BBN VISER TRECVID MED 11 System

Outline

- **Overview**
- **Feature Extraction**
  - Low-level Features
  - High-level Features: Objects and Concepts
  - Automatic Speech Recognition (ASR) Features
  - Videotext OCR
- **Event Detection**
  - Kernel-based Early Fusion
  - System Combination
- **Salient Waypoint Experiments**
- **MED’11 Evaluation Results**
- **Conclusion**

BBN MED’11 Team

- BBN Technologies
- Columbia University
- University of Central Florida
- University of Maryland

Feature Extraction

Outline

- Low-level Features
- Compact Representation
- High-level Visual Features
- Automatic Speech Recognition
- Video Text OCR

Low-level Features

Low-level Features

- **Considered 4 classes of features**
  - *Appearance Features*: Model local shape patterns by aggregating quantized gradient vectors in grayscale images
  - *Color Features*: Model color patterns
  - *Motion Features*: Model optical flow patterns in video
  - *Audio Features*: Model patterns in low-level audio signals
- **Explored novel feature extraction techniques**
  - *Unsupervised feature learning* directly from pixel data
  - *Bimodal features* for modeling correlations in audio and visual streams

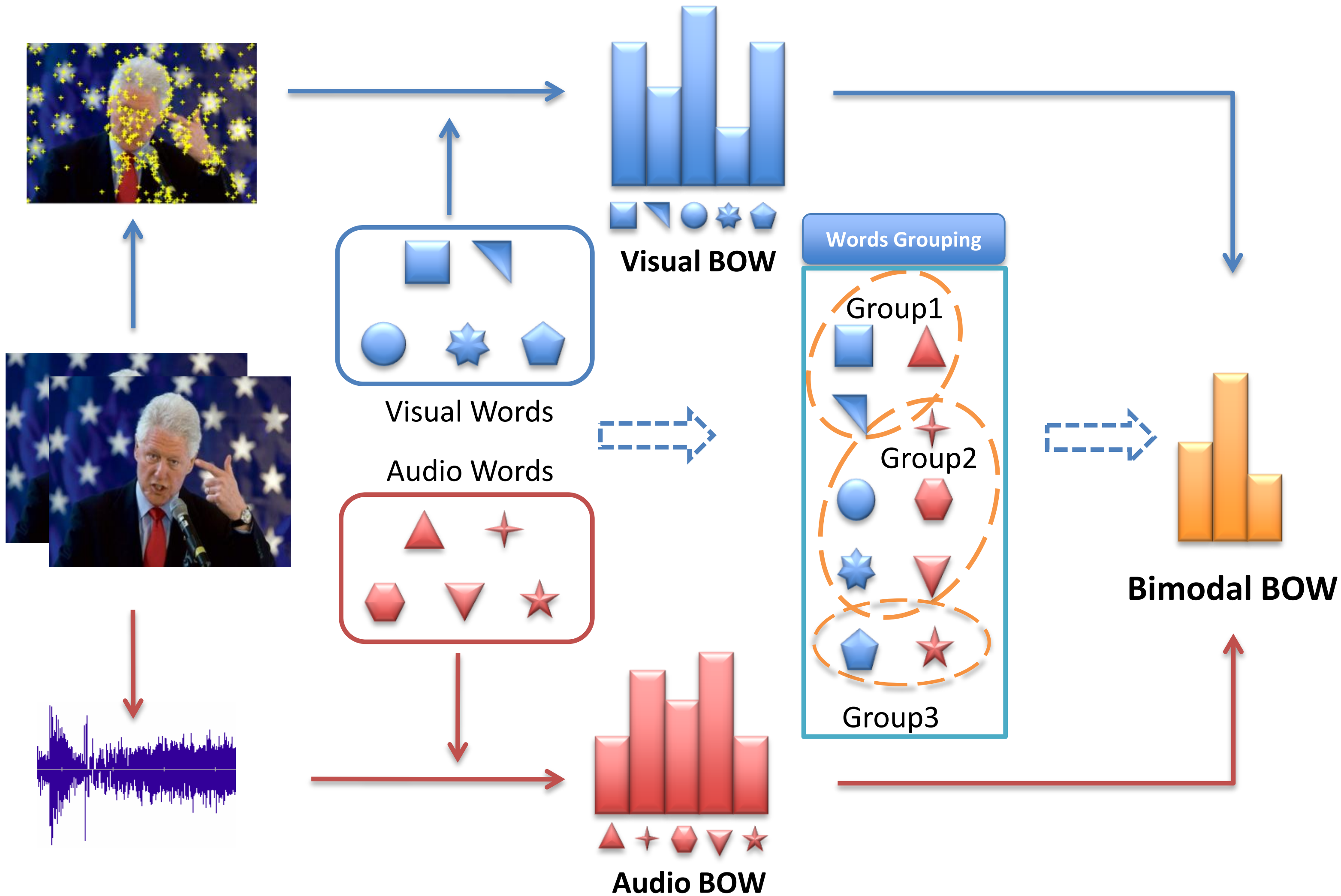
Unsupervised Feature Learning

- **Visual features like SIFT, STIP are in effect hand coded to quantize gradient/flow information**
- **Explored use of independent subspace analysis (ISA), to learn invariant spatio-temporal features from data**
- **Method was tested on UCF 11 dataset**
  - Produced 60% accuracy on UCF11 set, with block size of 10×10×16 and 16×16×20 for the first and second ISA levels
  - Produced similar results with block size of 8×8×10 and 16×16×15
  - When the two systems were combined, accuracy improved to 72%

Bimodal Audio-Visual Words

- **Joint audio-visual patterns often exist in videos and provide strong multi-modal cues for detecting events**
- **Explored joint audio-visual modeling to discover audio-visual correlation**
  - First, apply bipartite graph to model relations between the audio and visual words
  - Then apply graph partitioning to construct bi-modal words that reveal the joint patterns across modalities
- **Produced 6% MAP gain over Columbia's baseline MED10 system**

Bimodal Audio-Visual Words Model Illustration



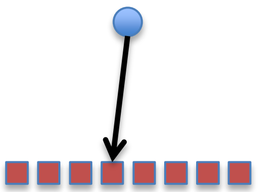
Compact Representation

Compact Feature Representation

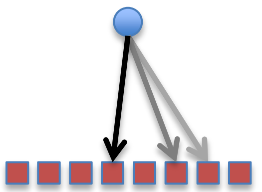
- **Two-step process**
- **Step 1: Coding to project extracted descriptors to a codebook**
- **Step 2: Pooling to aggregate projections**
  - Explored several spatio-temporal pooling approaches to model relationships between different features e.g. spatio-temporal pyramids

Coding Strategies

- **Hard Quantization**
  - Assign feature vector to nearest code-word
  - Binary



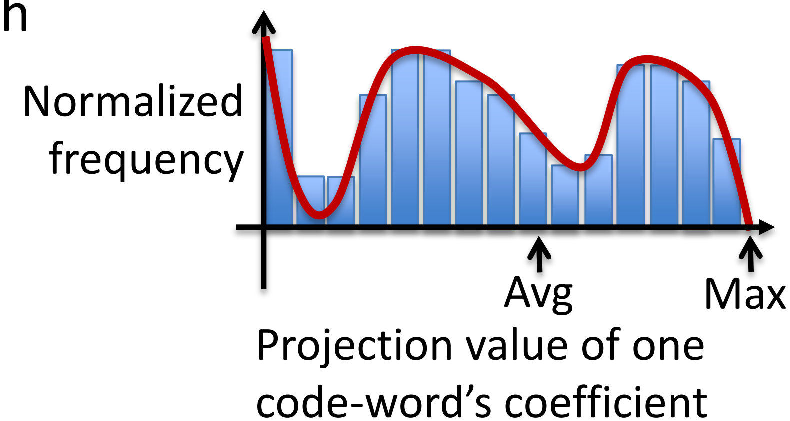
- **Soft Quantization**
  - Assign feature vector to multiple code-words
  - Soft assignment determined by distance



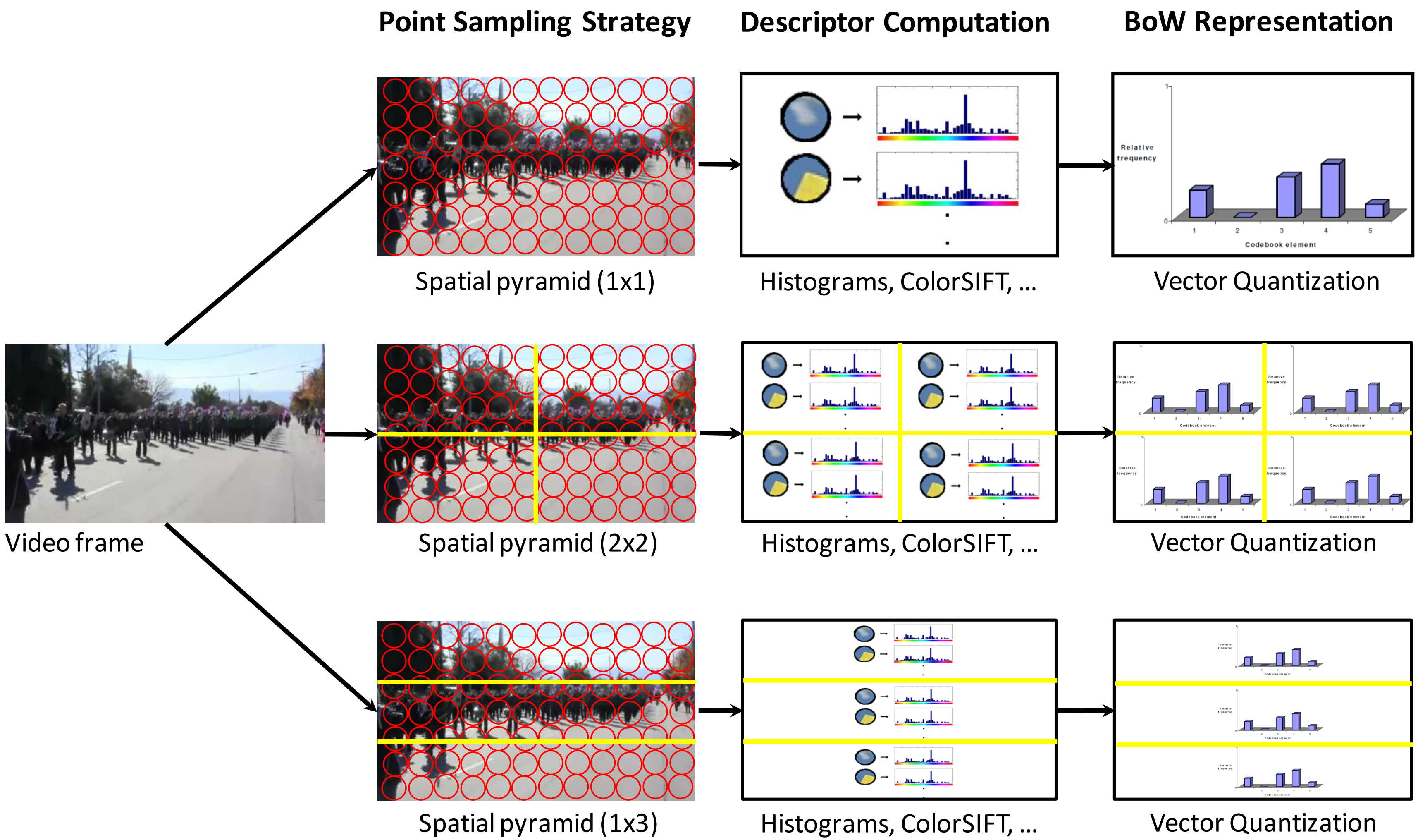
- **Sparse Coding**
  - Express feature vector as a linear combination of code-words $x_i = \Phi \alpha_i$
  - Enforce sparsity – only k non-zero coefficients

Pooling Strategies

- **Average pooling**
  - Average value of projection for each code-word
- **Max pooling**
  - Maximum value of projection for each code-word
  - Shown to be effective for image classification
- **Alpha Histogram**
  - Histogram of projection values for each code-word
  - Captures distribution of projections
  - Experiments indicate utility for video analysis



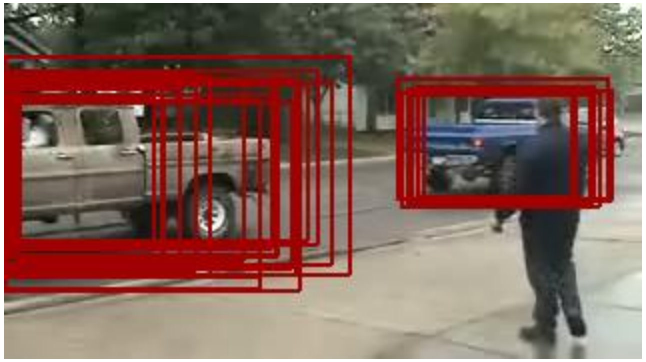
Spatio-temporal Pooling



High-level Visual Features

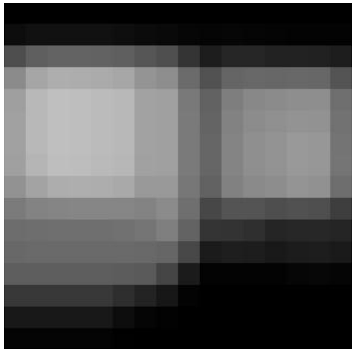
Object Concepts for Event Detection

- **Desirable properties**
  - Object should be semantically salient and distinctive for the event
    - E.g., Vehicle is central to "vehicle getting unstuck"
  - Accurate detection
    - Car detection has been studied extensively, e.g. PASCAL
  - Compact and effective representation of statistics
- **We employed a modified version of U. of C. object detector**
- **For each video frame, compute a mask from the bounding boxes**
- **Average over the duration of video**



Example of car detection in video frame

Accumulate over time



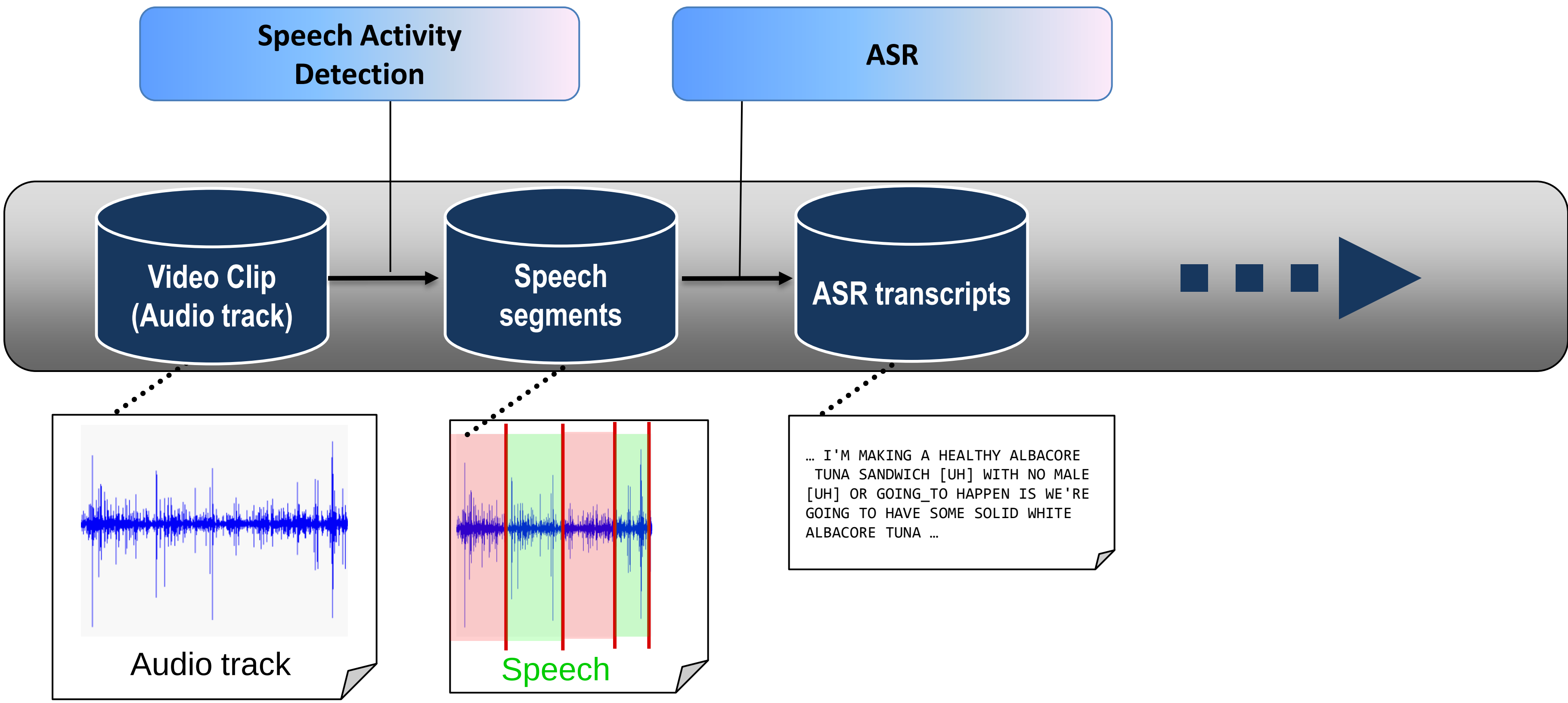
Spatial probability map of car detections mapped to a 16x16 grid

Concept Detection

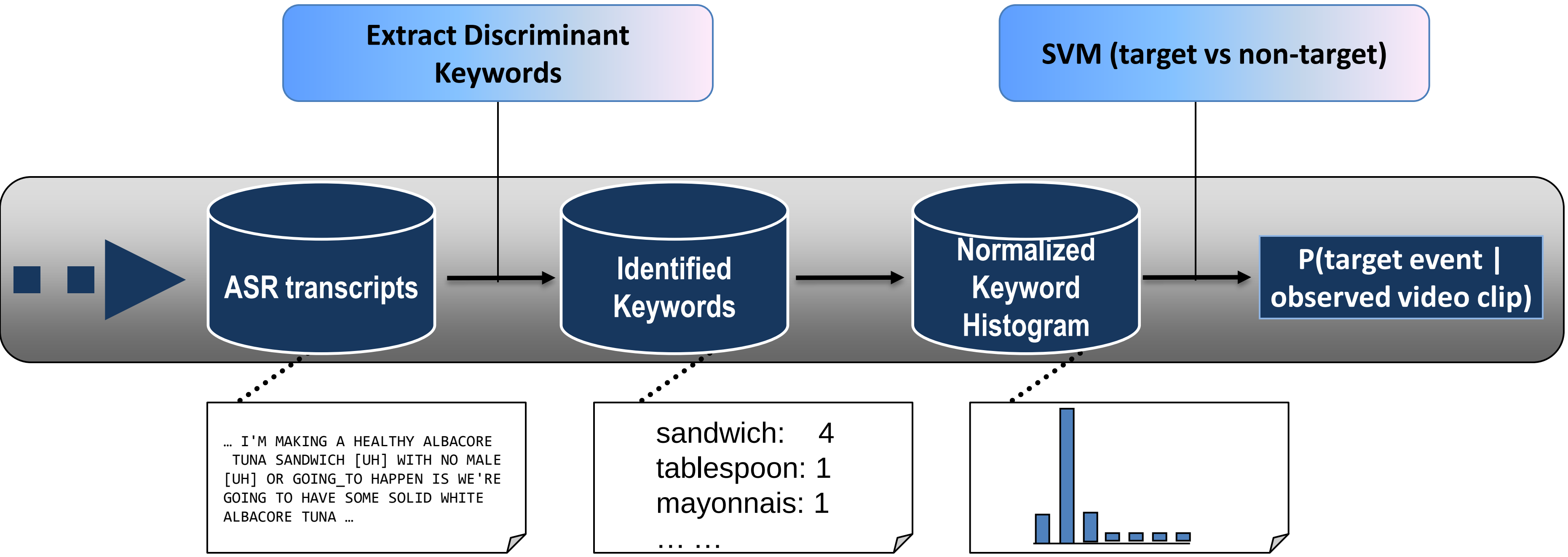
- **Preliminary investigation of concept features**
  - LSCOM: multimedia concept lexicon of events, objects, locations, people
- **Generated mid-level concept features from large LSCOM concept pool**
- **Trained the Classemes model provided in [Torresani et al. 2010]**
  - The concept scores generated by the classifiers were used as features for final event detection
- **Conclusions**
  - Concept-features < SIFT < SIFT + concept-features
  - Continue investigation in year 2

Automatic Speech Recognition

Getting Speech Content

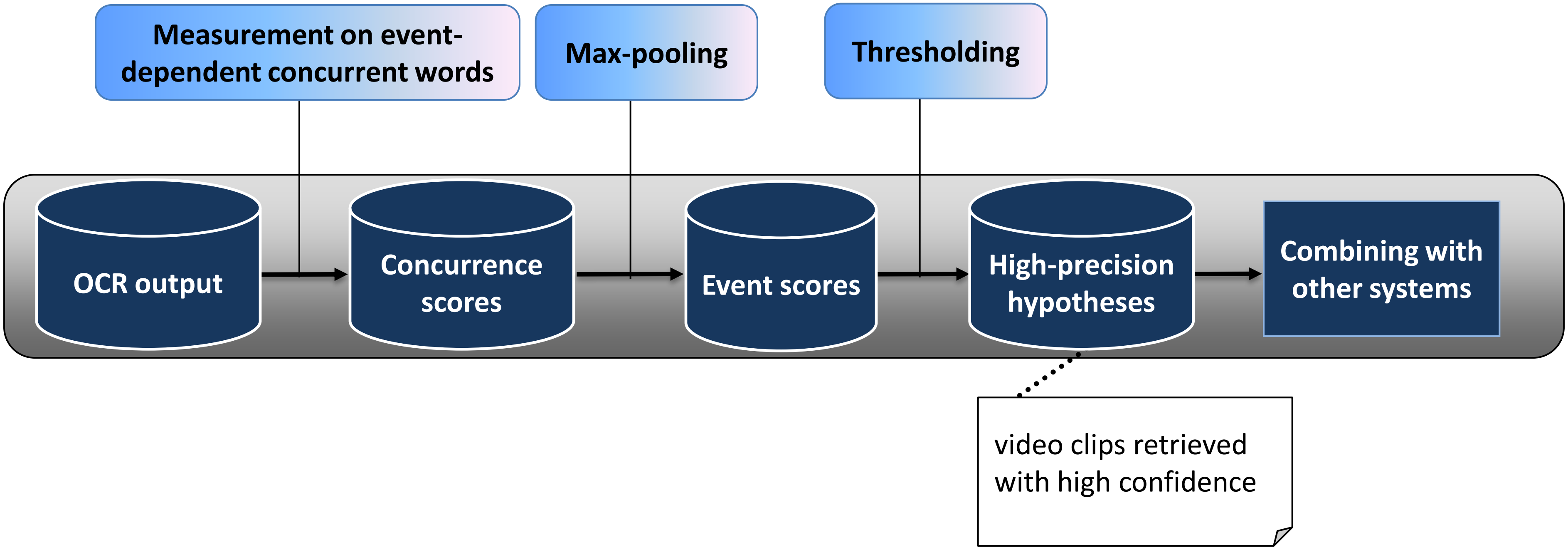


Event Detection Using Speech Content

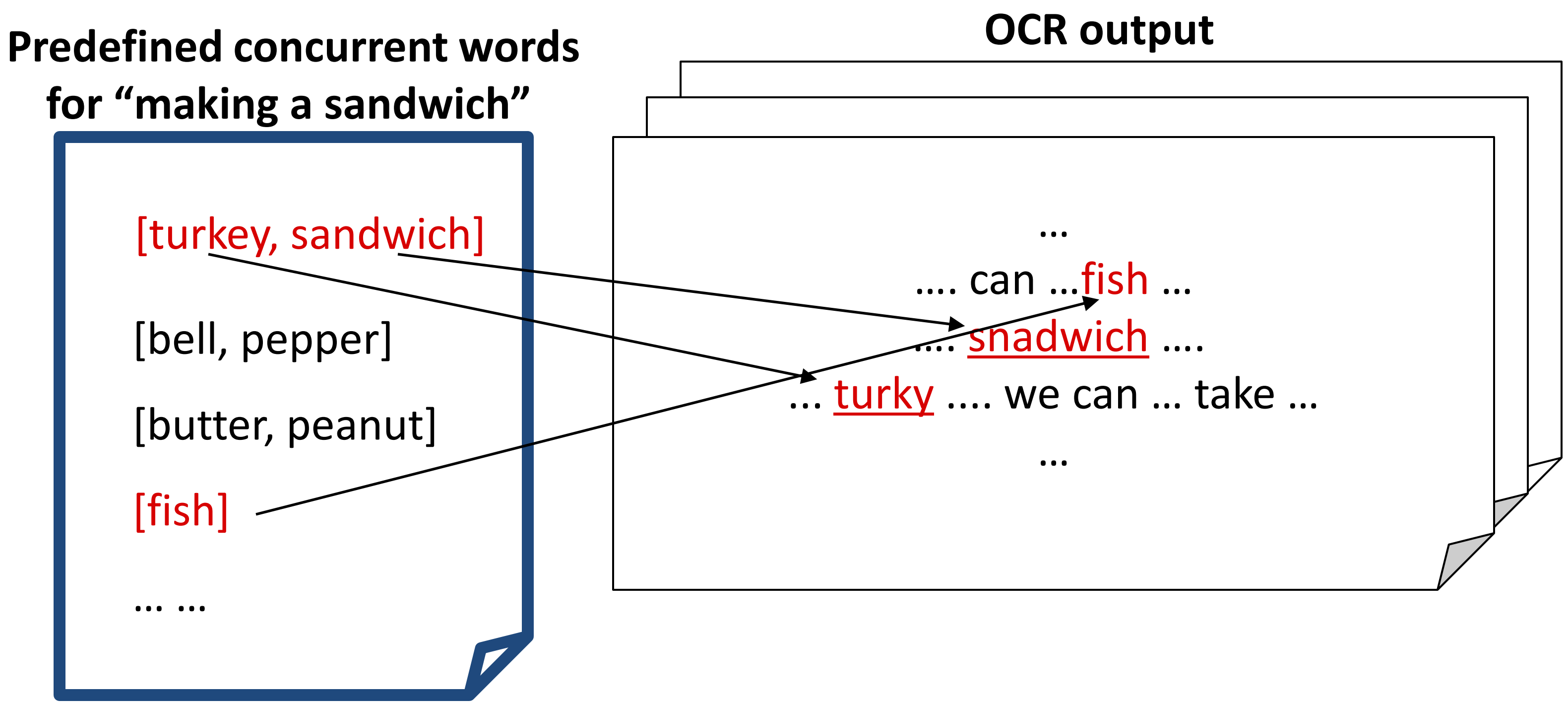


Video Text OCR

Using Video Text OCR



OCR-based Event Score for a video clip



Concurrence scores are converted to **OCR-based event score** by max-pooling over different dictionary entries and different frames.

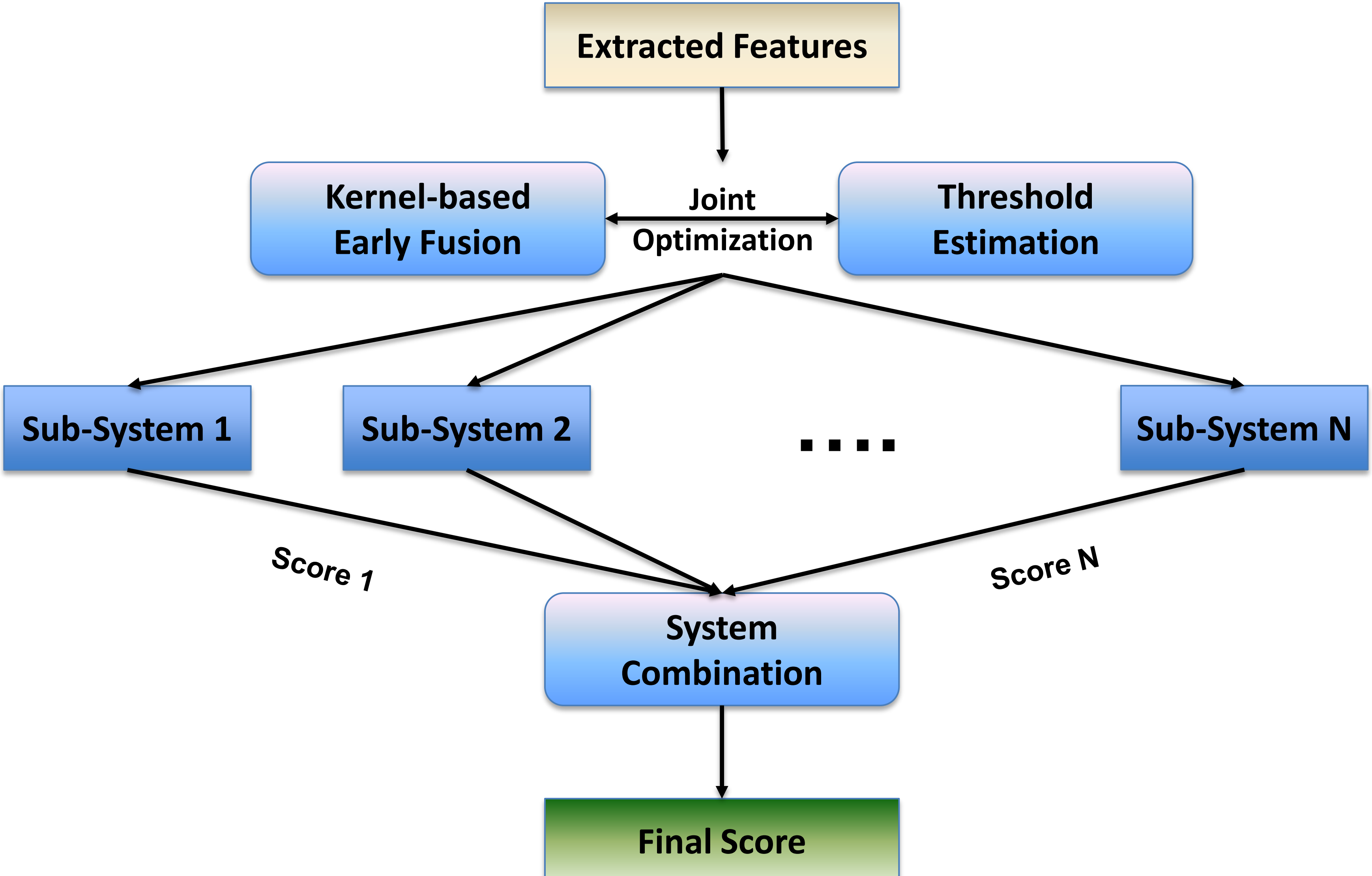
Event Detection

Outline

- **Event Detection Overview**
- **Kernel-based Early Fusion**
- **Detection Threshold Estimation**
- **System Combination**
  - BAYCOM
  - Weighted Average Fusion

Raytheon BBN Technologies

Event Detection Overview



Threshold Estimation Procedure

- Classifiers produce probability outputs, need to select a threshold for event detection
- Perform 3-fold validation on training set, generate DET curve of false alarm vs. missed detection for every threshold
- Select threshold to optimize for NDC/Missed detection rate on curve for each fold
- Average thresholds over each fold and apply estimated threshold on test set

System Combination: BAYCOM

- Bayesian approach, selects the optimal hypothesis according to:

$$c^* = \underset{c \in C}{\operatorname{argmax}} P(c \mid r_1, \ldots, r_n)$$

- Factorize assuming independence of system hypotheses

$$P(c \mid r_1, \ldots, r_n) = P(c) \prod_{i=1}^{N} P(s_i \mid c_i, c) P(c_i \mid c)$$

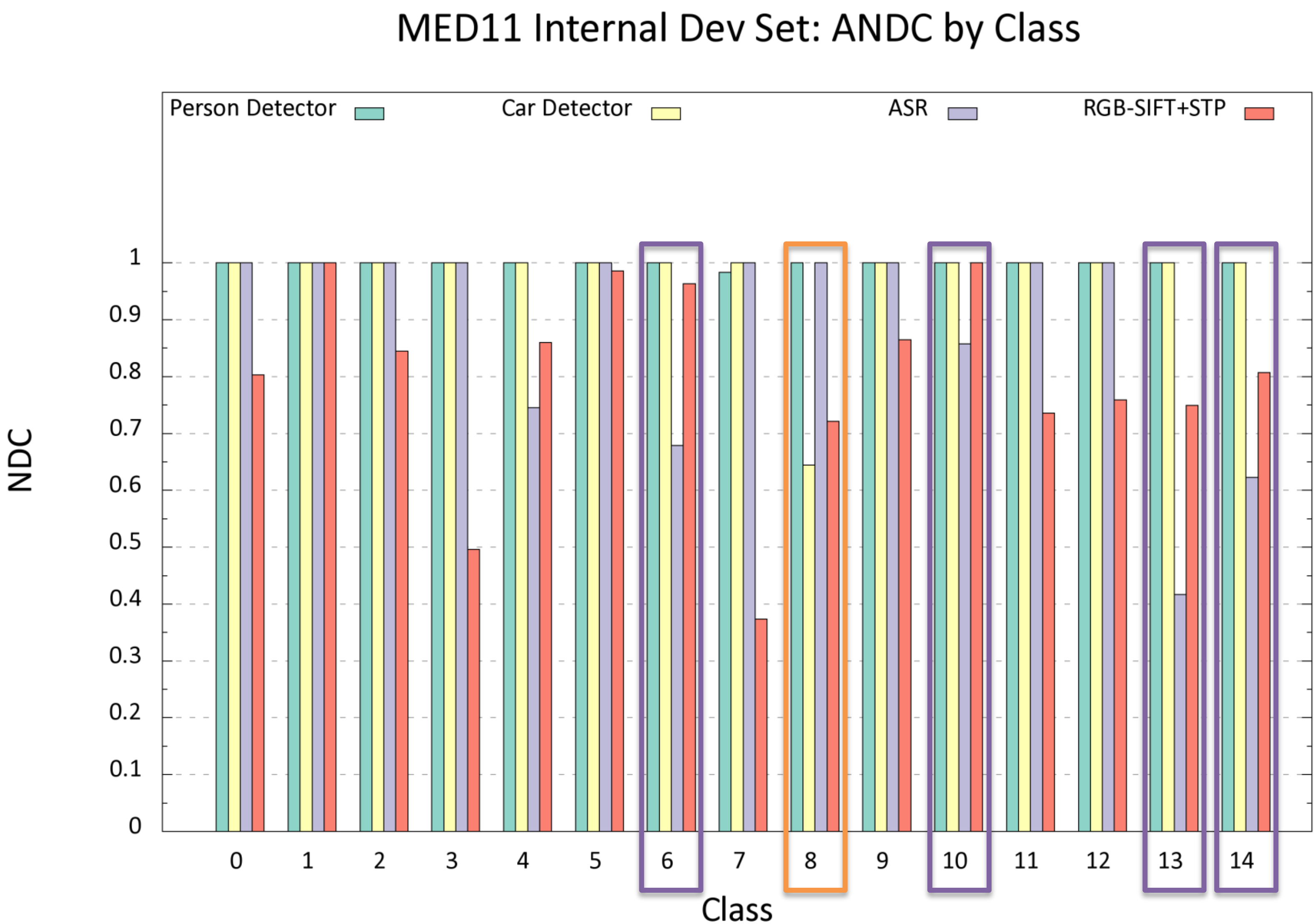
- Probabilities estimated from system performance relative to threshold
- Apply smoothing of conditional probabilities with class independent probabilities to overcome sparseness

Salient Waypoint Experiments

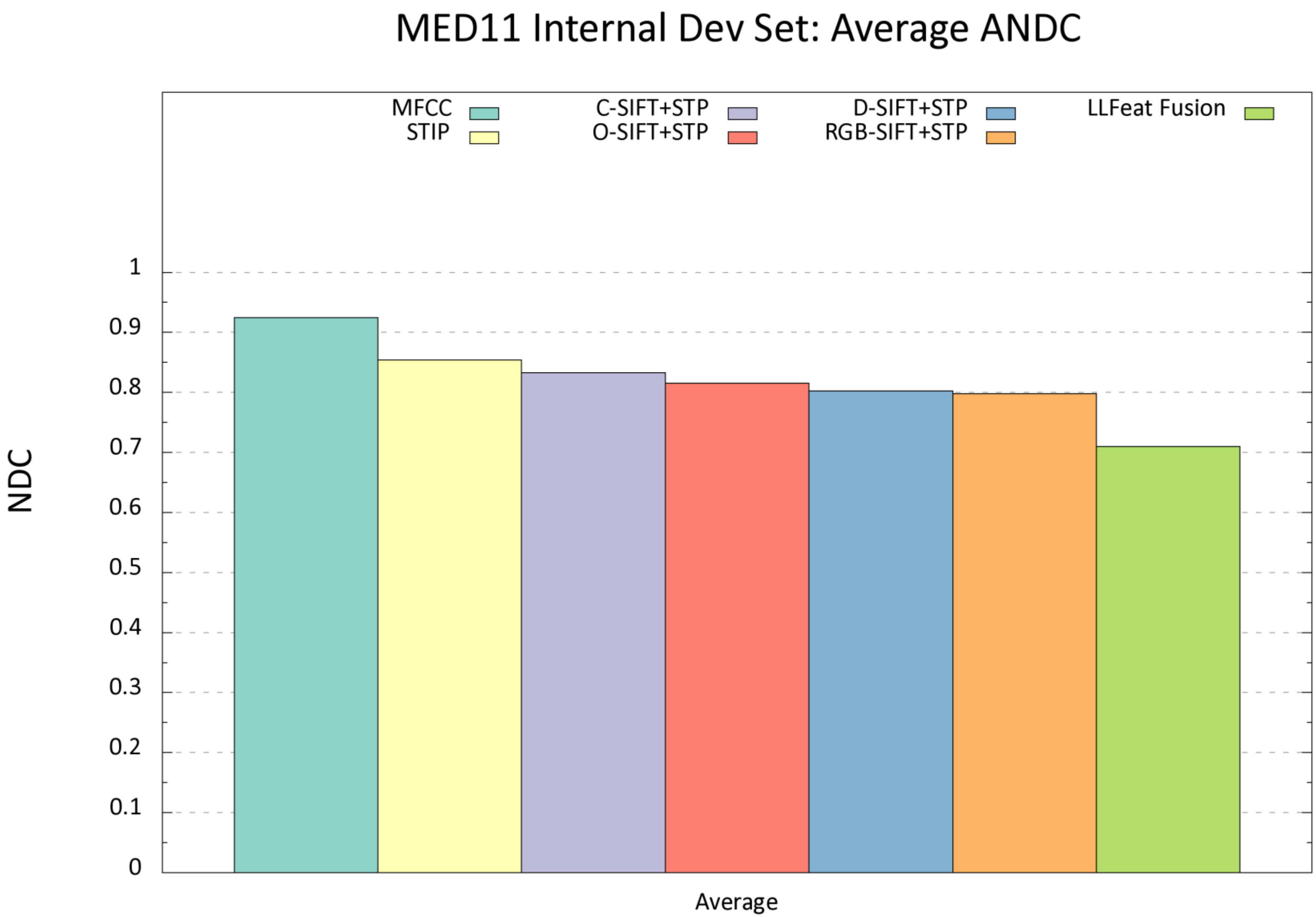
Experimental Setup

- **Event Kits and Dev-T are split into Train, Dev and Test partitions**
  - Train: for training initial models
  - Dev: for parameter optimization, fusion experiments
  - Test: to validate adjustments on the dev set
- **5 training events in Event Kits are split in Train and Dev, to simulate evaluation submission where all event kit videos are used for classification**
- **Positives in Dev-T set for the 5 training events placed into Test partition**
- **Setup may be sensitive to unlabeled positives in the negative Dev-T videos**

High Level Features



MKL Based Early Fusion



Late Fusion

Approach	Dev Set			Test Set		
	Avg. P_{MD}	Avg. P_{FA}	Avg. ANDC	Avg. P_{MD}	Avg. P_{FA}	Avg. ANDC
Min	0.5060	0.0154	0.6979	0.4950	0.0139	0.6686
Max	0.3606	0.0272	0.6999	0.3436	0.0263	0.6721
Voting	0.4161	0.0178	0.6383	0.3881	0.0154	0.5796
Average	0.3555	0.0230	0.6432	0.3219	0.0217	0.5925
BAYCOM	*0.5008*	*0.0068*	**0.5855**	*0.5105*	*0.0080*	0.6109
Weighted Average	0.3873	0.0166	0.5951	0.3599	0.0159	**0.5583**

MED'11 Evaluation Results

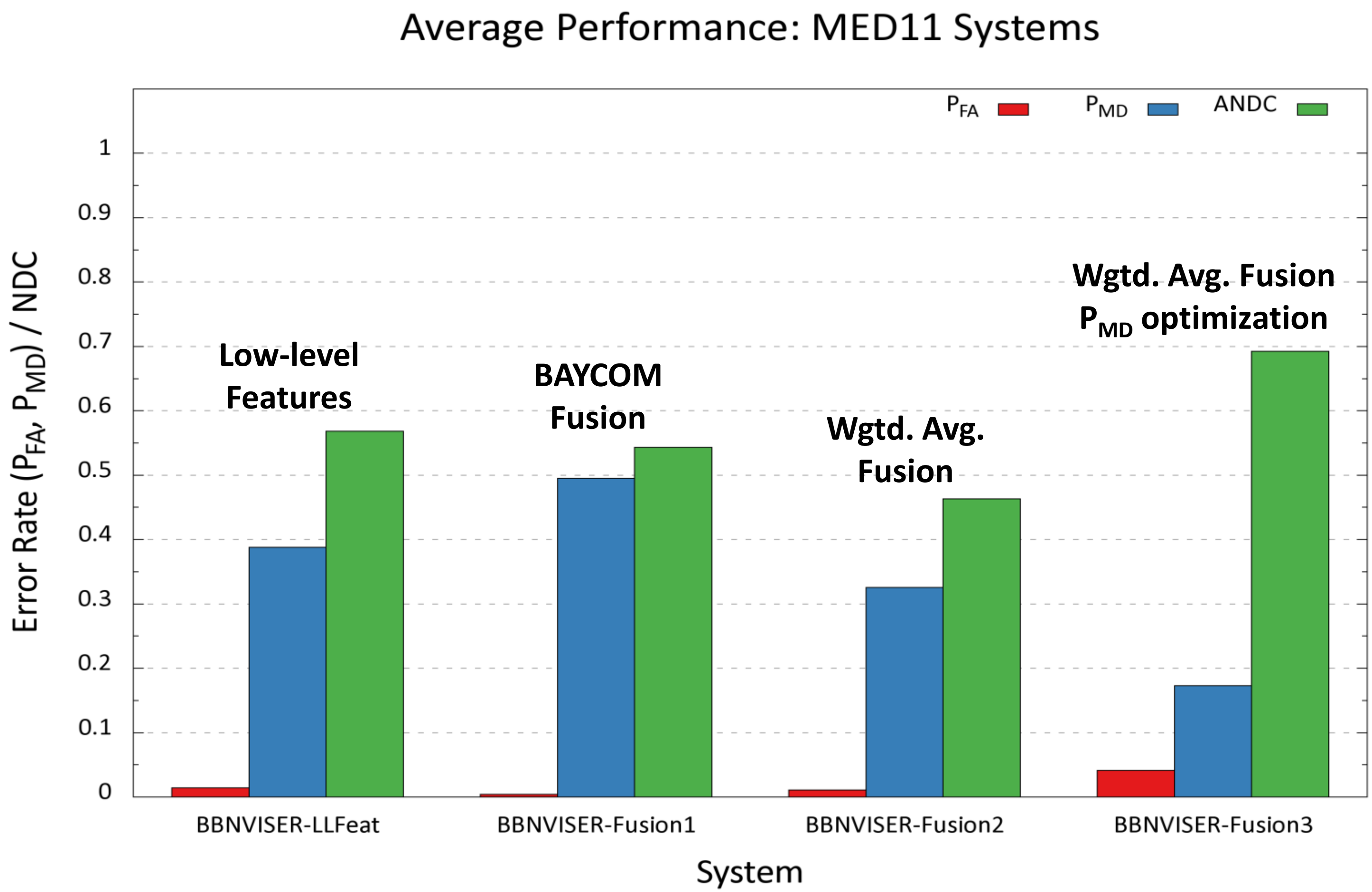
System Description

- **BBNVISER-LLFeat**
  - Combination of appearance, color, motion based, MFCC, and audio energy using MKL-based early fusion strategy
  - Threshold estimated to minimize the NDC score
- **BBNVISER-Fusion1**
  - Combines several sub-systems, each based on different sets of low-level features, ASR, and other high-level concepts using BAYCOM
  - Threshold estimated to minimize the NDC score

System Description

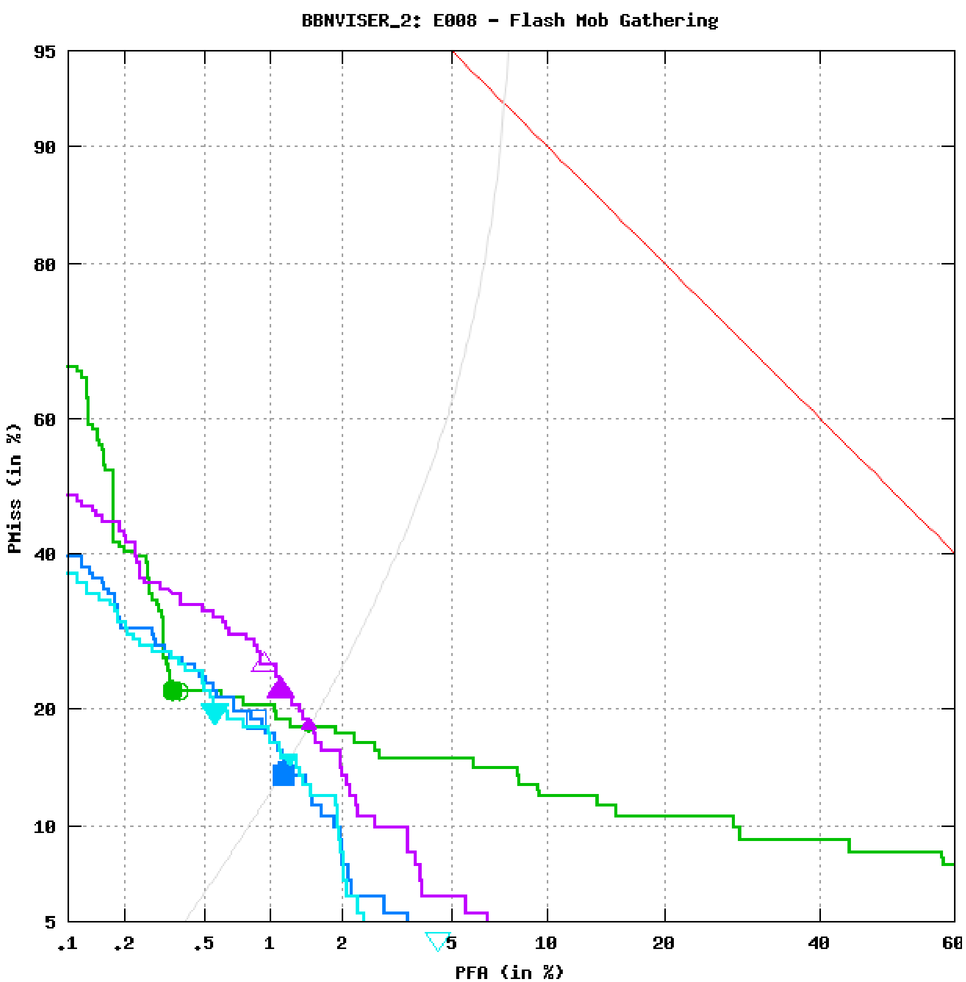
- **BBNVISER-Fusion2**
  - Combines same set of subsystems as BBNVISER-Fusion1 using weighted average fusion
  - Threshold estimated to minimize the NDC score
- **BBNVISER-Fusion3**
  - Combines all the sub-systems used in BBNVISER-Fusion3 with separate end-to-end systems from Columbia and UCF using weighted average fusion
  - Threshold estimated to minimize the probability of missed detection in the neighborhood of 6% false alarm rate ceiling

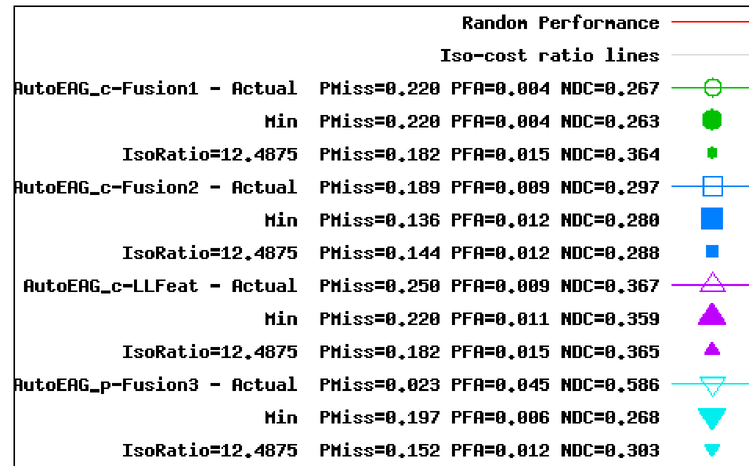
Summary of Performance



- Both early fusion of features and late fusion of systems are important
- High-level information from ASR, object/scene concepts, and video text OCR produces significant gains

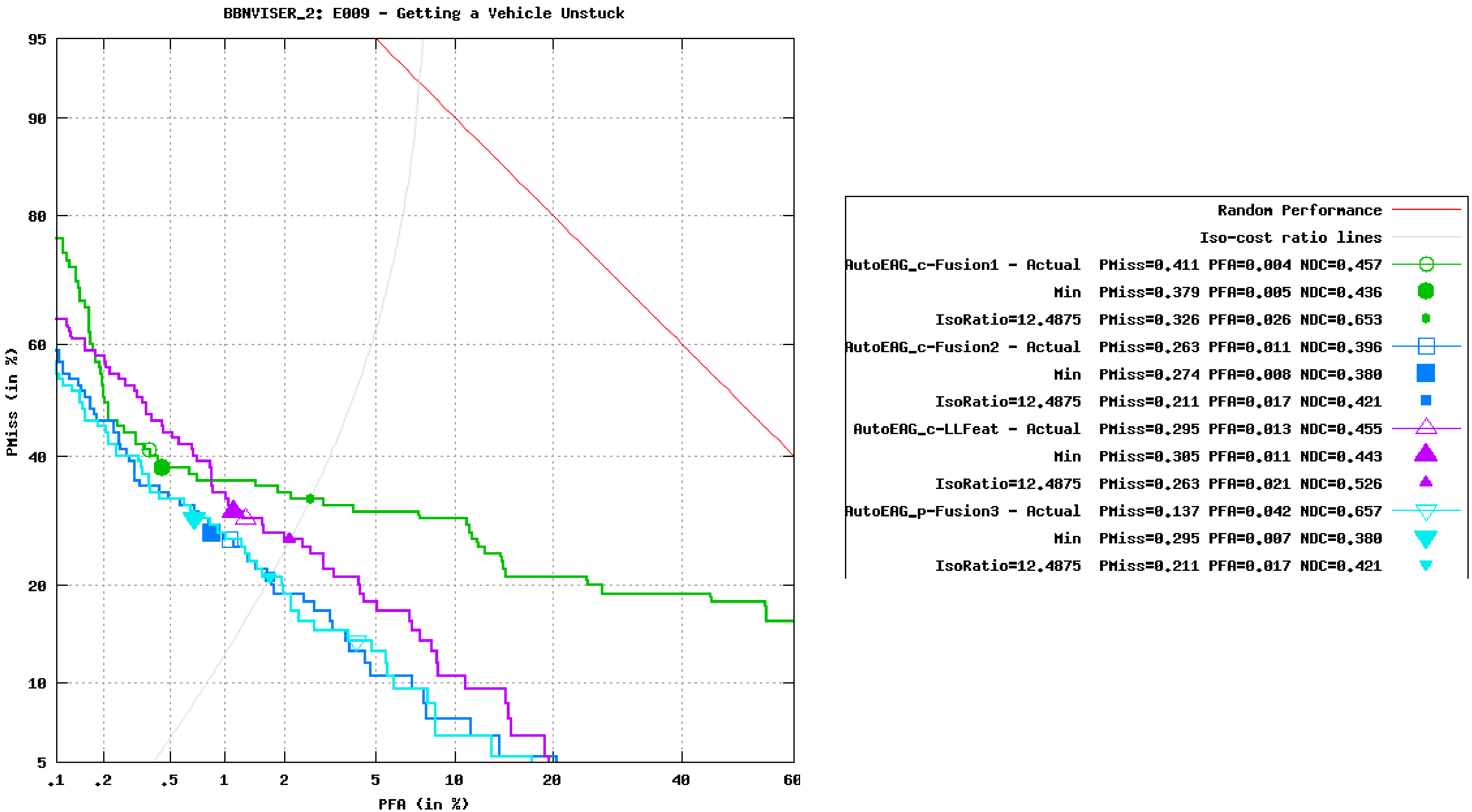
Performance Analysis (Flash Mob Gathering)



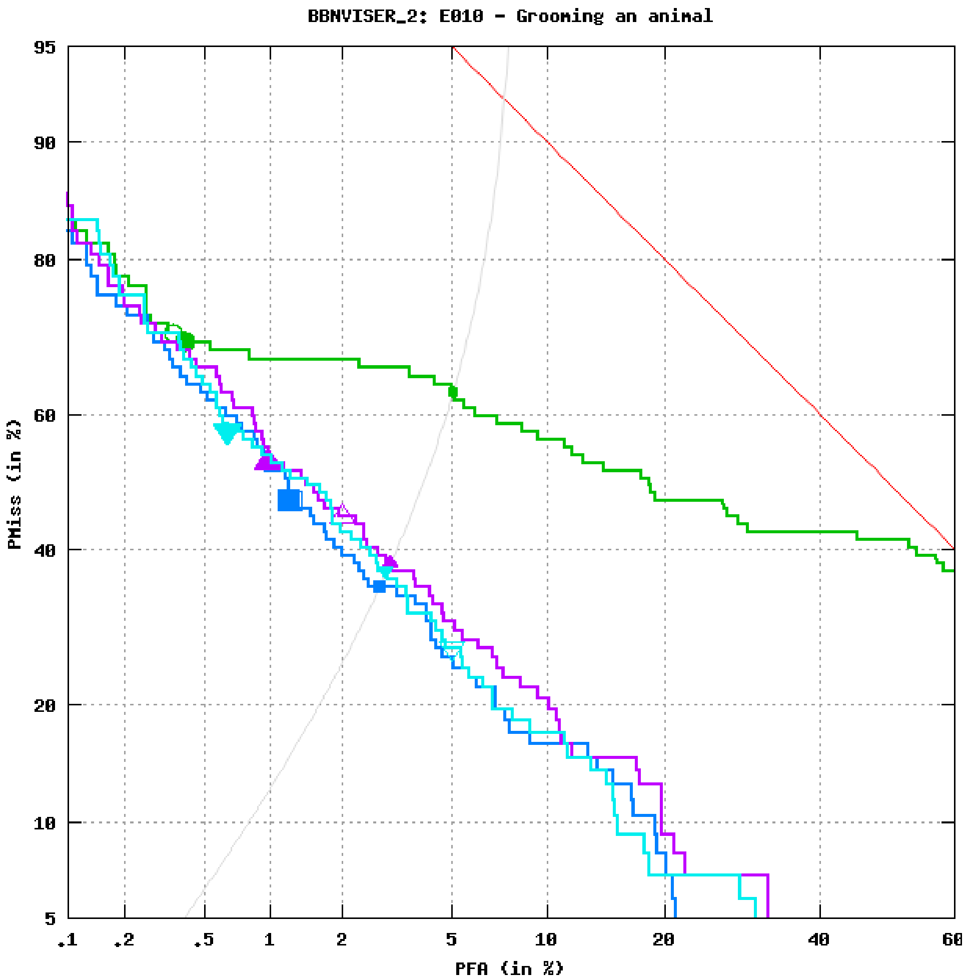


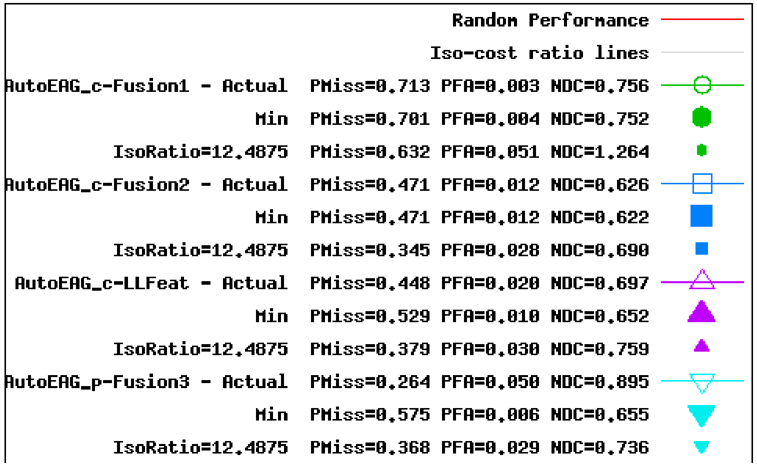
- High-level features provide significant gains
- BAYCOM optimizes the performance at a single point on the DET curve (detection threshold) and is sub-optimal at other points
- Weighted average fusion strategy improves performance over the entire DET curve

Performance Analysis (Getting Vehicle Unstuck)



Performance Analysis (Grooming an Animal)





- **The gain from high-level features is minimal**
  - Most of the videos did not have any associated audio or text information for ASR or videotext OCR to work
  - Scene and object concepts were not helpful either

Conclusions

- **Low-level features demonstrate strong performance and form the core of the system**
- **Speech and Video-text OCR provide significant performance gains**
- **Object and scene concept detection are promising, but gains are not consistent**
- **MKL fusion of even similar features produce gains, while diverse feature combinations produce largest gains**
- **Late fusion of multiple systems produces consistent gains**
  - Video-specific weighted averaging has the best performance