habibian, Dennis C. Koelma, Zhenyang Li, Masoud Mazloom, Silvia-Laura Pintea, Koen E.A. van de Sande, Arnold W.M. Smeulders
- **University of Southern California**
  - Ram Nevatia, Sung Chun Lee, Pramod Sharma, Chen Sun, Remi Trichet

Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes nonwithstanding 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, or the U.S. Government.